Proceedings of the

International Conference on Sequence Analysis and Related Methods (LaCOSA II)

Lausanne, Switzerland, June 8-10, 2016

Gilbert Ritschard and Matthias Studer (editors)

(June 8, 2016)

Foreword

The International Conference on Sequence Analysis and Related Methods (LaCOSA II) was held in Lausanne, June 8-10, 2016, four years after the Lausanne Conference on Sequence Analysis (LaCOSA). The conference brought together scholars using innovative methods for analyzing longitudinal data in social, managerial, political, population, psychological, health, and environmental sciences with developers of methods for longitudinal analysis.

Sequence Analysis (SA) has become a popular exploratory tool in social sciences since the pioneering contributions of Andrew Abbott and the recent release of powerful pieces of software. Nevertheless, SA remains essentially exploratory and needs to be complemented with other modeling tools, especially when it comes to testing hypotheses or studying the dynamics that drives the trajectories. In that perspective, this LaCOSA II conference did not limit itself to SA by also covered alternative longitudinal methods, such as survival and event history analysis, Markov-based and other longitudinal stochastic models. The aim was to debate how these different approaches can complement each other.

Alongside three keynote talks by Francesco Billari, Jeroen Vermunt and Aart Liefbroer, 57 papers were presented at the conference. The papers were selected among 79 propositions on the basis of reviews made by members of the scientific committee. In addition LaCOSA II featured also a longitudinal data analysis contest.

The present proceedings collect the papers presented at the conference. Some authors chose to not include their full paper for copyright reasons. In those cases only a—short or long—abstract is included.

We would like to warmly thank here all members of the Scientific Committee for their scientific support and their help in the reviewing process. Many thanks also to the Organizing Committee, as well as to Christelle Burri for her administrative assistance. We also acknowledge the financial support of the University of Geneva and its Geneva School of Social Sciences, the Foundation and the Institute of Social Sciences of

Organising Committee

Gilbert Ritschard, NCCR LIVES, U. of Geneva (chair)
André Berchtold, U. of Lausanne
Felix Bühlmann, U. of Lausanne
Jacques-Antoine Gauthier, U. of Lausanne
Jean-Marie Le Goff, U. of Lausanne
Emmanuelle Marendaz Colle, U. of Lausanne
Grégoire Métral, U. of Geneva
Emanuela Struffolino, WZB Berlin
Matthias Studer, U. of Geneva
Eric Widmer, U. of Geneva

Scientific Committee

Gilbert Ritschard, Switzerland (chair)
Andrew Abbott, USA
Silke Aisenbrey, USA
André Berchtold, Switzerland
Torsten Biemann, Germany
Francesco Billari, UK
Ivano Bison, Italy
Philippe Blanchard, UK
Göran Broström, Sweden
Christian Brzinsky-Fay, Germany
Felix Bühlmann, Switzerland
Benjamin Cornwell, USA
Daniel Courgeau, France
David Edwards, Denmark
Cees Elzinga, Netherlands
Anette Fasang, Germany
Alexis Gabadinho, Portugal
Jacques-Antoine Gauthier, Switzerland
Brendan Halpin, Ireland
Dominique Joye, Switzerland
Jean-Marie Le Goff, Switzerland
Laurent Lesnard, France
Tim Liao, USA
John J. McArdle, USA
Melinda Mills, UK
Madalina Olteanu, France
Raffaella Piccarreta, Italy
Gary Pollock, UK
Nicolas Robette, France
Patrick Rousset, France
Fiona Steele, UK
Katherine Stovel, USA
Emanuela Struffolino, Germany
Matthias Studer, Switzerland
Eric Widmer, Switzerland
Frans Willekens, Germany

iv

Contents

x

Keynotes

Discovery and explanation in demography and life course research

Francesco C. Billari
Oxford University

Simple and advanced latent Markov modeling: A flexible probabilistic approach to sequence analysis

Jeroen K. Vermunt
Universiteit Van Tilburg

Abstract: Latent Markov modeling – also referred to as hidden Markov, Markov switching, regime switching, or latent transition modeling – has become a quite popular tool for longitudinal data analysis in social, behavioral, and biomedical research. In this talk, I will introduce the latent Markov model as a tool for sequence analysis. The key feature of this probabilistic model is that it takes into account that true and observed sequences may not match perfectly. More advanced versions allow among others dealing with multivariate, parallel, and correlated sequences, with various forms of observed and unobserved heterogeneity, with nested/multilevel sequences, with incompletely observed sequences, and with an unequal spacing of the measurements in time. I will illustrate these simple and more advanced versions of the latent Markov model with an empirical example and discuss their implementation in the Latent GOLD 5.1 software.

Using sequence snalysis to understand the family-life course: Developments and future perspectives

Aart C. Liefbroer
Netherlands Interdisciplinary Demographic Institute (NIDI)

Session 3A: Epistemology

Do different approaches in social science lead to divergent or convergent models?

Daniel Courgeau

Institut National d'Etudes Démographiques (INED)

Abstract   Sequence analysis is essentially an exploratory tool and needs to be complemented with other modelling approaches when it comes to testing hypotheses or studying the dynamics that drives the trajectories. This paper will first explore some of these tools: event-duration models which lead to event history analysis; event-sequences models which lead to sequence analysis; multiple level models which lead to multilevel analysis; social network models which lead to multilevel social-network analysis; models based on individual agents which lead to agent-based analysis. It then shows that these models can be classified under some more general concepts: the *statistical individual* concept covers event history and sequence analysis; the *statistical network* concept covers multilevel and social-network analysis. Only the agent-based analysis seems to escape from these concepts as it models theoretical ideas rather than data. However as it remains at the individual level it is too reductionist to explain social behavior. It seems then necessary to set up a more robust research program for demography. This research program may follow the induction's way given by Bacon in searching for the structure of the studied phenomena and the interactions between the networks created by people. Such a program will be able to lead to a convergence of these different models.

1 Introduction

From it's inception by Graunt in 1662, the scientific study of population called by Petty (1690) *political arithmetick,* paved the way for around 200 years for demography, epidemiology, political economics, and more generally for population sciences. During this period a *cross-sectional* approach was followed for which social facts of a period exists independently of the individuals who experience them, and can be explained by various characteristics of their society. After the end of World War II, population scientists took a new view on these facts, which introduced the individual's lived time. This *cohort analysis* approach considered that the occurrence of a given event, during the life of a generation or a cohort, can be studied in a population which preserves all its characteristics and the same charac-

2

teristics for as long as the phenomenon manifests itself (Courgeau, 2007). This approach was however submitted to very restrictive conditions (Courgeau and Lelièvre, 1994), and leads to the more recent approaches which we will present more thoroughly in this paper.

Social scientists today use various methodological approaches that perform often complementary but sometimes divergent tasks. We shall first briefly describe the main methods used in population science, emphasizing their potential convergences as well as their divergences.

From this comparison, we shall try to identify the conditions that would allow a synthesis of the approaches through an analysis of a more epistemological nature regarding an inductive construction in the Baconian sense (1620). This method of induction[1] consists of discovering the principles of natural or social processes by way of experimentation and observation. It rests on the requirement that without these principles the properties observed would be different (Franck, 2002).

2 Research Areas

We shall present and mainly discuss here the five main approaches by event duration, event sequence, multiple levels, network and agent-based decisions, used in population science.

2.1 An approach by event duration

This first approach made its debut in social sciences in the early 1980s, more than thirty years after the introduction of longitudinal analysis. However, it was already in circulation earlier, particularly among statisticians. We can trace its origin to the notion of martingale, used by Ville in 1939 and Doob in 1953. In 1972, Cox proposed the joint use of life tables and a regression model. In 1975, Aalen suggested the use of counting process theory for the joint analysis of several events that an individual could experience simultaneously. In 1980, the analysis by Aalen et al. of the interaction between events in an event history introduced the approach into the field of population sciences.

This approach rests on robust mathematical and statistical foundations, which permit to establish risk factors and to treat censored observations. They are presented in statistical books by Kalbfleisch and Prentice (1980), Cox and Oakes (1984), Andersen et al. (1991), and Aalen et al. (2008). They make it possible to analyze changes of state, however diverse, and to demonstrate the role of many individual characteristics that can change over time during such transitions. The

---

[1] Induction is not taken in the sense of Mill (1843) and his followers, i.e. generalization from particular facts. In Bacon's sense, induction designates the complete research process.

3

application of the method in demography (see for example Courgeau and Lelièvre 1992) brought fresh progress in that field. Many other social sciences adopted it as well, including epidemiology, biostatistics, sociology, econometrics, actuarial sciences, and medicine.

The event-history approach eliminates the need for the overly restrictive hypotheses of longitudinal analysis while maintaining the individual point of view. Individuals can be tracked over a part of their entire lifetime, typically by means of a retrospective or prospective survey. It focuses on the duration between different events occurring in a person's life, and its application requires special surveys. For example, in 1981 the "triple event history" survey (currently called 3B, see Courgeau, 1999) allowed the simultaneous analysis of family events, occupational events, and migration events occurring over a lifetime up to the survey date for cohorts born between 1911 and 1936. As censored observations can be treated without problem by this approach, the persons who were always in the labor force at the time of the survey (four fifths) can be studied for their occupational history in the same manner than those who were retired (one fifth).

It basically relies on semi-parametric methods, which, while preserving a non-parametric vision of the duration between events, use parameters to describe the effects of personal characteristics (Courgeau and Lelièvre, 1992).

However, the event-history approach did pose a certain number of problems, to which we now turn.

The first problem is that of unobserved heterogeneity. How does unobserved heterogeneity affect the estimation of parameters of observed characteristics? To help us answer the question, we have an important result obtained by Bretagnole and Huber-Carol in 1988 but overlooked by some users of these models. The two authors showed that, in a Cox model, when the omitted characteristics are independent of the observed characteristics, the omission has no impact on the sign of the estimated parameters, reducing only their absolute value. Thus, if the effect of a characteristic is found to be fully significant, the introduction of unobserved characteristics will merely strengthen that effect. Conversely, a characteristic that does not have a significant effect may have one when the omitted characteristics are introduced. We need to be aware of this risk.

When observed and omitted characteristics are connected, the situation is more complex. It may be tempting to introduce this heterogeneity as a particular type of distribution, which Vaupel et al. called *frailty* in 1979. When we have information on the distribution, its introduction is entirely legitimate. The problem is that we typically do not know this distribution, and that it is often chosen for no other valid reason than convenience. In such circumstances, some estimates may even change the sign of certain parameters, as Trussell and Richards showed in 1987, while a model without frailty avoids this problem.

We therefore totally agree with Aalen et al. (2008), who, in their extensive studies on stochastic processes, have tried to identify individual frailty:

> As long as there is no specific information about the underlying process and observations are only made once for each individual, there is little hope of identifying what kind of process is actually driving the development.

4

Indeed, for the analysis of non-repetitive events, there is only one model without observed heterogeneity, but an infinity of models with unobserved heterogeneity. Their estimates differ, but they display an identical fit with observed data (Trussell, 1992). By contrast, if we are analyzing repetitive events—such as successive births or migrations—we have the option of estimating multilevel models that allow the introduction of unobserved heterogeneity, which reflects the multiple events experienced by every individual. We shall present these multilevel models later.

The second problem concerns the concept of probability used. Apart from Kalbfleisch and Prentice, most of the earlier-mentioned statisticians who developed the method chose an objective probabilistic approach, which places certain constraints on the expected results of an analysis. Could an epistemic approach enable us to lift many of these constraints? We cannot give a full description of the probabilistic approach here, such as in Courgeau (2012), but we can elaborate on the constraints linked to statistical inference.

The purpose of statistical inference is to optimize the use of the incomplete information available in order to take the best decision. Statistical inference will therefore consist in providing an analysis of a past phenomenon and a prediction of a similar phenomenon to come. The first point is important for sciences such as demography or epidemiology, which must analyze human behavior. The second point is crucial for sciences, such as medicine, or those focusing on public health which aim to produce the best possible forecast of the outcome of a treatment course or a decision on the best policy to implement. Statistical inference notably leads to testing various hypotheses about the phenomena studied.

Objectivist methods, also called frequentist methods, seek to verify whether a given factor does or does not affect the phenomenon studied, and this brings us to the notion of statistical test. This means treating the sample under analysis as one possible selection from an infinity of other samples that we extract from a population also assumed to be infinite. When we assign a confidence interval of, say, 95% to a parameter estimated on this sample, we might conclude that the probability of the unknown parameter lying in the interval is 0.95. In fact, however, the objectivists tell us that this conclusion is wrong. All we can state is that if we draw an infinity of new samples, then the new estimated parameters will lay in that interval 95% of the time. As Jeffreys wrote in 1939, when examining various definitions of objective probability:

> The most serious drawback of these definitions, however, is the deliberate omission to give any meaning to the probability of a hypothesis. All they can do is to set up a hypothesis and give arbitrary rules for rejecting it in certain circumstances.

That is exactly what happens with statistical tests. Similarly, the use of frequentist methods for prediction will consist in taking the parameters estimated, for example, by means of maximum likelihood and introducing them into the distribution function of the new observation. But this will not allow us to factor in the uncertainty of the parameter estimation, and will lead to an under-estimation of the variance of the predicted distribution.

5

That is why Jeffreys himself showed that if we accept that a probability is never a frequency—in other words, if we adopt the epistemic framework—then a 95% confidence interval truly means an interval in which the statistician rightly believes that the unknown parameter may lie with a probability of 0.95. Moreover, this approach enables us to solve the prediction problem, for which the objective approach could provide only an approximate solution. All we need to do is calculate the "*posterior predictive distribution*" of a future observation from the initially observed data, which are known. What we obtain is not a value, as with the objectivist method, but a distribution whose variance will now be calculated correctly.

We have not described in detail all the advantages of using an epistemic method, but they have led a number of authors to propose it for event-history analysis, especially when the sample studied is small: see for example the book published by Ibrahim et al. in 2001. However we will see in *2.3* that another way of making statistical inference is possible with epistemic probability.

The last problem we would like to address is that of the risk of *atomistic fallacy* involved in this approach. If we can draw on all individual characteristics to explain a behavior, we shall overlook the context in which the behavior occurs. In fact, when using a cross sectional approach the researcher introduced only the characteristics of the society to explain social facts. This aggregate approach was on the contrary under the risk of *ecological fallacy*, as Robinson (1950) so clearly demonstrated: he showed that the correlations between two characteristics measured in binary mode on individuals, or by proportions applied to different geographic segmentations, generally diverged. We will se later how to solve this difficulty.

2.2 An approach by event sequences

We can trace the origin of sequence analysis in computer science as used by Levenshtein (1966); then in molecular biology for the study of DNA and RNA sequences as used by Levitt (1969). It was introduced later in the social sciences, with the work of the sociologist Abbott (1983, 1984) in order to study social processes which occur by whole sequences generally during a long period of time.

However, this approach in social science rests on less robust mathematical and statistical foundations than event history analysis. Its main object is to describe whole sequences (ordered list of elements) in terms of types that reflect socially meaningful trajectories experienced by subjects (individuals or more general entities like stimuli in psychology or artifacts in archaeology). It follows a two-step approach. First it tries to compute a distance between sequences under some operations (insertions / deletions called "indels" or substitutions) with a given cost for each operation. The main used metric is called Optimal Matching (OM), but we will see later that many other methods to compute these distances may be used. Then in a second step, using cluster analysis, it is possible to detect types of sequences, regrouping the whole set of subjects into exclusive and mutually exclu-

6

sive categories. Cornwell (2015) gives a more detailed description of these methods. A great number of social sciences adopted it: sociology, demography, psychology, economics, anthropology, political science, linguistics, etc.

The sequence approach permits to turn from the cross-sectional research of causes in Durkheim's sense (1895), to an emphasis on contexts, connections and events which Abbott (1995) called a quiet revolution in social science. The surveys used to track subjects over their life time are very similar to event history surveys, with an emphasis on the observation of whole processes without censoring. Their goals, however, are very different: while event history analysis seeks the causes of the studied phenomena, sequential analysis explores the paths followed without offering reasons for the underlying processes that generate them (Robette and Bry, 2012). So that individual characteristics need not to be recorded for this analysis, out of their event sequences and their characteristics before the analyzed sequences. For example the 2001 "Event histories and contact circle" survey made by Lelièvre, on a sample of cohorts born between 1930 and 1950, following the example of the "triple event history" survey but more detailed, permitted to apply sequence analysis to the professional trajectories of mothers and their daughters in order to compare them (Robette et al. 2012).

Contrary to event history relying mainly on semi-parametric methods, it basically relies on non-parametric methods which make no assumption about the process underlying the life course. Its aim is to explore and describe the course of events as a whole, without trying to focus on the risk of experiencing events and their determinants. There are also some recent Bayesian extensions of social sequence analysis (Bolano, 2014), through Hidden Markov models similar to biological approaches (Liu and Logvinenko, 2003).

However, this analysis poses some new problems, different from those posed by event-history analysis, to which we now turn.

The first problem lies in the metric used, particularly in the use of OM methods for social sciences. As we have said this approach was imported from information theory and molecular biology. For these disciplines the hypotheses lying in their foundations have been shown to be plausible and of wide applicability: the model proposed by Levitt (1969) for transfer ribonucleic acid confronted with experimental observations was in good agreement with them and theories about chemical processes give a strong support to these methods. However in social sciences the structure of sequences appears to be much more complex. As Wu (2000) said:

> Part of my skepticism stems, in part, from my inability to see how the operations defining distances between trajectories (replacements and indels) correspond, even roughly, to something recognizably social.

For example giving the same cost to a transition from unemployment to employment than from employment to unemployment seems highly implausible. This skepticism was more clearly demonstrated by Bison (2009), while the use of OM techniques has multiplied. He clearly shows, by means of simulations, that varying the substitution and indels costs may produce inconsistent results. This may lead to find regularities even when they do not exist (Bison, 2014).

In order to solve this problem a number of generalizations of OM method were proposed: variable substitution costs, different distance measures, spell-adjusted measures, non-alignment techniques, monothetic divisive algorithm (MDA), etc. See Cornwell (2015) for more detail on these improvements.

However, while the number of distances and costs measurements increases the problem of their comparability becomes more and more important. While comparisons exist between few different metrics, using empirical data, the only study comparing a large number of metrics, using a reasoned set of artificial sequences, was made by Robette and Bry (2012). They did not try to find the best metric but "rather to unravel the specific patterns to which each alternative is actually more sensitive". Even if they found some differences between the results of these metrics, "the main patterns they conceal will be uncovered by most of the metrics". However the differences exist and the inconsistent results found by Bison let the problem of the used metric largely unsolved.

The second problem lies in the use of cluster analysis for detecting classes of sequences. This method of classification was used long before sequence analysis, as it was already the title and the subject of a book written by a psychologist (Tryon, 1939) for manual calculation. When computers developed, they permitted not only an increase in the use of cluster methods but also a development of an increasing number of techniques to detect these groups. Simultaneously a great number of problems associated with clustering techniques appeared.

A paper by Everitt (1979) developed some of them, which we will present here shortly. One of the most important criteria for a good cluster solution lies in the choice of the number of groups that should exist in a given study. Unfortunately when the classification criterion is plotted against the number of groups, in the majority of cases, no "sharp step" permits to determine the best number of classes which remains entirely subjective. Other attempts to solve this problem let it unresolved. The assessment of the validity and stability of the clusters found by different techniques poses also problems. As there are many reasons leading different analyses to arrive at different sets of clusters, it is important to show the validity of such analyses and more importantly the validity of the hypotheses lying behind them. Unfortunately there are few validity tests of these different approaches, and fewer tests of their social meaning (Cornwell, 2015). We can cite Byrne and Uprichard recent comment (2012) on these problems: "Although written in the late 1970s, actually many of the 'unresolvable problems' raised in Everitt's article are still problems today".

The emphasis on context, connections and events leads sequence analysis to abandon regression methods and to consider the research of causes as obsolete. This leads to a third problem: "could clusters be an artifact of not controlling, say, for an observed variable?" (Wu, 2000). If the problem of unobserved heterogeneity was important for event history analysis, here even observed heterogeneity leads to difficulties. While sequence analysis attempts to approach trajectories as a whole, it is only possible to introduce characteristics measured before the starting of the analyzed trajectory. Introducing characteristics measured later or time dependent ones will lead to many conceptual issues and these characteristics are very rarely incorporated. However we will see in the part on synthesis that new at-

8

tempts to combine event-history and sequence analysis may permit to solve this difficulty (Studer et al., 2016; Rossignon et al., 2016).

A fourth problem is linked to the fact that sequence analysis cannot handle censored observations, contrary to event-history analysis: it views its subject of analysis as a single unit at its completion and can only analyze fully observed trajectories, leaving aside the partially observed ones. Such a limitation will let aside incomplete trajectories and will only permit a study of the past. For example, as the age at retirement was 65 years at the time of the 3B survey, if we want to make a sequence analysis of professional life history, we could only be able to make it on people born between 1911 and 1916 only, while the survey covers people born between 1911 and 1936. Like the previous problem similar authors are trying to circumvent this difficulty, accentuating the similarity between the analyses of event duration and event sequences.

Sequence analysis permits to describe the trajectories in terms of types or classes which are considered to reflect socially meaningful patterns experienced by subjects. However, the meaning of these patterns appears to be not so clear. First as one individual is allocated to one and only one type, this leads to a very narrow classification, while we know that an individual may in fact be allocated to a great number of groups such as family, business firms or organizations, contact circles, etc. These groups are real entities while the types given by a sequence analysis may be questioned. Second what are the grounds to believe in the existence of such types? Abbott and Tsay (2000) argue that sequence methods "would find this particular regularity because people in particular friendship networks would turn up in grouping of similar fertility careers". Their argument however presumes that data on friendship networks are available simultaneously with data on the fertility history of the same people. Unfortunately as far as I know, we have no examples showing the congruence of cluster results with friendship networks.

More recently a number of authors have similarly argued that network analysis may be a valuable tool to solve a number of these problems. For example Bison (2014) proposes to convert individual sequences into network graphs. Even if this method permits "to bring out career patterns that have never previously been observed", it has important limitations. As he said, the main one

> that creates methodological and philosophical problems is the annulment of individual sequences. … Everything is (con)fused to form a different structure in which the individual trajectories disappear to make space for a 'mean' trajectory that describes the transitions between two temporally contiguous points.

If we want to remain with the fundamental description of sequence analysis given previously, this point is really confusing. However Cornwell (2015) goes further and devotes a whole chapter on *Network methods for sequence analysis*. Even if some methods used in network analysis may be useful in sequence analysis, it is important to say how the object of each approach is different. For sequence analysis as we have already said its main object is to understand a life history as a whole and to identify regularities and structures. For network analysis, as we will see in *2.4*, the main object is to understand the relations between entities (individ-

9

uals, or more general levels of collective agency) and to see how changes at each level drives the evolution at other levels. We will try to find a solution to this problem in the final synthesis in part *3* of this paper.

We will have to see now how to introduce a more complex approach.

2.3 From a contextual to a multilevel approach

While the two preceding analyses operated at a given aggregation level, contextual and multilevel analyses introduce the effects of different levels on human behavior. It derived from the hierarchical models used in biometrics and population genetics since the late 1950s (Henderson et al., 1959). Their application and generalization to the social sciences came later—in sociology with Mason et al. (1983) and in education science with Goldstein (1986).

The simplest solution for introducing a contextual dimension is to incorporate into the same model the individual and aggregate characteristics of the groups involved, as Loriaux showed in 1989. We can now grasp the difference between this approach, which uses aggregate characteristics to explain an individual behavior, and the aggregate approach, which explained an aggregate behavior by equally aggregate characteristics.

We can thus eliminate the risk of *ecological fallacy*, for the aggregate characteristic will measure a different construct from its equivalent at the individual level. It no longer acts as a substitute, but as a characteristic of the sub-population that will influence the behavior of a member of that sub-population. Simultaneously, we remove the *atomistic fallacy*, as we take into consideration the context in which the individual lives. We may ask, however, if the inclusion of the aggregate characteristics provides an entirely sufficient representation of that context: as we shall see, it will be necessary to take further steps in a fully multilevel analysis.

In fact, the use of contextual models imposes highly restrictive conditions on the formulation of the log-odds (logarithm of relative risks) as a function of characteristics. In particular the models assume that the behaviors of individuals within a group are independent of one another. In practice, the risk incurred by a member of a given group more likely depends on the risks encountered by the group's other members. Overlooking this intra-group dependence generally biases the estimates of the variances of contextual effects, generating excessively narrow confidence intervals. Likewise, these log-odds, for individuals in different groups, cannot vary freely but have restrictive constraints imposed by the model used (Loriaux, 1989; Courgeau, 2004).

In our view, the solution to this double problem lies in multilevel analysis. It aims to introduce into a single model different aggregation levels. In addition to the individual random parameter, multilevel models include random parameters for the groups at different levels identified in the analysis. The basic assumption is that these randoms are normally distributed, so that the analysis will focus only on their variances and covariances, but may introduce individual or group characteristics at different levels.

10

Multilevel analysis no longer focuses on the group, as in the analysis on aggregate data, or on the individual, as in the event-history approach. Instead, it incorporates the individual into a broader set of levels. It thus resolves the antagonism between holism and methodological individualism. As Franck noted in 1995:

> Once we have admitted the metaphysical or metadisciplinary concept of hierarchy, it no longer makes sense to choose between holism and atomism, and—as regards the social sciences—between holism and individualism.

Figure 1 summarizes the connections between two levels, depending on whether we study them separately as in event-history and sequential models, or jointly as in a multilevel model.
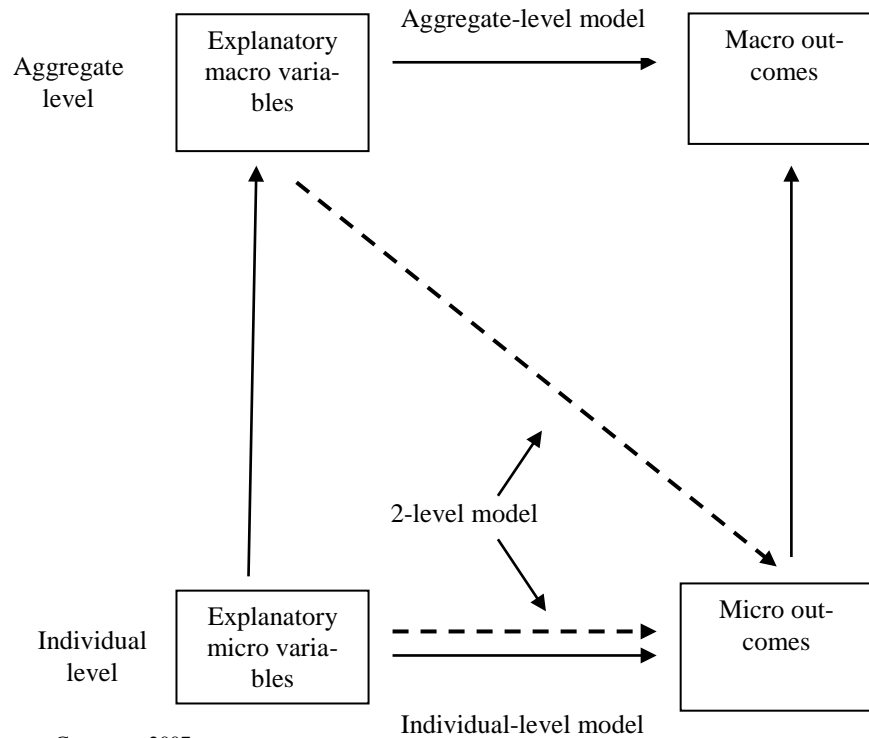
Source: Courgeau, 2007

Figure 1: Connection between different levels

This approach requires new types of surveys to capture and define the various levels to examine (Courgeau, 2007). It has been used in education science, demography, epidemiology, economics, ecology, and other disciplines to identify a multilevel structure of society. In particular, multilevel event history permits a synthesis of the two approaches (Courgeau, 2003). They are semi-parametric as before,

11

but in some cases they can take non-parametric forms, as in multilevel factor models (Goldstein, 2003).

It privileges also the use of the Bayesian paradigm in order to deal appropriately with nested or clustered data (Draper, 2008). However some other models use the frequentist paradigm. As discussed by Greenland (2000) the multilevel approach permits to unify these two paradigms, leading to empirical Bayes estimation encompassing the two approaches. About such a convergence the interested reader may read the recent book by Schweder and Hjort (2016) on statistical inference for epistemic probability understood as confidence distributions.

In its turn, the new approach encountered certain problems, which we shall now examine.

The first problem lies in the fact that it frequently uses, as group characteristics, mean values of each member's individual characteristics or even variances or co-variances. In fact there is a need to know more detailed characteristics of the aims and rules prevailing in a group and how to define them in order to explain a collective action. What are the mechanisms of social influence which permit the emergence of a collectively owned social capital in different social contexts, which "is more than the sum of the various kinds of relationship that we entertain" (Adler and Kwon, 2002)?

The second problem is that "independence among the individuals derives solely for common group membership" (Wang et al., 2013). In fact, the groups are generally more complex. For example a family, generally taken as a simple group, is in fact a more complex one where parents and children play very different and even conflicting roles. This dissymmetry of roles partly undermines the value of the family for multilevel analysis, in which we are looking for what unites group members rather than what divides them. Here again, we should take into account the interactions between group members and their changes over time in order to fully incorporate their social structure. This task will require new observation tools and new analytical methods. We will see in *2.4* how a multilevel network approach permits to avoid this problem.

The third problem lies on the difficulty to define valid groups and to use existing geographic or administrative groupings, which have little to do with their inhabitants' behaviors. Only the increase of observed existing networks in future more detailed surveys, such as those included in the *Stanford Large Network Data Set Collection*, will permit to avoid these unsatisfactory groupings.

Last, while multilevel analysis enables us to incorporate a number of known aggregation levels that constitute a society, it continues to focus only on one of these levels—an event, an individual or a group. So that this "approach assumes that links between groups are non existent" (Wang et al. 2013). Contrary to this idea it is important to take the analysis further by trying to identify the interactions that necessarily exist between the various levels. As Robert Franck wrote in 1995: "the point now is to determine how the different stages or levels connect, from top to bottom and from bottom to top". We must therefore develop a deeper study of the behaviors specific to each level but, above all, we must to try to connect the levels together in both directions.

12

We will see now how social networks permit to solve a major part of these problems.

2.4 From a network to a multilevel network approach

While earlier examples exist, research on social networks effectively began with the work of the sociologists Moreno and Jennings in the 1930's, particularly with a paper they wrote in 1938 in which they used the term 'network theory' and proposed statistics of social configurations. However almost up to the 1970's, if research teams in various social sciences worked on network analysis, no integrated cumulative effort resulted (Freeman, 2004). During the 1970s and 1980s social networks take off as a field, under the development of structural models inspired by White et al. (1976) and Freeman (1989), which examine the interdependent relationships between actors and the similar relationships between the positions of these actors in the different social networks.

This approach rests on robust mathematical foundations, which are however very different from the previous ones, as the assumption of independence of observations on individuals no more holds: network analysis argues that units are no more acting independently, but influence each other. The use of graph theory and matrix analysis is important in this field. These methods are described in detail in the book of the sociologists Wasserman and Faust (1994) as well as in the paper by the physicist Newman (2003). Many sciences, not only social, adopted this approach: information science, computer science, management, communication, engineering, economics, psychology, political science, public health, medicine, physics, sociology, geography, demography, etc.

More recently a multilevel network analysis was developed and permitted to make the link with multilevel analysis (Lazega and Snijders, 2016). While network theory is generally analyzing one given level, this approach is looking not only at the networks existing within different levels but also at the links existing between these levels. It leads to important extensions of existing models representing social structure, with networks as the dependent variable. A first kind of models tries to "reveal the interdependencies among the micro-, macro-, and meso-level networks", the meso level being here "defined between nodes of two adjacent models" (Wang et al., 2013). They generalize graph models for multiple-networks. A second kind of models "accommodate multiple partially exchangeable networks, as well as treatment effects and other covariate effects on network structure" (Sweet et al., 2013). They are often called hierarchical network models and are a generalization of multilevel models. A third kind of models "is to partition the units at all levels into groups by taking all available information into account and determining the ties among these groups" (Žiberna, 2014). It is a generalization of classical blockmodeling developed for single relations.

As for the multilevel approach, many of these models use Bayesian estimators, which have algorithmic advantages particularly for non-nested data structures, and Markov Chain Monte Carlo (MCMC) algorithms. As for multilevel models they

13

use simultaneously the frequentist paradigm and this common use may lead to more general empirical Bayes estimators (Greenland, 2000).

This approach requires surveys capable of capturing the different networks simultaneously. For example, the demographic survey on networks of relationships (Courgeau, 1972) captured the family, occupational, friendly, and community relationships of individuals living in a rural area. A network analysis of this survey by Forsé (1979) permitted to construct, from a complete diagram of acquaintance networks, "sociability" groups distinguished by social and demographic characteristics. Many other examples on more restricted networks include a biomedical research network, an isolated monastery, and so on (White et al., 1976), or on large scale networks such as those given in the *Stanford Large Network Data Set Collection* containing social, citation, collaboration, internet networks, etc. (see their use in Leskovec et al, 2009).

What new problems will this approach encounter?

A first problem lies in the surveys or on the existing data collections used in order to get the ties between individuals or between levels. They will never be exhaustive and, as many possibilities exist for their limitation, this may lead to important implications for their study. Very often in surveys, only a limited number of ties is asked and this number may vary from one survey to the other. There is also ambiguity about the qualification of these ties: the term "best friends" may have a different signification than "more frequently met" or 'more trusted" person. If a survey may ask for different kind of networks (family, friends, people at work, etc.), generally an existing data collection, like people on Facebook, will not permit this distinction. Even some persons may report more connections with popular, attractive or powerful persons than there are in reality.

A second problem lies in the fact that network clusters are generally created by the researcher rather than pre-existing to him. The way used to create them need many decisions that are difficult to pose in an entire scientific way. As Žiberna (2014) said:

> In conceptual terms, the main disadvantages are that there are no clear guidelines concerning what are the appropriate restrictions for ties between levels and what are appropriate weights for different parts of multi-relational networks, that is for level specific one-mode networks and for the two-mode networks.

Even if this citation is more linked to his blockmodeling approach, it is also true for a more general multilevel network approach. In every case decisions must be made by the researcher how to include or exclude people, merge or divide network clusters, etc. But this also permits the statistical analysis when these networks are clearly defined.

A third problem lies in the difficulty to introduce individual or network characteristics in the study of these networks. Only the use of hierarchical network models permit to introduce them. But, even in this case, there are few data sets which give measures of covariate effects on network structure (Sweet et al., 2013). These covariates may be individual, network, tie-specific, or a combination of the three.

14

A fourth problem lies in the introduction of time in these studies. Again, very few surveys permit to observe the changes occurring to networks through time. Some multi-wave surveys give at different times the structure of a network. Lazega et al. (2011) used a three wave survey in order to show that the structure of an organization remains the same regardless of the turnover of the members. However there is a need of more detailed longitudinal observations at multiple levels of analysis and of new methods in order to study the organizational mobility and relational turnover implied by the introduction if time in multilevel networks.

We can conclude this examination of the different problems and challenges encountered by network and multilevel network analysis, by the conclusion given by Lazega and Snijders in their 2016 book:

> Among the most difficult (challenges), we find combining network dynamics and multilevel analysis by providing statistical approaches to how changes at each level of collective agency drive the evolution of changes at other levels of collective agency. In all these domains, much remains to be done.

So that we can think that these problems are more a challenge for this approach than unsolvable ones.

We will now turn to the last approach presented in this paper.

2.5 An approach by agent-based decisions

Individual- or agent-based models[2] constitute an approach that differs much more from the previous ones. These models are derived from the analyses of simulation by the mathematicians Von Neumann and Ulam and the physicist Metropolis (1949). The economist Schelling (1971) suggested their use to study segregation processes and in 1972 the ecologists Botkin et al. proposed a computer model in order to predict the evolution of forest growth. During the 90's these models spread to different social sciences, taking often care not to consider each science separately but on the contrary to view them as a whole incorporating all the various social processes—demographic, economic, sociological, political, and so on. They are now largely used in many domains.

Rather than modeling specific data, this approach models theoretical ideas and is based on computer simulation. Its aim is to understand how the behavior of biological, social, or more complex systems arises from the characteristics of the individuals or more general agents making up these systems. As Billari et al. (2003) said:

---

[2] In general the ecologists prefer to speak of individual-based models, while the social scientists prefer the term agent-based, but the two denominations recover quite the same approach. We will use here the denomination of agent-based models usual for the social sciences.

15

> Different to the approach of experimental economics and other fields of behavioral science that aim to understand why specific rules are applied by humans, agent-based computational models pre-suppose rules of behavior and verify whether these micro based rules can explain macroscopic regularities.

So that this approach is bottom-up, with population-level behavior emerging from rules of behavior of autonomous individuals. These models are described in a number of books, as for example Epstein (2007) in social science or Railsback and Grimm (2012) in ecology. Many other natural and social sciences adopted it, including physics, ecology, archaeology, demography, sociology, computer science, economics, epidemiology, political science, etc.

This agent-based approach eliminates the need of empirical data on personal or social characteristics in order to explain a phenomenon as it is based on simple rules of decision followed by individuals, which can explain some real-world phenomenon. As Burch (2003) said:

> A model explains some real-world phenomenon if a) the model is appropriate to the real-world system ..., and b) if the model logically implies the phenomenon, in other words, if the phenomenon follows logically from the model as specified to fit a particular part of the real world.

Such a theoretical model cannot be validated in the same way than an empirical model, as the four previously presented approaches. About this approach Franck (2002) said: "... one has ceased to credit deduction with the power of explaining phenomena. Explaining phenomena means discovering principles which are implied by the phenomena." As it focuses on the mechanisms which drive the action of individuals or agents, it will simulate the evolution of such a population from simple rules of behavior. So that it may use game theory, complex system theory, emergence, evolutionary programming and, in order to introduce randomness, Monte Carlo methods. It may also use survey data, not in order to explain the studied phenomenon, but only to verify if the parameters used in the simulation lead to a similar observed behavior as in the survey. For example Heiland (2003) used an agent-based model in order to recover the observed distribution of migrants across different West German states and over a period of 9 years (1989-1997) from an Eastern state (Sachsen). With few theoretical assumptions about the decision to migrate the simulations indicate that heterogeneity in mobility can explain the observed decline in migration.

Again these agent-based models raised some new problems.

The first problem is that these models "are intended to represent the import and impact of individual actions on the macro-level patterns observed in a complex system" (Courgeau et al., 2016). This implies that an emergent phenomenon at the aggregate level can be entirely explained by individual behavior. However Holland (2012) said about agent-based models that they include "little provision for agent conglomerates that provide building blocks and behavior at higher level of organization". In fact a study by a multilevel model on the effects of an individual characteristic (being farmer) and the corresponding aggregate one (the proportion of farmers living in an area) on the probability of internal migration in Norway

16

(Courgeau, 2007) shows that these effects are opposite ones. It seems then diffi-cult to explain a macro-characteristic acting positively by a micro-characteristic acting negatively. Indeed, micro-level rules find hardly a link with aggregate-level rules, and I think that aggregate-level rules cannot be modeled with a micro-approach, since they transcend the behaviors of the component agents.

A second problem is that this approach is mainly bottom-up. As we have al-ready previously seen for multilevel network models it is important to consider simultaneously a top-down process from higher-level properties to lower-level en-tities. More precisely we will have to speak about a Micro-Macro link which "is the loop process by which behaviour at the individual level generates higher-level structures (bottom-up process), which feedback the lower level (top-down), some-times reinforcing the producing either directly or indirectly" (Conte et al., 2012). The bottom-up approach of agent-based models is unable to take into account such a Micro-Macro link.

A third problem lies on the validation of a given agent-based model. Such an approach is an attempt for imitation of human behavior using some well chosen generative mechanisms to produce it. It may be judged as successful when it leads to a correct reproduction of the structural characteristics of this behavior. The way to ascertain this judgment is however very far from usual tests used to verify the validity of the effects of different characteristics in the previous approaches. Such a test which can be made in natural science is less evident in social science. As Küppers and Lenhard (2005) said:

> The essential point is that (often) in the natural sciences one has a general theory about the objects and simulation models are used as instruments to generate data and to make predic-tions about the behaviour of these objects. On the other side, agent-based models are instru-ments to explore the theoretical structure of the data.

In order to see if such an exploration had been successful, we need to consider dif-ferent aspects. First, how to test that there are no other models able to explain bet-ter the observed phenomenon? Often the researcher tries different kind of models in order to permit to choose the one which give the better accord with empirical data. But this does not solve the problem, as there is an infinity of models which may predict the same empirical result as well or better. The agent-based approach gives no way to avoid this problem. Second, how to test that the chosen model gives a good fit to the observed data? Unfortunately, there are no clearly defined procedures for testing the fit of the simulation models, like goodness of fit proce-dures or tests of significance for the previous approaches. We can conclude that there are no clear verification and validation procedures for agent-based models in population science.

In consequence we will have to see in the next section how to try to overcome these problems, as well as those encountered with the four previous approaches.

17

3. Towards a synthesis

As we have already presented and criticized five different main approaches used nowadays in social sciences, we will have now to see if we can give a more synthetic view of them.

Let us first consider the two main concepts without which no population science would be possible.

The first one is the creation of an abstract fictious individual, whom we can call a *statistical individual* as distinct from an *observed individual*. As Aristotle (330 BC) said: "individual cases are so infinitely various that no systematic knowledge of them is possible", Graunt (1662) was the first to introduce the possibility of a population science letting aside the observed individual and using statistics on a few number of characteristics, leading to a statistical individual. As Courgeau wrote in 2012:

> Under this scenario, two observed individuals, with identical characteristics, will certainly have different chances of experiencing a given event, for they will have an infinity of other characteristics that can influence the outcome. By contrast, two statistical individuals, seen as units of a repeated random draw, subjected to the same sampling conditions and possessing the same characteristics, will have the same probability of experiencing the event.

The essential assumption permitting to use the theory of probability in this case is that of *exchangeability*[3] (de Finetti, 1937): *n* trials will be said to be exchangeable if the joint probability distribution is invariant for all permutations of the *n* units. We will use it here for the residuals given the explanatory characteristics measured on these individuals.

The second concept is the notion of a *statistical network*, different from the observed ones: it appeared more recently, for example with the work of Coleman in 1958. While *observed networks* may be as diverse as the infinite kind of ties existing between observed individuals, *statistical networks* may be more precisely defined with the use of statistics on ties and the choice of criteria to circumscribe them. Again the essential assumption permitting the use of the theory of probability is that, given the explanatory characteristics introduced at each level, the residuals are assumed to be exchangeable.

It is interesting here to compare these two concepts with the contexts proposed by Billari (2015) to explain population change: the micro- and the macro-level contexts. In fact he clearly recognized at the basis of micro-level context the abstract concept of *statistical individual*, the same that we propose here. However for macro-level context he only proposes to see how "population patterns re-emerge from action and interaction of individuals", without recognizing the abstract concept at the basis of this interaction: the *statistical network*, which permits to flesh out this macro-analysis. For example we have already seen how multilevel analysis permits to reconcile the macro- and micro-level results.

---

[3] In this first paper on this topic de Finetti called it *equivalence*.

18

Once these two main concepts defined, we can see that the study of event duration and the study of event sequences are directly connected to the same concept of *statistical individual*. Even if their approach of this individual is different, as we have already seen, they can be considered as two complementary ways to study him. In addition some more recent papers, as those presented in this conference by Studer et al. (2016) or Rossignon et al. (2016), combines the advantages of the two approaches modeling "the relationships between time varying covariates and trajectories specified as processes outcomes that unfold over time". The definition given by Courgeau and Lelièvre (1997) of event history analysis appears to be also valid for sequence analysis: "Throughout his or her life an individual follows a complex itinerary, which at a given moment depends on the life course to date and on the information acquired in the past". The itinerary is followed event after event, in the first analysis, and with more complex sequences of events in the second one.

Similarly we can see that the contextual, multilevel and network multilevel approaches are simultaneously connected to the same concept of *statistical network*. They appear as complementary in its study. We can say that contextual and multilevel analysis focuses on attributes while network multilevel analysis focuses on relations, combining the different levels.

It may also be interesting to see that contextual and multilevel analysis may be seen as complementary of event history analysis, introducing the effects of network membership on individual behavior. Similarly network multilevel analysis may also be seen as complementary to sequence analysis. This proximity may explain why Cornwell (2015) tries to introduce network methods in sequence analysis: however sequence methods remain at the statistical individual level, while multilevel networks methods introduce statistical networks.

The different problems encountered when using one of these four approaches may largely disappear when considering simultaneously the statistical individual and the statistical network under a more general *biographical multilevel network analysis*. As we already said such an approach is able to avoid the risks of atomistic or of ecological fallacy through the use of a synthesis of holism and individualism. It may also avoid the problems linked to the choice between Bayesian or frequentist probability through the use of a more general compromise on confidence distributions (Schweder and Hjort, 2016), which opens to a better statistical inference. It permits to answer to some problems posed by unobserved heterogeneity, while introducing networks which permit to have a better understanding of human behavior. We can also think that a number of problems encountered by sequence analysis (metric used, cluster analysis and artifacts) may be solved by undertaking more complex surveys on social networks, which may permit to replace theoretical clusters by real networks of individuals linked together by existing social forces. Similarly the main problems encountered by multilevel analysis may largely be solved by multilevel network analysis, such as: the use of a Multilevel Social Influence (MSI) model (Agneessens and Koskinen, 2016) to explain the emergence of a social capital; the use of Exponential Random Graph Models (ERGM) to show that within-level network structure are interdependent with network structures of other levels (Wang et al., 2016); etc. Last we think that the problems re-

cently posed by multilevel network analysis are more a challenge for future research in this field, than unsolvable problems. For example such an analysis will reach its full potential when longitudinal observations at multiple levels of analysis will be available, by providing a combination of an event history of networks with a multilevel analysis (Lazega and Snijders, 2016).

The situation is more complex for agent-based approach. While it apparently resembles event-history approach in its focus on individual behavior alone, it seeks however to explain collective behavior with the aid of individual behavior. This gives it some affinity with the multilevel network approach. The main question is: how to generate the macroscopic regularity from the bottom-up, using simple local rules? The difficulties encountered with such an approach are clearly described in Conte et al. (2012):

> First, how to find out the simple local rules? How to avoid *ad hoc* and arbitrary explanations? As already observed[4], one criterion has often been used, i.e., choose the conditions that are sufficient to generate a given effect. However, this leads to a great deal of alternative options, all of which are to some extent arbitrary.

As we have already shown, in social science we cannot obtain the macro-level patterns by simply aggregating the micro-level outcomes, so that local rules are not sufficient to explain a complex social behavior. It is then necessary to introduce theories of decision making to get more valuable models: however the number of options for modeling decision making is almost infinite (Klabunde and Willekens, 2016). As the choice of a decision theory is driven by the researcher background, an economist, a demographer, a geographer, a psychologist, etc., may reach quite different results for the same studied phenomenon.

In our view, further work is needed to go over these contradictions and place agent-based analysis in a broader setting and a more explicit theory-founded model.

4 Conclusion

By restricting ourselves to defining a scientific method solely by its methods, we condemn ourselves to taking a partial view of the core scientific approach. We need to set up a more robust research program for demography and, more generally, the social sciences—a program that converges with the now well established program of the physical and biological sciences. The source for this program lies in Bacon's work (1620):

> There are and can be only two ways of searching into and discovering truth. The one flies from the senses and particulars to the most general axioms, and from these principles, the truth of which it takes for settled and immovable, proceeds to judgment and to the discovery

---

[4] See Conte (2009) p. 29.

20

> of middle axioms. And this way is now in fashion. The other derives from the senses and particulars, rising by a gradual and unbroken ascent, so that it arrives at the most general axioms last of all. This is the true way, but as yet untried.

Bacon calls the second approach *induction*, not in the meaning later given to the term by the empiricist tradition of Hume and Popper—i.e., the generalization of observations—but in the sense of the search for the structure of observed phenomena. That is how Galileo, Newton, Graunt, Einstein, Darwin, and others developed their approach to the study of phenomena—whether physical or social.

It is important for the social sciences to start with the observation and measurement of facts, for this measurement, far from being secondary, makes it possible to assess the "potentialities" of a social fact (Courgeau, 2013). Next, instead of relying on often arbitrary hypotheses, like in agent-based models, the modeling of observed phenomena should follow the method recommended by Bacon by analyzing the interactions between the networks created by people and seeking their structure (Franck, 2002; Courgeau et al., 2016).

Even if one can think that each individual has an unlimited and unknowable number of characteristics with his own freedom of choice, social science has to see that he is born in a given society with its rules and laws, which restrain his freedom, that he is submitted to biological laws, which are the same for all humans. So that a social science can exist which takes into account only a limited number of characters and which is based on a number of concepts without which the properties of these characters would be inconceivable or impossible (Franck, 2002).

A final point: We have often viewed the social sciences here as a whole to which certain approaches applied and not others. We must now consider that it is not by erasing the boundaries between disciplines that we can improve our knowledge (Franck, 1999). The boundaries are real, for each discipline endeavors to analyze different properties of human societies. However, we think that it is possible to construct a new formal object that can explain certain properties of human societies—an object that encompasses existing disciplines and allows their synthesis.

References

Aalen, O.O. (1975). *Statistical inference for a family of counting processes*. PhD thesis, Berkeley: University of California.

Aalen, O.O., Borgan Ø, Keiding, N., Thorman, J. (1980). Interaction between life history events. Nonparametric analysis for prospective and retrospective data in the presence of censoring. *Scandinavian Journal of Statistics*, 7, 161-171.

Aalen, O.O., Borgan, Ø, Gjessing, H, K (2008). *Survival and event history analysis. A process point of view*. New York: Springer.

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

Abbott, A. (1984). Event sequence and event duration: colligation and measurement. *Historical methods*, 17 (4), 192-204.

21

Abbott, A. (1995). Sequence analysis: new methods for old ideas. *Annual Review of sociology*, 21, 93-113.

Abbott, A., Tsay, A. (2000). Sequence analysis and optimal matching methods in sociology: review and prospect. *Sociological Methods and Research*, 29, 3-33.

Agneesens, F., Koskinen, J. (2016). Modeling individual outcomes using a Multilevel Social Influence (MSI) model: individual versus team effects of trust on job satisfaction in an organisational context. In T. Lazega, T. Snijders (eds) (2016). *Multilevel network analysis for the social sciences,* Methodos Series, vol.12, pages 81-105, Dordrecht / Boston / London: Springer

Alder, P.S., Kwon, S.-W. (2002). Social capital: prospects for a new concept. *Academy of Management*, 27 (1), 17-40.

Andersen, P.K., Borgan, Ø, Gill, R.D., Keiding, N. (1993). *Statistical models based on counting processes*. New York: Springer-Verlag.

Aristotle (around 330 B.C.). *Rhetoric*. The Internet Classic Archive: translated by Roberts, W.R.

Bacon, F. (1620). *Novum Organum*, London: J. Bill.

Billari, F. (2015). Integrating macro- and micro-level approaches in the explanation of population change, *Population Studies*, 69 (S1), S11-S20.

Billari, F., Prskawetz, A., ed. (2003). *Agent-based computational demography: using simulation to improve our understanding of demographic behaviour.* Heidelberg, New York: Physica-Verlag.

Bison, I. (2009). OM matters: the interaction effects between *Indel* and substitution costs. *Methodological Innovations Online*, 4 '2), 53-67.

Bison, I. (2014). Sequences as network: an attempt to apply network analysis to sequence analysis. In P. Blanchard, F. Bühlman, J.-A. Gauthier (eds), *Advances in sequence analysis: theory, method, applications,* pages 231-248, Dordrecht: Springer.

Blanchard, P., Bühlman, F., Gauthier, J.-A. (eds) (2014). *Advances in sequence analysis: theory, method, applications*. Dordrecht: Springer.

Bolano, D. (2014). Hidden Markov models: an approach to sequence analysis in population studies. *Annual meeting of the Population Association of America*, presented in Poster Session, paa2014.princeton.edu/papers/141879, 22 p.

Botkin, D.B., Janak, J.F., Wallis, J.R. (1972). Some ecological consequences of a computer model of forest growth. *The Journal of ecology*, 60, 849-872.

Breiger, R.L. (1974). The duality of persons and groups. *Social Forces*, 53, 181-190.

Bretagnole, J. & Huber-Carol, C. (1988), Effects of omitting covariates in Cox's model for survival data, *Scandinavian Journal of Statistics*, 15, 125-138.

Burch, T.K. (2003). Data, models, theory and reality: the structure of demographic knowledge. In F.C. Billari, A. Prskawetz (eds), .*Agent-based computational demography*, pages 19-40, Heidelberg New York: Physica Verlag.

Byrne, D., Uprichard E. (2012). Introduction; (useful) hey texts. In D. Byrne, E. Uprichard (eds), *Cluster analysis*, vol. 2, pages vii-xii, London: Sage Publications Ltd.

Coleman, J.S. (1958). Relational analysis: the study of social organizations with survey methods. *Human Organization*, 17, 28-36.

Conte, R. (2009). From simulation to theory (and backward). In F. Squazzoni (ed), *Epistemological aspects of computer simulation in the social sciences*, pages 29-47, Berlin Heidelberg: Springer.

Conte, R., Gilbert, N., Bonelli, G., Cioffi-Revilla, C., Deffuant, G., Kertesz, J., Loreto, V., Moat, S., Nadal, J.-P., Sanchez, A., Nowak, A., Flache, A., San Miguel, M., Helbing, D. (2012). Manifesto of computational social science. *The European Physical Journal Special Topics*, 214, 325-346.

Cornwell, B. (2015). *Social sequence analysis*. New York: Cambridge University Press.

Courgeau, D. (1972). Les réseaux de relations entre personnes. Etude d'un milieu rural. *Population,* 27 (4-5), 641-684.

22

Courgeau, D. (1999). L'enquête « Triple biographie : familiale, professionnelle et migratoire ». In Groupe de réflexion sur l'approche biographique, *Biographies d'enquêtes*, pages 59-74, Paris: INED.

Courgeau, D., ed. (2003). *Methodology and epistemology of multilevel analysis: approaches from different social sciences.* Methodos Series, vol. 2, Dordrecht / Boston / London: Kluwer Academic Publishers.

Courgeau, D. (2007). *Multilevel synthesis: from the group to the individual.* Series on demographic methods and population analysis, vol. 18, Dordrecht: Springer.

Courgeau, D. (2012). *Probability and social science. Methodological relationships between the two approaches.* Methodos Series, vol. 10, Dordrecht / Boston / London: Kluwer Academic Publishers.

Courgeau, D. (2013). La mesure dans les sciences de la population. *Cahiers Philosophiques*, 135 (4), 51-74.

Courgeau, D., Bijak, J., Franck, R., Silverman, E. (2016). Model-based demography: towards a research agenda. In A. Grow, J. Van Bavel, (eds), *Agent-based modelling in population studies,* Series on demographic methods and population analysis, vol. 41, Dordrecht: Springer.

Courgeau, D., Lelièvre, E. (1992). *Event history analysis in demography.* Oxford: Clarendon Press.

Courgeau, D., Lelièvre, E. (1997). Changing paradigm in demography. *Population, An English Selection*, 9, 1-10.

Cox, D.R. (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society*, 34 (2), 187-220.

Cox, D.R. & Oakes, D. (1984). *Analysis of survival data*. London, New York: Chapman and Hall.

de Finetti, B. (1937). La prévision: ses lois logiques, ses sources subjectives. *Annales de l'Institut Henri Poincaré*, 7 (1), 1-68.

Degenne, A., Forsé, M. (1999). *Introducing social networks*. London: Sage.

Doob, J.L. (1953). *Stochastic processes*. New York: Wiley.

Durkheim, E. (1895). *Les règles de la méthode sociologique*. Paris: Alcan.

Epstein, J. (2007). *Generative social science: studies in agent-based computational modelling*. Princeton: Princeton University Press.

Everitt, B.S. (1979). Unresolved problems in cluster analysis. *Biometrics*, 35, 169-181.

Fararo, T.J., Doreian, P. (1984). Tripartite structural analysis: generalising the Beiger-Wilson formalism. *Social Networks*, 6, 141-175.

Forsé, M. (1979). Les réseaux de sociabilité dans un village. *Population*, 36 (6), 1141-1162.

Franck, R. (1995). Mosaïques, machines, organismes et sociétés. *Revue Philosophique de Louvain*, 93 (1-2), 67-81.

Franck, R. (1999). La pluralité des disciplines, l'unité du savoir, et les connaissances ordinaires. *Sociologie et Sociétés*, 1, 129-142.

Franck, R., Ed. (2002). *The explanatory power of models: bridging the gap between empirical and theoretical research in the social sciences.* Boston / Dordrecht / London: Kluwer Academic Publishers.

Freeman, L.C. (1989). Social networks and the structure experiment. In L. C. Freeman, D. R. White and A. K. Romney (eds), *Research Methods in Social Network Analysis*, pages 11-40, Fairfax: George Mason University Press.

Freeman, L.C. (2004). *The development of social network analysis: a study in the sociology of science*. Vancouver: Booksurge Publishing.

Goldstein, H. (1986). Multilevel mixed linear model analysis using iterative generalized least-squares. *Biometrika*, 73, 43-56.

Goldstein, H. (2003). *Multilevel statistical models.* London: Edward Arnold.

Granger, G.-G. (1994). *Formes, opérations, objets*. Paris: Librairie Philosophique Vrin.

Graunt, J. (1662). *Natural and political observations mentioned in a following index, and made upon the bills of mortality.* London; Tho. Roycroft.

23

Greenland, S. (2000). Principles of multilevel modelling. *International Journal of Epidemiology*, 39, 156-167.

Heiland, F. (2003). The collapse of the Berlin wall: simulating state-level East to West German migration patterns. In F.C. Billari, A. Prskawetz (eds), *Agent-based computational demography: using simulation to improve our understanding of demographic behaviour*, pages 73-96, Heidelberg, New York: Physica-Verlag,.

Henderson, D.F., Kempthorne, O., Searle, S.R., Von Krosigk, C.M. (1959). The estimation of environmental and genetics trends from records subject to culling. *Biometrics*, 15, 192-218.

Holland, J.H. (2012). *Signals and boundaries. Building blocks for complex adaptive systems*. Cambridge London: The MIT Press.

Holland, P.W., Leinhardt, S. (1976). Local structure in social network. *Sociological Methodology* 7, 1-45.

Ibrahim, J.G., Chen, M.-H., Sinha, D. (2001). *Bayesian Survival analysis*. New York: Springer-Verlag.

Jeffreys, H. (1939). *Theory of probability*. Oxford: Clarendon Press.

Kalbfleisch, J.D., Prentice, R.L. (1980). *The statistical analysis of failure time data*. New York, Chichester, Brisbane, Toronto: John Wiley and Sons.

Klabunde, A., Willekens, F. (2016). Decision-making in agent-based models of migration: state of the art and challenges. *European Journal of Population*, 32, 73-97.

Küppers, G., Lenhard, J. (2005). Validation of simulation: patterns in the social and natural science. *Journal of Artificial Societies and Social Simulation*, 8 (4), 1-13.

Lazarsfeld, P.F., Menzel, H. (1961). On the relation between individual and collective properties. In R. Boudon (ed.), *Complex organizations: a sociological reader*, pages 172-189, Chicago: University of Chicago Press.

Lazega, T., Sapulete, S., Mounier, L. (2011). Structural stability regardless of membership turnover? The added value of blockmodelling in the analysis of network evolution. *Quality & Quantity*, 45, 129-144.

Lazega, T., Snijders, T., eds., (2016). *Multilevel network analysis for the social sciences*. Methodos Series, vol.12, Dordrecht / Boston / London: Springer.

Leskovec, J., Lang, K., Dasgupta, A., Mahoney, M. (2009). Community structure in large networks: natural cluster sizes and the absence of large well-defined clusters. *Internet Mathematics*, 6 (1), 29-123.

Levenshtein, V.I. (1966). Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10, 707-710.

Levitt, M. (1969).Detailed molecular model for transfer ribonucleic acid. *Nature*, 224 (5221), 759-763.

Liu, J.S., Logvinenko, T. (2003). Bayesian methods in biological sequence analysis. In D.J. Balding, M. Bishop, C. Cannings (eds), *Handbook of statistical genetics*, 3rd Edition, pages 67-96, Chichester: John Wiley & Sons, Ltd.

Loriaux, M. (1989). L'analyse contextuelle : renouveau théorique ou impasse méthodologique? In J. Duchêne, G. Wunsch, E. Vilquin (eds), *Explanation in the social sciences. The search for causes in demography*, *Chaire Quetelet '87*, pp. 333-368, Bruxelles: Editions Ciaco.

Mason, W.M., Wong G.W., and Entwistle B. (1983), Contextual analysis through the multilevel linear model, in *Sociological Methodology 1983-1984*, Leinhart S. ed., San Francisco: Jossey-Bass, pp. 72-103.

Metropolis, N., Ulam, S. (1949). The Monte Carlo method. *Journal of the American Statistical Association*, 44 (247), 335-341.

Mill, J.S. (1843). *A system of logic, ratiocinate and inductive, being a connected view of the principles of evidence, and the methods of scientific investigation*, vol. I, London: Harrison.

Moreno, J.L., Jennings, H.H. (1938). Statistics of social configurations. *Sociometry*, 1 (3/4), 342-372.

Newman, M.E. (2003). The structure and function of complex networks. *Siam Review*, 45 (2), 167-256.

24

Petty, W. (1690). *Political arithmetick*. London: Robert Clavel & Hen. Mortlock.

Railsback, S.F., Grimm, V. (2012). *Agent-based and individual-based modelling*. Princeton: Princeton University Press.

Robette, N., Bry, X. (2012). Harpoon or bait: a comparison of various metrics to fish for sequence patterns. *Bulletin de Méthodologie Sociologique*, 116 (1), 15-24.

Robette, N., Lelièvre, E., Bry, X. (2012). La transmission des trajectoires d'activité : telles mères, telles filles. In C. Bonvalet, E. Lelièvre (eds), *De la famille à l'entourage*, pages 395-418, Ined, Paris.

Robinson, W.S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review*, 15, 351-357.

Rossignon, F., Studer, M., Gauthier, J.-A., Le Goff, J.-M. (2016). Childhood family structure and home-leaving. A combination of survival and sequence analysis. *LaCOSA II Conference*, 11 pages, Lausanne.

Rozenblat, C., Melançon, G. (2013). *Methods for multivariate analysis and visualization of geographical networks*. Methodos Series, vol. 11, Dordrecht / Boston / London: Springer.

Schelling, T. (1971). Dynamic models of segregation. *Journal of mathematical sociology*, 1, 143-171.

Schweder, T., Hjort, N.L. (2016). *Confidence, likelihood, probability: statistical inference with confidence distributions*. Cambridge: Cambridge University Press.

Studer, M., Struffolino, E., Fasang, A. (2016). A new tool for old questions: the sequence-analysis multistate model to study relationships between time-varying covariates and trajectories. *LaCOSA II Conference*, 31 pages, Lausanne.

Sweet, T.M., Thomas, A.C., Junker, B.W. (2013). Hierarchical network models for education research: hierarchical latent space models. *Journal of Educational and Behavioral Statistics*, 38, 295-318.

Trussell, J. (1992). Introduction. In J. Trussell, R. Hankinson, Tilton, J. (eds), *Demographic applications of event history analysis*, pages 1-7, Oxford: Clarendon Press.

Trussell, J. Richards, T. (1985), Correcting for unmeasured heterogeneity in hazard models using the Heckman-Singer procedure, in N. Tuma (ed.), *Sociological Methodology*, pages 242-276. Jossey-Bass, San Francisco (CA).

Tryon, R. (1939). *Cluster analysis*.Ann Arbor:Edwards Brothers

Vaupel, J.W., Manton, K.G., Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*, 16, 439-454.

Ville, J.A. (1939). *Etude critique de la notion de collectif*. Paris: Gauthier-Villars.

Wang; P., Robins, G., Pattison, P. Lazega, E. (2013). Exponential random graph models for multilevel networks. *Social Networks*, 35 (1), 96-115.

Wang, P., Robins, G., Matous, P. (2016). Multilevel network analysis using ERGM and its extensions. In T. Lazega, T. Snijders (eds) (2016). *Multilevel network analysis for the social sciences,* Methodos Series, vol.12, pages 125-143, Dordrecht / Boston / London: Springer.

Wasserman, S., Faust, K. (1994). *Social network analysis: methods, and applications*. Cambridge: Cambridge University Press.

Wasserman, S., Iacobucci, D. (1991). Statistical modelling of one-mode and two-mode networks: simultaneous analysis of graphs and bipartite graphs. *British Journal of Mathematical and Statistical Psychology*. 44, 13-44.

White, H.C., Boorman, S.A., Breiger, R.L. (1976). Social structure from multiple networks. I Blockmodels of roles and positions. *American Journal of Sociology*, 81 (4), 730-780.

Žiberna, A. (2014). Blockmodeling of multilevel networks. *Social networks*, 39 (1), 46-61.

Mapping the field of sequence analysis

Philippe Blanchard, University of Warwick

Updated extended summary, 27.05.2016

To those who make history, writing it may seem to be unnecessary. When it comes to science, some who make it believe they are to advance knowledge, not contemplate the story of how they are able to do so. However, this communication is taking the opposite stance. As other cases of history of science show, scholars at a given time are prone to see their field as an ideal, global college of individual minds, interacting freely and efficiently, and their field's story as a linear upward curve. This view bears the risk of forgetting about multiple factors, some structural, some contingent, heteronomous to intellectual life, that influence research: the endogenous sociological logic of research communities; personal affinities with topics and tools; journals' editorial lines; existing paradigm wars along which one has to align oneself; the agenda of funding organisations; the availability of data and software; and the inertia of individual and collective agendas, even if logical reasons would push to innovation. Naturally, researchers constantly update their knowledge and assessment of other works in the field, contributing to a Darwinian-like selection of best ideas. But this selection takes place in settings that are constrained technically, economically, sociologically, institutionally, politically or ideologically. These settings have the capacity to make scholars forget how multicausal, and sometimes arbitrary, the directions they take can be. There is no reason to imagine, as gentle the community of sequence analysts may be, that it would escape these contingencies.

Looking back at the progress made collectively since the introduction of sequence analysis (SA) in the social sciences enables, not only to feel good about the achievements, but also to: 1. Detect trends and turning points; 2. Spot the factors of methodological change, both internal or external to the intellectual sphere; 3. Understand why some paths have been taken, and others sealed off; and 4. Encourage alternative options. This communication aims to observe SA as a research programme (Lakatos and Musgrave 1970), taking the view of an outsider to the community of sequence analysts, yet using an insider's knowledge to decipher it. To this purpose, it finds inspiration in pioneer theories of how scientific innovation sometimes happens, and sometimes not: K. Popper's (1934) assertion of the falsificationist internal logic of scientific discovery; T. Kuhn's (1962) more sociological and historical view on the contingent factors that trigger scientific revolutions; I. Lakatos and A. Musgrave's (1970) position, centred on "research programmes", intermediate between all-logical and all-contingent theories; and J.-M. Berthelot's (1996) demonstration of the irreducible but fruitful pluralism of social scientific demonstrations.

Here SA is defined as a set of concepts and tools designed (or redesigned from other disciplines) for the study of series of social events, or states along social trajectories. This method, or approach, has been specific in several respects: the nodal role of

2

A. Abbott, the most cited author in the field, although he played the role of an intermediary with other scientific fields as much as the one of a creator (Gauthier, Buhlmann and Blanchard 2014); the importation and adaptation of core tools from computer science and genetics; the dependence on the availability of adequate sequential data, computing power and software developments; the diversity of geographic, institutional and disciplinary loci of development; the laborious, still challenging competition with established longitudinal statistical methods (time series, duration models, Markov models, timed regression models); and, consequence of the latter, the fact that core elements of the method have been and are still disputed, in a constructive manner, although sometimes radically (Abbott and Tsay 2000; Robette, Bry and Lelièvre 2015). These conditions of development came to be met and articulated with each other only in the 2000s. This, besides strictly internal (intellectual) factors, explains the near-exponential rise of publications from this period onward, after a more stagnant time, marked nonetheless by the works of A. Abbott and his colleagues.

Following a short track of previous literature reviews (Aisenbrey and Fasang 2010; Blanchard 2011; Cornwell 2015; Lesnard 2006), this communication will be based on an extensive review of publications making use of SA in the social sciences, including: empirical applications; methodological and software developments; literature reviews; introductions and textbooks; comparisons with other approaches to longitudinal data; as well as critics and alternative propositions. Taking a mapping approach to the history and sociology of science (Blanchard, Rihoux and Alamos-Concha 2016), I will apply multivariate data analysis and network analysis to this corpus in order to reveal trends and events, convergences and divergences, imitations and distinctions, and other aspects of the historical dynamic of innovation in the field. This will enable to: contrast distinct visions of SA; contrast mainstream case studies and more marginal ones; trace the evolution of competing data and methods over time; spot converging and diverging methodological options, schools and authors.

Bibliography

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

Abbott, A. (2000). Reply to Levine and Wu. *Sociological Methods and Research*, 29 (1), 65-76.

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

Aisenbrey, S. and Anette E. Fasang (2010) New Life for Old Ideas: The "Second Wave" of Sequence Analysis. Bringing the "Course" Back Into the Life Course. *Sociological Methods and Research*, 38(3), 420–462.

Berthelot, J.-M. (1990). *L'intelligence du social : le pluralisme explicatif en sociologie* [*Understanding Society: Pluralism in Sociological Explanation*]. Paris: Presses universitaires de France

3

Blanchard, P. 2011. Sequence Analysis for Political Science. Working Papers of the Committee on Concepts and Methods, International Political Science Association

Blanchard, P. (forthcoming) Les vicissitudes de l'innovation méthodologique. "Validity, falsifiability, parsimony, consistency, precision, and so on" [The challenges of methodological innovation: "Validity, falsifiability, parsimony, consistency, precision, and so on"], *in* Morgan Jouvenet and Didier Demazière, eds, *La sociologie d'A. Abbott*. Paris : EHESS.

Blanchard, P., Rihoux B. and Alamos Concha P. (2016 - Online first). Comprehensively Mapping Political Science Methods: An Instructors' Survey. *International Journal of Social Research Methodology*.

Cornwell, B. 2015. *Social Sequence Analysis*. Methods and Applications. New York: Cambridge Univerity Press

Elzinga, C. H. (2015). Comment: On the Association between Sequences in GIMSA. *Sociological Methodology,* 45, 45-51.

Gauthier, J.-A., Bühlmann F. and Blanchard P. (2014). Introduction: Sequence Analysis in 2014 1-17. In P. Blanchard, F. Bühlmann and J.-A. Gauthier (Eds.), *Advances in Sequence Analysis: Theory, Methods, Applications*. London: Springer.

Kuhn, T. (1970). *The structure of scientific revolutions*. Chicago: University Press of chicago.

Lesnard, L. (2006). Optimal matching and social sciences. *Document de travail du CREST 2006-01*, INSEE, Paris.

Levine, J. (2000). But What Have You Done for Us Lately? Commentary on Abbott and Tsay. *Sociological Methods and Research*, 29 (19), 34-40.

Popper, K. (1959 [1934]). *The Logic of Scientific Discovery*. London: Hutchinson

Robette, N., Bry X., and Lelièvre E. (2015). A "Global Interdependence" Approach to Multidimensional Sequence Analysis. *Sociological Methodology,* 45, 1-44.

From 07.00 to 22.00: a dual-earner typical day in Italy.

Old questions and new evidences from social sequence analysis.
Ivano Bison & Alessandro Scalcon
Dep. of Sociology and Social Research,
University of Trento,
via Verdi, 26, I-38100 – TRENTO (Italy).
ivano.bison@unitn.it

Abstract

The paper analyses the daily activities of dual-earner couples in Italy. The goal is to discover how Italian dual-earner couples organize their daily activities (sleep, personal care, work, moving, housework, free time), during a typical work day from Monday to Friday. The analysis, carried out on data from the 2008 Italian Census on Time Use (the last one available), involves all the 873 couples that filled in their diaries on the same day.

Using the binary index (Bison, 2006, 2011a, 2011b, 2014; Bison, Rettore, Schizzerotto 2010; Franzosi, Bison 2010), we conduct a 'multichannel' analyses on the dual-earner couple's activities from 7.00 to 22.00.

Quite strong relations with socio-demo-geographic conditions emerge from these analyses. Hence there is a strong relation of time packaging and the time spent on the various activities according to where couples live with respect to both geographical area (North, Centre and South & Islands) and the size of the town (metropolitan, more than 50,000 residents, from 10,000 to 50,000, fewer than 10,000).

Strong relations also emerge with the level of education, the social class and the occupational sector of Him and Her. Relations with the presence of children are observed mainly at the beginning and the end of the day. At the same time, the different time packaging profile of the dual-earner couple that emerges from the k-means cluster analysis seems to have a direct effect on His and Her level of satisfaction.

All the preliminary analyses seem to confirm the idea that dual-earner couples package their life time mainly in accordance with their jobs. Moreover, the analyses show that this time packaging changes in relation to the kind of job (social class) and the occupational sector. Secondly, the time spent on each activity changes according to the level of education of Him and Her, but there is an additional effect due to the social and cultural level of the area where they live.

1. Introduction

Since the 1960s and 1970s, the daily use of time has radically changed in industrialized countries. For instance, the average work time per person has declined while the leisure time has generally increased (Gershuny, 2000); high-skilled workers have progressively worked longer hours compared to unskilled and low-status ones, inverting the traditional work hours/social class gradient (Warren 2003; Lesnard et al., 2009); the overall gender gap in non-paid work has being partially filled – even if to a lesser extent than expected (Gershuny et al., 1988) – as a consequence of the increase of women's participation in the labor market (Hook 2006; Raley et al., 2012). Factual changes in daily behaviors follow new gender values, like the diffusion of more 'career-oriented' attitudes among women (Hakim, 2003) and the parallel rise of more conciliatory and intimate fatherhood (Naldini et al., 2011).

Dual-earner couples experience strong time constraints and need constantly to negotiate their time use by dealing with the family's time scarcity (Saraceno, 2012). They may try to be more or less aligned during the day, according to their 'production' and 'consumption' complementarity strategy (Mansour et al., 2013). This happens especially on weekdays, where the combined

1

exposure of time pressure increases due to the work time of each spouse. Thus, the spouses are required to 'find time' to spend *with* the family as well as *for* the family and these collective needs are followed by individual and private ones as well. The two individual careers have to coexist with a third one – that of 'family life' – which today seems to be equally important for both the spouses (Levner, 2000 in Haddock et al., 2006). It is for this reason that the daily work-family balance becomes a crucial dimension for the quality of life among dual-earner couples.

A dual-earner couple's daily strategy is obviously constrained to the 'work' activity, which represents a totalizing and exclusive time. Regarding workdays, we may say that free time and household care – as well as travel/moving, sleep and personal care – must be primarily managed according to work times by filling the gaps in working schedules. This is why «not only the total amount of work time but also its scheduling are very significant» in understanding dual-earners' entire workdays (Lesnard et al., 2009:3). The non-work activities, in fact, are expected to be affected by work in terms of both quantity and timing.

According to their work-family strategies and time constraints, couples may prefer higher or lower levels of synchronization in their daily working schedules. Nock and Kingston (1984) found that when both the spouses work, the time for family may be more desynchronized than that of single-earner couples. This happens because their working schedules may not overlap, thus reducing the time that those spouses could spend together on other joint activities (Nock et al., 1984). Moreover, it has been argued that a certain degree of 'off-scheduling' would be preferred by some particular couples, especially those with children, and that – although fathers seem to be less used to altering their hours of labor force participation in favor of childcare (Raley et al., 2012) – both the spouses might want to reduce the overlap of their working schedules in order to maximize the potential time for childcare (van Klaveren et al., 2011).

In general, the literature shows that dual-earner couples that are more desynchronized in working schedules share household and family duties more equally, and that this could be a desirable solution for them (Presser, 1994; Chenu et al., 2002; Lesnard, 2008; Naldini et al., 2011). At the same time, scholars have pointed out that dual-earner couples always seem to prefer a certain level of synchronization in their working schedules – with an increase in the overlap between them – in order to maximize a shared conjugal leisure time (Hamermesh, 2002; Lesnard, 2008).

Here, the point is that, independently of their most desired work-family solution, dual-earner couples must deal with the rigidity of the time constraints imposed by working schedules. On the basis of their 'time sovereignty' over their working schedules, spouses could better align their preferences with factual time-use behaviors. However, the literature shows that this capability is strongly related to the overall job commitments of the spouses, and these are associated with their individual occupational class and the more general social ladder (Warren, 2003). Thus, basically, workday schedules and the constraints for the couples' daily activities is not random (Warren, 2003).

Several papers have pointed out the externalities of such a class-related 'wealth' of time among couples (Warren, 2003), analyzing the impact of different socio-economic conditions on daily working time. Indeed, the higher the spouses' social position, the greater their bargaining power with employers for the purpose of daily time and schedules management (Lesnard, 2008; Warren 2003). However, high-skilled and high-status jobs are likely to require long work hours (Gershuny, 2000; Warren, 2003; Lesnard et al., 2009). Thus, more freedom in the organization of the working schedule may be associated with greater work hour's commitments.

Another point concerns education. For instance, we know that higher-educated people have more autonomy in determining their work schedules (Voorpostel et al, 2010). At the same time, they are expected to have higher earnings on average than lower-educated individuals, and for wealthy couples we know that they consume more synchronous leisure time (Hallberg, 2002). On the other hand, lower-wage workers are more likely to work evening shifts (Hamermesh, 2002) and some studies have pointed out a negative association between evening working hours and some

2

family togetherness moments regularly scheduled at the end of the day, such as shared meals, television watching, or leisure activities (Nock & Kingston, 1988; Lesnard, 2008). Finally, being more educated helps couples to increase the availability of shared conjugal time.

As said, that of family solidarity and togetherness is a non-reducible valuable dimension for couples' everyday life (Hamermesh, 2002; Lesnard, 2008). Scholars have underlined the importance of the total amount of time spent jointly in the same place by the spouses. However, even during time spent together, significant gender differences have been identified, especially in the configuration of free/leisure time and household task boundaries.

For instance, it has been noticed that even if spouses spend free time together in the same place, the woman is more likely to do unpaid work simultaneously, as a 'secondary activity'. This difference in multi-tasking allows men to spend their leisure time in blocks, while that of women is more likely to be interrupted and reduced by household care tasks (Bittman et al., 2000; Mattingly et al., 2003; Kilkely et al., 2010; Naldini et al., 2011). Women's multi-tasking may undermine their daily quality of life (Offer et al., 2011) by negatively affecting the free time quantity and continuity.

Moreover, the risk of a free time deprivation for women increases by the kind of household care activities. For instance, the contribution of men to house care occurs mainly for less routine tasks – i.e. pet care, maintenance of the garden, repairs, care of adults – while the more routine, essential and demanding activities – i.e. cooking, cleaning and laundry – still seem to be a women's responsibility (Kan et al., 2011; Moreno-Colom, 2015).

Something similar has been observed for gender segregation in childcare time. In fact, fathers do less child care than mothers, even if this gap is lower than in the past (Raley et al., 2012; Craig et al., 2014). Moreover, fathers tend to leave mothers alone in the most stressful, ordinary, and onerous tasks: they are more likely to provide childcare simultaneously with the mother, rather than alone; and they are generally more likely to be engaged in less routine and more desirable tasks (Budig et al., 2004; Craig, 2006; Raley et al., 2012).

Finally, women are 'care managers' (Naldini et al., 2011) engaged in the most onerous tasks in terms of time and energy, while fathers seem to fill the timetable gaps with the more pleasant and 'desirable' ones. The main outcome is that women's unpaid work throughout the day is more constant and repetitive (Wajcman, 2008; Kilkely et al., 2010) while that of men is typically sporadic and delimited in both time spent and variety of tasks (Kilkely et al., 2010:245). All these findings depict a well-recognized and documented scenario: within the spouses' daily life there is a 'leisure gap' in favor of men even if their partner works: «in most industrial countries […] employed women work longer hours (paid and unpaid) than employed men» (Mattingly et al., 2003; Beblo et al., 2008:281) and the gradual gender convergence in the housework time allocation of recent decades (Raley et. al, 2012; Craig, 2006; Kilkey et al., 2010) seems to be not enough for the protection of women's free time.

Whether the spouses are spending time together or they are performing more or less similar activities at a certain point in time, the issue of 'being (de)synchronized' matters. In this sense, that of off-scheduling becomes a crucial concept. However, the main problem is how to measure it.

When off-scheduling has been measured by means of time use diaries,[1] the main approach has been to count the slots in which both the spouses have worked or not, obtaining work synchronicity ratios or percentages. In this way, the (de)synchronization is seen as the quantity of time in which both the spouses work. However, nothing is known about 'when' the work schedules are overlapped and when they are not. This is a crucial limitation for two main reasons. First, time is socially structured as well as social rhythms and social constraints. Hence, being at work simultaneously at 10 a.m. or 10 p.m. has radically different impacts on a couple's daily life (Lesnard, 2008). Second,

---

[1] We shall not discuss the other older and less precise techniques to measure working time, like those of surveys in which two different questions – when did you start / when did you finish – are posed. We focus on modern time use surveys.

take the case of a full-time shift perfectly synchronized with a part-time afternoon shift: by considering only the duration of the overlap, we will mistakenly classify it as a highly desynchronized working schedule. However, such kinds of '*structural desynchronizations*' – due simply to differences in duration – should not be compared with hypothetical others with the same off-scheduling amount but different organization during the day (Nock et al., 1984).

Moreover, most of the above scholars have shown an important limitation regarding the mainstream time approach. As suggested by Lesnard (2005), if «we know little about how family time is daily balanced with work time for both spouses», this is mainly due to the limits of the dominant time-budget perspective, «an individualistic approach and a simplification of time» (Lesnard 2005:2). In fact, scholars have underestimated the importance of daily scheduling, while paying more attention to total amounts of time (Lesnard, 2008). They have traditionally acquired time budget information related to different daily activities, but these should be seen in a holistic perspective that makes it possible to study the couple's days as a whole, avoiding the manipulation of time as if it were clay.

According to Hallberg (2002), «while the traditional time allocation model typically studies the total time spent in, e.g., market work, over a day or a week, it provides little or no insight into the temporal pattern of time-use and therefore, potentially, misses a vital part of the mechanisms underlying empirical observations» (Hallberg, 2002:2). On the other hand, if schedules are studied as sequences (Lesnard, 2004), new insights could emerge on the interdependence and the synchronization within different daily activities scheduled by both the spouses. A sequence analysis of time-use would point out the routine aspect of the daily life, as well as the couples' *projects*[2] performed to perceive their strategies across several daily constraints and unexpected events (Hägerstrand, 1982; Hellgren, 2014). Finally, the analysis of the time-use temporal patterns seems more relevant – instead of time budgets – in the study of the daily strategies and behaviors of a couple (Hallberg, 2002).

Lesnard's works pave the way to the solution of these distortions by considering working *schedules as sequences* (Lesnard, 2004), overcoming the time budget framework's limits and measuring (de)synchronizations in an integrated quantity-timing perspective (Lesnard 2004, 2005, 2008; Lesnard et al. 2009). Unfortunately, and despite this fundamental contribution, the entire complexity of daily schedules have been basically reduced to working schedules. Thus, the crucial dimension of the spouses' daily (de)synchronization-in-time – i.e. their combination of activities at each point-in-time – has been studied only for the work activity. For instance, we do not know how leisure and household activities interact with the working schedules during the whole day by filling the out-of-work time of the spouses. For these activities, we have only information on different time budgets.

The point is that we are not able to locate these different out-of-work activities in the different parts of the day and we don't know how the different activities of the day are combined by the spouses at each point-in-time. Somehow, by considering multiple activities simultaneously, the problem of being able to recognize structural desynchronizations and those that are not is crucial (Nock et al., 1984). In a sense, we might say that the same gendered time budget gap in house care could be more or less impactful between two different couples, according to His and Her structure of the day in terms of both duration and sequence of activities.

Finally, in current research on couples' daily work-family balance strategies, there are several shortcomings in the implementation of a holistic approach of time integrity. Garcia et al. (2016) underline that: *a)* not many studies have adopted a real couple-level-approach[3] (Lesnard, 2008;

---

[2] The word *project* is a key concept in the 'time-geography' perspective, founded by Torsten Hägerstrand (1970). This concept «was meant to tie together into a whole all those 'cuts' in evolving situations that an actor must secure in order to reach a goal» (Hägerstrand, 1982: 324)

[3] Even among the more recent and important time-use surveys, the provision of information about both the partners is a rare feature – i.e. the American Time Use Survey (ATUS) does provide it. In this regard, the Italian survey 'Uso del

Craig et al., 2011 in Garcia et al., 2016); *b)* few others have focused on how working schedules are related to multiple daily activities (Wight et al., 2008; Lesnard, 2008; Garcia et al., 2016). We add that *c)* only Lesnard has studied working schedules from a sequence analysis perspective (Lesnard 2004). Moreover, to our knowledge, *d)* no one has considered the entire complex of daily activities of Him and Her as a unique sequence, neither at an individual nor at a couples' level.

Thus, in order to understand the complexity of work-family balance strategies, it is necessary to study the couple's daily time-use pattern as a whole and in a more holistic way. In what follows, we aim to contribute to overcoming these limitations with a new viewpoint based on a 'multichannel' sequence analysis. We believe that new useful evidence may emerge from old questions if we start to consider couples' work-family strategies as an overall temporal pattern of combinations of multiple activities. And the first is to wondering: how to measure it?

2. Order as distance

> "All things whatever stand to each other in some relation of time. Every phenomenon, when considered in connection with any other, must be cognized either as occurring before it, as being simultaneous with it, or as occurring after it. But all objects of thought, and, among others, relations of time, admit of being compared, and their likeness or unlikeness recognized. The time-relation of events that occur simultaneously, is manifest different from the time-relation that occur one after the other. Two sequences are alike in so far as they are sequences, and each of them is unlike a coexistence. Hence, if there are time-relations so completely alike as to be indistinguishable, they may properly be called equal." (Spencer 1855, p.119).

One of the main problems of current techniques used to compute the distances among sequences derives from the way in which similarity between two sequences is defined. Abbott& Forrest (1986), Elzinga (2003) and Dijskra (1997) agree that two sequences are equal when they comprise the same elements in the same order. They also agree, although they take different approaches to the problem, that two sequences are maximally different when they have no elements in common (as regards order and type of elements). In other words, imagine that four women are being observed for a period of length 2, and that whether or not they give birth to a child in each time interval is recorded, thus obtaining the following four different sequences: A {0,0}; B {0,1}; C {1,0} and D {1,1}. With the methods proposed, the distances between sequences BC and AD are greater than the distances between the sequences AB, AC, DB and DC. In fact, A and D do not have elements in common; nor do – in the same order – B and C.

But what is meant by 'maximally dissimilar'? Whilst it is evident that if two sequences share the same number of elements in the same order they are maximally similar, it is less clear what 'maximally dissimilar' signifies. In other words, is the number of shared elements the only possible way in which we can establish the distance between two sequences?

In the above example, woman A had no fertility events during the observation period; woman B had a fertility event immediately before the conclusion of the observation period; woman C had a fertility event at the beginning of the observation period; and woman D had two fertility events, one at the beginning and one at the end of the observation period. It is clear that in this case is difficult to say that the women B and C are maximally distant: both had just one child, the only difference being that they did so at different moments of the observation period, while it is reasonable to believe that the distance between A and D is maximum. The problem therefore resides in the importance given to the temporal order of the events, their numerousness, and the presence of shared elements when the distance is calculated.

Tempo' (ISTAT, 2011) gathers such information, which is essential for understanding both how spouses coordinate their family time and whether men and women differ in their time-use patterns.

The principal focus of the above-cited studies is the search by means of pair wise comparison for common elements, but this may not be only the way to establish distances (similarities or differences) between sequences. If the sequences are ordered according to the number of observed events and, all of observed events being equal, an order is established along the temporal axis in which the events have occurred, then woman B, who had her fertility event immediately before conclusion of the observation, follows (is logically closer to) woman A, who had no fertility event during the observation period;[4] woman C follows A and B in that she also had only one child, but did so before B, and obviously before A. Finally, comes, woman D, who had two fertility events and is therefore more distant from the woman who did not have any, but also from those who had only one. Hence, on considering numerousness and the order of occurrence of the events, one obtains a unilinear structure with the order A,B,C,D of the four sequences. The similarity between sequences and their proximity thus results from the order in which the events occur, not from the common elements. The order itself exhibits that there are more elements there are in common, the closer the two sequences, without requiring complicated and not always clear measures to compute the weights and contributions for comparison. The problem is therefore how to find a simple and rapid way to compute this order.

3.0 A first definition of lexicographic index

We can define a generic sequence as a list of episode-states observed in a particular time interval. Ideally, but also graphically, this list develops along a single dimension, that of time. For instance, suppose that one is observing labor market participation by subject A for six months. Assume that at any particular time subject A can exhibit only one of the followings three states: $1=$ employed; $2 =$ unemployed; $3=$ inactive. At the end of the six months of observation, sequence A is [123321]. Graphically, this sequence is a list of the episodes experienced by the subject in a one-dimensional space. At the beginning of the observation, A spends the first month in employment and in the second month is unemployed. At the end of the second month, s/he exits the labor market and only returns in the fifth month to seek a job. Finally, in the sixth month s/he is once again employed. Now, the question is: can we reduce a multinomial sequence to a one-dimensional list of different episodes linked in time? The answer is no. Indeed, a multinomial sequence cannot be represented by any list at all. A multinomial sequence does not exist as such; rather, it results from the co-action of the states of which it is composed. A multinomial sequence is a point in the space-time defined by the states space. The problem now is how to represent a multinomial sequence.

It said above that every state has its own generating mechanism, which operates independently of the others. In geometric terms this signifies that each individual state defines an axis of a q-dimensional space, where q is the size of the state space, i.e. the set of values defined in the state space that the generic sequence can assume. Each axis in its turn represents the state space of the set of all the possible orders in which a state can occur in a sequence of length t. Put otherwise: a space will have as many dimensions, each orthogonal to the others, as there are states defined in q. Taken individually, each dimension of this space will represent the set of all the possible realizations with which every state may occur in a binary sequence of length t.

Now suppose that it is possible to attribute a value to all the elementary sequences observed so that they can be arranged along their axis. Suppose also that it is possible to draw as many orthogonal straight lines as there are axes starting from each point defined by each elementary sequence making up the multinomial sequence. The point in space defined by the intersection of all these straight lines will be the multinomial sequence. A multinomial sequence is therefore a point in the q-dimensional space of the states, and its coordinates are the values of the individual elementary sequences of which it is composed.

---

[4] Although this is not to say that in the following, not immediately observed, period also A will have a fertility event.

In this way, three things are obtained. The first is that a multinomial sequence consisting of manifold states is reduced to a single point in space. The second is that the multinomial sequence is no longer a series of realizations along a one-dimensional line in which the different mechanisms that have produced it are con-fused in a whole. Rather, each individual state, each individual event, is free to define the form and the length of the individual lines, which in turn are free to interweave with each other to form the complicated plot of a story which has a logical narrative. Finally, and perhaps most interestingly, the distance between two multinomial sequences is the distance between two points in a Euclidean space. It will be only necessary to decide what method one wants to adopt for calculating the distance.

Moving from a multinomial sequence to its elementary sequences is straightforward. Just as a qualitative variable of *k* modality can be represented by *k*-dummy variables, so a multinomial sequence of *q*-states can be represented by *q* binary sequences. For example, the sequence A = [123321] can be represented into the following three sequences A_1 = [100001]; A_2 = [010010]; A_3 = [001100].

Still to be defined is a method to calculate the coordinates, and therefore a method to attribute univocal values to the elementary sequences. The next section provides a possible solution.

3.1 The lexicographic index

The index now introduced derives from the example in the previous section. The goal is instead to attribute a univocal value to each different binary sequence of length *t*. The intention is also for the index in question to have the properties of triangularity, symmetry and positivity proper to a distance; and even more importantly for it to take account of time; that is, of the different ways in which a state can come about in time within a binary sequence. For these various purposes, we must impose an order of all the possible binary sequences of length t. However, the problem is deciding what order to impose.

Take, for instance, the following binary sequence relative to labor market participation by A = [0101] in a period of four months. This sequence gives us two items of information. The first is that A was employed for two months. The second is that A was employed at the time t_2 and at the time t_4. A binary sequence, therefore, responds to two distinct ways of observing time. The first concerns the quantity of time and answers the question 'how long'. The second concerns the moments in which the states are realized and answers the question 'when', 'at what time'. This twofold nature of the binary sequence forms the basis of the sorting order introduced here.

The index is based on the sorting order of these two different modes of observing time. The first order is given by the quantity of time and is therefore based on the number of times that state *q* is observed in the sequence. The second order is given – the quantity of observed time being equal – by the 'moment' or 'moments' in the sequence when state *q* occurs. This second sorting order is also based on a twofold order. The first is the reverse order in which the events occur in time. It thus puts first the events that occurred last and then the events that occurred first. This solution is adopted in light of the discussion in Section 3. It will be recalled that woman B with sequence [0,1] followed woman A with sequence [0,0]. The two sequences were considered to be closer to each other than the others because B's fertility event occurred immediately before the end of the observation period, with it being hypothesised that A would have had her own fertility event immediately after conclusion of the observation. The second order is a direct consequence of the first. The events that occur last will vary more slowly in the order than those that occurred first.

Because the nature of the sorting order is double, also the proposed index consists of two distinct parts. The first part, ranging from 0 and 1, takes account of the different amount of time/realization recorded in each sequence and it is:

$$d'(x_i) = \frac{u}{T} \quad \text{for } u > 0 \text{ and } 0 \text{ for } u = 0. \tag{1}$$

The second part, ranging from 0 and 1, which takes account of the different numbers of combinations displayed by the sequences with variation in the amount of time. Calculation of this part is slightly more elaborate.

Tab.3.1State space of the sequences length T = 4 ordered by the lexicographic index (d'_i) and (d''_i) of lexicographic index $>_{b2}$.

i.	t1	t2	t3	t4	(d'_i)	(d''_i)
1	0	0	0	0	0	0.0000
2	0	0	0	1	0.25	0.2000
3	0	0	1	0	0.25	0.4000
4	0	1	0	0	0.25	0.6000
5	1	0	0	0	0.25	0.8000
6	0	0	1	1	0.50	0.1429
7	0	1	0	1	0.50	0.2857
8	1	0	0	1	0.50	0.4286
9	0	1	1	0	0.50	0.5714
10	1	0	1	0	0.50	0.7143
11	1	1	0	0	0.50	0.8571
12	0	1	1	1	0.75	0.2000
13	1	0	1	1	0.75	0.4000
14	1	1	0	1	0.75	0.6000
15	1	1	1	0	0.75	0.8000
16	1	1	1	1	1.00	0.0000

Suppose that there is a binary sequence x_i containing the observations of T time periods. Observation x_i can assume only two modalities represented by the numbers 0-1. These modalities we shall call absence/presence. Consider the case in which exactly u realizations equal to 1 occur. There are obviously several sequences that have this characteristic. For example (Table 1.) sequences from 6 to 11 have $u = 2$.

The problem now is to allocate to each sequence x_i a number (x''_i) representing its position, normalized between 0 and 1, in the sorting order of the sequences. An example will aid understanding of the computational procedure. The set of the possible binary sequences x_i of length T=4 are exactly 2^4=16, and they are represented in Table 1.

Consider, for simplicity, only the $x_i = \binom{T}{u} = \binom{4}{2} = 6$ sequences for $u = 2$, which are denoted in Table 1 by the numbers from 6 to 11. Following the chronological order, the first and second realization (the 1s) of the sequence are called s_1 and s_2. We calculate for every realization u of the sequence x_i three values $s_k\{A_k, B_k, C_k\}$, where:

A_k is the exact position of s_k in the sequence. For instance, in the sequence 6, for s_1: $A_1 = 3$, and for s_2: $A_2 = 4$; in the sequence 13, s_1: $A_1 = 1$, s_2: $A_2 = 3$, s_3: $A_3 = 4$;

B_k is the maximum position that s_k can occupy within the sequence. For example, for the sequences 6, 7, 8, s_1 can occupy at most position t3 because position t4 is occupied by s_2, so that $B_1 = 3$, while for sequences 9 and 10, $B_1 = 2$, because s_2 now occupies position t3. Finally, for sequence 11, $B_1 = 1$. For s_2 in all six sequences considered, $B_2 = 4$;

C_k is the minimum position s_k that can occupy. In this case, for all six sequences considered $C_1 = 1$ and $C_2 = 2$.

At this point, the calculation of the second part of any binary sequence of length **t,** for any number of realizations u, will be:

8

$$d''(x_i) = \frac{\binom{T}{u}}{\binom{T}{u} + 1} * \frac{1 + \binom{T}{u} - \left[\binom{B_u}{C_u} - \sum_{s_k=1}^{s_k} \binom{B_k}{C_k} - \binom{A_k}{C_k}\right]}{\binom{T}{u}} \tag{2}$$

The (d''_i) part of the lexicographical index is one plus the difference between the set of all the possible realizations of a given value of u minus the difference between the set of the possible realizations for the maximum position $(\boldsymbol{s_k})$ reach of u minus the summation of the difference between the set of all the possible realizations of $\boldsymbol{s_k}$ taking account of the upper and lower limits within which $\boldsymbol{s_k}$ can occur and the position in which $\boldsymbol{s_k}$ is observed divided by the binomial coefficient of all the sequences that can be realized for a given value of u, normalize from 0 to 1.

The decision to normalize (d''_u) is taken because the number of sequences varies with the value of u. Thus, independently of the value of u, the distance between the first and the last sequence defined with the same u will be at most 1.

The solution thus is a measure of distance to an index formed by a couple of distances/coordinates on a Cartesian space (Graph 1.0), where $d'(x_i)$ concerns the quantity of time and $d''(x_i)$ concerns the moments (the timing) in which the states are realized.

The similarity/distance between two sequences (x_i, x_j) is the Euclidean distance between a couple of lexicographic indices '$>_{b2}$'.

$$r'(x_i, x_j) = \sqrt[2]{\sum_{q=1}^{Q} (d'(i) - d'(j))^2 + (d''(i) - d''(j))^2} \tag{3}$$
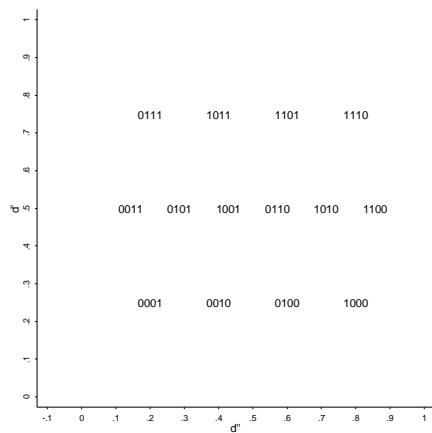
Fig.3.1. Plot of a binary sequence of length 4 according to the lexicographic index coordinates.

3.2 From binary to multinomial sequences

The next step is to pass from a binary sequence to a multinomial sequence: that is, the case in which there are more than two states (for example, 'employed', 'unemployed', 'never worked', etc.). In the above paragraph we have underlined that just as a qualitative variable of $\boldsymbol{k}$ modality can be represented by $\boldsymbol{k}$-dummy variables, so a multinomial sequence of $\boldsymbol{q}$-states can be represented by $\boldsymbol{q}$ binary sequences with values 0-1. So which element $x_{qi}(t) = 1$ if the ith unit assumes the qth modality in the tth instant, $x_{qi}(t) = 0$ otherwise. It is possible to apply both the lexicographical

9

indices to each of these sequences and compute the distance measure $r_q(x_i)$ or the coordinate/distance numbers $\{d_q'(x_i); d_q''(x_i)\}$. The multinomial sequence x_i is therefore described by a vector to real numbers. The distance function between two multinomial sequences x_i and x_j is the Euclidean[5] distance between their transformations r_i and r_j. Formally the coordinate/distance lexicographic index ($>_{b2}$) the distance is:

$$D'(x_i, x_j) = \sqrt[2]{\sum_{q=1}^{Q} \left(d_q'(i) - d_q'(j)\right)^2 + \left(d_q''(i) - d_q''(j)\right)^2} \qquad (4)$$

3.3 Some considerations on the index

We shall conclude this second part of the article by briefly discussing the index just presented. Firstly, the index is a measure defined *a priori*, independently of the sequences observed. Hence, it is not comparison between the sequences that defines their distance; instead, their distance is given by definition in an independent system of measurement. The index has a known beginning and end; each point of the measure is univocal and identifies one and only one combination of states in sequence. Two sequences which differ in the position of only one element will have different positions. Two sequences with the same number of realizations in the same order will share exactly the same point in the index. None of the information contained in the sequences is lost. From every point one can retrace the exact sequence that has produced it.

A second characteristic of the index concerns its output. With current methods, the output is a symmetrical matrix of distances. This matrix can be used only in statistical procedures based on matrixes of distance, like hierarchical clusters and multidimensional scaling. The output of the lexicographical index is a cases by variables format, where the cases are the sequences, and the variables are the lexicographical indexes of the states that make up a sequence. Each value of the index, in fact, can be conceived as a coordinate in the space of the multinomial sequence. This characteristic enables the researcher to adopt different methods to calculate the distance, but also to define forms of space other than Euclidean. Moreover, the index can also be used with other statistical analysis programs, like the k-mean cluster or the fuzzy k-means cluster. In this case, the matter is not only technical but also substantive. When applying a hierarchical cluster, one implicitly assumes that the phenomenon studied is organized into successive specializations. But this is only one of the possible ways in which a phenomenon may structure itself; the hierarchical model is only one of the possible ways in which a relationship model can be structured.

4. The data, their organization and the coding of the activities'

Analyses were conducted on 873 dual-earner couples[6] carry-out on data of the Italian Census on Time Use 2008/09 (the last disposable). The goal is to discover how the Italian dual-earner couples organize their daily activities (sleep, personal care, work, moving, housework, free time), during a typical work day from Monday to Friday.

From Him and Her time-use diaries were considered the dual-earner activities from 7.00 to 22.00. Each daily activity is observed every 10 minute, and the data files for the sequence analysis consisted of two pair sequences, one for Him and one for Her, with a total of 90 points in time.

---

[5] Given the metric nature of the two lexicographical indices, the Euclidean distance is only one of the possible ways of define the distance between two sequences. It is possible to adopt both different measures of Euclidean distance that to define other geometric spaces different from that Euclidean.

[6] Excluded from the sample were: (a) couples who living with other couples (parents or others); (b) couples that fill the questionnaire in different days, or fill the questionnaire in the week-end; (c) couples with incomplete information by one or both the spouses; and, (d) age of him or she greater than 65 year old.

Each couples of row of this file corresponded to a cohabitation, while each variable corresponded to 10 minutes of observation and each cell of the row/column intersection states the activities of the man and the woman at time **t**.

In order to simplify the analysis, the paper considered 6 different groups of activity: (a) sleep; (b) personal care, i.e. have a shower etc., eat (breakfast, lunch, dinner); (c) work; (d) moving (any kind); (e) house care, i.e. housework, children care, repair, etc.; (f) free time and other activities with or without others.

Once having defined the six daily macro-activities, the next step was establishing how to codify the day activities of the man and the woman in the couple. In this case, His activities and Her activities interact in time to give rise to the couple's daily activities. Taken individually, each of these two sequences takes the form of a series of mutually exclusive episodes. The problem is therefore how to codify two interacting sequences composed of a plurality of mutually exclusive events. All the solutions proposed to date (Abbott 1990b, Dijkscra 1995, Elzinga 2003, Gauthier et al. 2010) have been based on the generation of events combinations: that is, on the construction of a single sequence that combines the states of Him and Her.

Fig. 4.1. WSS, log(WSS), η2, and PRE for all twenty cluster solutions on the lexicographic index for the sequences from 07 to 22.
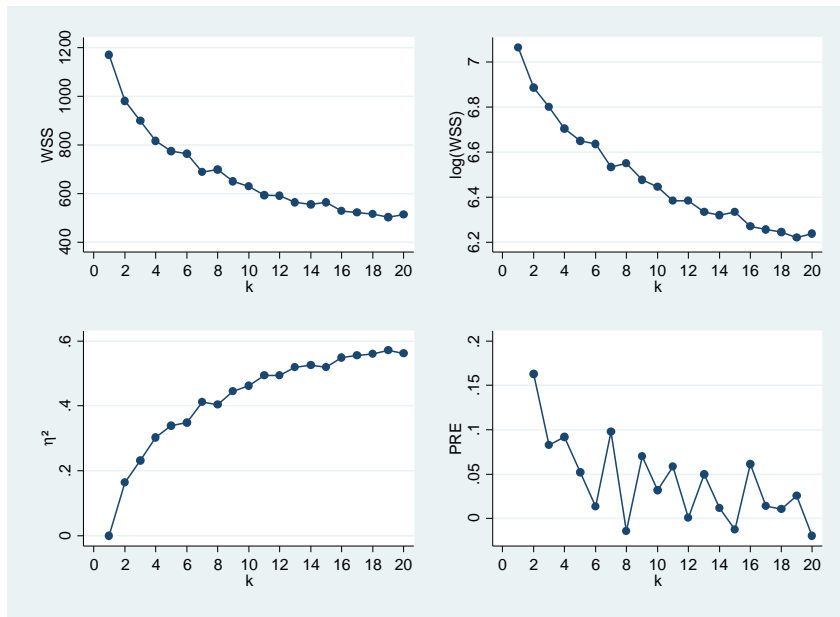
This operation has a number of consequences. Firstly, as Abbott pointed out, using combinations of events requires one to pay «… the price of losing all information about the temporal 'shape' of events – their duration and their intensity in terms of producing occurrence – in short their time horizon. (Abbott 1990b: 146)». Secondly, there is the risk that distinct time-use patterns will be tied together, although the order of causality may be bi-directional.

There are various reasons to believe that the day activities of Him and Her cannot be reduced to a simple combination of states. Internally, moreover, each sequence consists of states which themselves are regulated by their own mechanisms which operate differently in defining the timing and duration of each individual episode/state. By way of example: consider the mechanisms that underlie the regulation of the states of housework and free time. In the former case, it is the

11

educational level, for instance, that regulates the time spent in these two activities; in the latter, and there is an interaction between gender and educational level.

It is therefore possible to hypothesize that the sequences of Him and Her – and the states of which they are composed – have their own underlying generative mechanisms which establish the timing and duration of episodes. These generative mechanisms work independently of each other and interact in time: they stand in a coexistence relationship. The couple's daily activities are therefore the result of a complex process of co-action between two sequences, that of Him and that of Her, regulated by different generative mechanisms resulting from the co-action between different generative mechanisms underlying each state. Consequently, reducing everything to a combination of events is to lose large part of the variability inherent in each single time use sequence.

Couple's daily activities, or more correctly the couple sequences analyzed here, are therefore configured by co-action by two multinomial sequences composed of mutually exclusive episodes. On extending the applications proposed (Bison, 2011a, 2011b) of the lexicographical index ">b2", 12 binary sequences can be defined, six for His states and six for Her states, of equal length t = 90, that is, to the overall number of point of observation. The couple sequence is defined as a point in a 24-dimension space whose coordinates are the 12 binary sequences defining the respective sequences of Him and of Her. The distance between two couple sequences will be given by the Euclidean distance between the two points of the two sequences in the 24-dimension space.[7]

These 24 variable/coordinates defined for all the 873 couples were analyzed using a simple k-means cluster algorithm to find clusters of similar couple sequences. From the analyses it was decided to adopt a seven-cluster solution for the first analysis (Fig.4.1).

These clusters has been analyzed both as dependent variables in order to verify which features were most expressive of the individual patterns and subsequently as independent variables to estimate the effect on His/Her satisfaction of time and daily activities. For all these analyses we have used information relative to the educational qualification[8], job (sector, class position, full-time/part-time), area of residence[9], dimension of town[10], age, the presence of children.

5. From 7.00 to 22.00: a typical working day of a dual-earner couple in Italy

It is not news that the everyday life of a dual-earner couple is complex. It involves a long and difficult schedule of: waking up, having a shower, breakfast, taking the car-bus-train, going to work, beginning work, lunch, resuming work, coming back home, then housework and family/child care for Her, relaxation for Him, dinner, and at the end of the day, before they go to sleep, some leisure activity.

Data from the Italian Time Use survey demonstrate that this is the typical daily routine of a dual-earner couple. The time plot of activities of figure 5.1 does not show a clear difference in the behaviors of Him and Her or differences among couples. In the morning, at 7.00, 75.0% of couples are involved in personal care, are at work, or going to work. From 8:00 to 12:00 all the couples are at work. After that, they have a break for the lunchtime. From 1.30pm to 5.00pm they are again at work. Finally at home, they engage in housework and free-time activities, and then the day is ended.

Until noon the couple's everyday lives are perfectly synchronized. He and She show some differences in the afternoon, where fewer women than men are at work. The women seem to use

---

[7] It can be easily shown that information is not lost or altered with this coding scheme. Moreover, the procedure does not require the researcher to perform complicated combinational operations and make arbitrary coding choices.
[8] Educational qualifications were classified as: (a) compulsory (elementary school certificate (including no educational qualifications), lower-secondary school certificate (including 2-to-3 year vocational certificates), (b) upper-secondary school diploma (including post-secondary diplomas), and (c) degree (including postgraduate qualifications).
[9] The areas were (a) North; (b) Centre (c) South and Islands.
[10] The town dimension were classified as: (a) Metropolitan; (b) more than 50.000 people; (c) from 10.000 to 50.000 people; and, (d) less than 10.000 people

this time not spent on the job mainly for house care (housework, child care, etc.). This is not a new finding, and it is due to the unequal distribution of house care activities between genders in the couple.

Regarding to the total average, time spent on the single activities by men and woman during the day (Tab. 5.1), He and She devote more or less the same amount of time to personal care (i.e. showering, eating, etc.) and travel (any kind). Differences are observed in the time spent on work, house care, and free time. The men spend an average of 8 hours and 6 minutes on work, 1:06 on house care, and 2:02 on free time. The women spend 6:32 on work, 3:37 on house care (more than three times that spent by men), and only 1:22 on free time.

Then, on the one hand, the couple's daytime activities are synchronized with a common social rhythm; on the other, there are differences between men and women in the time spent on some specific activities. The questions that arise are the following. How do He and She organize their lives and how do they synchronize their daily activities? Do all dual-earner couples follow exactly the same pattern or are there different patterns, or no patterns? And if different patterns exist, are they due to internal and individual bargaining between him and her, are they the result of some external constraints, or are they a combination of both?

Fig.5.1. Tempograms: Daytime activities on total and by gender.

Sleep   Personal care

Work   Move

House/child/family care   Free time
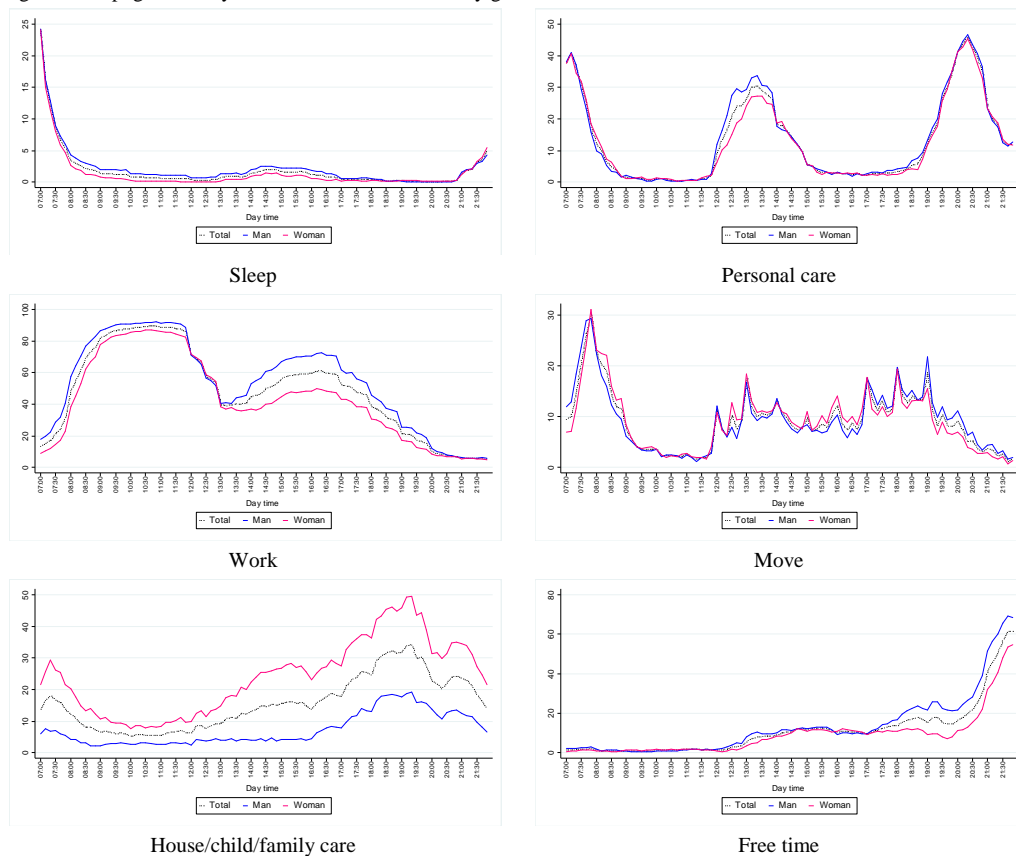
Given the evidence that job activities and work time play a central role in daily time organization, one wonders whether the dual-earner daytime is similar both within the couple,

13

between Him and Her, and among couples, or whether there are different patterns of daytime activities within and among dual-earner couples. The hypotheses that derive from this are the following:

(a) Among couples, job activities and working time have an effect not only at the beginning of the day but also on the organization of large part of the daytime. We expect to find an effect on lunchtime, on free time, and on evening activities;

(b) Within dual-earner couples we expect to observe a different timing and packaging of daytime activities according to the different job characteristics of Him and Her. Moreover, we expect to observe that the different distribution of house care not only depends on the educational/cultural level of Him and Her (mainly to configure a different timing at the beginning and end of the day) but is also strictly connected with the different worktime and work schedules (working hours, hour of starting and finishing work) of Him and Her;

(c) Social rhythms and social constraints have an effect both among and within dual-earner couples. We expect to find that the timing of a couple's activities varies according to the social and the cultural context in which it lives.

To falsify these hypotheses, we start by observing the entire daytime of dual-earner couples from 7.00 to 22.00.

The results of the k-means cluster analysis carried out on the lexicographic index >b2 show that a seven cluster solution (Fig.4.1) is acceptable (but not optimal).

Considering each cluster in detail, however, we can observe some systematic differences among clusters. Moreover, these differences seem to describe clear and reckonable patterns in the day organization of both individuals and couples. Finally, the strong relations with social-cultural-demographic features support the idea that these behaviors and patterns are the results of two factors: one is the internal organization of each couple conditioned by the social-cultural-economical characteristics of the two partners; and the other is the external social rhythms constraint.

There follow brief biographical sketches for each cluster.

Fig.5.2. Daily activities of men and women: stacked percent.
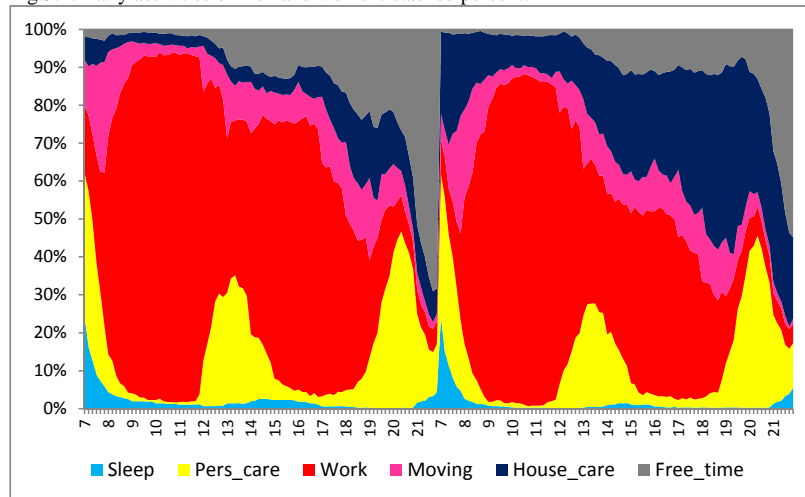
The dual-earner couples in **cluster 1** wake up very early in the morning (Fig.5.3, Tab. 5.2). At 7:00am both are awake (Tab.5.1). Both work for a long time: He for 8 hours and 32 minutes, and She for 7 hours and 28 minutes. Both spend more than one hour and a half on travel/moving. Their

14

lunch breaks begin at different times (Tab.5.3): for Him at 13:20, for Her at 13:30. Both have very little free time. However, between the two, She has less free time, only 49 minutes in the entire day, compared to His 1:42 hours.

The sequence index plot shows that these couples have an unequal behavior in the division of housework. He spends no more than one hour a day on house care; moreover, He does not do anything in the evening. She not only spends more than three hours a day on house care, but this activity is mainly carried out in the evening when He is relaxing.

These dual-earner couples live mainly in metropolitan areas or in medium-sized towns (10-50,000 residents) in the Centre of Italy (Tab. 5.5). They have one or more children aged between 0 and 14 years old. Both are in the middle class (IIIa), the only difference being that He works in private services and She in industry (Tab. 5.6 & 5.8).

Tab.5.1. Mean (in hours/minutes) of the times spent in each activity by the clusters. (F. Test)

Clu.num	Sleep		Personal Care		Work		House Care		Moving		Free Time		N.
	M	W	M	W	M	W	M	W	M	W	M	W	
1	0:01	0:00	2:02	1:50	8:32	7:28	1:08	3:15	1:35	1:38	1:42	0:49	133
2	0:04	0:03	1:50	1:49	7:26	6:30	1:51	3:51	1:23	1:18	2:27	1:28	158
3	0:03	0:38	2:09	2:01	8:38	6:07	0:50	3:35	1:18	1:17	2:02	1:23	99
4	1:04	0:44	2:02	1:58	7:27	6:09	1:10	3:27	1:17	1:19	2:00	1:24	108
5	0:51	0:02	2:10	1:53	7:46	6:58	1:05	3:25	1:14	1:24	1:55	1:18	129
6	0:08	0:08	2:08	1:51	8:45	6:07	0:30	3:59	1:24	1:78	2:05	1:37	179
7	0:14	0:11	2:06	2:01	7:58	6:16	1:13	3:30	1:42	1:26	1:47	1:36	67
Total	0:19	0:13	2:03	1:54	8:06	6:32	1:06	3:37	1:24	1:23	2:02	1:22	873
F.	31.5**	42.5**	3.6**	1.6	12.9**	8.6**	18.9**	2.8*	3.1**	3.9**	4.3**	6.9**	

Sign: (*) p.<0.05; (**) p.<0.01

Both are very dissatisfied with the time available for the partner, children, and relaxation (Tab. 5.7). He is much stressed, highly dissatisfied with daily life and with the time that he has for himself. Finally, both have many difficulties in reconciling daily tasks with the opening hours of office, shops, and leisure center.

The dual-earner couples in **cluster 2**, like those in the previous cluster 1, wake up very early in the morning. At 7:00 am both are awake; 37.1% of men are at work or are about to reach the workplace. She starts work slightly later than He; only 24.7% of women are at work or going to work at 7:00 am.

Compared with the other couples, the couples in this cluster spend less time on personal care (1 hour and 50 minutes on average) and on work: He 7 hours and 26 minutes, and She 6 hours and 30 minutes. Work activities are undertaken mainly in the morning. Both devote a great deal of time to house care.

Tab.5.2. What does He and She do at 7:00am. Activities conducted from 7.00am to 7:10am by gender and cluster.

Clu.num	Sleep		Personal Care		Work		House Care		Moving		Free Time		N.
	M	W	M	W	M	W	M	W	M	W	M	W	
1	0.8	0.0	52.6	51.9	18.1	6.8	7.5	30.1	20.3	10.5	0.8	0.8	133
2	1.3	0.6	43.7	44.9	26.6	15.8	15.8	29.1	9.5	8.9	3.2	0.6	158
3	1.0	90.9	62.6	4.0	16.2	1.0	3.0	3.0	16.2	1.0	1.0	0.0	99
4	87.0	91.7	5.6	3.7	5.6	2.8	0.9	1.9	0.9	0.0	0.0	0.0	108
5	81.4	0.8	4.7	53.5	10.9	10.9	0.8	27.1	1.6	7.0	0.8	0.8	129
6	1.7	2.2	50.8	43.6	22.9	13.4	4.5	29.6	16.8	10.6	3.4	0.6	179
7	9.0	22.4	41.8	50.8	17.9	4.5	6.0	14.9	19.4	6.0	6.0	1.5	67
Total	24.3	24.1	38.0	37.7	17.8	9.1	6.0	21.7	11.9	7.0	2.1	0.6	873

Among the men, He is the one that devotes most time to house care (1 hour and 51 minutes), while among women, She is the one that devotes most time to house care (3 hours and 51 minutes). Lunchtime for Him is between 12:30 and 13:30, while for Her it is at either 13:10 or at 14:10.

Dinner is for both at 20:10. After dinner both spend their time on relaxing and leisure activities. Moreover, among men, He is the one that has most leisure activities (2:27).

Tab.5.3. What does He and She do at 22:00. Activities conducted from 21.50pm to 22.00pm by gender and cluster.

Clu.num	Sleep		Personal Care		Work		House Care		Moving		Free Time		N.
	M	W	M	W	M	W	M	W	M	W	M	W	
1	3.8	0.8	9.0	9.0	1.5	3.0	13.5	61.7	0.8	0.0	71.4	25.6	133
2	4.4	8.2	8.2	10.8	3.8	3.8	0.0	1.9	2.5	0.0	81.0	75.3	158
3	5.1	0.0	12.1	14.1	3.0	5.1	3.0	23.2	3.0	4.0	73.7	53.5	99
4	0.0	0.0	9.3	5.6	13.9	10.2	7.4	30.6	2.8	2.8	66.7	50.9	108
5	0.0	5.4	9.3	9.3	14.0	5.4	4.7	23.3	0.8	0.8	71.3	55.8	129
6	8.4	11.7	8.9	8.9	0.0	1.7	6.7	0.0	0.6	0.6	75.4	77.1	179
7	9.0	9.0	53.7	38.8	11.9	11.9	16.4	25.4	7.5	4.5	1.5	10.5	67
Total	4.4	5.5	12.7	11.8	6.0	5.0	6.6	21.5	2.1	1.4	68.3	54.8	873

These couples live mainly in large towns (over 50,000 residents) in center and northern Italy. They have children. He and She belong to the middle class (IIIa) or the working class (VI+VIIab), and they both work in the public sector.

Tab.5.4. When do He and She eat? Lunch and dinner hour by gender and cluster

cs7	Lunch		Dinner		
	men	women	men	women	Freq.
1	13:20	13:30	20:10	20:08	128
2	13:27	13:33	20:00	19:53	151
3	13:08	13:21	19:58	20:03	91
4	13:26	13:34	20:20	20:10	97
5	13:22	13:39	20:14	19:57	119
6	13:03	13:30	19:51	19:47	169
7	13:24	13:36	20:43	20:21	61
Total	13:18	13:32	20:07	20:00	816
F.	4.15**	1.22	12.81**	10.04**	

Sign: (*) $p < 0.05$; (**) $p < 0.01$

He is quite satisfied with the time for himself and with daily life. He is slightly more satisfied than Her with the time available for the partner, children, and relaxation. It is mainly She that has difficulties in reconciling daily tasks with office opening hours; while neither of them has particular difficulties in reconciling daily tasks with the opening hours of leisure centers and shops.

In **cluster 3**, He wakes up before She. At 7:00 in the morning 62.6% of men are having a shower and/or breakfast, and another 32.4% are at work or about to reach the workplace. She wakes up on average 38 minutes after Him. Both spend more time, compared to the overall mean, on personal care during the day. Lunchtime for Him is at 13:27; for Her it is at 13:33. Dinner for Him is at 20:00 while for Her it is at 19:53.

In this dual-earner couple a first main difference concerns work time. He spends 8 hours and 38 minutes on average at work. By contrast, She spends only 6 hours and 7 minutes. Moreover, compared with the other women, she is the one that spends less time at work. The second main difference concerns the time that He and She devote to house care. He is occupied for only 50 minutes a day in this activity; She 3 hours and 35 minutes at day. The third main difference concerns free time in terms of duration and moment (when) they relax. On average, He has 2 hours and 2 minutes of free time, while She has only 1 hour and 23 minutes. For Him, free time starts at 21:00 while for Her it starts later. In these couples, at 22:00 only 53.5% of women are in leisure activities against the 73.7% of men.

These couples reside in medium-sized towns in northern and central Italy. They have no children or have children aged over 14. Both are entrepreneurs, professionals or managers (I+II). He works in industry or the private services sector; She works in the public service sector.

Tab.5.5. Cluster distribution by geographic area, municipality size, children, couple's educational level, and couple's social class. (row percentage)

| | Clusters | | | | | | |
	1	2	3	4	5	6	7
Geographical area (X^2=22.8; Pr = 0.03)							
North	13.6	19.1	11.7	12.1	14.9	22.9	5.6
Center	19.0	18.4	12.9	15.1	11.2	17.3	6.2
South and Islands	15.5	16.0	9.5	10.8	17.2	18.1	12.9
Municipalities (X^2=33.6 Pr = 0.01)							
Metropolitan	17.7	16.6	11.6	12.7	20.4	13.8	7.2
over 50.000 residents	14.1	24.7	11.2	15.9	11.2	14.7	8.2
10.000-50.000 residents	18.5	13.7	12.8	10.9	15.6	21.3	7.1
lower than 10.000 residents	12.2	18.3	10.3	11.3	12.9	27.0	8.0
Children (X^2=23.5 Pr = 0.02)							
No child	11.4	14.4	13.8	19.2	13.8	16.8	10.8
Children 0-14	17.2	19.0	9.6	11.0	16.3	19.2	7.8
Children 14+	14.5	19.0	12.6	10.4	13.0	24.9	5.6
Couple school degree (X^2=17.5 Pr = 0.13)							
University	14.8	14.3	9.8	17.6	14.8	19.7	9.0
Secondary	15.9	21.0	11.8	10.2	14.5	18.3	8.3
Compulsory	14.8	17.5	12.1	10.5	15.2	24.5	5.5
Couple Social Class (X^2=35.3 Pr = 0.01)							
I+II	15.7	11.4	14.6	15.0	13.9	22.1	7.1
IIIa	16.9	23.2	9.4	10.2	16.3	16.9	7.2
IVabc	10.4	14.8	10.4	16.5	13.0	22.6	12.2
VI+VIIab	13.8	21.6	10.3	8.6	13.8	25.9	6.0
Total	15.2	18.1	11.3	12.4	14.8	20.5	7.7

Tab.5.6. Cluster distribution by education level, social class and sector of man and woman. (row percentages)

| | Cluster | | | | | | |
	1	2	3	4	5	6	7
Man school degree (X^2=18.6 Pr = 0.10)							
University	14.3	16.8	9.9	19.3	16.8	16.2	6.8
Secondary	15.1	21.5	12.0	11.0	13.9	17.7	8.8
Compulsory	15.7	16.0	11.4	10.6	14.7	24.6	7.1
Woman school degree (X^2=12.8 Pr = 0.38)							
University	15.8	16.3	7.6	16.3	14.7	20.7	8.7
Secondary	16.1	19.2	13.0	12.2	13.6	17.5	8.5
Compulsory	14.0	17.9	11.6	10.5	16.1	23.6	6.3
Man social class (X^2=44.2 Pr = 0.00)							
I+II	15.0	11.3	13.8	14.6	15.0	22.5	7.9
IIIa	19.8	24.1	8.2	11.6	15.1	14.7	6.5
IVabc	13.2	11.2	11.2	17.1	15.8	20.4	11.2
VI+VIIab	12.5	23.3	12.1	8.0	13.7	24.1	6.4
Woman social class (X^2=33.5 Pr = 0.02)							
I+II	17.0	9.4	16.0	18.9	9.4	17.9	11.3
IIIa	18.0	19.7	11.1	11.3	14.9	18.9	6.2
IVabc	9.3	16.1	8.5	17.0	14.4	25.4	9.3
VI+VIIab	11.6	20.2	11.1	8.6	17.7	22.7	8.1
Man sector (X^2=53.7 Pr = 0.00)							
Industry	14.0	19.6	12.2	8.0	11.6	27.4	7.1
Private services	16.7	11.8	12.8	16.7	18.3	15.9	7.9
Public services	14.6	28.7	6.4	11.7	13.5	17.0	8.2
Woman sector (X^2=44.6 Pr = 0.00)							
Industry	17.7	18.8	9.7	3.8	18.3	25.8	5.9
Private services	16.1	14.4	13.1	18.1	13.6	16.1	8.6
Public services	12.4	22.8	10.0	10.0	14.1	23.1	7.6
Total	15.2	18.1	11.3	12.4	14.8	20.5	7.7

Tab.5.7. ANOVA: mean score, F.test and probability of the level of satisfaction, level of stress, difficulties in reconciling daily tasks, according to husband and wife by cluster types.

	Cluster									
	1	2	3	4	5	6	7	Mean	F.test	Pr.
Satisfied with the time for:										
He: Yourself	1.63	1.85	1.84	1.86	1.63	1.89	2.16	*1.81*	3.16***	0.005
She; Yourself	1.54	1.56	1.76	1.72	1.52	1.70	1.68	*1.63*	1.35	0.234
He: Partner	1.62	1.97	1.87	1.81	1.74	1.91	2.03	*1.84*	2.48**	0.022
She: Partner	1.74	1.82	1.81	1.92	1.74	1.86	2.11	*1.84*	1.39	0.217
He: Children	1.54	2.05	1.81	1.80	1.68	1.69	2.03	*1.77*	4.23***	0.000
She Children	1.73	2.07	2.05	2.08	1.93	2.06	2.30	*2.00*	2.94***	0.008
He: Parent	2.05	1.85	1.99	1.82	1.91	1.68	1.79	*1.86*	2.02*	0.061
She: Parent	1.88	1.70	2.00	1.86	1.87	1.76	1.78	*1.83*	0.98	0.435
He: Job	2.49	2.62	2.51	2.46	2.58	2.44	2.57	*2.52*	2.14**	0.046
She: Job	2.61	2.53	2.61	2.56	2.53	2.70	2.61	*2.60*	1.77*	0.100
He: Friends, social relations	1.51	1.93	1.78	1.53	1.66	1.94	1.80	*1.75*	4.65***	0.000
She: Friends, social relations	1.62	1.70	1.76	1.81	1.76	1.74	1.70	*1.73*	0.51	0.799
He: Leisure, hobby	1.42	1.55	1.57	1.37	1.48	1.43	1.60	*1.47*	1.08	0.371
She: Leisure, hobby	1.47	1.42	1.48	1.59	1.48	1.43	1.53	*1.47*	0.61	0.726
He: Relax	1.65	2.00	1.91	2.00	1.75	1.60	2.10	*1.81*	4.86***	0.000
She: Relax	1.51	1.61	2.07	1.75	1.67	1.79	1.57	*1.71*	4.16***	0.000
Are you stressed?										
He	2.27	2.09	2.17	2.19	2.12	2.24	2.01	*2.17*	1.48	0.182
She	2.35	2.32	2.29	2.24	2.32	2.62	2.34	*2.38*	3.97***	0.001
Difficulty to reconciling daily tasks with:										
He: Work time of partner	2.29	2.16	2.26	2.31	2.22	2.43	2.27	*2.29*	1.25	0.280
She: Work time of partner	2.34	2.16	2.17	2.20	2.27	2.04	2.24	*2.19*	1.55	0.160
He: Children school hours	2.39	2.16	2.39	2.38	1.97	2.32	2.22	*2.26*	1.80*	0.097
She: Children school hours	2.40	2.09	2.20	1.97	2.18	2.21	2.00	*2.17*	1.90*	0.079
He: Office opening hours	2.69	2.24	2.34	2.39	2.23	2.55	2.24	*2.41*	4.08***	0.001
She: Office opening hours	2.45	2.24	2.17	2.26	2.44	2.16	2.21	*2.28*	1.78*	0.100
He: Leisure center opening hours	2.53	1.95	2.23	2.08	2.07	2.26	2.30	*2.20*	4.36***	0.000
She: Leisure center opening hours	2.31	1.90	1.98	2.07	2.18	1.81	2.09	*2.03*	4.34***	0.000
He: Shops opening hours	2.34	1.86	2.17	2.13	2.09	2.41	2.13	*2.18*	5.15***	0.000
She: Shops opening hours	2.33	1.95	1.98	2.05	2.12	1.80	1.98	*2.02*	4.47***	0.000
Satisfied of daily life:										
He	2.74	2.87	2.80	3.01	2.77	2.71	3.08	*2.82*	6.22***	0.000
She	2.82	2.89	2.90	2.94	2.82	2.75	2.89	*2.85*	1.88*	0.082
Satisfaction for labor division with the partner regarding:										
He: House care	3.20	3.17	3.18	3.27	3.17	3.33	3.17	*3.22*	1.38	0.219
She: House care	2.58	2.70	2.68	2.68	2.59	2.40	2.68	*2.59*	2.21**	0.041
He: Children care	3.15	3.22	3.18	3.23	3.35	3.17	3.41	*3.23*	2.14**	0.047
She: Children care	3.02	3.12	2.95	3.07	2.94	3.03	2.92	*3.02*	0.89	0.499

Legend: (*) p<0.1; (**); p<0.05; p<0.01

Both are only slightly more satisfied than couples in cluster 1 with the time that they have for the partner and children. She is very satisfied with the amount of time for relaxation: She records the highest value among women. Moreover, She does not signal particular difficulties in reconciling daily tasks with the opening hours of shops and leisure centers. By contrast, He seems to suffer more than She does. He is rather stressed and not very satisfied with daily life. He states that he has more problems in reconciling daily tasks with the opening hours of offices, shops and leisure centers.

The dual-earner couples in **cluster 4** are characterized by a quite synchronized and relaxed day. Both wake up one hour after the majority of members of other couples. 87.0% of men and 91.7% of women in this cluster are asleep at 7:00. Like the couples in cluster 2, those in cluster 4 spend a relatively short amount of time at work: He 7 hours and 27 minutes, She 6 hours and 9

minutes. Both have lunch at the same time (He at 13:26; She at 13:24). In the evening, She has dinner slightly earlier (20:10) than He (20:20).

Fig.5.3. Sequence index plot and stacked area plot of the seven clusters.

Cluster 1
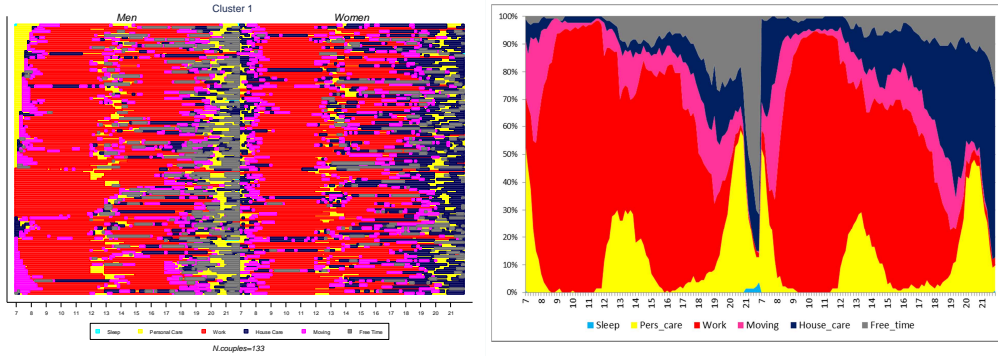
Cluster 2
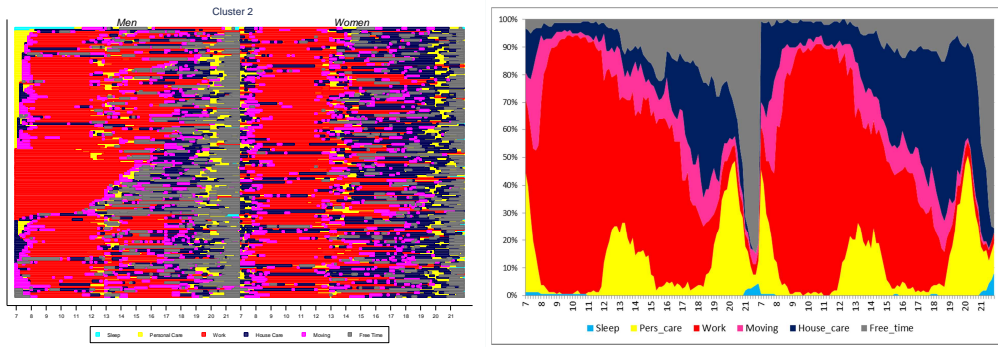
Cluster 3
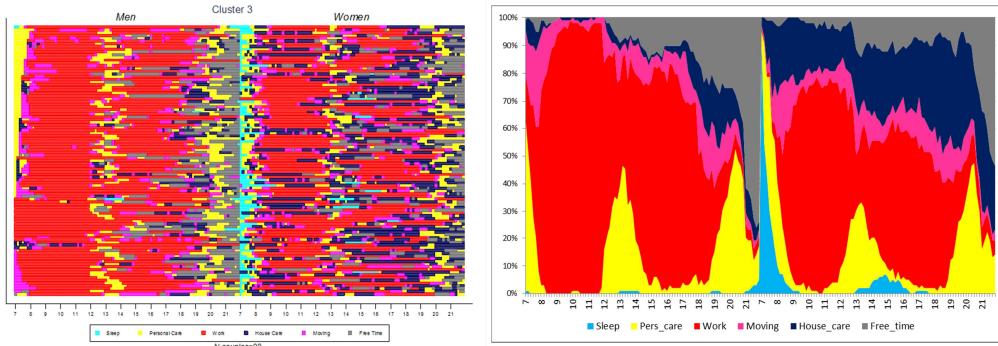
Regarding the time for house care, neither is this equally distributed. He devotes slightly more time than the overall mean of men to this activity (1 hour and 10 minutes). She devotes slightly less

time than the overall mean of women to this activity (3 hours and 27 minutes). In the evening, after 21:00, 66.7% of men and 50.9% of women are in leisure activities.

Fig.5.3. Sequence index plot and stacked area plot of the seven clusters.

Cluster 4
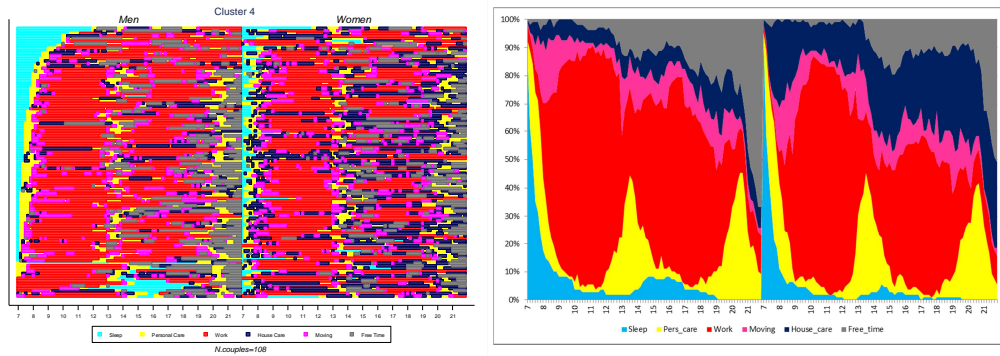
Cluster 5
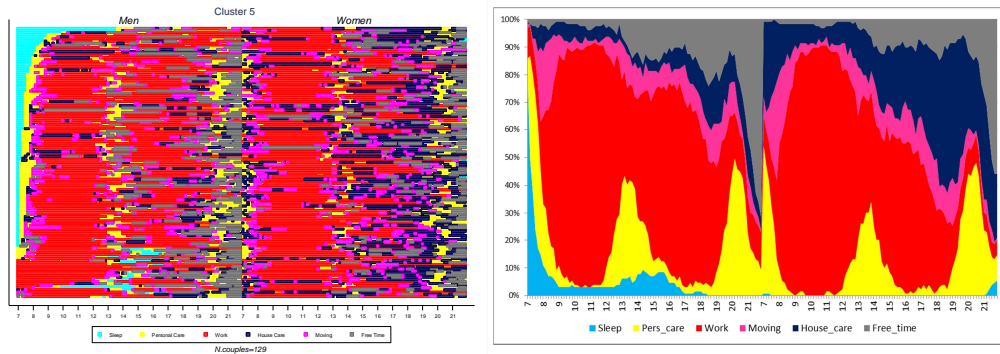
Cluster 6
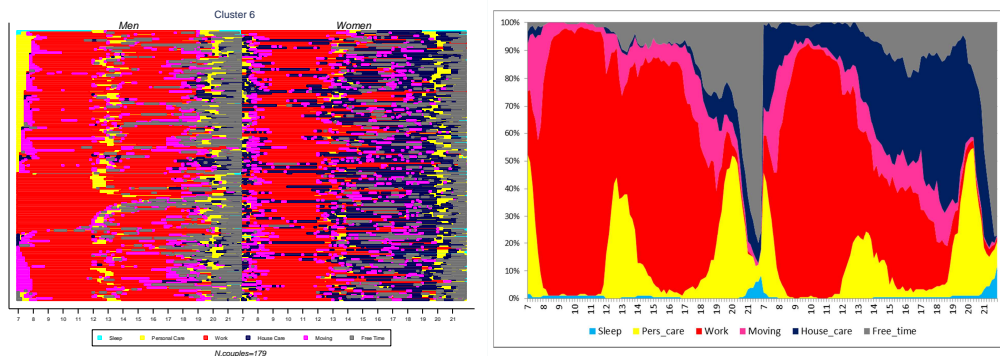
These couples live mainly in large towns of central Italy. They do not have children. Both partners have a university degree, are entrepreneurs, professionals, managers (I+II) or are members of the petty bourgeoisie (VIab) and work in the private service sector.

20

He is very satisfied with daily life and with the amount of time that he has for relaxation. Both have some difficulties in reconciling daily tasks with the opening hours of offices, shops, and leisure centers.

Fig.5.3. Sequence index plot and stacked area plot of the seven clusters.
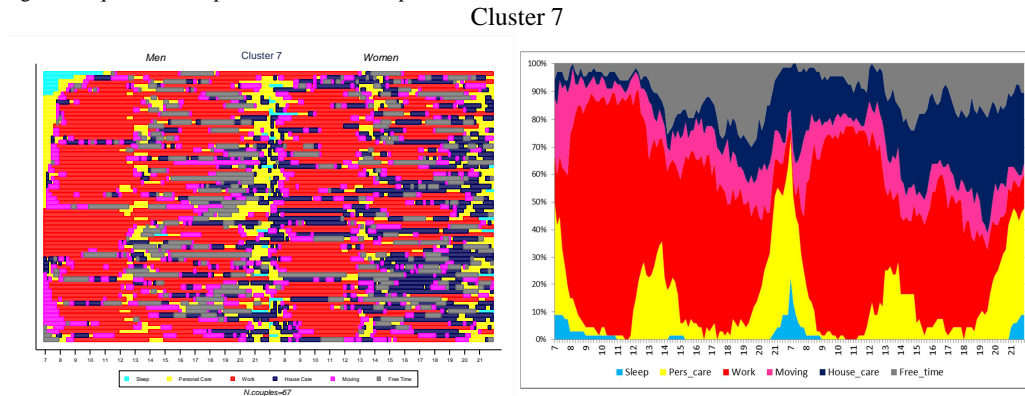
Cluster 7

In **cluster 5**, She wakes up earlier than Him: on average, 50 minutes before He does. She starts work earlier (8:00) than Him (9:00). Among women, the women in this cluster are those who work longer (6 hours and 58 minutes). Conversely, the men in this cluster work less than the overall mean of men (7 hours and 46 minutes).

He spends more time than She does on personal care: 2 hours and 10 minutes compared with 1 hour and 50 minutes. He has lunch at 13:22 and She at 13:39. Moreover, in the evening He has dinner at 20:14 and She at 19:57.

These couples record one of the most unequal time distributions of house care between men and women: He devotes 1 hour and 5 minutes to house care, while She devotes 3 hours and 25 minutes. Moreover, these couples are those that exhibit the lowest amounts of time for leisure. In total, He has only 1 hour and 55 minutes of free time, while She has only 1 hour and 18 minutes. They seem to have free time only in the evening. After dinner, 71.3% of men and 55.8% of women relax.

These couples reside in the metropolitan areas of Italy's North and South and Islands. They have one or more children aged between 0 and 14. He is a member of the petty bourgeoisie (IVab) while She is a member of the working class (VI+VIIab) or middle class (IIIa). He works in the private service sector and She in industry.

Both are very dissatisfied with the time available for the partner, children, and relaxation. Both have difficulties in reconciling daily tasks with the opening hours of offices, shops, and leisure centers. Finally, He is very unsatisfied with the amount of time that he has for himself and in general with daily life.

Both members of the dual-earner couples in **cluster 6** wakes up earlier in the morning. Half of the men are at work or are about to reach the workplace, while one third of women are involved in house care.

Like the men in cluster 3 also the men in cluster 6 spend a long time at work: 8 hours and 45 minutes. Also the women in this cluster are similar to those in cluster 3: in fact, these women spend an average of 6 hours and 7 minutes at work. Lunchtime for Him is 13:03, while for She it is around 13:30, but there is no clear time: probably she eats when she can. For both, dinner is at 19.50.

As regards house care, the unequal distribution of time between genders is maximum in this cluster. He spends only 30 minutes a day on house care, while She spends 3 hours and 59 minutes.

These couples reside in small towns (fewer than 10,000 residents) in North Italy. They have children aged over 14. Both have a low level of education (compulsory), and both are members of the working class (VI+VIIab) or petty bourgeoisie (IVab). Men work in industry, while women work in both industry and the public service sector.

They are very dissatisfied with the time available for the partner, children, and relaxation. Both have many difficulties in reconciling daily tasks with the opening hours of offices, shops, and leisure centers. Moreover, He is very unhappy about the amount of time that he has for himself and with daily life in general.

The last cluster, **the seventh**, comprises dual-earner couples who are awake at 7:00 and having breakfast. 41.8% of the men are having breakfast, and another 37.3% are at work or are about to reach the workplace. For the women in this cluster, 50.8% are having breakfast, 14.9% are cleaning the house, and another 22.4% are still asleep.

He works an average of 7 hours and 58 minutes; She works an average of 6 hours and 16 minutes. For both, their jobs are concentrated mainly in the morning. Only forty percent of women and fifty percent of man work in the afternoon.

The distinctive characteristics of these dual-earner couples are the following: (a) free time for both in the afternoon; (b) a constant proportion of men and women engaged in house care throughout the day; (c) a fragmented packaging of activities; (d) She shows one of the largest amounts of free time among women (1:36), while he has one of the smallest amounts among men (1:47).

Another characteristic of these couples is the timing of lunch and dinner, which both eat very late. He has lunch at 13:24 and She at 13:36; while He has dinner at 20:43 and She at 20:21.

These couples reside in large towns of Italy's South and Islands. They do not have children. They have a medium/high level of education (tertiary or secondary). He is a member of the petty bourgeoisie (IVab), and She is a member of the bourgeoisie (entrepreneurs, professionals, and managers). She is employed in the private service sector.

These couples show the highest level of satisfaction with the time available for the partner, children, and relaxation. They record the lowest level of difficulties in reconciling daily tasks with the opening hours of offices, shops, and leisure centers. He is very happy with the amount of time for himself, and he is not stressed.

At the end of this part of the paper, it is evident that the sequences analysis of time use diaries provides a quite clear and meaningful representation of the main patterns of daytime organization of Italian dual-earner couples. In the next paragraph, we deeply enter on generative mechanism that acts on give shape and relevance at each pattern and in defining different forms of (de)synchronization.

6. (De)synchronization. The dual-earner strategies to combine work and couple's life

We have now outlined the seven patterns of dual-earner couples' daily time organization. The foregoing discussion has singled out the different forms of daily activities' organization and their specific combinations among the spouses. It has been noticed that the distribution of activities across the day, the timing with which they are carried out, and the amount of time devoted to each activity systematically varies both between different patterns and within spouses of the same pattern. Finally, it seems clear that dual-earner couples perform same common daily strategy.

Hence, the spouses' daily life seems to develop with socially shared, recognized, and identifiable patterns of combined time use. This insight raises two further questions. The first is how these patterns result from a complex process of adaptation to both work-social-family

constraints and individual needs. The second question concerns how the daily times are combined by the spouses, and how their performed combinations are random instead of being regulated by common generative mechanisms.

Tab.6.1. Multinomial logistic regression on the seven clusters by age of woman, geographic area, presence of children, sector, level of education and social class of the man and woman. (Weight parameter)

	Cluster 2		Cluster 3		Cluster 4		Cluster 5		Cluster 6		Cluster 7	
	β	$\sigma(\beta)$	β	$\sigma(\beta)$	β	$\sigma(\beta)$	β	$\sigma(\beta)$	β	$\sigma(\beta)$	β	$\sigma(\beta)$
Age												
Woman	0.08***	0.031	0.03	0.030	0.04	0.029	0.03	0.030	0.04	0.028	0.05*	0.030
Area												
North[+]												
Center	-0.20	0.365	-0.60	0.408	-0.08	0.405	-1.08***	0.412	-0.23	0.424	-0.30	0.532
South & Island	-0.37	0.373	-0.33	0.392	0.07	0.388	-0.18	0.360	-0.29	0.352	0.89**	0.424
Children												
No child[+]												
Children 0-14	-0.43	0.455	-1.40***	0.462	-1.28***	0.436	-0.41	0.448	-1.04**	0.454	-0.80	0.510
Children 14+	-0.95*	0.589	-1.54***	0.594	-1.60***	0.581	-1.00*	0.607	-1.13**	0.542	-1.92***	0.701
He sector												
Industry[+]												
Priv. services	-0.55	0.373	0.09	0.394	0.15	0.405	0.24	0.377	-0.11	0.384	-0.34	0.446
Public services	0.43	0.435	-0.76	0.524	0.26	0.554	-0.11	0.468	0.07	0.481	0.01	0.602
She sector												
Industry[+]												
Priv. services	0.30	0.405	0.41	0.452	1.51***	0.529	0.01	0.423	0.16	0.418	0.48	0.532
Public services	0.77*	0.430	1.02**	0.486	1.15**	0.600	0.52	0.446	1.07**	0.469	0.99*	0.580
He school degree												
University[+]												
Secondary	-0.30	0.437	-0.33	0.590	-0.68	0.493	-0.49	0.454	0.42	0.481	0.40	0.728
Compulsory	-1.16**	0.551	-0.27	0.637	-0.35	0.581	-0.50	0.534	0.24	0.564	0.22	0.834
She school degree												
University[+]												
Secondary	0.07	0.396	1.02*	0.579	0.00	0.480	0.02	0.429	-0.21	0.431	-0.26	0.632
Compulsory	0.09	0.528	1.25**	0.646	-0.08	0.564	0.10	0.549	-0.86*	0.532	-0.27	0.851
He social class												
I+II[+]												
IIIa	0.36	0.426	-0.84*	0.508	-0.65	0.458	-0.13	0.414	-0.92**	0.418	-0.54	0.547
IVabc	0.54	0.571	-0.13	0.572	0.29	0.530	0.33	0.570	0.02	0.504	0.70	0.579
VI+VIIab	1.23***	0.481	-0.36	0.503	-0.55	0.530	-0.06	0.500	0.38	0.451	0.10	0.565
She social class												
I+II[+]												
IIIa	0.73	0.584	-0.89*	0.496	-0.56	0.516	0.01	0.566	-0.05	0.490	-0.82	0.538
IVabc	1.52**	0.758	-0.62	0.716	0.00	0.673	0.47	0.771	1.41**	0.648	0.22	0.734
VI+VIIab	1.73**	0.695	-0.38	0.608	0.19	0.644	0.90	0.676	1.57**	0.630	0.64	0.733
Constant	-4.05**	1.444	-0.56	1.433	-1.14	1.447	-0.59	1.421	-1.13	1.276	-2.52*	1.557

Legend: (+) reference category; (*) p<0.1; (**) p<0.05; (***) p<0.01
Reference cluster (1)
Pseudo R2 = 0.09

In this regard, our hypothesis is that these time-use patterns result from the complex co-action of individual, family and social factors whose combination defines the relevance and the shape of patterns. The time balance within His and Her activities, as well as its configuration across the day, is not random; rather, it changes according to multiple latent factors. The first of these factors is work and its schedules, and therefore mainly the type of job and the economic sector (Hamermesh, 2002; Warren, 2003; Lesnard, 2008). The second is the spouses' and family's socio-demographic features – like the age of the spouses, the presence of children, and the geographic area of residence. The third is the cultural level of the cohabitation in terms of both the educational level of the

spouses and the level of predisposition toward egalitarian gender attitudes (Hakim, 2003; Oláh et. al, 2014). We maintain that all these dimensions contribute to defining the patterns of couples' daily activities.

Tab.6.2. Predicted probability of the multinomial logistic regression on the seven clusters. Probability change over the value of age of woman, geographic area, presence of children, sector, level of education and social class of the man and woman at mean of the others parameters. (Weight parameter)

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7
Age Woman							
30	24.5	9.4	11.4	11.5	16.4	21.1	5.7
40	17.2	14.9	11.4	12.2	15.0	22.4	6.9
50	11.5	22.2	10.7	12.2	12.9	22.5	8.0
60	7.2	31.4	9.5	11.5	10.5	21.2	8.6
Area							
North[+]	14.9	15.6	12.6	10.9	17.9	22.5	5.5
Center	20.5	17.6	9.5	13.8	8.3	24.6	5.6
South & Island	16.3	11.7	9.9	12.8	16.3	18.4	14.6
Children							
No child[+]	8.3	12.1	16.9	17.4	11.7	25.1	8.5
Children 0-14	18.3	17.3	9.1	10.6	17.0	19.4	8.4
Children 14+	24.6	13.9	10.7	10.3	12.8	24.0	3.7
He sector							
Industry[+]	16.2	17.4	12.3	10.5	13.2	22.5	7.8
Priv. services	17.2	10.6	14.2	13.0	17.8	21.4	5.9
Public services	15.3	25.2	5.4	12.8	11.2	22.8	7.4
She sector							
Industry[+]	24.0	15.0	9.7	5.9	18.1	21.3	5.9
Priv. services	17.4	14.6	10.6	19.2	13.2	18.2	6.8
Public services	11.4	15.4	12.9	8.9	14.4	29.6	7.5
He school degree							
University[+]	13.6	23.1	11.8	15.0	18.1	14.0	4.4
Secondary	15.9	19.9	9.9	8.8	12.9	24.8	7.8
Compulsory	18.2	9.7	12.0	14.1	14.6	23.9	7.4
She school degree							
University[+]	16.0	13.8	4.5	12.0	13.5	32.0	8.2
Secondary	15.8	14.6	12.2	11.8	13.7	25.7	6.3
Compulsory	17.3	16.3	16.8	12.0	16.2	14.6	6.8
He social class							
I+II[+]	15.9	8.2	15.4	15.8	14.2	23.8	6.7
IIIa	23.3	17.2	9.7	12.0	18.3	13.8	5.7
IVabc	13.0	11.6	11.1	17.2	16.2	19.9	11.1
VI+VIIab	13.3	23.5	9.0	7.7	11.2	29.1	6.2
She social class							
I+II[+]	18.1	6.2	23.1	16.8	12.0	14.2	9.6
IIIa	22.7	16.0	11.9	12.0	15.2	16.9	5.3
IVabc	11.0	17.1	7.6	10.2	11.7	35.2	7.3
VI+VIIab	8.8	16.9	7.7	9.9	14.4	33.4	8.9
At Mean	16.7	15.4	11.4	12.2	14.9	22.5	7.0

We have already pointed out the limitations of a time-budget approach to the study of couples' daily time-use strategies. At the same time, we noted that the sequence analysis approach has been underestimated in the literature (Lesnard, 2004): to our knowledge, no efforts have been made to study the daily scheduling of multiple activities from a holistic and integrated perspective. In other words, we do not know how different daily activities are integrated into a single schedule and how spouses combine their activities each other. If we want to study daily work-family strategies, we need to preserve the integrity of the whole process and the interaction of different activities among the spouses. Indeed, from a time-integrity perspective, the study of the differences among couples'

time-use organization requires the ability to observe the activities' combination of Him and Her for each point-in-time and simultaneously.

In order to investigate the complex process of adaptation to both work-social-family constraints and individual needs of dual-earner couples' daily time organization we performed two separate analyses on the seven time-use patterns. The first analysis used a multinomial logistic regression model to verify if these patterns resulted from working-social-familial and individual constraints (Tab.6.1 & 6.2). The second consisted of a graphic representation of the change in the combination of His and Her modal state/activity during the entire day, at each point-in-time (10-minute intervals).

For this graphic analysis, we considered all the 36 possible combinations of activities (six for each spouse). For each cluster and for each point-in-time, the most frequent activities combination was identified. On this criterion, only 16 of all the possible combinations was found to be frequently performed by the couples, suggesting a certain routine by couples in everyday life (Hägerstrand, 1982; Hellgren, 2014). The graphs (Fig.6.1 – Fig.6.7) depicted the sequence of activities' combination for each cluster of dual-earner couples across the entire day (07.00-22.00).

A jointly reading of the modal multichannel sequence graphs and the multinomial logistic regression parameters quite clearly shows what are the (de)synchronization strategies adopted by couples and what may be the hidden generative mechanisms (Hallberg, 2002).

In particular, three different forms of time-use organization are highlighted by the graphs. The first is characterized by a general *synchronization* of the different spouses' activities during the days. We recall that some scholars suggest (Hamermesh, 2002; Lesnard, 2008) that this solution is expected to be the one preferred by dual-earner couples. However, the synchronization of the spouses' activities may be considered as one of the most complex forms of time organization, since it requires the alignment of social-work-individual times and constraints at a couple-level. Thus, even if it is the most desirable solution, not all the spouses are able to synchronize themselves during the day.

Couples in clusters 2 and 4 are associated with the highest synchronization levels. These dual-earner households share some specific features (Tab.6.1 & 6.2): the high level of education for Him – tertiary education and the medium-lower level of education for Her. Moreover, men in clusters 2 and 5 are employed in the services sector. Furthermore, couples from these two clusters mostly live in Central Italy.
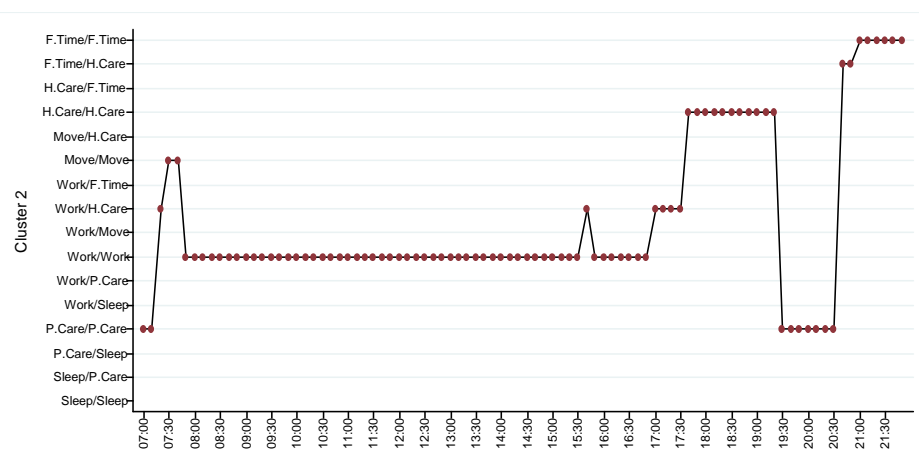
Fig.6.1. Modal (de)synchronized couples activities: cluster 2

25

What seems to distinguish the two clusters is that the probability of being a member of cluster 2 increases with the woman's age and with the presence of young children (0-14), while for the couples in cluster 4 the woman's age does not matter and they are more likely not to have a child. There are also differences of occupational sector and class between the couples in clusters 2 and 4. While both the spouses of cluster 4 are more likely to be employed in the private service sector, those of cluster 2 mostly work in the public sector or, secondly, in industry. As regards differences in occupational class, we can say that men in cluster 2 are white collar (IIIa) or workers (VI + VIIab), those in cluster 4 are self-employed (I+II and IVabc). At the same time, women in cluster 2 are mainly clerical workers (IIIa), while in cluster 4 they are more likely to be self-employed (I + II).

This particular combination of characteristics – and constraints – creates synchronized couples' sequences, as mentioned before (Fig.6.1). However, there are some substantive differences. In the case of cluster 2, he and she seem to have breakfast together before going to work, and they start working synchronically. They both stop working quite early in the afternoon, probably favored by their kind of job and the economic sector in which they are employed (mainly in the public services). At 17:00 as later she is back at home, while he follows her shortly thereafter, at 17:40. From that moment, both the spouses spend the rest of the day at home and engage in child care, before having dinner together and, finally, enjoying their free time to relax at the end of the day. In the second part of the day, the only moment in which spouses are not synchronized is immediately after dinner, when she postpones her free time for 20 minutes in order to do some rapid house care.
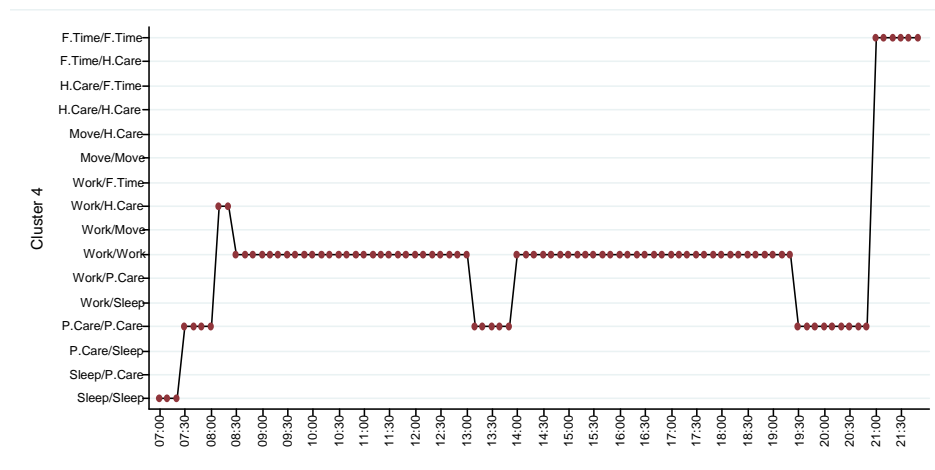
Fig.6.2. Modal (de)synchronized couples activities: cluster 4

In cluster 4 the absence of children and the type of work induces the couples to start their day in a different way compared with the others (Fig.6.2). Both the spouses wake up together, and they do so later than the other couples. They also have breakfast at the same time. Then, He leaves the house while She quickly tidies up before going to work. Job commitments equally fill most of their daily time. Moreover, they are synchronized in both their lunch and dinner times. Finally, and because of the pressure and extent of job commitments, spouses in cluster 4 jumps directly to free time and leisure, frequently avoid any kind of house care task. In general, house care activities seem to be almost absent within this daily time-use pattern. However, to be noticed is that this it doesn't mean that spouses haven't done housework, but rather that they are more likely to do it non-

26

regularly, during brief and scattered moments of spare times; although they may not necessarily do the housework every day, maybe postponing these tasks to the weekend.

Also cluster 7 falls – although not completely – into the synchronized time-use pattern (Fig.6.3). On the one hand, the strategy of dual-earners in this cluster has some elements in common with those of cluster 2 and 4 (Tab.6.1 & 6.2). In particular, as for cluster 2, couples are more likely to be part of cluster 7 with the increasing age of Her, and Her employment in the public sector. Moreover, like those in cluster 4 these couples share the feature that both spouses are mostly self-employed (He IVab; She I + II). Like both the previous clusters, they also do not have children or, at least, have children younger than 14. On the other hand, a distinctive feature of this group – compared to the other two with a synchronized time-use pattern – is the strong presence of couples from the South of Italy.
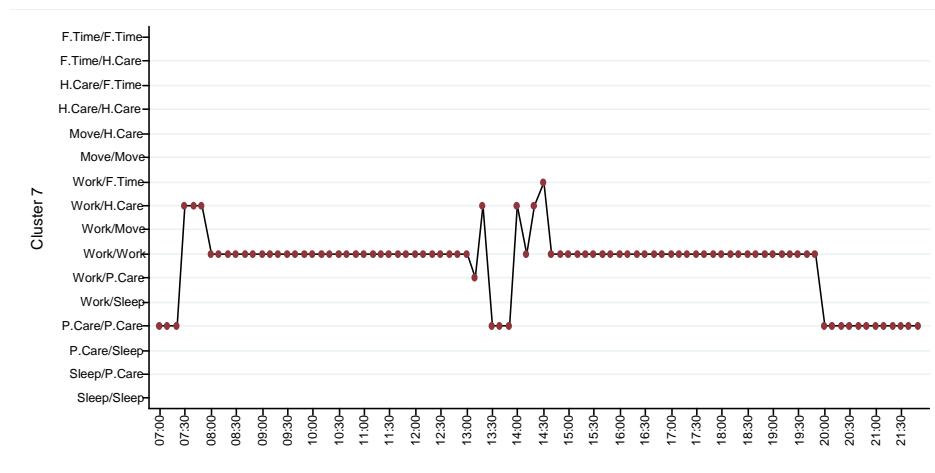
Fig.6.3. Modal (de)synchronized couples activities: cluster 7

Cluster 7 has some characteristics in common with the 2nd and the 4th also in terms of daily time organization. Both the spouses have breakfast together, then He goes to work while She quickly rearranges the house: the same dynamic is exhibited by cluster 4, only that it is shifted earlier in the morning because of their different wake-up times. Also in the evening, clusters 7 and 4 are similar in that both these couples synchronically return home later. However, the 7th time-use pattern ends with a longer tail of synchronized personal care: spouses may still be having dinner together at the end of the observation (22:00). In any case, what really makes cluster 7 unique is the time organization around lunch. While for cluster 2, there is no specific time for lunch, and for cluster 4 the time interval for lunch is well defined between two work 'segments', for cluster 7 the break from work is longer for Her; and around a certain synchronized lunch-time, there is a certain desynchronization due to His work commitments and Her house care tasks. Finally, before going back to work, She is even able to spend a short time relaxing. Here, the sequence of activity combinations over time is much more chaotic, fragmented and socially desynchronized compared with the other patterns. However, this desynchronized part of the day seems functional to the production of a certain form of a general, mostly synchronized, daily couple strategy.

Among all the seven clusters, finally, dual-earner couples in the 2nd, 4th and 7th are the only ones to report little difficulties in balancing daily activities. They also have relatively low levels of stress, being more satisfied with their daily life and the division of house and child care demands with the partner.

Alongside the *synchronized patterns* there emerge other desynchronized daily time-use patterns. These strategies of desynchronization seem to be specialized into two forms, on the basis of the kind of tasks sequentially performed and combined by the two spouses during the day. As expected by previous scholars, women are more engaged in household care activities (Gershuny et al., 1988; Raley et al., 2012; Craig et al. 2014;), and there are important differences in how they organize and distribute their care tasks during the day, according to their time and that of the partner.[11] Here, an important role is played by the kind of job and the work hours of the woman.[12]

The first strategy is a *functionally desynchronized* pattern of the spouses' different activities during the days. This is apparently more desirable than others. Here, gender differences in activities-in-time seem to be an adaptation to the *structural desynchronization* (Nock et al., 1984) of couples' working schedules. The difference in work duration between men and women appears to produce a counterbalancing force by which – at the end of the work day – the woman compensates the different spread of work commitments of man's with house care in a quite calibrated way that preserves the free time of both the spouses.
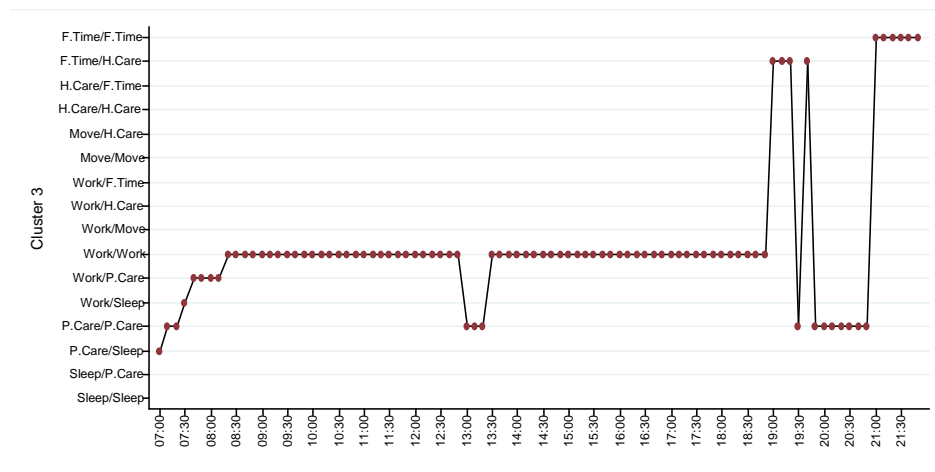
Fig.6.4. Modal (de)synchronized couples activities: cluster 3

Couples in clusters 3 and 5 are associated with the highest *functionally desynchronized* pattern levels. These dual-earner households share some specific features. They have similar educational levels: the men are either graduates or have a compulsory education, while the women have a medium-low level of education, and therefore upper-secondary at most. Both the spouses belong to the middle class or bourgeoisie. They both live mostly in the North of Italy, even if some of the couples in cluster 5 are also from southern metropolitan areas. Moreover, there are two substantial differences between the compositions of these two clusters. The first is the presence of young children within the household of cluster 5. The second is the different time at which She starts work in the morning and, consequently, the time when She comes back home in the evening.

---

[11] The problem is that not enough attention has been paid to their partners and what He does while She performs care activities. In fact, being at work instead of watching television while She is dealing with housework or child care is substantially different in terms of gender inequalities.
[12] The literature shows that, frequently, women are employed in shorter or facilitated work hours so that they can devote themselves more – in spite of their wishes – to household care (Bernardi, 1999a). Previous scholars have shown that women's permanence in the labor market in Italy, even after the birth of children, is often higher in the public sector. This may be seen as the best solution for them to combine work hours with their social role as mothers and care givers (Bernardi, 1999a)

In cluster 3, He starts work much earlier than Her. On the other hand, She prepares herself calmly before going out to work. That is possible also because of the absence of children care. At the end of the workday, the spouses come back home later and synchronically. Once at home, they desynchronize themselves again; while He takes a break to relax, She does some housework. It seems as if there is some sort of compensation of daily time activities. He started work much earlier than Her in the morning, and once returned home He perhaps believes that he has the right to get back the free time that She gained in the morning. Finally, they both eat and relax together.
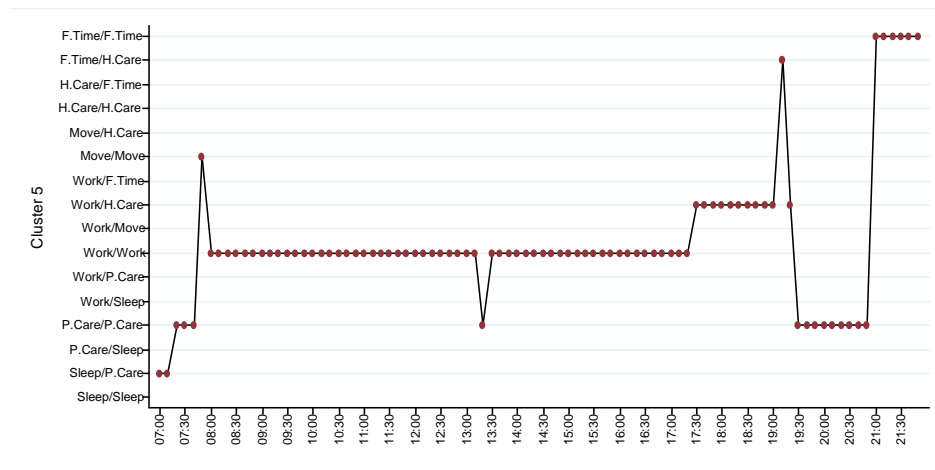
Fig.6.5. Modal (de)synchronized couples activities: cluster 5

In the cluster 5 couples' time-use pattern, She wakes up a little before Him, probably because of the presence of child care demands. They then have breakfast together before going to work, and they start working synchronically. In the afternoon, She leaves the workplace much earlier than Him, perhaps in orders to devote herself to house and child care. Soon after His return from work, they eat together, before spending synchronous free time. Compared to cluster 3, the clear non-cooperation of Him in the household tasks – among the spouses of cluster 5 – may be due to the different spread of work commitments during the whole day.

For these two last clusters, the household activities' overload for Her and the less time spent 'doing the same things' have an effect on the satisfaction expressed by the spouses. Compared to the well-synchronized couples, for those in cluster 3 and 5 we notice a reduction in the levels of satisfaction, as well as an increased difficulty of balancing the work-family activities. However, the reported levels of stress for women in cluster 3 and 5 are lower than the overall mean and slightly higher than those expressed by the wives of synchronized clusters.

The second *desynchronizations* strategy is a *dysfunctional* pattern of the different spouses' activities during the day. Here, the couple's distribution of activities during the day does not seem to follow any compensatory mechanism. The overall day desynchronization is less structural and due to working schedule commitments. It seems to be more weakly linked to the spouses' different time constraints: conversely, it appears to be an outcome of more traditional and less equal gender attitudes. Here, the result is a marked overload in paid + unpaid work for the women (Mattingly et al., 2003), with stronger evidence of the leisure gap (Beblo et al., 2008).

Couples in clusters 1 and 6 are associated with the highest *dysfunctional desynchronizations* pattern of the different spouses' activities during the day. Also these patterns share some specific features. The men in clusters 1 and 6 have low levels of education, mainly compulsory level, and spouses are parents of at least one child and that they live mostly in Central Italy.

These two clusters partially differ for the occupational class of the spouses. Those in cluster 1 are both from the middle class (IIIa), She is employed in the industry sector, while He works in the private service sector. Spouses in cluster 6 are mostly workers (VI + VIIab) or self-employed (He: I + II; She: IVabc). Women in the two clusters also differ in their educational level: those in cluster 1 have a lower-secondary education, while those in cluster 6 have mostly an upper-secondary or tertiary education.
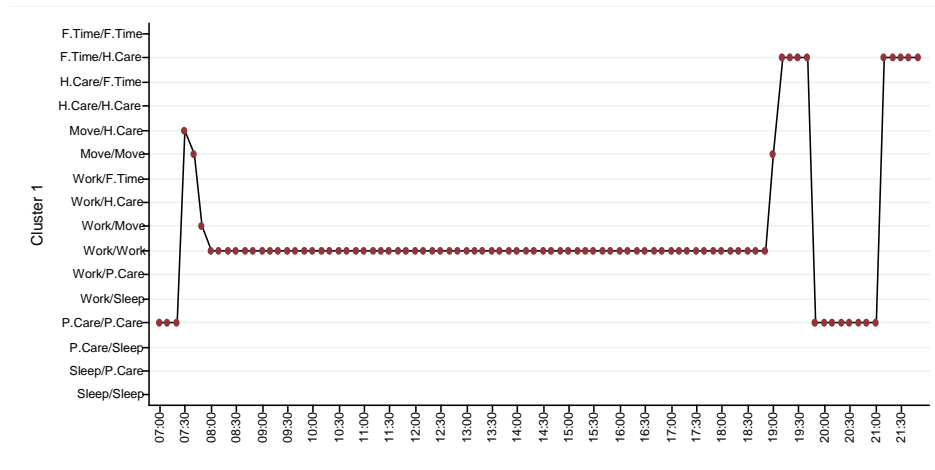
Fig.6.6. Modal (de)synchronized couples activities: cluster 1
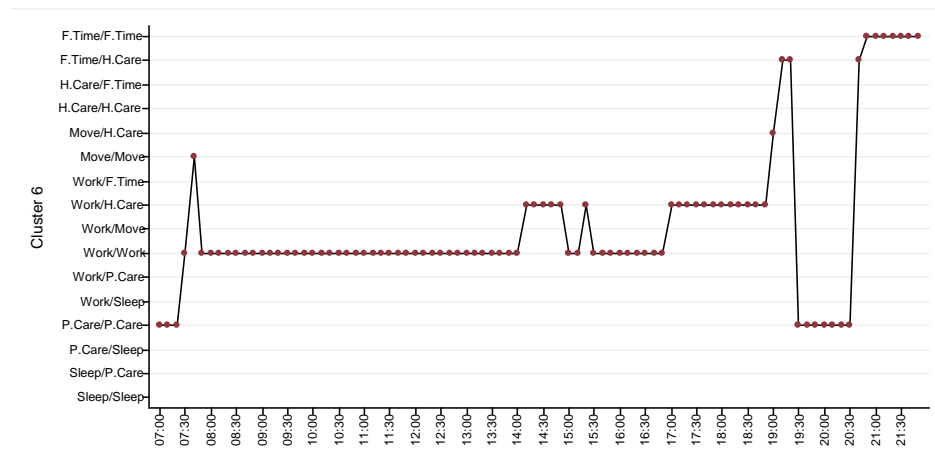
Fig.6.7. Modal (de)synchronized couples activities: cluster 6

In some way, the time-use pattern of cluster 6 is apparently similar to that of cluster 5. In fact, She comes back home before Him, dealing with house care activities. However, compared with cluster 5 we notice a greater extension of Her household commitments, from the early afternoon until the evening, when He has already returned home from work. Thus, if on the one hand the desynchronization is functional for the long time spent by Him at work, on the other, this couple's

time-use pattern does not show any cooperative or compensatory forms of time-use organization among the spouses.

Last but not least, cluster 1 is certainly the maximum expression of what we call 'dysfunctional desynchronization'. The relative time-use pattern describes a couple in which everything is on the shoulders of the woman. The delay of the exit from home is followed by a double move, probably due to the fact that – before going to work – She takes the children to school. Then, she continues to work until the late afternoon. Finally, when both the spouses return home, He takes a break and rests, while She continues to do housework and child care. The only synchronized moment in the final part of this couple's pattern is when they have dinner. Among all the time-use patterns, this is certainly the one with the highest level of gender inequality in regard to the daily work-family balance challenge.

There are clearly some differences between these two last clusters. However, they are both characterized by the total absence of His cooperation in the house and child care demands. Thus, strong implications regarding the levels of satisfaction and the ability to combine different daily activities are expected. Not surprisingly, both cluster 1 and 6 present the highest levels of difficulty in reconciling daily activities. They also have the lowest level of satisfaction with regard to the division of care tasks and the lowest levels of satisfaction with daily life as a whole. Finally, women in cluster 1 present the lowest levels of satisfaction and the greatest daily difficulties in all the areas investigated.

7. Conclusions

In this paper we study the workdays of dual-earner couples. As already argued, such a two-fold choice has important substantive implications for the exploration of combined time-use patterns: considerable parts of these days are *obligated* by both the amount of work time and its scheduling – an unavoidable constraint for the other activities in terms of quantity and timing as well (Lesnard et al., 2009). We may say that couples' daily strategies in regard to work-family balance are more likely to be settled around the working schedule's obligations by organizing the other activities within out-of-work times.

What happens around the working schedules is not expected to be random, and the activities of the spouses should not be randomly combined during the non-work parts of the day. We have argued that daily strategies and 'projects' follow certain routine pathways (Hägerstrand, 1982; Hellgren, 2014) and that temporal patterns in time use may spotlight the hidden generative mechanisms behind couples' strategies (Hallberg, 2002). Thus, in the study of workdays and dual-earner couples' dynamics, a crucial point is to find regularities.

Scholars have pointed out an important discriminating factor for dual-earners' work-family balance: that of being (de)synchronized. However, by adopting a time-budget perspective, we are unable to capture the timing dimension of (de)synchronizations. Within this framework, we may know the total duration of (de)synchronized times, but we cannot assess 'when' spouses have done the same activity. We may know the amount of time couples spent together in the same place doing different or similar activities, but, again, we do not know 'when' they did it. Thus, although this approach may allow very detailed descriptions of several sub-activities and time quality, this focus on quantities is not informative about the overall strategies of daily time organization performed by couples.

According to Lesnard (2004), a consistent alternative is to preserve the time dimension by considering daily *schedules as sequences* (Lesnard, 2004). The problem is that from a sequence analysis perspective, the entire complexity of daily schedules has been basically reduced to working schedules. Consequently, spouses' (de)synchronization-in-time has been treated as a matter of being both working or not working for each point-in-time. Here, the possible combinations of states are reduced to a tripartite scheme: 1) spouses are both working, 2) only one is working, 3) neither is

31

working. On this view, we can only assess the synchronization for one activity, that of work, by treating the possible (de)synchronizations on other activities-in-time as vague residuals. Moreover, we are not able to distinguish the different combinations of activities when the spouses are being desynchronized. This, in our view, is a strong weakness: it is – at least – reasonable to suppose that different combinations of activities are more or less desirable at different times, according to the overall context of previous and next performed activities. It is possible that some desynchronized daily schedules may be more or less complementary and *functional*, according to both the spouses' roles and commitments during the whole day.

Finally, we should move to a more complex framework. We must be able to capture both the timing of multiple synchronizations within different activities and the experienced variety of activity combinations for desynchronized time intervals.

The main contribution of this paper is its use of a 'multichannel' sequence analysis approach for the simultaneous exploration of multiple-activity schedules at a couple level. We analyze typical daily schedules of dual-earner Italian couples during a weekday, from 07.00 to 22.00, by jointly considering the combinations-in-time of activities within six different domains: 'work', 'sleep', 'personal care', 'moving', 'house care' and 'free time'. Thus, a whole view of the couple's daily organization is proposed. In fact, previous findings may help us with deeper interpretation of certain daily phenomena among couples. At the same time, we believe that a multichannel sequence analysis can yield new insights by itself.

In our analysis seven different clusters of couple's time-use patterns have been identified. What clearly emerges from the analysis is a time use organization among dual-earner couples that describe a more complex reality than that we have been used to point out. Where three main types of time use pattern have been found: *synchronization*; *functional desynchronization; dysfunctional desynchronization*. These patterns describes a variegate set of work-family balance strategies performed by dual-earner couples, with reasonable different expected levels of desirability. Moreover, these patterns are associated with socio-demographic, educational, cultural, and work characteristics of both the spouses – thus, with different latent mechanisms of time constraints. Finally, these particular couple's solutions in daily scheduling affect spouses' level of satisfaction as an outcome of the daily life quality.

References

Abbott, A.
1983 *Sequences of Social Events,* in «Historical Methods», 16, 129-47.
1988 *Transcending General Linear Reality,* in «Sociological Theory», 6, 169-86.
1990a *Conceptions of Time and Events in Social Science Methods,* in «Social Science History», 23, 140-50.
1990b *A Primer on Sequence Methods,* in «Organization Science», 1, 373-92.
1992 *What Do Cases Do?* Pp. 53-82 in *What Is a Case?* a cura di C. Ragin H. Becker, Cambridge, Cambridge University Press.
1995a *Comment on Dijkstra and Taris: The Scope of Alignment Methods,* in «Sociological Methods & Research», 24, 232-43.
1995b *Sequence Analysis,* in «Annual Review of Sociology», 21, 93-113.
1997a *Of Time and Place,* in «Social Forces», 75, 1149-82.
1997b *On the Concept of Turning Point,* in «Comparative Social Research», 16, 89-109.
1998 *The Causal Devolution,* in «Sociological Methods & Research», 27,148-81.
1999 *Temporality and Process in Social Life,* in *Social Time and Social Change*, a cura di F. Engelstad and R. Kalleberg, Oslo, Scandinavian University Press, 28-61.
Abbott, A e Barman, E.

1997 *Sequence Comparison via Alignment and Gibbs Sampling* in «*Sociological Methodology*», 27, 47-87.

Abbott, A. & DeViney, S.

1992 *The Welfare State as Transnational Event* in «*Social Science History*», 16, 245-74.

Abbott, A. & Forrest, J.

1986 *Optimal Matching Methods for Historical Data* in «*Journal of Interdisciplinary History*», 16, 473-96.

Abbott, A. e Hrycak, A.

1990 *Measuring Resemblance in Social Sequences* in «*American Journal of Sociology*», 96, 144-85.

Abbott, A. e Tsay, A.

2000 *Sequence analysis and optimal matching methods in sociology* in «Sociological Methods and Research», **29**,1, 3-33.

Beblo, M.; Robledo, Julio R.

2008 *The wage gap and the leisure gap for double-earner couples*, Journal of Population Economics, 21(2)

Becker, G. S.

1964 *Human Capital: A Theoretical and Empirical Analysis*, New York, NBER.

1981 *A Treatise on the Family*, Cambridge (Mass.), Harvard University Press.

1991 *A Treatise on the Family: Enlarged Edition*, Cambridge (Mass.), Harvard University Press.

Bernardi, F.

1999a *Does the Husband Matter ? Married Women and Employment in Italy*, in «European Sociological Review», 3, 285-300.

1999b *Donne fra famiglia e carriera. Strategie di coppia e vincoli sociali*, Milano, Franco Angeli.

2000 *The Employment Behaviour of Married Women in Italy*, in *Careers of Couples in Contemporary Society*, a cura di H.-P. Blossfeld e S. Drobnic, Oxford, Oxford University Press.

Bernasco, W.

1994 *Coupled Careers : Effects of Spouse's Resources on Success at Work*, Amsterdam, Thesis Publishers.

Bernasco, W.; De Graff, P.M. e Ultee, W.C.

1998 *Coupled Careers : Effects of Spouse's Resources on Occupational Attainment in the Netherlands*, in «European Sociological Review», 1, 15-31.

Bianchi, Suzanne M.; Lesnard, L.; Nazio, T.; Raley, S.

2014 *Gender and time allocation of cohabiting and married women and men in France, Italy, and the United States*, Demographic Research, 31(8)

Bielby, W.T. Bielby, D.D.

1986 *Family Ties: Balancing Commitments to Works and Family in dual Earner Household.* American Sociological Review, 54 (5), 776-789

Bison I,

2014 *Sequence as Network: an attempt to apply network analysis to sequence analysis*. in Philippe Blanchard, Felix Bühlmann and Jacques-antoine Gauthier (a cura di), *Advances in Sequence Analysis: Methods, Theories and Applications*: Springher International Publishing,

2011a '*Lexicographic index: A new measurement of resemblance among sequences*', in M. Williams and P. Vogt (eds), The SAGE Handbook of Innovation in Social Research Methods, London, SAGE, pp. 422-441

2011b, *Education, Social Origins and Career (Im)Mobility in Contemporary Italy: A holistic and categorical approach*, in European Societies, v. Volume 13, n. Issue 3 (2011), p. 481-503

2009 *OM matters: the interaction effects between indel and substitution costs* in Methodological Innovation On Line, 4 (2), p. 53-67

2006 *When She Helps Him to the Top*, Intergenerational transmissions: cultural, economic or social resources? RC28 Spring meeting Nijmegen, 11-14 May, 2006.

Bison I, E. Rettore, A. Schizzerotto,

2010 *"La riforma Treu e la mobilità contrattuale in Italia. Un confronto tra coorti"* in D. Checchi (a cura di), Immobilità diffusa - Perché la mobilità intergenerazionale è così bassa in Italia, Bologna: Il Mulino, p. 267-296

Bison, I.; Pisati, M. e Schizzerotto, A.

1996 *Disuguaglianze di genere e storie lavorative*, in *Genere. La costruzione sociale del maschile e del femminile*, a cura di S. Piccone Stella e C. Saraceno, Bologna, Il Mulino, 1996.

Bison, I; Esping-Andersen, G. K.

2000 *The Life Sequences in Italy*. "Polis meeting", Juan March Institute, Madrid, 10–11 March.

Bittman, M.; Wajcman, J.

2000 *The rush hour: The character of leisure time and gender equity*, Social Forces, 79(1)

Blossfeld H.P. (eds.)

1995 *The new role of women : family formation in modern societies* Boulder, Colo. [etc.] : Westview press

Blossfeld, H.P. Drobnic, S. (eds.)

2003 *Careers of Couples in Contemporary Society: From Male Breadwinner to Dual-Earner Families*, Oxford, Oxford University Press.

Blossfeld, H.P. Hakim C.

1997 *Between equalization and marginalization. Part-time working women in Europe and the United States of America*, Oxford University Press, Oxford

Budig, Michelle J.; Folbre, N.

2004 *Activity, proximity or responsibility? Measuring parental childcare time*, in Family time: The social organisation of care, edited by N. Folbre and M. Bittman, London: Routledge

Chenu, A.; Robinson, John P.

2002 *Synchronicity in the Work Schedules of Working Couples*, Monthly Labor Review, 125(4)

Craig, L.

2006 *Does father care mean fathers share? A comparison of how mothers and fathers in intact families spend their time with children*, Gender & Society, 20(2)

Craig, L; Powell, A.; Smyth, C.

2014 *Towards intensive parenting? Changes in the composition and determinants of mothers' and fathers' time with children 1992-2006*, British Journal of Sociology, 65(3)

Dijkstra,W. Toon T.

1995 *Measuring the Agreement Between Sequences.* Sociological Methods & Research 24 (2): 214-31

Drobnic, S.

1997 *Part-Time Work in Central and East European Countries.* In Blossfeld, H.P. Hakim C. *Between equalization and marginalization. Part-time working women in Europe and the United States of America*, Oxford University Press, Oxford, 71-89

Elzinga C. H.

2003 *Sequence Similarity, A Nonaligning Technique*, Sociological Methods & Research, 32, (1):. 3-29

England, P. and G. Farkas

1986 *Households, Employment, and Gender: A Social, Economic, and Demographic View.* New York: Aldine.

Esping-Andersen, G.

2002 (a cura di) *A New Gender Contract*, in *Why We Need a New Welfare State?*, a cura di G. Esping-Andersen, Oxford, Oxford University Press.

Ferre, M. M.

1987 *Family and Job for Working-Class Women: Gender and Class Systems Seen from Below*. In Gerstel N. and Gross H.E (eds.), *Families and Wok*. Philadelphia Penn, Temple University Press, 289-301.

Franzosi R., Bison I,

2010 *Temporal Order: Sequence Analysis* in R. Franzosi (a cura di), Quantitative Narrative Analyis, Thousand Oaks, California: SAGE Publication, Inc., 2010, p. 118-123

Garcia, P.; Kalmijn, M.

2016 *Parents' family time and work schedules: the split-shift schedule in Spain*, Journal of Marriage and Family, 78(2)

Gershuny, J.

2000 *Changing times: Work and leisure in postindustrial society*, Oxford: Oxford University Press

Gershuny, J.; Robinson, John P.

1988 *Historical changes in the household division of labor*, Demography, 25(4)

Gauthier, J-A, Widmer, E.D., Bucher, P., Notredame, C.

2010 *Multichannel Sequence Analysis Applied to Social Science Data*. Sociological Methodology. 40:1-38.

Haddock, Shelley A.; Zimmerman, Toni S.; Ziemba, Scott J.; Lyness, Kevin P.

2006 *Practices of dual earner couples successfully balancing work and family*, Journal of Family and Economic Issues, 27(2)

Hägerstrand, T.

1970 *What about people in regional science?* Papers in Regional Science, 24(1)

1982 *Diorama, path and project*, Tijdschrift voor Economische en Sociale Geografie, 73(6)

Hakim, C.

2003 *A new approach to explaining fertility patterns: Preference Theory*. Population and Development Review, 29(3)

Hallberg, D.

2002 *Synchronous leisure, Jointness and Household*. Labour Economics, 10(2)

Hamermesh, Daniel S.

2002 *Timing, Togetherness, and Time Windfalls*, Journal of Population Economics, 15(4)

Hellgren, M.

2014 *Extracting more knowledge from time diaries?* Social Indicators Research, 119(3)

Hook, Jennifer L.

2006 *Care in context: Men's unpaid work in 20 countries, 1965-2003*. American Sociological Review, 71(4)

ISTAT

2011 *Indagine multiscopo sulle famiglie - Uso del tempo Anno 2008-2009. Manuale utente*, Roma: ISTAT

Kan, Man Y.; Sullivan, O.; Gershuny,J.

2011 *Gender convergence in domestic work: Discerning the effects of interactional and institutional barriers from large-scale data,* Sociology, 45(2)

Kilkey, M.; Perrons, D.

2010 *Gendered divisions in domestic work time. The rise of the (migrant) handyman phenomenon*, Time & Society, 19(2)

Lesnard, L.

(2004), *Schedules as sequences: a new method to analyze the use of time based on collective rhythm with an application to the work arrangements of French dual-eaner couples*, Electronic International Journal of Time Use Research, 1(1)

2005 *The effects of the family work day on family time*, Working paper: Série des Documents de Travail du CREST, Institute National de la Statistique et des Etudes Economiques, No 2005-25

2008 *Off-scheduling within dueal-earner couples: An unequal and negative externality for family time*, American Journal of Sociology, 114(2)

Lesnard, L; Kan, Man Y.
2009 *Two stage optimal matching analysis of workdays and workweeks*, SociologyWorking Papers, 2009-04, University of Oxford

Makles, A.
(2012), *How to get the optimal k-means cluster solution*, The Stata Journal 12, Number 2, pp. 347–351

Mansour, H.; McKinnish, T.
2013 *Couples' time together: complementarities in production versus complementarities in consumption*, IZA Discussion Paper, No. 7848

Mattingly, M.; Bianchi, Suzanne M.
2003 *Gender Differences in the Quantity and Quality of Free Time: The US Experience*, Social Forces, 81(3)

Moen, P.
1992 *Women's two roles*. New York: Auburn House.

Moreno-Colom, S.
2015 *The gendered division of housework time: Analysis of time use by type and daily frequency of household tasks*, Time & Society, 0(0)

Naldini, M; Saraceno, C.
2011 *Conciliare famiglia e lavoro. Vecchi e nuovi patti tra sessi e generazioni.* Bologna: Il Mulino

Nock, Steven L.; Kingston, Paul W.
1984 *The Family Work Day*, Journal of Marriage and the Family, 46(2)

Offer, S.; Schneider, B.
2011 *Revisiting the gender gap in time-use patterns: Multitasking and well-being among mothers and fathers in dual-earner Families*, American Sociological Review, 76(6)

Oláh, L.Sz., Richter, R.; Kotowska, I.
2014 *State-of-the-art report: The new roles of men and women and implications for families and societies*, FamiliesAndSocieties Working Paper 11(2014)

Presser, Harriet B.
1994 *Employment Schedules among Dual-Earner Spouses and the Division of Household Labor by Gender*, American Sociological Review, 59(3)

Raley, S.; Bianchi, Suzanne M.; Wang, W.
2012 *When do fathers care? Mothers' economic contribution and fathers' involvement in child care*, American Journal of Sociology, 117(5)

Reyneri, E.
2001 *Sociologia del ML*, Bologna, Il Mulino.

Saraceno, C.
2003 *La conciliazione di responsabilità familiari e attività lavorative in Italia: paradossi ed equilibri imperfetti*, in «Polis», 2, 199-228.
2012 *Coppie e famiglie. Non è questione di natura*. Milano: Feltrinelli

Spencer H.
(1855), *The Principles of Psychology*, London, Longman, Brown, Green and Longmans

Stanley, S.C., Hunt, J.G., Hunt L.L.
1986 *The relative deprivation of Husbands in Dual-Earner Households*. Journal of Family Issues, 7, 3-20.

van Klaveren, C.; van den Brink, H.M.; van Praag, B.M.S.
2011 *Intra-household work timing: the effect on joint activities and the demand for child care*, Tinbergen Institute Discussion Paper, No. 11-065/3

Voorpostel, M.; van der Lippe, T.; Gershuny, J.

2010 *Spending Time Together. Changes Over Four Decades in Leisure Time Spent with a Spouse*, Journal of Leisure Research, 42(2)

Warren, T

2003 *Class- and gender- based working time? Time poverty and time division of domestic labour*, Sociology, 37(4)

Wight, V.R.; Raley, S.B.; Bianchi, S.M.

2008 *Time for children, one's spouse and oneself among parents who work nonstandard hours*, Social Forces, 87(1)

Wu, L.L.,

2000 *Some comments on "Sequence Analysis and Optimal Matching Methods in Sociology: Review and Prospect*, in Sociological Methods and Research, **29**,1, 41-64.

Zimmermann H.J.

1990 *Fuzzy set theory and its application*, Kluwer, Dordrecht.

Session 3B: Life sequences of disabled

Depressive symptom trajectories across working life and workload in paid and unpaid work among Swedish men and women

Paraskevi Peristera, Hugo Westerlund, Linda Magnusson Hanson

Stress Research Institute, Stockholm University

Abstract Long working hours and high work load are possible risk factors for depressive symptoms, but relatively few prospective studies have been conducted thus far. Little is also known about the influence of the work environment on the course of depressive symptoms. This study aimed to use trajectories of depressive symptoms across working age to examine whether work load from paid and unpaid work predict these trajectories. The study was based on data from 4 waves of the Swedish Longitudinal Occupational Survey of Health (SLOSH, 2008-2014). We applied latent-class growth modelling in order to identify groups with similar development of depressive symptoms (depression trajectories) between ages 20 to 65. We further studied workload from paid and unpaid work as risk factors for certain pattern of symptoms and time-varying modifiers of trajectory level within each group. Six groups were identified with symptoms: 'stable very low', 'stable low', 'mild decreasing', 'recurrent mild-moderate', 'stable moderate-high through midlife', 'stable high'. More hours in unpaid work in 2008 increased the risk of belonging to the groups with 'stable moderate-high through midlife' and 'recurrent mild-moderate'depressive symptoms. When treating workload from unpaid and paid work as time-varying covariates we found that workload from unpaid work was associated with higher depressive symptoms although the effect was not statistically significant for all groups. On the other hand higher workload from paid work in terms of long working hours resulted in higher depressive symptoms for all groups except for individuals belonging to the group 'stable moderate-high through midlife'. Females predicted depressive symptoms in all groups when compared to the 'stable very low' group. This study supported that are heterogeneous individual patterns of depressive symptoms over the working life. The results also indicate that workload both from paid and unpaid work influences the evolution of depressive symptoms across age.

1 Introduction

Mental health problems and especially depression are of major public health concern, owing to the high prevalence and substantial negative consequences on personal functioning but also work productivity. It is well known that onset of depression may vary between individuals and that symptoms occasionally recur over the life course. However, relatively few studies have investigated the course of depressive symptoms in the general population (Steinert et al 2014). Little is also known in about individual trends in the development of depressive symptoms since most longitudinal studies focus on the analysis of the mean trend of depressive symptoms over time by using growth curve modeling techniques or hierarchical linear modelling (Castelao and Kröner-Herwig 2013). Advanced statistical approaches in growth modelling, such as latent class growth analysis, however, provide a flexible and easily applied method for identifying distinctive clusters of individual trajectories within the population and for profiling the characteristics of individuals within the clusters (Nagin 2005). These models can thus be used to identify distinctive trajectory groups of people with different long-term patterns of depressive symptoms, some of which may be associated with more functional limitations.

It has also been suggested that different risk factors may be associated with different long-term patterns of depressive symptoms (Colman, 2010; Papachristou 2013), but very few studies thus far have investigated how environmental factors influence the development of depressive symptoms over the long-term. It has been suggested that factors such as sex, socioeconomic status, history of depression or psychopathology, and stressful life events relate to different courses of development (Musliner et al 2016). One study by Hardeveld et al (2010) also suggested that stress-related factors, including certain work stressors, predicted recurrent major depressive episodes. However, little is still known about the influence of potentially modifiable factors such as work-related characteristics on different long-term patterns of depressive symptoms. Work stressors such as high work demands or high demands in combination with low control (job strain) have generally been implied as risk factors for depressive symptoms (Bonde, Netterstrom 2008, Theorell et al 2015). Bannai and Tamakoshi (2014) concluded that also long working hours is associated with depressive states. However, non-work stressors may also be associated with poorer mental health (Clark et al 2012). Especially a "double burden" of work and non-work related responsibilities could have a negative health effect. This kind of double burden or high total workload has been suggested to contribute to common physical and mental symptoms (Krantz and Ostergren 2001), but more knowledge is needed on both changes in stress from workload in paid and unpaid work, and the interplay between these stressors, over the life course, and how they influence health among men and women (Payne and Doyal 2010).

3

The aim of this study was to identify different developmental courses /trajectories of depressive symptoms over the working life of Swedish men and women, and examine how work load in terms of long working hours and hours spent in unpaid work are associated with different trajectories.

2 Methods

2.1 Data

Study Population The study population consisted of participants from SLOSH (Swedish Longitudinal Occupational Survey of Health) study, a longitudinal cohort study which focuses on work life participation, social situation, and health/wellbeing with repeat self-reported measures every second year (starting in 2006) on an originally representative sample of the Swedish working population.

Study Sample The present study is based on those who responded to the SLOSH questionnaire and were in paid work more than 30 % of full time in all four waves with start year 2008 (6291 participants).

Time Axis We applied an accelerated longitudinal design (Raudenbush and Chan 1992), using age rather than calendar years as the time variable. This allowed assessment of depression trajectories over an age range between 20 and 70 years based on a 6 year period (2008-2014). However, we limited our analyses to the individuals aged 20-65 years of age, the typical working age.

2.2 Measures

Outcome: Depressive symptoms were measured with a brief subscale from the Hopkins Symptom Checklist (SCL-90), the SCl-CD6 (Magnusson Hanson et al 2014) which assesses perception of being troubled by: Feeling blue; Feeling no interest in things; Feeling lethargy or low in energy; Worrying too much about things; Blaming yourself for things; and Feeling everything is an effort, quantified on a five-category scale from 0=Not at all to 4=Extremely. The six items represent core symptoms, selected based on principals of clinical validity. The scale has validated and was found to have good psychometric properties and results have showed that the items are suitable to add into a composite score indicative of depression severity (Magnusson Hanson et al 2014).

Workload: Workload was measured repeatedly 2008-2014 by a modified version of a measure developed and psychometrically evaluated by Mardberg et al (1991). In addition to hours/week in paid employment and overtime at work, constituting a measure of workload from paid work, the instrument covers unpaid work activities, such as household duties (mending, sewing, laundry, gardening), childcare (homework/teaching, care-taking, playing) and other unpaid duties (voluntary work in unions and organizations, care of sick or elderly relatives). Hours/week spent on household duties (shopping, cleaning, cooking, mending, sewing, laundry and gardening) and on childcare (homework/ teaching, care taking, playing) were added to a measure of unpaid work. In this study workload from paid work was divided into 4 categories: <40, 40-49, 50-59, 60+ working hours/week. An exception was in SLOSH 2010, then the corresponding categories were <41, 41-50, 51-60, 61+. Workload from unpaid work was also divided into 4 categories: <8, 8-11, 12-20, 21+ hours on average/week). The total number of hours spent on paid and unpaid work constituted the total workload measure also divided into the following 4 categories: <58, 58–67, 68–80, and >80 h/week.

2.3 Statistical Analysis

Group-based trajectory modelling (GBTM) was used in order to identify distinctive groups of individuals who can be classified into groups with similar developmental trajectories over time, and examine whether work load predicts the development of depressive symptoms.

Group-based trajectory modelling (GBTM) is a semi-parametric model-based clustering technique that is mostly applied for the identification of latent groups of individuals following a similar progression of an outcome over time (Nagin 2005). Model estimation produces posterior probabilities of membership in each trajectory group for each participant. These probabilities are then used to assign individuals to the trajectory group to which they have the highest probability of belonging.

An accelerated longitudinal design was adopted in our analysis in order to use age rather than calendar years as the time variable. The first step before the GBTM analyses was to rearrange data from the 4 SLOSH waves (years 2008 to 2014) so as to follow a longitudinal design covering an age span ranging from 20 to 65 years of age.

In order to select the best model we followed Nagins' recommended two-step procedure for model selection. Firstly the number of latent trajectories is selected based on fit indices. Subsequently the order of the polynomials describing the level and shape of the latent trajectories is determined. We first considered with a single trajectory model for depression described by a cubic polynomial. In the

5

next step we revised the model by increasing the number of groups using fit criterions. Trajectory models with two to eight groups and varied shapes were compared. Estimation of depression trajectories was accomplished using the censored normal model (CNORM), which is appropriate for continuous data.

We relied on several criteria to choose the best model. The Bayesian Information Criterion (BIC) is one of the most commonly used fit statistics to determine the number of subgroups with the model with value closer to zero providing the best fit (Nagin 1999, 2005). In general BIC measures improvement in model fit gained by adding more parameters (e.g. more groups and more complex trajectory shapes) but also emphasizes model simplicity by applying a penalty for complex models (Kass and Raftery, 1995) We further considered the Akaike's Information Criterion [AIC]) , the significance of polynomial terms (at the confidence level alpha 0.05), the values of group membership probabilities and of average posterior probability (entropy) (Nagin and Odgers, 2010; Andruff et al., 2009). Entropy is an index used in classification accuracy based on posterior probabilities with higher values denoting better classification. A value greater than 0.7 for all groups is generally recommended since it indicates that the trajectory encompasses individuals with similar patterns of change and discriminates between those with dissimilar patterns of change (Nagin, 2005). Although there is a big discussion in the literature on which measures to use for the selection of the best model, there is no commonly accepted single standard model fit statistic but there are several suggestions from existing simulation studies (Henson et al. 2007, Nylund et al. 2007). The magnitude of difference in BIC, the Bayes factor as well as the BIC-based probability approximation were used to choose between more complex and simpler models (Nagin, 1999; Nagin and Odgers, 2010; Jones, Nagin and Roeder, 2001). A Bayes factor is the ratio of the probability of model 1 being the correct model to the probability of model 2 being the correct model (Nagin, 2005). If two models have equal probability of being correct the Bayes factor would be one. Values than 1 favor model 2 whereas values greater than one imply that model one has a higher probability of being the correct model.

After selecting the optimal model in terms of fit we selected the shape for each of these groups. The trajectory model was fitted using maximum likelihood methods that allow for incomplete data and assumes that missing data were missing at random.

After finding the optimal trajectory model we included risk factors i.e. factors influencing the probability of membership of a particular trajectory group in 2008, and covariates. Workload from paid work and from unpaid work were considered simultaneously relative to the defined trajectory groups as time-invariant covariates (risk factors) and time varying covariates (TVCs) measured in 2008-2014). Coefficients for risk factors indicate the increase in relative odds of being in a trajectory (relative to the lowest group) per unit change in the risk factor (Nagin, 2005). Coefficients for TVCs can be interpreted as: given membership in a trajectory group, how much higher (if coefficient is positive) or lower (if coefficient is negative) is the depression trajectory for a unit increase in the covariate. Unadjust-

ed models as well as models adjusted for sex (0 men, 1 women), civil status (0 not married, 1 married or cohabiting) and socioeconomic status (Statistics Sweden, 1982) were presented. We further conducted the same analyses for men and women separately. In addition, we have considered the potential influence of life stage at baseline by adjusting for three age groups: young adulthood (20-34), midlife (35-49) and mature adulthood (50-65).

The GBTM analyses were conducted using the PROC TRAJ procedure developed by Jones and Nagin, which can be downloaded from http://www.andrew.cmu.edu/user/bjones (Jones et al., 2007; 2001) in the SAS software (version 9.4; SAS Institute).

3 Results

Our fit evaluation of the different models resulted in the selection of a model with six trajectories with a linear order for four groups suggesting a linearly decreasing or increasing trajectory, and a cubic order for two of the groups suggesting a trajectory where there are two turning points (inflections), a maximum and a minimum in depressive symptoms. Figure 1 depicts the shapes of the six trajectories obtained from the GBTM analysis as well as their class sizes. The pattern of symptoms over the working life could be described as either "stable very low", "stable low", "mild decreasing", "recurrent mild-moderate", "stable moderate-high through midlife" and "stable high". The trajectories named 'stable very low', "stable low" and 'stable high' followed a slight downward linear trend as people got older, but remained either at a low to mild level (<12 on the depression scale) or at a high level (>16 on the depression scale). The majority of individuals were classified in either the 'stable low' (43.8%) or 'mild decreasing' (25.3%) group. The 'stable high' group represented the smallest group (2.1%). The 'stable moderate-high through midlife' group (8.8 %) followed a quadratic trend with two turning points. In this group a moderate decrease in symptoms was evident during young adulthood (until about 35 years of age), after which symptoms remained on a moderate level throughout midlife, followed by a steeper decrease during mature adulthood (from ages 50 and onwards). For the 'recurrent mild-moderate' group (8.2 %) the pattern was also described by a quadratic trend with a decrease in symptoms from mild to low during young adulthood, but with an increase in symptoms to moderate in midlife. The former group thus appeared to have recurrent symptoms over the working life, while two groups the 'stable very low' and the stable low' group, appeared to be free of limiting symptoms over the working life. Altogether there were 3 favorable (mild decreasing, stable low and stable very low) and 3 unfavorable (recurrent mild-moderate, stable moderate-high through midlife, and stable high) trajectories.

7
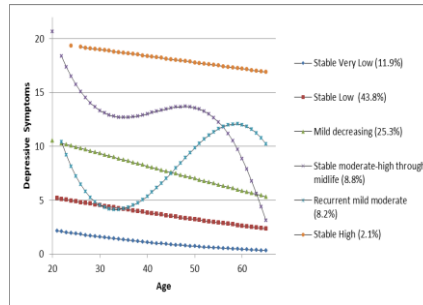
Fig. 1: Depression trajectories (20-65 years), (six-group model, no-time invariant covariates or TVCs), SLOSH, 2008-2014; standard errors are given in parentheses

A description of the demographic characteristics and the distribution of paid-unpaid workload in the year 2008 for the six trajectory groups are given in Table 1. The mean age was lowest in the 'stable moderate-high through midlife' trajectory (46.4 years), followed by 'very low stable' trajectory and 'mild decreasing' (Table 1). The proportion of women varied from 46.6 % in the 'very low stable' to 69.4% in the 'stable moderate-high through midlife' and 65% in the 'stable high' trajectory whereas proportion of married individuals varied from 83% in the 'stable very low' trajectory to 50.7% in the 'stable high' trajectory. Overall there was a higher proportion of women in the unfavorable trajectories. The highest proportion of unskilled manual workers was in the 'stable high' group, while the highest proportion of skilled manual workers was in the 'stable very low' trajectory. Assistant non-manual employees represented the highest proportion in the 'stable moderate-high through midlife ' group while the 'mild decreasing' group was represented by a high proportion of other non-manual employees. The higher proportion for self-employed appears to be in the 'stable very low' trajectory. Generally it also seemed as a higher proportion of individuals spent high number of hours on unpaid work in the unfavorable trajectory than in the favorable trajectory groups.

Table 1. Characteristics of the trajectory groups in terms of demographic characteristics and workload from paid and unpaid work in Wave 2008

		Stable very low	Stable Low	Mild decreasing	Stable moderate-high through midlife	Recurrent mild-moderate	Stable High
Mean Age		49.82	52.29	50.17	46.37	56.14	50.67
Sex (% females)		46.64	55.83	62.35	69.46	63.82	68.32
Civil status (% married)		83.13	80.91	78.46	73.50	79.02	65
SEI	Unskilled Manual Workers	16.07	14.18	15.49	14.88	13.81	20.83
	Skilled Manual Workers	18.21	14.40	12.93	14.05	15.02	13.54
	Assistant Non-Manual Employees	11.07	12.40	13.06	12.12	15.02	13.54
	Intermediate Non-manual Employees	29.64	31.45	32.59	31.96	30.63	26.04
	Higher non-manual employees	17.68	20.95	21.35	20.11	19.22	18.75
	Self-employed	7.32	6.62	4.58	6.89	6.31	7.29
Work Load paid work (hrs per week)	(% <40 or 41)	28.93	30.60	29.29	33.33	36.14	30.77
	(% 40/41-49/50)	50.49	50.76	50.80	51.07	44.91	51.28
	(% 50/51-59/60)	13.20	12.04	12.28	9.48	12.28	12.82
	(% 60/61+)	7.38	6.60	7.63	6.12	6.67	5.13
Workload unpaid Work (hours per week)	<8	24.15	20.16	17.74	10.37	18.58	14.49
	8-11	21.79	22.43	21.45	17.39	16.21	21.74

9

12-20	20.94	22.04	22.03	19.73	26.88	18.84
21+	33.12	35.37	38.78	52.51	38.34	44.93

When examining the role of workload as risk factor for trajectory group, higher workload from paid work at the baseline measurement (2008) were not clearly associated with a higher or lower likelihood of a certain trajectory in the model adjusting for sex, civil status and socioeconomic status. Individuals with higher workload from unpaid work, however, had increased likelihood of being in the three unfavorable trajectory groups and the mild decreasing trajectory compared to the very low stable trajectory group, but the results were statistically significant only for 'mild decreasing', 'stable moderate-high through midlife', and 'recurrent mild-moderate ' groups in the adjusted model.

Sex, civil status and socioeconomic status were included in the adjusted models as potential predictors (risk factors) of the trajectory group membership. Females were a statistically significant predictor (results not shown) of membership in all the groups compared to the 'stable very low' group. Being married or cohabiting predicted the membership in the 'stable moderate-high through midlife' and 'high stable' groups (statistically significant at 5%) compared to the 'very low stable group'. Similar results were found for the unadjusted models although there were some differences in terms of the statistical significance of the estimates. We further adjusted for three age groups. In that case we found that people in the age group (35-49) had a lower likelihood for being in most of the groups compared to the 'stable very low' group but the results were only statistically significant for those with 'stable low' symptoms. Regarding the age group (50-65) we found that people in this group showed higher likelihood of being in the stable moderate-high through midlife group and lower likelihood in all other trajectory groups with statistically significant effects only for group 'stable low'. We also tested for interaction effects between the different age groups and workload from paid/unpaid work but there was no indication of statistically significant interaction effects.

Table 2. Increase in relative odds of being in a certain trajectory group as compared to being in the very low stable group, according to workload at baseline

	Coefficients (SE) Unadjusted model		Coefficients (SE) Model adjusted for sex, civil status and SEI	
	Workload Paid Work	Workload Unpaid Work	Workload Paid Work	Workload Unpaid Work
Stable very low	-	-	-	-
Stable low	-0.02 (0.01)	0.20(0.09)*	-0.02 (0.01)	0.14 (0.09)
Mild decreasing	-0.01 (0.01)	0.14 (0.08)	-0.00 (0.01)	0.04 (0.08)*

Stable moderate-high through mid-life	-0.01 (0.01)	0.39 (0.09)*	-0.01 (0.01)	0.33 (0.09)*
Recurrent mild-moderate	-0.03 (0.01)*	0.43 (0.10)*	-0.02 (0.02)	0.32 (0.10)*
Stable High	-0.01(0.02)	0.38 (0.18)*	0.005 (0.02)	0.31 (0.18)
* p value < 0.05				

These analyses were also stratified by sex. However, there were no major differences between men and women. If anything, workload due to unpaid work seemed a stronger predictor of group membership in the 'recurrent mild-moderate' group for women, but the power was insufficient for further comparisons (data not shown).

Both paid work and unpaid work over the study period were associated with depressive symptoms across the working age but the results varied in terms of direction and statistically significance for the various groups (Table 3). An increasing level of workload due to paid work was associated with higher depressive symptoms (when adjusted for sex, civil status and socioeconomic status) in all groups except for the 'stable very low' group. Higher work load due to unpaid work (adjusted model) was statistically significantly related to an increase in depressive symptoms in the ' stable low', 'recurrent mild-moderate' groups but associated with a downward trend in depressive symptoms in the 'stable moderate-high through midlife' group (statistically significant at 5%).

Table 3. Influence of workload on the trajectory level within each group (standard errors are given in parentheses; * statistically significant at 5%)

	Coefficients Unadjusted model		Coefficients Model adjusted for sex, civil status, SEI	
	Workload Paid Work	Workload Unpaid Work	Workload Paid Work	Workload Unpaid Work
Stable very low	-0.04 (0.03)	0.26 (0.18)	-0.01 (0.03)	0.15 (0.17)
Stable low	0.03 (0.01)*	0.12 (0.06)*	0.07 (0.02)*	0.49 (0.11)*
Mild decreasing	0.05 (0.01)*	0.34 (0.08)*	0.03 (0.01)*	0.09 (0.06)
Stable moderate-high through midlife	-0.04 (0.03)	0.14 (0.13)	-0.12 (0.04)*	-1.35 (0.35)*

11

Recurrent mild-moderate	0.82 (0.10)*	4.63 (0.82)	0.08 (0.04)*	1.60 (0.35)*
High stable	0.05 (0.04)	0.16 (0.21)	0.10 (0.04)*	0.19 (0.22)

Analyses of the total work load (measured as <58hrs/week, 58-67 hrs/week, 68-80 hrs/week, >80hrs/week) finally showed that a higher total workload increased the likelihood of being in the 'mild decreasing' trajectory and tended to increase the likelihood of being in the 'high stable' group, although the latter risk estimate was not statistically significant in the adjusted model. Furthermore, no significant interaction was observed between the workload paid work and workload unpaid work for any of the six trajectories.

4 Discussion

The current study examined patterns of depressive symptoms throughout ages 20-65 years for an initially representative sample of working men and women as part of the Swedish Longitudinal Occupational Survey of Health (SLOSH) study (waves 2008-2014), and the influence of workload due to paid and unpaid work.

The first objective was to determine whether there are distinct trajectories of depressive symptoms in the working population. This work identified six distinct depression trajectories between ages 20 and 65.

In accordance with data on major depressive disorder in different ages (Ferrari et al 2013), we observed higher levels of symptoms around the ages 45-54 in one of the groups, although the peak occurred little later in one of the groups. However, a relatively high proportion, around 5 % of the population, already experience depression around 20-24 years of age (Ferrari et al 2013). It has also been found that a relatively high proportion of people in the general population have a recurrent episode of depression and that a considerable proportion of the general population experience a stable recovery after a depressive episode (Steinert 2014). It is therefore not surprising that we also found several groups with a decreasing symptoms levels. A smaller proportion could also be expected to have a more chronic course as observed in the present study.

Studies on individual trajectories of depressive symptoms in the general adult populations are still relatively rare. Musliner at al, however, reviewed mainly population based studies in different age periods and concluded that stable patterns were common such as stable low symptoms. A small group with stable high symptoms as in our study have also commonly been found in other studies. Patterns with varying degree of symptoms over time were more seldom observed. Howev-

er, in line with some previous work on depressive and anxiety symptoms (Colman et al 2007) we also identified trajectory profiles with decreasing symptoms in adulthood, repeated mild-moderate symptoms, moderate symptoms starting in midlife. In contrast to Colman et al (2007) studying depressive and anxiety symptoms up to age 53, we did, however, not observe any group with onset of severe symptoms during the study period.

Our next goal was to determine whether work load influenced the above trajectories. Consistent with other reports (Musliner et al 2016) sex predicted depressive symptoms in all trajectory groups but neither civil nor socioeconomic status was associated with depression for the majority of the six groups. Some other studies did, on the other hand, find that educational level or income predicted trajectories with high symptom burden (Musliner et al 2016) as well as occupational grade (Melchior et al 2013). Moreover, it has previously been indicated that factors such as problems with peers and parents, alcohol/tobacco/drug use, parental history of depression and negative cognitive styles could lead to worse depression trajectories over time among children/adolescents (Musliner et al 2016). Among older people, poor self-rated health, past history of somatic illness, functional and cognitive impairment and low social support have been associated with negative development of symptoms. Stressful life events has also a found to be a predictor of poor depressive symptom trajectories (Musliner et al 2016). This work suggest that also workload is associated with unfavorable long-term trajectories of depressive symptoms. Both workload from paid work and unpaid were associated with depression over age progression. Analyses of work load as a time-varying covariate showed that workload from both paid and unpaid work was generally associated with an increase in depressive symptoms supporting some previous findings (Virtanen et al 2011, 2012, Shields 1999, Krantz Ostergren 2001, Krantz et al 2005). It is possible that this increase may lead to a rise in symptoms from moderate level to severe in some cases.

Individuals with higher work load from unpaid work at baseline were also at increased risk of belonging to the 'mild decreasing', 'recurrent mild-moderate', and 'stable moderate-high through midlife' groups compared to the 'stable very low' group. Higher workload from paid work was, however, not clearly associated with a certain trajectory. The study further indicated that a higher total (double) burden increased the likelihood of one of the most unfavorable trajectories.

Some strengths of the study are that it was conducted in a sample from the general working population, and with measures of depressive symptom every second year over a period of 8 years. In contrast to studies of depression trajectories in clinical population, studies in the entire population may more accurately represent the true underlying continuum of the disorder (Musliner et al 2016). The accelerated design also allowed us to characterize depressive symptoms trajectories over the entire working age. Dropout from the study, however, may have restricted the possibility to detect heterogeneity. Given that the subjects were originally working and repeatedly in paid work for more than 30 % over the study period, the sample

13

is probably characterized by relatively healthy individuals with high educational level etc.

We based the trajectories on four time points, which is above the minimum required for estimating quadratic trajectories, but the more time points to estimate cubic trends the better (Andruff et al., 2009). We further used an accelerated design to make the most of these data. This allowed us to model symptom trajectories over a much longer time period. There are, however, potential problems associated with this approach. One is inferential ambiguity due to an age-cohort interaction (Raudenbush and Chan 1993). We tested and found no evidence of an age interaction in the main models but this may be associated with uncertainty due to a relatively small number of overlapping time points in adjacent cohorts. A longer time series of measurement may also have contributed to more power for identifying heterogeneity in symptoms over time and potential determinants of different trajectories.

Another limitation of the present study is that we were not able to measure symptoms in childhood. Some individuals may therefore already have had high symptoms in childhood making it difficult to disentangle onset of moderate or severe symptoms. Data were also missing on family history of mental health problems and childhood/adolescent characteristics, which may be predictors of depressive symptoms trajectories and workload and may confound the relationships of interest. Finally, when relating time-varying covariates to the trajectory it is possible that the depressive symptoms over time may influence later workload measures, thus we are not able to draw any causal conclusions based on these analyses. It is less likely though that the depression trajectory influences the initial measurement of workload suggesting that there may be a causal association between unpaid workload and depression trajectories over the working life.

References

Andruff, H., Carraro, N., Thompson, A., Gaudreau, P., Louvet B. (2009) Latent class growth modeling: a tutorial, *Tutorials Quantitative Methods Psychology*. **5**: 11-24.

Bannai, A. and Tamakoshi, A. (2014) The association between long working hours and health: a systematic review of epidemiological evidence, *Scand J Work Environ Health.* **40**(1),5-18. doi: 10.5271/sjweh.3388.

Bonde, J.P. (2008) Psychosocial factors at work and risk of depression: a systematic review of the epidemiological evidence, *Occup Environ Med.* **65**(7), 438-45.

Castelao, C.F. and Kröner-Herwig B. (2013) Different trajectories of depressive symptoms in children and adolescents: predictors and differences in girls and boys, J Youth Adolescence. **42**, 1169-1182.

Clark, C., Pike, C., McManus, S., Harris, J., Bebbington, P., Brugha, T., Jenkins, R., Meltzer, H., Weish, S., Stanfeld, S. (2012) The contribution of work and non-work stressors to common mental disorders in the 2007 Adult Psychiatric Morbidity Survey, Psychol Med. **42**(4), 829-42.

Colman, I., Ataullhjan, A. (2009) Life course perspectives on the epidemiology of depression, *Can J Psychiatry*. **55**(10), 622-32.

Colman, I., Ploubidis, G.B., Wadsworth, M.E., Jones, P.B., Croudace, T.J. (2007) A longitudinal typology of symptoms of depression and anxiety over the life course, *Biological Psychiatry*. **62**,1265–1271.

Ferrari, A.J., Charlson, F.J., Norman, R.E., Patten, S.B., Freedman, G., Murray, C.J, Vos, T., Whiteford, H.A. (2013) Burden of depressive disorders by country, sex, age, and year: findings from the global burden of disease study 2010, PLoS Med. **10**(11), e1001547. doi: 10.1371/journal.pmed.1001547.

Hardeveld, F., Spijker, J., De Graaf, R., Nolen, W.A., Beekman, A.T.F. (2010) Prevalence and predictors of recurrence of major depressive disorder in the adult population, *Acta Psychiatrica Scandinavica*. **122**, 184-191.

Henson, J.M., Reise, S.P., and Kim, K.H.(2007) Detecting mixtures from structural model differences using latent variable mixture modelling: A comparison of relative model fit statistics, *Structural Equation Modeling*. **14**, 202-226.

Jones, B.L., Nagin, D.S., Roeder, K. (2001) A SAS procedure based on mixture models for estimating developmental trajectories, Socio Methods Res. **29**, 374-393.

Jones, B.L., Nagin, D.S. (2007) Advances in Group-Based Trajectory Modeling a SAS Procedure for estimating them, *Sociological Methods and Research*. **35**(4), 542-571.

Kass, R.E. and Raftery, A.E. (1995) Bayes Factors. *Journal of American Statistical Association*. **90**, 773-795.

Krantz, G., Berntsson, L., Lundberg, U. (2005) Total work load, work stress and perceived symptoms in Swedish male and female white-collar employees, *European Journal of Public Health*. **15**(2), 209-214.

Krantz, G., Ostergren, P.O. (2001) Double Exposure: The combined impact of domestic responsibilities and job strain in employed Swedish women, *European Journal of Public Health*, **11**, 413-419.

Magnusson Hanson, L.L., Westerlund, H., Leineweber, C., et al. (2014) The Symptom Checklist-core depression (SCL-CD6) scale: Psychometric properties of a brief six item scale for the assessment of depression, *Scand J Public Health*. **42**(1),82-8. doi: 10.1177/1403494813500591. Epub 2013 Aug 27. 2013.

Mardberg, B., Lundberg, B., Frankenhaeuser, M. (1991) The total workload of parents employed in white-collar jobs. Construction of a questionnaire and a scoring system, *Scandinavian Journal of Psychology*. **32**(3), 233-239.

Melchior, M., Chastang, J., Head, J., Goldberg, M., Zins, M., Nabi, H., Younes, N. (2013) Socioeconomic position predicts long-tern depression trajectory: a 13-year follow-up of the GAZEL cohort study, *Mol. Psychiatry*. **18**(1), 112-121. http://dx. doi.org/10.1038/mp.2011.116.

Musliner, K.L. Mink-Olsen, T. Eaton, W.W. Zandi, P.P. (2016) Heterogeneity in long-tern trajectories of depressive symptoms: Patterns, predictors and outcomes, *Journal of affective disorders*. **192**, 199-211. DOI: http://dx.doi.org/10.1016/j.jad.2015.12.030.

Nagin, S.D. (1999) Analyzing Developmental Trajectories: A Semiparametric Group-Based Approach, *Psychological Methods*. **4**(2), 139-157.

Nagin, D.S. (2005) *Group-based modeling of development*. Cambridge, MA: Harvard University Press.

Nagin, D.S., Odgers, C.L. (2010) Group-based trajectory modeling in clinical research, *Annu Rev Clin Psychol*. **6**: 109-38.

Netterstrom, B. et al. (2008) The relation between work-related psychosocial factors and the development of depression, *Epidemiol Rev*. **30**, 118-32.

Nylund, K.L., Tihomir, A., Muthen, B.O. (2007) Deciding on the number of classes in latent class analysis and growth mixture modeling: A monte carlo simulation study, *Structural Equation Modeling*. **14**, 535-569.

Papachristou, E.S., Frangou, E.S., and Reichenberg, A. (2013) Expanding conceptual frameworks: life course risk modelling for mental disorders, *Psychiatry Res*. **206** (2-3), 140-5.

15

Payne, S. and Doyal, L. (2010) Older Women, work and health, Occup Med(Lond). **60**(3), 172-177.

Raudenbush, S.W., Chan, W-S. (1992) Growth Curve Analysis in Accelerated Longitudinal Designs, *Journal of Research in Crime and Delinquency*. **29**(4), 387-411.

Shields, M. (1999) Long working hours and health, *Health Rep*. **11**,33–48.

Statistics Sweden (1982) Socioeconomisk indelning (SEI) [Socioeconimic classification]. Retrieved from Stockholm.

Steinert, C., Hofmann, M., Kruse, J., Leichsenring, F. (2014) The prospective long-term course of adult depression in general practice and the community. A systematic literature review, *Journal of Affective Disorders*. **152–154**, 65–75.

Theorell, T., Hammarström, A., Aronsson, G., Träskman Bendz, L., Grape, T., Hogstedt, C., Marteinsdotter, I., Skoog, I., Hall, C. (2015) A systematic review including meta-analysis of work environment and depressive symptoms, *BMC Public Health*. **15**, 738. DOI**:** 10.1186/s12889-015-1954-4.

Virtanen, M., Ferrie, J.E., Singh-Manoux, A., Shipley, M.J., Stansfeld, S.A., Marmot, M.G., Ahola, K., Vahtera, J., Kivimäki, M. (2011) Long working hours and symptoms of anxiety and depression: a 5-year follow-up of the Whitehall II study, *Psychol Med*.1-10, doi: 10.1017/S0033291711000171.

Virtanen, M., Stansfeld, S.A., Fuhrer, R., Ferrie, J.E., Kivimäki, M. (2012) Overtime Work as a Predictor of Major Depressive Episode: A 5-Year Follow-Up of the Whitehall II Study, *PLoS One*. 7(1), e30719.

Labor market integration of adolescents with mental disorders

Isabel Baumann, Sibylle Juvalta, Szilvia Altwicker-Hàmori, Niklas Baer, Ulrich Frick, and Peter Rüesch

Abstract Earlier research has shown that mental health problems at early stages of life negatively affect the individuals' long-term employ-ment prospects and occupational attainments. Accordingly, the chances of individuals with mental disabilities of becoming economically independent are often constraint and their social integration is hampered. This situation impairs their well-being which in turn is likely to exacerbate their mental health. One of the reasons for the limited occupational achievements of individuals with mental disorders is their lack of educational de-grees. They may have dropped out of school, have achieved only low levels of education or have graduated from schools for children with special needs. Their educational attainments thus often do not allow them to take up standard jobs. Our paper addresses the question how mental health interventions in children with mental disabili-ties are linked to their labor market outcomes. We analyze whether there are educational and vocational programs or psychiatric interventions that mitigate the negative impact of mental disorders on educational and occu-pational outcomes. Moreover, we examine which consecution of educational and psychotherapeutic measures enhances young adults' chances to be in em-ployment at age 21. We analyze a dataset on 500 young adults in Switzerland who have either received a disability pension or rehabilitation measures be-

Isabel Baumann
Swiss National Centre of Competence in Research NCCR LIVES
e-mail: bams@zhaw.ch

Isabel Baumann · Sibylle Juvalta · Szilvia Altwicker-Hàmori · Peter Rüesch
University of Applied Sciences ZHAW, Department of Public Health, Winterthur, Switzerland

Niklas Baer
Psychiatric Clinic Baselland, Office for Psychiatric Rehabilitation, Liestal, Switzerland

Ulrich Frick
University of Applied Sciences HSD, Department of Applied Psychology, Cologne, Germany

tween 2010 and 2013. We find that the earlier in life interventions—such as rehabilitation measures—are implemented, the higher is the chance that individuals are in an economic activity in early adulthood. Our analysis shows that individuals who entered disability insurance system relatively late and exhibit frequent interruptions in their educational trajectory are likely to be without training or employment at age 21. This finding is in line with the life course paradigm that suggests that individuals who differ with respect to a particular characteristic at a young age – for instance mental health – manifest much stronger differences in this characteristic later in life. Moreover, focusing on rehabilitation measures, we find that a large share of individuals has experienced continuous psychiatric treatment. The implications of our findings are that there is a need for early intervention, particularly in school settings for children with mental disorders. The earlier the intervention takes place, the better there seems to be the outcome for the affected individuals.

Opportunities of Work and Family in Young Disabled People's Lives

A Comparative Study of Disabled and Non-disabled Young Adults in Nineteenth-century Northern Sweden Using Sequence Analysis

Helena Haage, Erling Häggström Lundevaller and Lotta Vikström

Abstract This study focuses on young adults with disabilities and their pathways towards work and family in past society. The aim is to explore their life trajectories and compare them to a non-disabled group of people who experienced the same time-space context, represented by the 19th-century Sundsvall region, Sweden. We employ sequence analyses on a series of demographic events that were to occur in the life of young adults: first occupation, marriage and parenthood. We also check for the events of death and out-migration. Disability studies show that disabled people were often subject to stigmatization caused by their impairment and prevailing perceptions about normalcy in in society. This would have limited their opportunities of work and family compared to non-disabled persons. Individual-level data consisting of parish registers digitized by the Demographic Data Base (DDB), Umeå University, Sweden, allow sequence analysis that helps to answer the questions of whether and how disability influenced people's life trajectories. We obtain a holistic picture of how their life developed that suggests that disability substantially limited people's opportunities to find job, marry and form a family. This indicates that a stigma was associated with disability beyond the impairment itself and worked to add to disabled individuals' difficulties in both the labor market and marriage market.

Helena Haage (first author)
PhD Student in History
Department of Historical, Philosophical and Religious Studies
Centre for Demographic and Ageing Research (CEDAR)
Umeå University, SE-901 87 Umeå, Sweden
helena.haage@umu.se

Erling Häggström Lundevaller (second and corresponding author)
Associate Professor in Statistics
Umeå School of Business and Economics, Department of Statistics
Centre for Demographic and Ageing Research (CEDAR)
Umeå University, SE-901 87 Umeå, Sweden
erling.lundevaller@umu.se

Lotta Vikström (third and corresponding author)
Professor of History
Department of Historical, Philosophical and Religious Studies
Centre for Demographic and Ageing Research (CEDAR)
Umeå University, SE-901 87 Umeå, Sweden
lotta.vikstrom@umu.se

1 Introduction: Background, Aims and Rationales

Historical research shows that getting an occupation, marrying and giving birth to children were common events in the transition to adult life for nineteenth-century young people. These events and the ordering and timing of them were governed by contemporary norms in society and encouraged by institutions such as the church and state. For instance, job and employment were to occur in young people's life before they married and established a family on their own, and children were to be expected upon marriage, not before. Across north-western Europe, this pathway was also the result of an economy based on agricultural production. It promoted a life cycle servant system according to which young people were to be hired as maidservants and farmhands in other households than their parents' (Dribe 2000; Lundh 1999; Lundh 2003; Whittle 2005; Harnesk 1990). This made young women and men move between different employers to secure work and income to gather the skills and material resources required to unite a spouse and set up a household. In this way, the servant system worked to structure the opportunities of work and family among young nineteenth-century individuals. Even though most of them followed this expected pathway, little is known about whether individuals with disabilities did. From disability studies we know that they were often subject to a stigma caused by the impairment and prevailing perceptions about normalcy in society (Kudlick 2003; Susman 1994; Oliver 1996; Barnes et al. 2010; Goffman 1972; De Veirman 2015; Haage et al 2016). If disability limited young people's opportunities for work and family, it would become evident from their life trajectories when compared with those of non-disabled individuals.

This study aims to do that and thus it fills in the gap of knowledge we have on past people with disabilities. We provide novel results based on sequence analysis on a series of demographic events that occur in the life of young adults, such as first occupation, marriage and parenthood. All events are consistently compared between disabled men and women and in relation to their non-disabled counterparts, all of whom resided in the Sundsvall region, Sweden. During the 19th century, this region witnessed a fast population growth from about 13,000 inhabitants in the early 1800s to 18,793 in 1840 and to 46,418 inhabitants in 1880 (Alm Stenflo 1994). This growth was due to the expansion of the sawmill industry and a large influx of migrants in combination with the mortality decline.

Accounting for several events and not just one, our sequence analysis and longitudinal data (see Sect. 2) make us obtain a more holistic picture of disabled people's life (Aisenbrey et al 2010; Abbot et al 2000). Only being able to conduct such analysis of individuals with disabilities living 150-200 years ago makes this study special in its approach, as quantitative methods are rare in disability studies and social history. The demographic experiences of disabled people are important to uncover not only because they constitute a minority long hidden in history, but because their life trajectories reflect their living conditions and were also shaped by attitudes that prevailed among the majority population.

2 Longitudinal Data and Methods: Definitions of Disabilities, Events and Sequences

The data consist of parish registers, digitized and stored by the DDB at Umeå University, Sweden.[1] These registers are based on original records for parishioners' birth, baptism, marriage, out- or in-migration, death, burial and the catechetical examination records, and they also report occupations. The DDB registers are made up by chosen parishes in Sweden during the 18th and 19th centuries and are linked on an individual level, which gives a demographic description about each parishioner over lifetime (Vikström et al. 2006). The catechetical examination records were collected on yearly basis due to the obligation for the ministers to keep records of the parishioners' knowledge of the catechism and their reading skills, first stated in the Church law of 1686 (Nilsdotter Jeub 2009). In these registers the ministers also made marks of impairments – *lytesmarkeringar* – that indicate disabilities among the parishioners. Those who the ministers labeled disabled are categorized as such in our dataset and analyses (Tab. 1). Those who did not have any of these impairments in the parish registers we regard as non-disabled (Haage 2012). Representing the average life trajectories of the young population in the Sundsvall region, they help us answer whether finding a job or a spouse to marry were less present events in the life of disabled persons.

The dataset consists of observations from 8,874 unique 15-year-old individuals born between 1820 and 1860 out of whom 117 had marks of impairments (Tab. 1). Their impairments were reported before the age of 15 or at the 15th birthday at latest. The reason for chosen young individuals is that they are in the beginning of their transition to adulthood associated with the events under study. We begin to follow them at their 15th birthday to identify the timing and ordering of three events of primary interest to us: first occupation, first marriage and birth of the first child. Events of secondary interest are death and out-migration from the parish. At the longest, all individuals are under study for maximum 18 years (between age 15 and 33), which covers the phase in life when the events of primary concern to us occur.

In this study a sequence is defined as a list of events. As an event gives rise to a new state in the sequence, a sequence can consist of different states. If an event occurs several times in the life trajectory, the sequence is recurrent, but we only consider when an event first happens and treat it as non-recurrent (Abbot 1995; Abbot et al. 1986). Sequence analysis further enables us to detect the states each individual experienced at each time point, which in this study is age (Gabadinho et al 2011). This means that a state describes a person according to certain events at one particular year of his/her life. The length of each state is consequently one year and the whole sequence is maximum 18 years. The first state starts at the 15th birthday and the second state at the 16th birthday etcetera.

---

[1] www.cedar.umu.se/english/ddb/databases/

Table 1. Disabled and non-disabled individuals by gender in the dataset (Digitized parish registers, the Sundsvall region, DDB, Umeå University)

Disability category	Men N	Women N	Total N
Blind	9	5	14
Deaf mute	23	13	36
Crippled	19	5	24
Mental disabilities	20	15	35
Multiple disabilities	4	4	8
Summary disabled	75	42	117
Non-disabled	4,385	4,372	8,757
Total summary	4,460	4,414	8,874

Notes: The individuals were born 1820-1860 and 15 years old at start of observation. We use the terms and concepts the ministers used even if some of these words today are derogatory.

3 First Findings on how Disability Influenced the Life Trajectories of Young Adults

To find whether and how disabled people's life trajectories differed from non-disabled individuals, we performed the analyses in two steps. First, we explored in which order the events of primary interest to us occurred in the individual life trajectory. Second, the entire sequences were investigated according to disability and gender distributed by age, and also with regard to the two events of death and out-migration from the parish.

Analyzing the order of the events of the life trajectories of people observed over the whole period (18 years) helps to picture the events they experienced (Fig. 1), disregarded the timing between these events. Among both genders the differences between the disabled and non-disabled were statistically significant below the 5 %-level.[2] Figure 1 shows that more than 30 % of the disabled men and women did not experience any of the events under study during the entire observation (state 0). Hence, they did not have any occupation, did not marry or got a child. Among the non-disabled people not even 10 % ended up in that state (0). Far more men than women (both disabled and non-disabled) took up an occupation, but stayed unmarried and childless (state 1). Getting a job, marrying and starting a family with at least one child (state 111) was the most common male trajectory, especially for non-disabled men, 58 % compared to 40 % of the disabled men. Even though a considerable share of the women ended up in the state of 110, which means that they married and gave birth to a child but never hold any occupation during observation, this was particularly the case for non-disabled women, 40 %, and not their disabled sisters, 20 %. Among the latter, about four women in ten represented a life trajectory showing no events at all (state 0). While almost 20 % of the disabled women are found to have experienced the event of giving birth to illegitimate offspring (states 10 and 11) only about 7 % of the non-disabled did.

---

[2] Pearson's Chi-squared test: P-value < 0.001 for both men and women
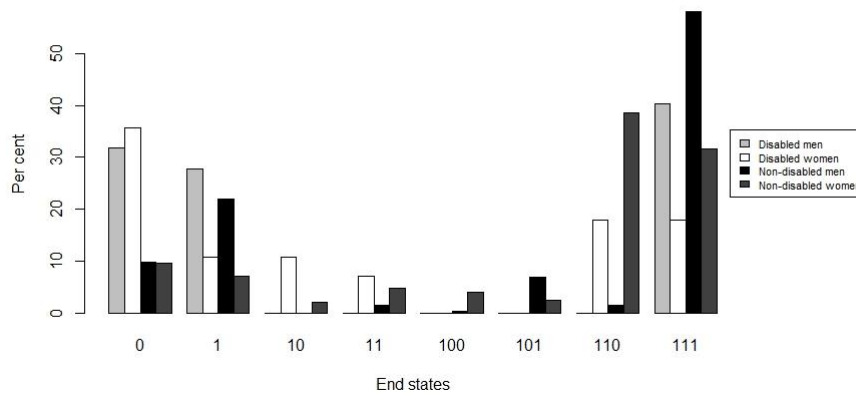
Fig. 1: Percentage distribution of entire sequences (end states) for individuals followed across 18 years: a comparison between disabled and non-disabled men and women (Digitized parish registers, the Sundsvall region, DDB, Umeå University)

Note: Pearson's Chi-squared test: p-value <0.001. The figure shows the eight most frequent orders of states in the dataset, defined as at least one group in the type has a proportion above 10 %. These types are as follow:

0 = No occupation/No marriage/No child   100 = No occupation/Marriage/No child
1 = Occupation/No marriage/No child   101 = Occupation/Marriage /No child
10 = No occupation/No marriage/Child   110 = No occupation/Marriage/Child
11 = Occupation/No marriage/Child   111 = Occupation/Marriage/Child

Sequence analysis further enables a graphical view of the state distributions by time points that displays a general pattern for all individuals' life trajectories by group (Gabadinho et al 2011). Figure 2 shows these distributions by time points (here age) per gender and disability. The proportion of disabled men who did not get any occupation and remained unmarried and did not get any child (green state) is greater than for the non-disabled men. The proportion decreases in a slightly slower pace, and delayed in time, for disabled men, which means that those who left the "green" state and got an occupation did so later in life than did non-disabled men. At the age of 28, the slope for disabled men levels off at about 20 % while it for non-disabled men continues to decline. The share of men who got an occupation without experiencing marriage and parenthood (yellow state) is similar despite disability. The only state where the disabled men were proportionally greater than for the non-disabled is that of death (orange state). The development over time for the state equal to having an occupation, being married and forming a family (purple state) shows no large difference between non-disabled and disabled men.
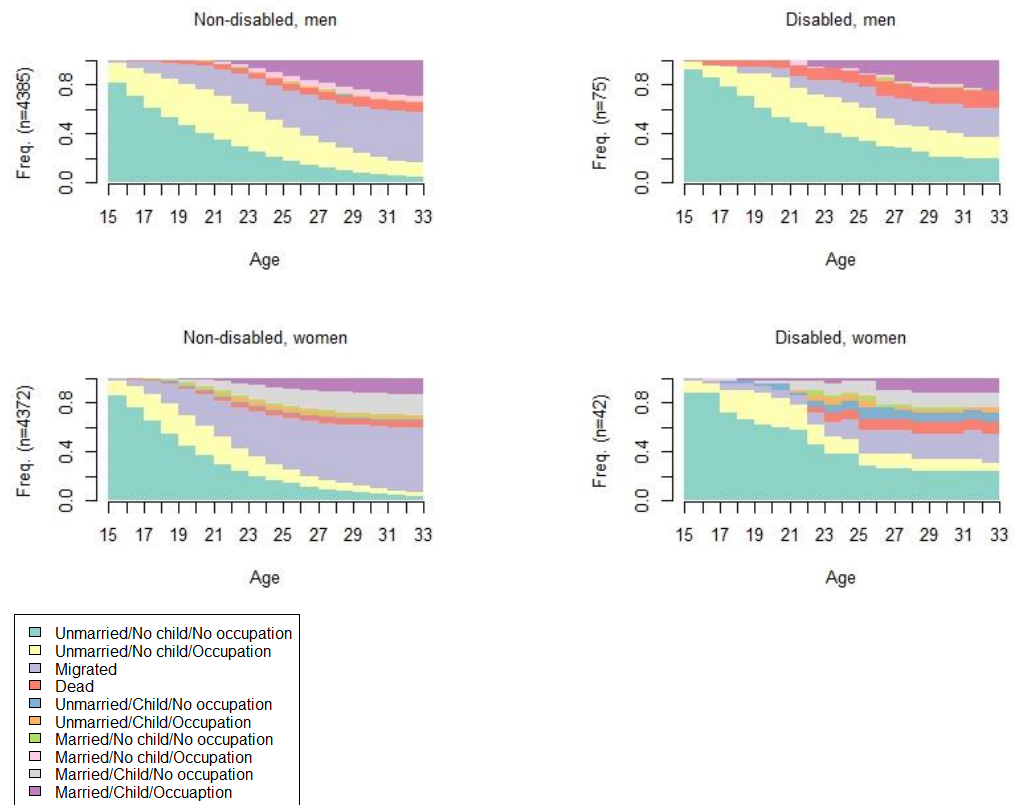
Fig. 2: State distribution by time points per gender and disability; a comparison between disabled and non-disabled men and women (Digitized parish registers, the Sundsvall region, DDB, Umeå University)

The differences between disabled and non-disabled women is clearer than among the men. The proportion of women who experienced no occupation, remained unmarried and childless (green state) is greater among the disabled women than their non-disabled sisters. However, women who attained an occupation, marriage and a child (purple state) have a similar trajectory over time despite disability. The two states where the trajectories differed substantially between disabled and non-disabled women are the state equal to getting a child without marriage and occupation (blue state), and the one equal to death (orange state). Among those who migrated from the parish, we find a larger proportion of non-disabled individuals of both genders. In all, the above sequence patterns show more similarities between non-disabled men and women and between disabled men and women respectively. Consequently, disability influenced the life trajectories of both genders.

4 Concluding Discussion

The findings of our sequence analyses provide a comparatively complete picture of young disabled individuals as they were beginning to seek their livelihood as adults in 19th-century Sweden. With regard to opportunities for work and family, we find substantial differences between how their trajectories developed when compared to non-disabled individuals. Whereas the latter moved through many events while under observation (from 15 years of age to maximum 33) and experienced these events in the expected ordering (first job, then marriage and parenthood), disabled people did not to the same extent. Even if it was not impossible for them take up work, which was key to marry and form a family, considerably fewer of them found a job or married a spouse compared to non-disabled people. In all, the sequence analyses provide results that clearly demonstrate that disability limited people's opportunities in both the labor market and marriage market. This is one major conclusion to be drawn; yet another one is that these results would hardly been obtained without employing sequence analysis. We view this tool as beneficial for identifying the impact disability had on people's opportunities in past society, here to work and family, and for two reasons. First, sequence analysis covers a longer portion of people's pathways than for instance Cox regression models do, in estimating the propensity to experience one single event, such as getting a job or marrying. Second, and although the explanatory power of sequence analysis is limited, it holds a pedagogical advantage in describing individual life trajectories as it does. The outcome is relatively comprehensible, even for scholars less familiar with statistical measures, many of whom are found in the fields of social history or disability studies, for example.

However, why do we come across the above results? Although impairment certainly brought difficulties in life, it must not per see have ruined people's opportunities to find some work and access to income or a spouse to marry. It takes more to explain why these typical events were less present in the life of disabled people. Beside the impairment itself, we argue that a stigma associated with being labeled disabled adds to explain the results. If individuals' appearance, behavior or inability were perceived as deviant according to prevailing perceptions about normalcy in society, this would render them exclusion from social and working life. We think that these labeling circumstances and negative attitudes contribute explanation to the limited opportunities to work and family among the disabled people we study. Less work or no job, cut their income and subsequently the marital prospects, as we have seen from the above sequence analyses. This suggests that disabled individuals accumulated disadvantages over lifetime that is also evident from their higher mortality level during observation compared to non-disabled people.

Acknowledgements

References

Archival sources:

Demographic Data Base (DDB), Umeå University, Sweden
Digitized parish registers and catechetical examination records from the following parishes:

Alnö	1803–1894
Attmar	1814–1896 (deficient records 1860–1868)
Hässjö	1814–1901
Indal	1814–1900
Ljustorp	1803–1901
Njurunda	1816–1891
Selånger	1813–1894
Skön	1803–1893 (Skönsmon included until 1883)
Sundsvall	1803–1892
Sättna	1806–1899
Timrå	1803–1895 (incl. *mantalsregister* 1852–1865)
Tuna	1804–1896
Tynderö	1811–1900

Literature

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

Abbot, A. and J. Forrest (1986). Optimal Matching Methods for Historical Sequences, *The Journal of Interdisciplinary History* 16(3), 471-494. DOI: 10.2307/204500

Abbot, A. and Tsay, A. (2000) Sequence Analysis and Optimal Matching Methods in Sociology. Review and prospect, *Sociological Methods & Research* 29(1), 3-33. DOI: 10.1177/0049124100029001001

Alm Stenflo, G. (1994) *Demographic Description of the Skellefteå and Sundsvall Regions during the 19th century.* Umeå: Demographic Database.

Aisenbrey, A and Fasang, A. E. (2010) New Life for Old Ideas: The "Second

Wave" of Sequence Analysis Bringing the "Course" Back Into the Life Course, *Sociological Methods & Research* 38(3), 420-462. DOI: 10.1177/00491241093557532

Barnes, C. and Mercer, G. (2010) *Exploring Disability. A Sociological Introduction – Second Edition.* Cambridge: Polity Press.

De Veirman, S. (2015) *Breaking the silence: the experiences of deaf people in East Flanderns, 1750-1950: a life course approach.* Dissertation. Ghent University.

Dribe, M. (2000) *Leaving Home in a Peasant Society. Economic Fluctuations, Household Dynamics and Youth Migration in Southern Sweden, 1829-1866.* Dissertation, Lund University.

Gabadinho, A., Ritschard, G., Müller, N. S. and Studer, M. (2011) Analyzing and Visualizing State Sequences in R with TraMineR, *Journal of Statistical Software* 40(4), 1-37. DOI: 10.18637/jss.v040.i04

Goffman, E. (1972) *Stigma. Den avvikandes roll och identitet* (trans. Richard, M). Stockholm: Prisma. English edition: Goffman, E. (1963). *Stigma, Notes on the management of spoiled identity*. Engelwood Cliffs, New Jersey: Prentice Hall Inc.

Haage, H. (2012) Identifying Disability using Nineteenth Century Parish Registers." Paper presented at the Workshop on Categories and Concepts in Health, Medicine and Society, Umeå, March 15-17, 2012.

Haage, H., Häggström Lundevaller, E. and Vikström, L. (2016) Gendered Death Risks among Disabled Individuals in Sweden: A Case Study of the 19th-century Sundsvall Region, *Scandinavian Journal of History,* 41:2, 160-184. DOI: 10.1080/03468755.2016.1155859

Harnesk, B. (1990) *Legofolk.* Stockholm: Almqvist & Wiksell International.

Kudlick, C. J. (2003) Disability History: Why we Need Another Other, *The American Historical Review* 108, 763-793. DOI: 10.1086/529597

Lundh, C. (1999) Servant Migration in Sweden in Early Nineteenth Century. *Journal of Family History,* 24(1), 53-73. DOI: 10.1177/036319909902400104

Lundh, C. (2003) Life Cycle Servants in Nineteenth Century Sweden – Norms and Practice. Lund: Lunds Papers in Economic History, No 84.

Nilsdotter Jeub, U. (2009) Parish Records, Digitalised Material from the Demographic Data Base. http://www.cedar.umu.se/digitalAssets/81/81164_parishrecords.pdf. Accessed December 17, 2015.

Oliver, M. (1996) *Understanding Disability. From Theory to Practice.* New York: St. Martins Press cop.

Susman, J. (1994) Disability, Stigma and Deviance. *Social Science and Medicine* 38, 15-22.

Vikström, P., Brändström, A. and Edvinsson, S. (2006) Longitudinal Databases: Sources for Analyzing the Life Course, History *and Computing* 14, 109-128.

Whittle, J. (2005) Servants in rural England c.1450 – 1650: hired work as means of accumulating wealth and skills before marriage, in M. Ågren and A. L. Erickson, A. L. (eds), *The Marital Economy in Scandinavia and Britain 1400 – 1900,* pages 89-107, Aldershot: Ashgate.

Session 4A: Social Policy

Employment security in non-traditional careers: Exploring the dynamic of long-term work trajectories in thirteen European countries

Maxim Kovalenko and Dimitri Mortelmans

Abstract *In this article, we approach the debate on the so-called flexibility-security nexus from the long-term perspective of career studies, exploring empirical evidence on whether the increased labor market flexibility in the past decades has also led to more employment insecurity. We look at lifelong career sequence patterns in 13 European countries, with three objectives in mind. First, we examine to what degree empirical career patterns correspond to the career types postulated in theoretical career literature, namely the traditional and the 'new' career. Second, we look at whether these career patterns have changed in their relative prevalence over time. This analysis provides evidence for the discussion on whether career patterns in Europe have changed in the past decades. Third, we consider how employment security has evolved over time, depending on the career type. Our analyses indicate that there are slight changes in career structure in several European countries, and that some of the highly mobile career patterns are characterized by a higher degree of precariousness, a dynamic often overlooked in mainstream career studies. In terms of policy, our results suggest the need to combine labor market flexibilization with individual supportive measures for some workers.*

1 Introduction

The debate on whether the increasing flexibility of employment relationships in the past few decades implied a lower degree of labor security, is prominent in the contemporary labor literature (Muffels and Luijkx, 2008; Standing, 1999; Heery and Salmon, 2000; DiPrete et al., 2006; Kalleberg, 2009; Cappelli, 1999; Barbieri, 2009; Howell, 2004). The same debate is much less well developed in the

Maxim Kovalenko
University of Antwerp, Belgium e-mail: maxim.kovalenko@uantwerpen.be

Dimitri Mortelmans
University of Antwerp, Belgium

field of career studies, which approaches the flexibility problematic from an explicitly long-term perspective and considers how individual working histories evolve over time (see e.g. Arthur and Rousseau, 1996; Baruch and Bozionelos, 2010; Sullivan, 1999). Despite several calls to pay more attention to the risks associated with the flexible career (e.g. Van Buren, 2003), the subject remains essentially understudied. Yet given the mainstream view of public policy makers in the Western countries that labor market flexibility is an essential tool for driving up economic competitiveness in the globalizing economy and for reducing unemployment, a proper understanding of these risks is crucial for all actors involved. Neglecting to account for the 'dark side' of flexibility may all too easily help install or strengthen the mechanisms of stratification on the labor market, and expose many workers to the risk of precarious work and poverty.

In this article, we empirically examine several assumptions that underpin mainstream theories on careers in flexibilizing labor markets, and look at the evolution of labor security over time for different career patterns in 13 European countries. These two interrelated issues form the backbone of our analysis.

In regard to the first issue, the central thesis of career theories dealing with the consequences of labor market flexibilization pertains to what can be termed 'the career transformation': the demise of the traditional career pattern with life-long employment in one or two organizations (Sullivan, 1999), and its replacement by the 'new' career, which is often characterized by higher external mobility (Sullivan and Arthur, 2006). Much research has been inspired by the assumption that such transformation took place, yet the empirical evidence on the extent of the change is far from being conclusive (e.g. see Gunz et al., 2011; Clarke, 2013). The first possible reason for this is that insufficient attention has been paid so far to the national variation of how career patterns have evolved over time. A lot of research on the 'new' career originates in the Anglo-Saxon countries, where labor markets are much less regulated than in Continental Europe. At the same time, it has been convincingly demonstrated that national institutions can significantly impact labor market outcomes (Muffels and Luijkx, 2008; DiPrete et al., 2006), which raises the question if the same claim about the career transformation can be applied to much more strictly regulated European labor markets, and to what degree. Second, the existing findings on the career transformation are predominantly based on short- or medium-term data, whereas research based on life-long data, so crucial for the understanding of careers in their totality, is very scarce.

Concerning the second issue that pertains to changes in labor security, it remains an open question as to what impact career flexibilization in Europe, whatever its extent, has had on the course of the individual career. Much of empirical research in the contemporary career literature has focused on factors that help workers navigate flexibilizing and uncertain employment environments, resting on the assumption that the risks associated with the increasing labor market flexibility can be tackled on the individual or organizational level, e.g. through investments in employability and self-directedness in career management. This approach implicitly leaves the insecurity-generating

factors on the macro level outside its field of vision, therefore potentially misrepresenting the total outcome of the flexibility-insecurity dynamic; and possibly suggesting a more positive view on career flexibilization than the reality would warrant (Guest et al., 2006).

In this article, we aim to contribute to a better understanding of career flexibilization and its consequences in Continental Europe. We start by describing the concerns related to the increasing labor insecurity that are voiced by some strands in economic and sociological literature on the subject of labor market transformations in the context of economic globalization and neoliberal marketization. Subsequently, we explore how these concerns connect with the prevailing theoretical insights within the contemporary career studies and provide empirical evidence that helps transpose the debate on the flexibility-insecurity controversy to the long-term career perspective. Our analyses shed empirical light on the dominant assumptions of the theories dealing with the allegedly 'new' career types, and provide insights into long-term changes that have occurred in careers of European workers in the past decades, pertaining both to shifts in career structure and to the dynamic of employment security. In addition, our analyses address a lacuna in research pertaining to the evolution of careers in Europe, given its characteristic regimes of labor market regulation.

2 Theoretical background

2.1 The debate on flexibility-security nexus

The last three decades of the twentieth century were marked by major transformations in the Western economies, spurred, among other factors, by an unprecedented rate of economic globalization, technological advances in transport and communication, the nascence of the Internet, increasing global competition and labor redistribution (Beck, 2000; Reich, 2008; Standing, 1999, 2009). These transformations went hand in hand with a Copernican shift in the dominant socio-economic paradigm, where the Keynesian model gave place to the neoliberal modes of economic thought and policy action (Harvey, 2011). Labor market deregulation took place in most Western countries, eroding to a large degree the traditional securities built during the Golden Era of full employment (Harvey, 2011; Standing, 1999). The 'new' economy, featuring higher levels of competition, market variety and technological complexity, required more flexibility from both organizations and the workers (Kalleberg, 2009). A result was the proliferation of various forms of flexibility, including increased job mobility, fixed-term contracts, subcontracting, freelance work, as well as other forms of 'atypical' working arrangements. There is an on-going debate in the sociological, economic and public policy literature about how the increasing flexibility in the labor markets has impacted workers (Muffels and Luijkx, 2008). Generally speaking, two perspectives can be discerned (see Tregaskis et al., 1998).

4   Maxim Kovalenko and Dimitri Mortelmans

The first perspective, sometimes labeled as 'neo-Fordist' or even 'neo-Marxist', assumes a negative view on flexibilization. It emphasizes an increasing labor insecurity due a plethora of macro-level factors inherent to the 'new' economy and its mainstream neo-liberal policies. Labor markets have become much less regulated, driven to a larger extent by the pure market logic of supply and demand, whilst disembedding themselves from the social institutions designed to protect weaker labor market groups. Welfare provisions have been trimmed down substantially, and the responsibility for one's employment has been transferred onto the workers, the trend finding a reflection in the re-orientation of labor market policies towards activation and stricter controls (Harvey, 2011). The use of non-standard employment relationships, such as temporary work, fixed-term work and self-employment (of the real or the bogus kind) has increased in many countries (Standing, 2009; Tregaskis et al., 1998). Flexible working arrangements of these and other types allow companies to make quick adjustments to the business cycle and to changes and shocks on the volatile and competitive markets. At the same time, it is feared that such arrangements do not offer the same degree of social and economic security as the standard full-time employment (Van Buren, 2003). In addition, trade union power has been in decline as well in the past decades, and along with it the traditional means of reducing inequalities between employers and employees (Standing, 2009). Finally, the ability of capital to move quickly and the global nature of economy allows firms, especially multinationals, to traverse international borders to find the economic regime most favorable to business, which usually comes at the cost of lower labor protection.

All in all, this perspective posits that as a consequence of these changes, many workers in weaker positions on the labor market, such as lower educated workers or migrants, are now at higher risk of precarious employment, unemployment and poverty (Standing, 2011; DiPrete et al., 2006; Rodrigues and Guest, 2010). Some authors go even further, proposing that insecurity is inherent to advanced capitalist societies in general, without restriction to specific population strata (Beck, 2000).

The second perspective, sometimes labeled as 'post-Fordist', assumes a positive stance towards flexibility, focusing on the mechanisms that allow labor flexibility whilst maintaining labor security. In this view, the relationship between flexibility and security is not a trade-off, but rather that of mutual reinforcement (Muffels and Luijkx, 2008). According to this perspective, not an erosion of labor security takes place, but rather a shift towards new forms of security. Thus, management literature focuses on how individuals can adapt to uncertain employment conditions of the flexible economy, emphasizing the importance of continuous investment in skills as well as adaptability to the shifting demands of the labor market (e.g. Clarke and Patrickson, 2008; Van Buren, 2003). Proactive and 'self-directed' behavior on the labor market, self-knowledge and employability are key concepts in this 'new employment relationship' approach. On the macro-level, the 'flexicurity' paradigm, popular in the European policy circles (European Commission, 2007; Wilthagen and Tros, 2004), supports the positive view on the flexibility-security nexus. It focuses on the role of institutions in facilitating employability and activating workers towards labor market participation.

While the debate on the consequences of flexibility in the global economy is well-developed in relation to labor in general, this is hardly the case on the level of career studies. The dominant paradigms pertaining to the dynamic of careers in the 'new' economy are by and large aligned with the 'new employment relationship' approach, focusing on the individual capacities to cope with uncertain employment environments. They often ignore the discussion of the mechanisms on the global scale that systematically generate insecurity which may impact career trajectories in the long run (Inkson et al., 2012; Tregaskis et al., 1998). As consequence, these theories can be dangerously open to ideological interpretation (Roper et al., 2010).

2.2 Flexibilization and the new career

The process of globalization has also had important consequences for the structure of career opportunities and career enactment (Arnold and Cohen, 2008). Changes in the domain of work have spurred a new branch in career research and theory that came to focus on what is often described as 'the new career' (Arthur et al., 1999). Different aspects of the 'new career' have been captured in various theoretical frameworks, most popular being the 'boundaryless career' (Arthur and Rousseau, 1996) and the 'protean career' (Hall, 1996, 2004). Both frameworks postulate a transformation of how careers unfold in the context of flexibilizing labor markets. This transformation allegedly entailed a move away from the traditional organizational career within one or two organizations (Sullivan, 1999), and to a modern career type that is characterized by a higher degree of psychological and physical mobility (Sullivan and Arthur, 2006), as well as by a weaker dependence on a particular organization in terms of its development.

Several authors have noted that the discourse of the 'new' career is inherently positive, and dangerously aligned with the currently mainstream neo-liberal view on labor, which valorizes individualism, self-reliance and the ability to accept responsibility for own actions (Guest et al., 2006; Roper et al., 2010; Inkson et al., 2012; Zeitz et al., 2009). It paints modern workers as 'career capitalists' (Inkson and Arthur, 2001), that must assume control over their working lives and market themselves in an entrepreneurial fashion (Zeitz et al., 2009), leveraging their human and social capital to traverse organizational, cultural and occupational boundaries (Pringle and Mallon, 2003).

These commentators have raised concerns that the 'new' career discourse ignores the increasing insecurity and uncertainty that typifies the flexibilizing labor market (Tams and Arthur, 2010; Inkson et al., 2012; Zeitz et al., 2009). According to this view, the freedom of individual action can be severaly constrained by contextual factors, which can result in a 'bifurcation in the labor market between those in a position to reap the benefits of the new, flexible career environment and those less able to gain a foothold' (Arnold and Cohen, 2008, p. 4).

Despite these calls to explore the 'dark side' of the modern career, empirical evidence is still scarce, especially for European contries, concerning how employment security has evolved in typical career patterns, given the vast changes in the economic order in the past several decades. This scarcity leaves the debate on the impact of labor market flexibilization on labor security, considered from the long-term career perspective, rather one-sided and predominantly geared towards its positive interpretation.

2.3 Hypotheses

In our analyses we focus on two separate, but closely interrelated sets of research questions. The first set pertains to the empirical validation of the assumptions that the 'new' career theory makes in postulating the shift from the 'traditional' to the 'new' career type, as described hereinabove.

These assumptions are largely accepted in the contemporary career literature, yet the empirical evidence that supports them is limited at best (Rodrigues and Guest, 2010; Clarke, 2013), especially in the European context (Kattenbach et al., 2014). More specifically, there exist substantial doubts as to whether external job mobility, a cornerstone element of the 'new' career, has effectively increased (Chudzikowski, 2012; Soens et al., 2005; Rodrigues and Guest, 2010; Kattenbach et al., 2014). Similar claims have been made in regard to career behavior and perceptions in general (Jacoby, 1999; Elchardus and Smits, 2014), implying that the traditional career is alive and well, certainly in the context of regulated European labor markets. In respect to the changes in career type prevalence, postulated by the 'new' career theories, we formulate the following hypotheses:

Hypothesis 1a: the theoretically postulated 'traditional' and 'new' career types can be identified empirically;

Hypothesis 1b: a shift has occurred in the prevalence of the two career types, with a relative decrease for the 'traditional' type in favor of the 'new' career type.

Detecting the empirical career patterns that correspond to the theoretical ideal types of the 'traditional' and the 'new' career, enables us to trace changes in career security over time in each career type. Different strands of theoretical literature offer opposing views in regard to changes in employment security in careers over time. It remains an open question, from the empirical perspective, whether there was a change, and in which direction. Moreover, it remains unclear to which degree the direction and the extent of changes in employment security depend on the career type. For example, it is possible that the core employees, who are more likely to have stable traditional careers, could have experiences different from thos of the periphery workers, who are more likely to have more mobile careers (see Rodrigues and Guest, 2010). The following hypotheses can be formulated:

Hypothesis 2a: employment security in careers has decreased over time;

Employment security in non-traditional careers 7

Hypothesis 2b: changes in employment security in careers are contingent on the specific career types.

2.4 Defining and operationalizing employment security on career level

Employment security is one of the seven forms of labor security, along with (income security, labor market security, work security, job security, skills reproduction security and representation security (e.g. Standing, 2011). Following Muffels and Luijkx (2008) and Wilthagen and Tros (2004), we define employment security as staying in employment, but not necessarily in the same job with the same employer. On the level of career, we consequently operationalize employment security as the percentage of career time that was spent in unemployment.

2.5 Control variables

In our analyses we control for gender and educational level, as both have been shown to be related to employment security (Azmat et al., 2006; Núñez and Livanos, 2010). In addition, we control for career duration to account for the heterogeneity in this respect.

In our analyses we will also consider how between-country differences in labor market mobility affect employment security. Mobility on the labor market is considered as one of the macro-level flexibility indicators (Muffels and Luijkx, 2008; Klau and Mittelstadt, 1986), and is often touted in European policy circles as an instrument for tackling labor market rigidity. It is therefore important to establish what effect does macro-level labor market mobility have on employment security, and how does it relate to individual-level career mobility.

3 Data and Methods

3.1 Data and Sample

To answer our research questions, we use data from SHARELIFE release 1.0, as of November 24, 2010, or SHARE release 2.5.0, as of May 24, 2011 (see Börsch-Supan, Hank, & Jürges (2005)). SHARELIFE data were collected in 2009 in 13 European countries, and contain a full retrospective record of working career mobility and its timing. Countries in the analysis are: Austria, Germany, Sweden, the Netherlands, Spain, Italy, France, Denmark, Greece, Switzerland, Belgium, the Czech

Republic, and Poland. The main selection criterion for the respondents was to be of age 50 or older at the first wave of the survey in 2004; partners or spouses of the initially chosen respondents were selected, when available, without considering age. Complete career records for the majority of the respondents are therefore available. The complete sample consisted of 18,841 respondents in all countries. Having excluded cases with missing values on one or several model variables, 16,673 respondents were retained in the analytical sample.

3.2 Methods

3.2.1 Optimal Matching Analysis

In respect to Hypothesis 1a, we establish typical career patterns in each country. When careers are conceptualized as sequences of work-related statuses, it becomes possible to derive a distance matrix between such sequences using Optimal Matching Analysis (Anyadike-Danes and McVicar, 2010). In our case, the principal statuses were *inactivity*, *unemployment*,*employment*, and *retirement*, with an additional splitting of *employment* into statuses pertaining to the sequential number of jobs in a career (e.g., 1st job, 2nd job etc). Fifty years counting from career start, were considered in the sequence analysis. The Levenstein I distance algorithm was used, with indel cost set to 1 and substitution cost set inversely proportional to frequencies at which respective transitions between statuses occurred (Lesnard, 2010).

The distance matrix was supplied as input to the Ward clustering algorithm, which yielded most typical career patterns based on the statuses mentioned above. Classifications were carried out for men and women in each country separately, to preserve gender-specific career types.

3.2.2 Kolmogorov–Smirnov test

In respect to Hypothesis 1b, we use the two-sample Kolmogorov-Smirnov test, which allows to detect differences in two distributions. Using this method, we can observe whether the occurrence of the 'new' careers has increased over time, relative to its traditional counterpart. A separate test was carried out for each country.

3.2.3 Hierarchical modeling using MCMC

In respect to Hypotheses 2a and 2b, we use two-level hierarchical modeling of unemployment proportion in a career sequence, level one pertaining to individuals and level two to countries. Career

types obtained in the OMA analysis are entered into the model as predictors, along with a set of control variables, with the goal of comparing the evolution of employment security between the typical career types.

There is a technical caveat in this part of the analysis, related to the fact that the number of countries is relatively low (N = 13). This implies using maximum likelihood-based estimation is not feasible, due to likely distortions of the standard errors. At the same time, hierarchical modeling offers crucial analytical advantages, such as the estimation of individual variance, having controlled for variance between countries. This renders the method preferable to alternative approaches such as generalized estimating equations (GEE). The issue was resolved by using Bayesian estimation methods; namely, Markov Chain Monte Carlo (MCMC) estimation. Bayesian methods have gained substantial ground in the social sciences in recent decades, especially in hierarchical modeling due to the mathematical properties of the method (Lynch, 2007). Their efficiency is supported by simulation studies showing that the MCMC approach leads to more adequate model estimation in comparison with the traditional maximum likelihood-based algorithms in analytical situations with a low number of level-two units (Stegmueller, 2013).

In order to trace evolutions in employment security in career structure over time, we employ the following analytic strategy. Given our operationalization of employment security, career trajectories are treated as primary units of analysis. Therefore we do not trace changes *within* a career trajectory (as the sum of these changes already constitutes that trajectory), but we explore whether careers occurring later in time are characterized by higher or lower employment security. This achieved by taking the year in which a career has started (i.e. the respondent entered the labor market for the first time), as one of the predictors for employment security. Careers that have commenced at a later time, especially after 1960s, have been to a higher degree exposed to the pressures of economic globalization and labor market flexibilization, that have precipitated the career transformation postulated by the 'new' career literature. Any effect on the start year variable would therefore signify a potential change in career dynamics.

3.3 Measures

3.3.1 Dependent variable

Employment security in a career is measured as the number of years spent in unemployment divided by the total number of years in a career. This is expressed as a percentage and varies per definition from 0 to 100.

3.3.2 Predictors and control variables

Career type is a dichotomous variable comparing two career types: traditional and transitional. The variable will be defined based on the results of the Optimal Matching Analysis (OMA) of careers sequences; see the corresponding section below for details. Briefly, the traditional career type is characterized by a prolonged employment period with a single organization towards the end of a career, whether or not it is accompanied by one or several job-to-job transitions in the beginning of the career. The transitional career type is characterized by multiple job-to-job transitions throughout the entire career span.

Career start year is expressed as two last digits of the year in which a career has commenced. The variable was centered 60 (thus pertaining to the year 1960), the value approximating its mean. The distribution of the variable is presented in Figure 1.

Education is the number of years spent in full-time education in the initial educational trajectory. The variable was centered around its mean. *Labor market mobility* is a country-level predictor, expressed as an average number of jobs in a career for a given country. It reflects the general degree of mobility on the labor market. *Career length* is the number of years spent on the labor market. We introduce it as a control variable to correct for careers of unequal lengths. *Gender* is coded as 0 for men, 1 for women.

4 Results

4.1 Career typology

As the result of OMA, we have obtained seven main career types. Figure 2 presents the overall distribution of career statuses in time for each career type.

Two career types are present in all countries in the analysis except Greece, for men as well as for women: traditional career and transitional career. The *traditional* career (TD) is distinguished by having a prolonged period of employment within the same organization at the end of the career trajectory, regardless of whether it was preceded by one or several transitions in the beginning.

The *transitional* career (TS) is characterized by a high number of career transitions, generally remaining in employment. In contrast with the traditional career, career transitions continue after mid-career, whereas in the former type the worker came to be attached to a single organization. In terms of physical career mobility, this type is in line with the career trajectory postulated by the 'new' career theories.

Late mover (LM) and *inactive* (IN) career types are also present for both genders, but not in all countries. The late mover type resembles the traditional career in having low mobility and

Employment security in non-traditional careers   11
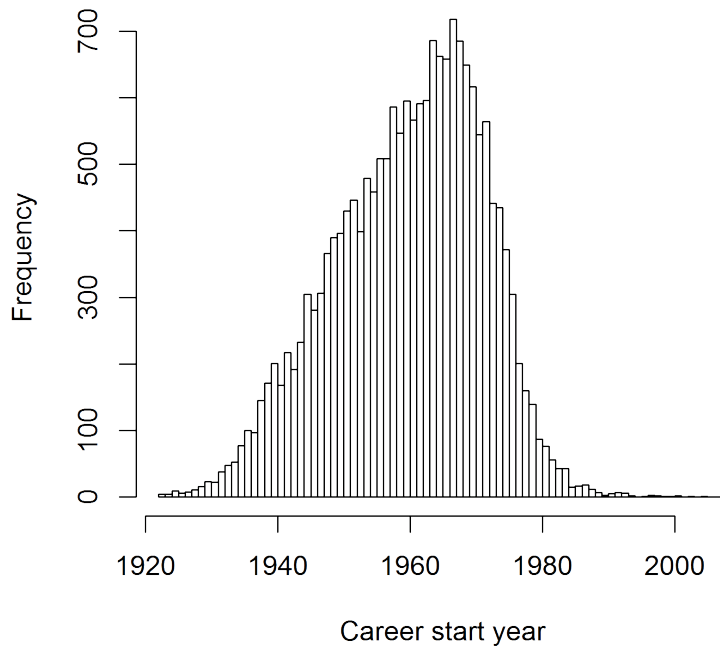
Fig. 1 Distribution of career start year in the study sample

prolonged attachment to the employing organization. The difference between the two is in the temporal placement of that attachment. For the late mover career, the period of attachment occurs in the beginning of the career, whereas for the traditional it is at the end. The inactive career type is marked by a relatively short period of employment, followed by labor market inactivity until retirement. This pattern is particularly characteristic for women, being present in all countries. For men, this career type is found only in Spain and Greece.

The remaining three career types are gender-specific. The *unemployed* career (UN) is characterized by prolonged periods of unemployment. It is typical for women only. In Belgium and the Netherlands the *intermittent* career (IT) type could be distinguished for men. Similar to the transitional career,

12   Maxim Kovalenko and Dimitri Mortelmans
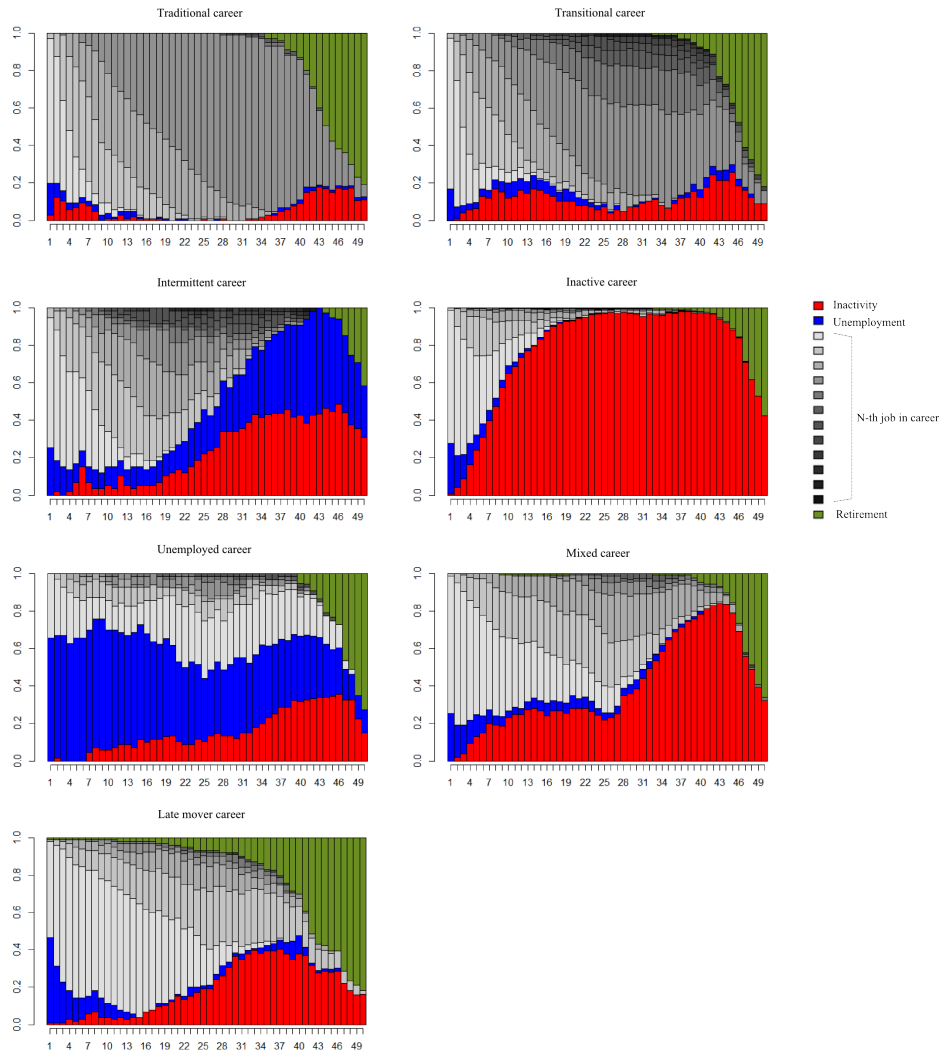
Fig. 2 Distribution of career statuses (y-axis) on career timeline (x-axis, years from career start). Source: Sharelife, own calculations.

Employment security in non-traditional careers 13

it is characterized by multiple career transitions throughout its entire course, yet the career is often interrupted by unemployment, especially in its second half.

The *mixed career* (MX) is typical for women in most countries. In this career type, periods of inactivity (roughly corresponding to the child rearing period in most cases) are combined with periods of employment. This type can be clearly distinguished from both the inactive and traditional types, being a combination of the two.

Table 1 presents the distribution of these careers types in the countries in the analysis.

Table 1 Career type distribution per country and gender (proportion and total)

Career type	TD	TS	LM	IN	MX	UN	IT	Total
Men								
Austria	0.88	0.05	0.07					332
Germany	0.75	0.19	0.06					854
Sweden	0.82	0.11	0.06					837
Netherlands	0.72	0.09	0.12				0.08	987
Spain	0.91	0.07		0.02				883
Italy	0.85	0.05	0.09					1113
France	0.80	0.10	0.10					1019
Denmark	0.74	0.26						940
Greece	0.86	0.03	0.10	0.02				1215
Switzerland	0.69	0.19	0.12					551
Belgium	0.90	0.05					0.05	1242
Czechia	0.72	0.14	0.14					784
Poland	0.72	0.12	0.16					817
Women								
Austria	0.47	0.15		0.35	0.03			438
Germany	0.53	0.13	0.11	0.23				941
Sweden	0.70	0.20		0.10				1013
Netherlands	0.39	0.08		0.41	0.11			1114
Spain	0.36	0.14		0.41	0.10			752
Italy	0.42	0.05		0.35	0.12	0.06		947
France	0.56	0.08		0.28	0.08			1233
Denmark	0.53	0.33		0.14				1137
Greece	0.73			0.13	0.14			928
Switzerland	0.48	0.23		0.29				697
Belgium	0.44	0.09		0.29	0.13	0.06		1334
Czechia	0.79	0.08	0.12	0.01				1061
Poland	0.56	0.09	0.07	0.23		0.05		951

These results support Hypothesis 1a. For tracing the evolution of employment security, we will retain only two types: traditional and transitional careers. These types correspond, in terms of mobility, to respectively the theoretical concepts of traditional and 'new' careers. Given how these two types are usually juxtaposed in theoretical literature, it is pertinent to contrast them in regard to employment security.

4.2 Shift in career type prevalence

Two-sample Kolmogorov-Smirnov test shows that the distribution of the traditional and the transitional patterns over time is different for the following six countries: Spain (p < 0.001), France (p = 0.023), Denmark (p = 0.04), Greece (p = 0.012), Czech Republic (p < 0.001), and Poland (p < 0.001). In all these countries except Greece the relative prevalence of the transitional pattern has increased (see Fig. 3), but not to the extent that would signify a replacement of the traditional type. For Greece, the shift has occurred in the opposite direction, the relative prevalence of the transitional pattern has increased over time. In other countries the test was insignificant, implying that no substantial shifts with respect to these types have occurred over time.

4.3 Evolution of employment security

The results for the hierarchical model of employment security evolution are presented in Table 2.

When interpreting the model, one needs to keep in mind that negative coefficients imply an improvement in employment security (decreasing the proportion of a career spent in unemployment), and vice versa. Most importantly, on the average, we observe an increase over time in employment security for the traditional career type, coupled with a decrease over time in employment security for the transitional career type. These findings support the Hypothesis 2b, and support Hypothesis 2a for the transitional career type, while rejecting it for the traditional type. The transitional type in itself is related to decreased employment security. Education, as expected, has a positive effect on employment security, whereas the female gender has a negative effect.

It is interesting to note that the contextual impact of the country-level mobility indicator, namely the average number of jobs in a career in a specific country, is positive in its main effect on employment security, but negative in its interaction with the transitional type. This implies that there is a differential contextual labor market effect for traditional and transitional careers. For both career types, being in a labor market with higher career mobility increases employment security, but more so for the traditional career type than for the transitional.

Employment security in non-traditional careers

15
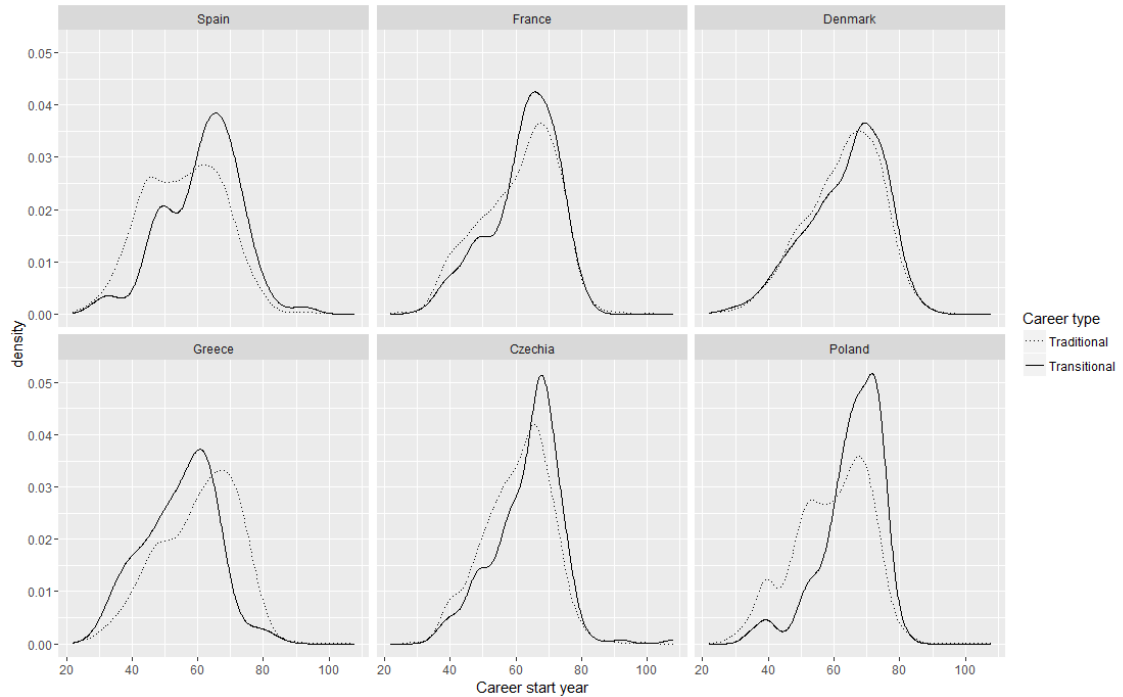
Fig. 3 Distribution of career start year density for countries with statistically significant differences between career types, based on two-sample Kolmogorov-Smirnov test.

An interesting observation can be made in regard to the differences between countries and individuals in how employment security has evolved in the two career types. Figure 4 shows predicted regression lines of how employment security changed for the traditional and the transitional career types. For the traditional career type, the estimations show a decrease of the career proportion spent in unemployment (i.e. increase in employment security) in all countries. For the transitional career the evolution has been more heterogeneous, some of the countries showing a positive trend, while the others—a negative one.

The contrast in the dynamic of employment security between the traditional and the transitional career remains even after the differences between the countries have been accounted for. Figure 5 shows the individual variance of employment security, plotted against career start time. The lower curve pertains to observations with the traditional career pattern, while the upper curve refers to observations with the transitional pattern. Both curves show an increase of variance, meaning that the internal differences within each career type have grown. Yet for the transitional career that

Table 2 Multilevel model of employment security

Fixed effects	Coef (SD)	95% CI		ESS
Constant	3.602***(0.523)	2.557	4.638	25000
Education	-0.146***(0.015)	-0.177	-0.116	23481
Transitional	1.482***(0.219)	1.053	1.912	25000
Gender	1.016***(0.122)	0.779	1.254	24418
Career start	-0.109***(0.024)	-0.157	-0.062	25000
Career start * Transitional	0.135**(0.044)	0.051	0.224	20581
Average jobs	-1.605**(0.610)	-2.817	-0.392	25000
Average jobs * Transitional	0.775**(0.240)	0.300	1.251	24378
Career length	-0.144***(0.021)	-0.184	-0.103	25000
Observations	16,673			
DIC	115589.102			
Note:		$^{*}p<0.05$; $^{**}p<0.01$; $^{***}p<0.001$		
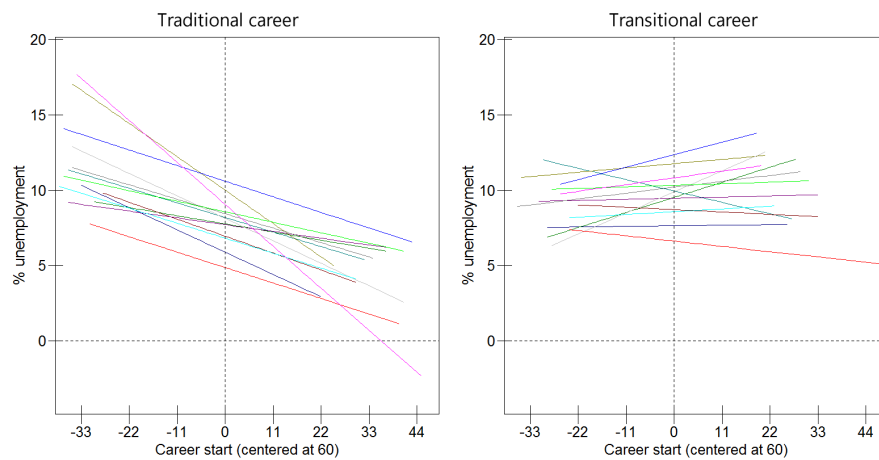
Fig. 4 Country-level evolution of employment security over time

Employment security in non-traditional careers   17

increase in variance is much more pronounced, hinting at a polarization in regard to employment security within that type.
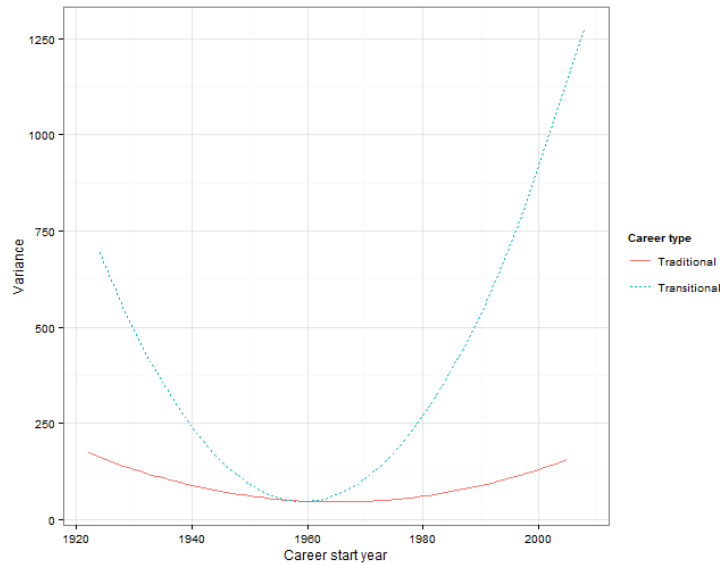
Fig. 5 Individual-level variance of employment security over time

5 Discussion and conclusion

In this article, we have primarily focused on the evolution of security in careers over time. The issue of labor security has been widely discussed in literature on labor, where two main and opposing views can be discerned (Tregaskis et al., 1998). Both views trace global changes in economy, technology and society, changes that may be summarily labeled as economic globalization or the 'new' economy, with its intensification of international trade, competition between individuals, firms and even countries, free flow of capital, labor market flexibilization and many technological advancements. The two views diverge, however, on the account of the implications that these changes have had on the labor security of individual workers (Reich, 2008). Our analysis contributes to the understanding of the

career dynamic that has accompanied the shift towards the 'new' economy in Europe, and to the evolution of employment security on the career level, from an explicitly long-term perspective.

We have started by looking at the basic assumptions of the mainstream career theories that pertain to the career transformations in the context of the 'new' economy. Our results go against the thesis of a major career transformation, which has been brought forward by these theories. Instead, we can observe minor relative changes in the prevalence of the traditional and the 'new' careers in some European countries, while the traditional type still remains dominant. An apparent question is, can we extrapolate this conclusion to the careers that are unfolding today? Is it perhaps the nature of our sample, largely comprised of completed careers, that masks the potential intensification of the shift towards the 'new' career in the third millennium? In the light of these questions, our results are best considered along with other recent research on the subject, based on shorter-term data. Kattenbach et al. (2014) arrives at conclusions that are similar to ours, finding that the 'new' career concepts, developed primarily in the Anglo-Saxon context, likely do not describe what is happening on German labor markets. Our results nuance this statement, in that there effectively exist career patterns that correspond to the 'new' career in terms of mobility, and, according to earlier research (Kovalenko and Mortelmans, 2014), are not necessarily characterized by poor career outcomes typical for precarious careers. This would imply that changes described by the 'new' career theories may be found in certain niches of the labor market. Our results are in line with the second statement of Kattenbach et al. (2014), namely that the career transformation can be perceived only to a certain extent. Elchardus and Smits (2014) echoes these findings, the focus being on young adults (18-36). The authors find that stable upwards trajectory is still the most popular career form, whereas the attraction for the flexible career wanes rapidly with age, a process that goes hand in hand with choosing to combine work and family. While 38% of respondents aged between 18 and 20 opted for the ambitious flexible career, only 20% had the same preference in the age group between 31 and 36. These results indicate, that the flexible career type is relatively unpopular not only in terms of objective prevalence, but also in terms of subjective preferences of young workers. Soens et al. (2005) also find that while there are minor shifts towards the transitional career pattern, the traditional career still holds its place firmly in Belgium. Heery and Salmon (2000) describe a similar dynamic for the UK, stating that aggregate job tenure declined only modestly in the UK, although for some groups more pronouncedly (cf. supra). Rodrigues and Guest (2010) report similar findings for several countries.

Our results demonstrate that there is a lot of variation between countries in terms of career type composition. For example, Denmark showed only a minor shift in relative prevalence of the transitional career, yet the country was already characterized by a higher share of the transitional career type (see Table 1). At the same time, other countries, such as Belgium, had a lower share of careers of the transitional type, without statistically significant changes over time. An important conclusion that becomes apparent from these results is that the processes of career flexibilization are

very diversified, can occur in different directions, as the example of Greece attests, and are highly dependent on the national context. The study of contemporary careers should take this into account, and explore the national variations, not abandoning the main thesis of the career transformation, but nuancing it. In other words, the career transformation may be seen as a spectrum that is embedded in a complex mosaic of the local socio-economic processes, rather than as a single form.

The main thrust of our analysis, pertaining to the decrease of employment security in careers in the aftermath of the economic globalization, has yielded several results. First, we've established that also in this regard the exact dynamic is contingent on the national context. Second, we've found no insecurity increase in the traditional careers, coupled with some increase thereof in some, but not all countries, for the transitional career type. This may suggest that while there has been some decrease in employment security, the process has been limited to some labor market strata, without becoming universal. Our results suggest that this process may have occurred in some transitional careers, as they indicated a possible polarization process within this type. This would imply that while some transitional careers remained stable or even improved in terms of employment security, in line with the positive interpretation of flexibilization, and possibly conform to the ideal type of the 'new' career, other have become more precarious, in line with the neo-Fordist view.

Several authors have, in fact, formulated the idea of labor market polarization as a consequence of flexibilization, (Clarke, 2008; Kim, 2013; Standing, 2011; Van Buren, 2003; Zeitz et al., 2009). It has been hypothesized that highly skilled and employable workers may benefit from operating in flexible labor markets, as it allows avoiding organizational bureaucracy, and enables them to utilize external job mobility to boost their careers. At the same time, those in weaker labor market strata, such as lower educated workers, or those without easily marketable skills, are the losers of the flexibilization, as it pushes them into poorly paid jobs without long-term security, easily discardable. Empirical evidence is acutely lacking in this respect, and while our results make a contribution, the polarization hypothesis remains a fruitful area for further research.

6 Limitations and directions for future research

As any empirical research, our is not without its limitations. First, the analysis of the reasons behind the between-country variations in both career type prevalence and changes in employment security over time was left outside the scope of this article. While we stressed observing the heterogeneity in the evolution of career dynamic over time, it would be interesting to examine the specifics of its interaction with the national context, related, for example, to the differences in labor market regulation, historical tendencies and so forth. Second, in constructing our career typology we focused on physical career mobility alone. At the same time, it can be argued, that changes in careers have occurred not only in terms of physical, but also in terms of psychological mobility (Sullivan and

Arthur, 2006). Bringing the subjective dimension into account may bring additional details into the spotlight, e.g. pertaining to the perception of insecurity, along with the stress and health effects associated with it.

7 References

Anyadike-Danes, M. and McVicar, D. (2010). My brilliant career: Characterizing the early labor market trajectories of British women from generation X. *Sociological Methods & Research*, 38(3):482 –512.

Arnold, J. and Cohen, L. (2008). The psychology of careers in industrial and organizational settings: A critical but appreciative analysis. In Hodgkinson, G. and Ford, J., editors, *International Review of Industrial and Organizational Psychology 2008*, pages 1–44. John Wiley & Sons, Ltd.

Arthur, M. B., Inkson, K., and Pringle, J. K. (1999). *The new careers: individual action and economic change*. SAGE.

Arthur, M. B. and Rousseau, D. M. (1996). Introduction: The boundaryless career as a new employment principle. In Arthur, M. B. and Rousseau, D. M., editors, *The boundaryless career: A new employment principle for a new organizational era*, pages 3–20. Oxford University Press.

Azmat, G., Güell, M., and Manning, A. (2006). Gender gaps in unemployment rates in OECD countries. *Journal of Labor Economics*, 24(1):1–38.

Barbieri, P. (2009). Flexible employment and inequality in europe. *European Sociological Review*, 25(6):621–628.

Baruch, Y. and Bozionelos, N. (2010). Career issues. In Zedeck, S., editor, *APA Handbook of Industrial and Organizational Psychology*, volume 2, pages 67–113. American Psychological Association, Washington, DC.

Beck, U. (2000). *The brave new world of work*. Wiley-Blackwell.

Cappelli, P. (1999). Career jobs are dead. *California Management Review*, 42(1):146–167.

Chudzikowski, K. (2012). Career transitions and career success in the 'new' career era. *Journal of Vocational Behavior*, 81(2):298–306.

Clarke, M. (2013). The organizational career: not dead but in need of redefinition. *International Journal of Human Resource Management*, 24(4):684–703.

Clarke, M. and Patrickson, M. (2008). The new covenant of employability. *Employee Relations*, 30(2):121–141.

Clarke, P. (2008). When can group level clustering be ignored? multilevel models versus single-level models with sparse data. *Journal of Epidemiology and Community Health*, 62(8):752–758. PMID: 18621963.

DiPrete, T. A., Goux, D., Maurin, E., and Quesnel-Vallee, A. (2006). Work and pay in flexible and regulated labor markets: A generalized perspective on institutional evolution and inequality trends in europe and the U.S. *Research in Social Stratification and Mobility*, 24(3):311–332.

Elchardus, M. and Smits, W. (2014). De levensloop van jongvolwassenen 18-36 jaar: Jongvolwassenen over hun loopbaan, arbeidsethiek, levenslang leren en tijdskrediet. Technical report, Vrije Universiteit Brussel.

European Commission (2007). *Towards Common Principles of Flexicurity: More and better jobs through flexibility and security SEC(2007) 861 SEC(2007) 862.*

Guest, D. E., Oakley, P., Clinton, M., and Budjanovcanin, A. (2006). Free or precarious? a comparison of the attitudes of workers in flexible and traditional employment contracts. *Human Resource Management Review*, 16(2):107–124.

Gunz, H., Mayrhofer, W., and Tolbert, P. (2011). Career as a social and political phenomenon in the globalized economy. *Organization Studies*, 32(12):1613–1620.

Hall, D. T. (1996). Protean careers of the 21st century. *The Academy of Management Executive*, 10(4):8–16.

Hall, D. T. (2004). The protean career: A quarter-century journey. *Journal of Vocational Behavior*, 65(1):1–13.

Harvey, D. (2011). *A brief history of neoliberalism.* Oxford, Oxford University Press.

Heery, E. and Salmon, J. (2000). The insecurity thesis. In Heery, E. and Salmon, J., editors, *The Insecure Workforce*, Routledge studies in employment relations, pages 1–24. Routledge.

Howell, D. R. (2004). *Fighting Unemployment: The Limits of Free Market Orthodoxy.* Oxford University Press.

Inkson, K. and Arthur, M. B. (2001). How to be a successful career capitalist. *Organizational Dynamics*, 30(1):48–61.

Inkson, K., Gunz, H., Ganesh, S., and Roper, J. (2012). Boundaryless careers: Bringing back boundaries. *Organization Studies*, pages 1–18.

Jacoby, S. M. (1999). Are career jobs headed for extinction? *California Management Review*, 42(1):123–145.

Kalleberg, A. L. (2009). Precarious work, insecure workers: Employment relations in transition. *American Sociological Review*, 74(1):1–22.

Kattenbach, R., Schneidhofer, T. M., Lücke, J., Latzke, M., Loacker, B., Schramm, F., and Mayrhofer, W. (2014). A quarter of a century of job transitions in Germany. *Journal of Vocational Behavior*, 84(1):49–58.

Kim, Y.-M. (2013). Diverging top and converging bottom: labour flexibilization and changes in career mobility in the USA. *Work, Employment and Society*, 27(5):860–879.

Klau, F. and Mittelstadt, A. (1986). Labour market flexibility. *OECD Economic Studies*, 6(745):87–126.

Kovalenko, M. and Mortelmans, D. (2014). Does career type matter? Outcomes in traditional and transitional career patterns. *Journal of Vocational Behavior*, 85(2):238–249.

Lesnard, L. (2010). Setting cost in optimal matching to uncover contemporaneous socio-temporal patterns. *Sociological Methods & Research*, 38(3):389 –419.

Lynch, S. M. (2007). *Introduction to applied Bayesian statistics and estimation for social scientists.* Springer, New York.

Muffels, R. and Luijkx, R. (2008). Labour market mobility and employment security of male employees in europe: 'trade-off' or 'flexicurity'? *Work, Employment & Society*, 22(2):221–242.

Núñez, I. and Livanos, I. (2010). Higher education and unemployment in europe: An analysis of the academic subject and national effects. *Higher Education*, 59(4):475–487.

Pringle, J. and Mallon, M. (2003). Challenges for the boundaryless career odyssey. *International Journal of Human Resource Management*, 14(5):839–853.

Reich, R. B. (2008). *Supercapitalism: The Transformation of Business, Democracy, and Everyday Life.* Vintage Books.

Rodrigues, R. A. and Guest, D. (2010). Have careers become boundaryless? *Human Relations*, page 0018726709354344.

Roper, J., Ganesh, S., and Inkson, K. (2010). Neoliberalism and knowledge interests in boundaryless careers discourse. *Work, Employment & Society*, 24(4):661–679.

Soens, N., De Vos, A., Buyens, D., Heylen, L., Kuppens, A., Van Puyvelde, I., and Mortelmans, D. (2005). *Belgische loopbanen in kaart: traditioneel of transitioneel? Eindrapport van het onderzoek naar het begrip Loopbaan.* Academia Press.

Standing, G. (1999). *Global Labour Flexibility: Seeking Distributive Justice.* Palgrave Macmillan, New York.

Standing, G. (2009). *Work after globalization: building occupational citizenship.* Edward Elgar, Cheltenham, UK; Northampton, MA.

Standing, G. (2011). *The precariat: the new dangerous class.* Bloomsbury Academic.

Stegmueller, D. (2013). How many countries for multilevel modeling? a comparison of frequentist and Bayesian approaches. *American Journal of Political Science*, 57(3):748–761.

Sullivan, S. E. (1999). The changing nature of careers: A review and research agenda. *Journal of Management*, 25(3):457–484.

Sullivan, S. E. and Arthur, M. B. (2006). The evolution of the boundaryless career concept: Examining physical and psychological mobility. *Journal of Vocational Behavior*, 69(1):19–29.

Employment security in non-traditional careers   23

Tams, S. and Arthur, M. B. (2010). New directions for boundaryless careers: Agency and interdependence in a changing world. *Journal of Organizational Behavior*, 31(5):629–646.

Tregaskis, O., Brewster, C., Mayne, L., and Hegewisch, A. (1998). Flexible working in Europe: The evidence and the implications. *European Journal of Work and Organizational Psychology*, 7(1):61–78.

Van Buren, H. J. (2003). Boundaryless careers and employability obligations. *Business Ethics Quarterly*, 13(2):131–149.

Wilthagen, T. and Tros, F. (2004). The concept of 'flexicurity': a new approach to regulating employment and labour markets. *Transfer: European Review of Labour and Research*, 10(2):166–186.

Zeitz, G., Blau, G., and Fertig, J. (2009). Boundaryless careers and institutional resources. *The International Journal of Human Resource Management*, 20(2):372–398.

Transitions, trajectories and the role of activation policies for young people

Margherita Bussi

ETUI Brussels, University of Brighton and University of Geneva (PhD candidate)

Abstract   The aim of this contribution is to understand to what extent activation policies are able to reduce the risk of transition into the labour market of young people unemployed or in social assistance. The relevance of the institutional structure in which activation policies are implemented and the importance of the nature of activation policies on young people's transitions are assessed by using optimal matching and event history analysis. The theoretical background underpinning the analysis is an original combination of the Capability Approach (CA)(Sen and Nussbaum 1993) and Transitional Labour Markets Approach (TLM)(Schmid and Gazier 2002; Schmid 2008). These theoretical frameworks, together, are believed to provide new perspectives on the assessment of activation policies. Their combination allows the researcher to assess whether the nature of activation policies shows different capacity of compensating for the lack of 'transitional skills' of the most disadvantaged young people (Bijwaard and Veenman 2008). The analysis is performed on a customized longitudinal administrative dataset and it focuses on the case study of the Brussels Capital Region (RBC).

1   Individuals and Institutions: the contribution of the TLM and CA

The individualisation of (new) social risks, the idea that the labour markets can no longer secure employment (Gazier 1999), the subsequent spread of activation, the increasing responsibilisation of individuals on welfare benefits, and the modified relationship between the state and its citizens were crystallised in a concise - though broad - concept: the concept of employability.
Gaining employability, being able to integrate the labour market and avoiding being excluded from the the labour market implies managing several risks which are embedded in changing status over the life-course (e.g. from student to worker, from employed to unemployed etc.).
Hence, employability can be understood as an empowering process where institutional and individual responsibilities are intermingled. On the one hand, I defined *active empowerment* as the active role played by the person through the control and management of the resources needed to reach valuable outputs. On the other hand, I define *passive empowerment* as the empowering action provided by institutions through the services and support to any person at risk of experiencing

a loss of employability. An example is the quality of training provided: (young) individuals who need to enhance their skills and enrol in training courses organised in the framework of active labour market policies. The quality of education and training provided is not something that is actively achieved or modified by the recipients, but is provided independently by the institution.

A fair balance of active and passive empowerment is essential in order not to blame the unemployed for their situation and avoid adopting a paternalistic attitude, which can easily happen namely with young people.

Investigating the capacity of activation policies – intended as the main tool for delivering employability to the inactive and unemployed - of providing *passive empowerment* - i.e. their capacity of reducing the risk of transitions for young people and the institution's role in shaping young people's trajectories – means using a theoretical background that investigates *both the institutional opportunity structure and its interactions with individual characteristics*.

Indeed, the risks of transitions cannot be only attributed to the lack of 'transitional skills' of (young) individuals but also to the demand side which is indeed responsible for the lack (abundance) of labour market opportunities. While not denying the importance of the demand-side in the outcome of labour market transitions of young people, this contribution only focuses on the efficacy of opportunity structure created to answer to the lack of transitional skills of among the young labour market entrants.

This is achieved by combining two theoretical frameworks deriving from different areas of study - labour economics and development economics – that are believed to offer a more comprehensive view of the kind of responses that take simultaneous account of the insecurity inherent in transitions to the labour market: the Transitional Labour Markets (TLM) and the Capability Approach (CA). Two sets of different statistical techniques are also deployed: sequence analysis and event history analysis.

1.1   The Transitional Labour Markets Approach

Going beyond the focus on the individual capacity of coping with these risks linked with life-course transitions (supply-side perspective), the Transitional Labour Markets approach encourages policy-makers to address labour market transitions through institutional forms of regulation. These forms of regulation – called 'transitional labour markets' – are provided by institutions designed to encourage risk-taking, while simultaneously providing workers with the prospect of relative economic security and employment satisfaction without confining them in dead-end jobs (Lassnigg et al 2007).

More in details, it is believed that TLM is particularly interesting for this analysis because it clearly indicates how institutional arrangements should work

in order to function as a bridge for individuals transitions towards labour market and social integration, thus answering to the need of rethinking institutions and their role in mitigating social risks. It does not overlook the relevance of empowering of individuals and supporting individual capacity for choice and also a capacity to reverse choices (Anxo and Erhel 2006), thus accounting for the individualisation of life trajectories. In addition, it adopts a dynamic perspective both of individuals' participation in the labour market and coordination of institutions by focusing on transitions rather than static labour market positions.

1.2   *The Capability Approach*

The Capability Approach, developed mainly by Amartya Sen and Martha Nussbaum (Nussbaum and Sen 1993; Sen 1999) has provided an umbrella theoretical framework for empirical qualitative and quantitative operationalization of well-being and agency concepts on development economics, social and psychological research (Bussi and Dahmen 2012).

In contrast with the TLM, the CA better accounts for the heterogeneity of individual needs and their relationship with the institutional, social and environmental context in which individuals act. This aspect of the CA has been applied to policy evaluation in diverse social policy fields, such as social work and employment policy, where policies are assessed in terms of their capacity of making people able to increase their opportunities and reach valuable outcomes.

The CA complements the institutional perspective of reducing risks of TLM by looking at individuals' potential to act and be, thus accounting for the individualisation focus of welfare policies. The CA provides analytical tools that allow the researcher to investigate the interactions between individuals and institutional arrangements. Thus it helps the researcher grasp the nature of the institutional answer and the nature of the contract between the citizen (in this case the young unemployed or social assistance beneficiary) and the State.

In light of the idea of passive empowerment, the research question can be expressed as follows *to what extent is the institutional set up at the local level able to compensate for lack of individual active empowerment of young disadvantaged people?*

1.3 The institutional mechanisms of smoothing transitions

From the combination of the two theoretical backgrounds, it derives that favourable transitional arrangements should include two important aspects through which institutions deliver passive empowerment and help prevent social exclusionary transitions: 1) the degree of empowerment of individuals and 2) a flexible coordination between levels of decision-making as to facilitate the adjustment of the institutional answer to individual needs and local circumstances.

Activation policies are meant to contribute to create favourable transitional arrangements insofar as they play an important role in empowering unemployed

or inactive young people. They do this by promoting training of basic and job-specific skills and contributing to reduce the risk of erosion and deterioration of human capital (Schmid and Gazier 2002). They support the acquisition and strengthening of employability through job-search support and work experience in order to reduce the negative signals to employers while helping (young) persons to define their own conception of a valued job (Bonvin and Farvarque 2006).

On the other hand, at an aggregate level, welfare and labour market systems that are equipped with institutions dealing with the variety of risks rising during transitions are more successful in reducing the risks over the (several) transitions in the life-course. Similarly, those systems providing a basic level of income insurance independent of actual work histories will perform well in reducing risks associated with transitions (Schmid 2008).

These two criteria will be investigated in the selected case study and target group: the young people in unemployed and in social assistance in Brussels Capital Region.

2   Young people's transitions and the case study

In this contribution, young people in their early careers are the group of interest. Young people are considered more at risk in the labour market compared to adults and are more likely to have unstable trajectories; thus at higher risks of exclusionary transitions. For this reason, the role of *passive empowerment* is most important.

Several quantitative studies on longitudinal data of school-to-work transitions have adopted an international comparative dimension (Brzinsky-Fay 2006; Wolbers 2007; Quintini and Manfredi 2009). Adopting a cross-national perspective allows the researcher to investigate at the macro level to what extent an institutional framework might be more empowering than others at the moment of crucial transitions in the labour market (Brzinsky-Fay 2006; Erhel et al 2014).

However, the heterogeneity of the institutional set up across countries overlooks the impact of single labour market institutions on different social groups as well as the diversity of the environment in which policies are implemented.

Focusing on a single case study and looking at the local level allows one to better grasp the complementarity of institutions and their impact on a variety of social groups known for their diverse active empowerment.

2.1 The choice of the Brussels Capital Region

The case study is the Brussels Capital Region (RBC). Various reasons encouraged this choice.
From a national perspective, the Belgian welfare system is not often included in international comparisons, because it is often associated with the continental Europe and close to its 'elder brother' (France).

Moreover, despite recent reforms, the incomplete transformation of the welfare system into an active welfare state makes Belgium and the Brussels Capital Region difficult to categorise it into the well-known "Work-First" of "Human Capital Development" (Clegg 2007). Its hybrid position makes it more interesting to define and investigate in terms what kind of passive empowerment institutions are able to provide.

Further, the complex governance structure makes the complementarity of welfare and labour institutions more challenging. The system also faces a high the heterogeneity of the youth population, which is particularly marked in the Brussels Capital Region, where a considerable share of young people is unemployed, inactive or living in poverty.

In addition, the focus on the Brussels Capital Region has been chosen because the decentralisation and local delivery of services for unemployment -

at the regional level- and for social assistance -at the municipal level- is likely to increase challenge the implementation of services across local entities.

These issues show the importance of social assistance (CPAS) and social insurance institutions (PES) and their activation policies in shaping and potentially reducing the risk of young people's first transitions into the labour market.

3   The hypotheses

The investigation of mechanisms (processes) and the effects (outcomes) of empowering individuals and of coordinated labour market and welfare institutions should be understood as complementary. The table below brings together the analyses while specifying the unit of analysis, the dependent and the independent variables used to answer the hypotheses as well as the methods.

The first hypothesis deals with the overall institutional structure and tries to understand how labour market and welfare systems intervene on individual trajectories.

The second hypothesis is connected and complements the first one insofar as it explores level the impact of single employability measures on individual transitions at a more disaggregated level. The scope of analysis is reduced to a single event occurrence and the interaction of individual and institutional variables is also accounted for.

Hypothesis	Dependent variable	Institutional-related variables	Individual-related variables	Methods
Looking at how the institutional coordination shapes individual trajectories *H1: Varied and complementary labour and welfare market institutions favour individual integrative trajectories into the labour market.*	Degree of risk of trajectories (cluster membership)	Time spent in certain labour market positions Probability of transition from unemployment or social assistance into employment or education	Age, gender, neighbourhood, education position in the household, origin, previous work experience	Optimal matching analysis, cluster analysis, transition matrices
Looking at the impact of single employability measures on individual trajectories *H2: ALMPs adopting an "enabling employability" approach will function more as a bridge leading to upwards transition into employment and compensate for lack of personal disadvantage.*	Probability of individual transition out of unemployment and social assistance	Employability programmes For CPAS (social assistance): supported employment (article 60) and education/training For PES (social insurance): CPP (action plan) and RAE (job-search programmes)	Age, gender, neighbourhood, education, position in the household, origin, previous work experience	Competing risks discrete-time event history analysis

Table 1 Investigating the coordination of the institutions and the empowering of individuals: two complementary levels of analysis

4   The choice of methods

If the life-course perspective of the TLM is taken seriously (Anxo and Erhel 2007), then research adopting this framework should consider the whole sequence of all transitions from one state to another during the life of an individual. This overall approach to individual trajectories is needed insofar as one should expect the system of institutions to work as an opportunity structure in a coherent and sequential manner for each type of different risk that may arise.

To this end, sequence analysis allows one to test whether the existence of "varied and complementary" welfare and labour market institutions enables trajectories with a low degree of risk. Sequence analysis is also relevant to

understand to what extent one trajectory differs from another (Abbott 1995; Studer and Ritschard 2015); therefore, this statistical technique can reveal whether the institutional complementarity works for different degree of individual risks.

In a second step, cluster analysis is used to reduced the complexity of the results of sequence analysis by clustering together the most similar trajectories. I also investigated trajectories by looking at the interactions between labour market institution systems (through transitions matrices) while also identified who is more likely to have risky trajectories (using regression analysis).

These steps identified the most vulnerable population (i.e. young people with a migrant background) that is investigated in the second hypothesis and from a micro dimension.

Beyond trajectories, there is a need to understand what determines the time spent in a specific state prior to the transition to another state (Abbott 1995). Hence, from an understanding of sequences as a whole unit, one needs to go down to the analysis of one single transition. The risk is then investigated in a step-by-step method and event history analysis is suitable for this aim, i.e. when there is a need of a fairly deep understanding of the relationship between the time spent in a (labour market) position and the subsequent state toward which the person goes (Haplin 2010).

What is interesting about this analysis is its implicit interest in risk: using event history models allows the researcher to answer the questions: "is the event under investigation going to take place? If so, when?" (Singer and Willett 2003). And "under these circumstances and rules, who is going to survive?" (Box-Steffensmeier and Jones 2004).

The 'rules and the conditions' I am interested in are the dichotomous, yet interactive, concepts of *passive* and *active empowerment* that help understand to what extent individuals' characteristics, institutional and environmental factors interact and what results they produce on transitions.

I also argue that analysing both trajectories *and* transitions[1] allows me to test whether an empowering approach to transitions favours this complementarity by making individuals *able* – and not only responsible – *to make a safe transition*. The cumulative and subsequent reduction of risk in (single) transitions should provide the reduction of risk on a longer temporal level – i.e. in trajectories.

5   The sample

A representative random sample of 3000 individuals drawn from a larger sample was used over an observation period of 6 years, from March 2005 to

---

[1] Based on the definitions coming from life-course analysis, I use "transition" to identify one single movement from one labour market position to another. The word "*trajectory*" is used to grasp several movements across labour market and embrace a longer time perspective. Mortimer, J.T. and M.J. Shanahan. 2006. *Handbook of the Life Course*: Springer.

December 2010 (quarterly observations). In March 2005 young people were aged between 15 and 30 years old and were all resident in the Brussels Capital Region. Seven positions in the labour market were identified as relevant: employed (including self-employed), in education, unemployed (both with and without work history but with unemployment benefits), social assistance, "*stage d'attente*" (i.e. unemployed without work history and waiting for unemployment allowances), employed with top-ups and other. Being in social assistance, unemployed, in "stage d'attente" or employed with top-ups are considered as bridging labour market positions in light of the theoretical background insofar as they are expected to function as stepping stones which empower young recipients and make them able to move upwards.

6   A summary of the results

6.1. *The quality of trajectories and the institutional complementarity*

After performing Optimal Matching Analysis (performed with the user-written ado SQ for STATA with "symmetric substitution costs matrix based on the mean of the transitions' probabilities (p) between every two neighbouring elements in the sequences", Brzinsky-Fay et al 2006), cluster analysis was performed (Ward's linkage). Four clusters (*Smooth* – with a prevalence of employment*; Trap* – with a prevalence of social assistance*; Failure* – with a prevalence of unemployment*; Other)* were identified[2].

The quality of each cluster is measured in terms of the 'degree of risk'. The degree if risk is the combination of high/low volatility and integrative capacity index (with some changes compared to Brzinsky-Fay 2007). Volatility is calculated as the standardised average of different positions of the trajectories of a cluster. Similarly, the integrative capacity is calculated out of transitions matrices and defined as the standardized ratio of between the probability going from a negative to a positive position and the probability of staying in a negative position (i.e. relative risk ratio). The degree of risks varies across clusters, with the *Smooth* cluster having a high but positive mobility (higher chances of spending time in positive labour market position) and the *Failure* and *Trap* clusters having, respectively, a negative mobility and negative immobility.

---

[2] The cluster *Other* is not reported in the analysis due to the lack of information on the dominant labour market position.

Table 2: Degree of risk by cluster.

Type of risk	Integrative capacity		
		High	**Low**
Volatility	**high**	positive mobility	negative mobility
		Smooth	*Failure*
	low	positive stability	Trapped
			Trap

Source: own elaboration

Transition probabilities matrices - often used in research adopting a TLM approach (Leschke 2006; Schmid 2008) - were used, at the cluster level, to investigate to what extent labour and welfare institutions work as bridges. Results show that being unemployed, in social assistance or "en stage d'attente" do not seem to keep their promises of working as bridges for those who are likely to be more vulnerable, i.e. young people in the cluster with a higher degree of risk.

Once established the four type of clusters, it is important to define which individual characteristics are more likely to predict trajectories with higher or lower degree of risk. This is crucial for the second part of the analysis where the capacity of compensating – through passive empowerment - for individual lack of individual active empowerment is assessed.

The results of the multinomial logistic regression of the cluster membership show that young people with lower "transitional skills" - defined as educational attainments, weak position in the household, short work experience, poor neighbourhood and a foreign nationality- are less able to have their positions into the labour market secured and less likely to enjoy favourable trajectories.

6.2. The results for event history analysis

Looking at single transitions means investigating each single activation measure and establish to what extent it is able to compensate for the lack of transitional skills, i.e. accelerating the process out of unemployment or social assistance.

Informed by the previous results, young people with a migrant background were chosen as vulnerable group on who the effect of compensating effect of activation policies could be tested. Having a migrant background was found closely associated with a wide range of disadvantages, particularly in the RBC, which cannot be grasped by single socio-demographic variables.

According to the hypotheses and the theoretical account, I expect that the activation policies that are known to be richer in *passive empowerment* will be

more likely to compensate for the lack of transitional skills. This means that they will predict a higher likelihood of transition out of inactivity or unemployment towards the labour market.

The variables available describing the type of employability measures were classified in ordinal variables according to their degree of intensity in passive empowerment. The example of the CPP (individual action plan) is provided below.

None	Low share of passive empowerment	
Independent job-search		
Action Plan + job guidance		
Training		
Action Plan + training	High share of passive empowerment	
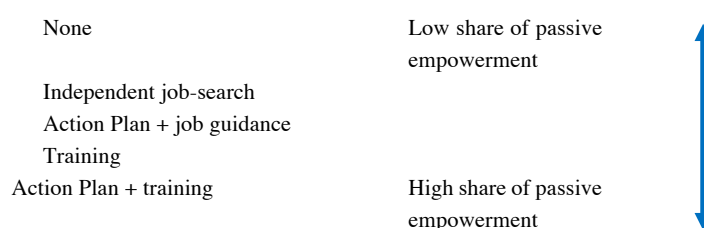
Fig. 1: Intensity of passive empowerment of CPP (individual action plan) measures.

These covariates are interacted with the variable for nationality in order to explore to what extent the compensation takes place when comparing young with a foreign background with young natives who received the same measure.

Discrete-time event history analysis (competing risks) was conducted separately for young people in social assistance (CPAS) and young people unemployed (PES).

For young social assistance recipients, I tested the compensating effect of attending full-time education while receiving social assistance allowances and being employed in supported employment (aka "article 60"). For young unemployed people, I tested the compensation capacity of two sets of of activation measures available namely linked with the individual action plan (CPP) and job-search activities (RAE).

Results for young people in social assistance show that interaction between the nationality and being in supported employment does not reveal any "compensation effect" for the young disadvantaged under the same type of contract as their native peers. In fact, difference between the probabilities of exiting towards employment or staying in social assistance are not different for natives and non-EU/EU employed in supported employment.

Similarly, the expected compensating effect of attending full-time education for young disadvantaged compared to natives does not seem to work. In fact, young people with a migrant background seem penalized from attending some education/training as they are twice as likely to exit towards unemployment compared to young Belgians.

As for young people in unemployment benefits, findings from the two sets of activation measures show the the compensatory effect of activation measures is limited.

For the job-search measures, the interaction between the variable of origin and the activation programme did not improve the model. Only conclusions related to the main effects of different types of programmes could be drawn. Findings show that young people who benefited from the empowering and bridging effect of an intense coaching (intense and individual) are much more likely to exit towards education than employment.

At the same time, when comparing exits towards employment and social assistance, young people who received an intense coaching are also more likely to end up in social assistance.

These results suggest that – although being conceived as leading to employment – more intense measures might bring to the surface other specific individual needs, such as lack of education and training or lack of adequate income which geared young people towards other services than the labour market.

As for the second set of measures linked with the action plan, results of the interactions between each of the measures and the variable for the origin show that, young people with a migrant background have higher likelihood of remaining in unemployment, and this independently from the type of measure undertaken.

7   Conclusions

From this contribution, two aspects are esteemed important.

First, the combined theoretical background is a first step towards a more holistic perspective on activation measures that include both an individual and institutional perspective. Although this work was limited to the identification of the passive empowerment, the definitions of passive and active empowerment open to a more interactive perspective of the impact of activation policies on young people's trajectories. The definition of passive empowerment has also allowed to classify the measures in terms of their capacity of providing empowerment in relation to individual needs and not in terms of their costs or length.

Moreover, the combination of sequence analysis and event history analysis has brought together two levels of analysis of the institutional action - yat the aggregate with the coordination of institutions and at individual level with the empowerment of individual - which are not often used together but that broaden the research agenda for life-course research (Aisenbrey and Fasang 2010).

References

Abbott, A. (1995). Sequence analysis: new methods for old ideas. *Annual review of sociology*, 93-113.

Aisenbrey, S., & Fasang, A. E. (2010). New life for old ideas: The" second wave" of sequence analysis bringing the" course" back into the life course. *Sociological Methods & Research*, *38*(3), 420-462.

Anxo, D., & Erhel, C. (2006). Irreversibility of time, reversibility of choices? The Life-Course foundations of the Transitional Labour Markets approach. in *Cahiers de la MES: Maison des Sciences Economiques*.

Clegg, D. (2007). Continental drift: on unemployment policy change in Bismarckian welfare states. *Social Policy & Administration*, *41*(6), 597-617. doi: 10.1111/j.1467-9515.2007.00574.x

Bijwaard, G., & Veenman, J. (2008). Unequal chances on a flexible labor market, The case of the Netherlands. *Tinbergen Institute Discussion Paper*.

Box-Steffensmeier, J. M., & Jones, B. S. (2004). *Event history modeling: A guide for social scientists*. Cambridge University Press.

Bonvin, J.-M., & Farvaque, N. (2007). A capability approach to individualised and tailor-made activation. In R. v. Berkel and B. Valkenburg, (Ed.), *Making it Personal: Individualising Activation Services in the EU*. Bristol: Policy Press.

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

Brzinsky-Fay, C. (2006). Lost in Transition: Labour Market Entry Sequences of School Leavers in Europe. *WZB Discussion Paper*.

Brzinsky-Fay, C. (2007). Lost in transition? Labour market entry sequences of school leavers in Europe. *European Sociological Review*, *23*(4), 409-422.

Bussi, M. & Dahmen. S. (2012). When ideas circulate. A walk across disciplines and different uses of the 'capability approach'. *Transfer: European Review of Labour and Research* *18*:91-95.

Erhel, C., Guergoat-Larivière, M., Trancart, D. (2014). Labour Market Mobility Patterns during the 2008 Crisis: Inequalities in a Comparative Perspective. *Document de travail du CEE*.

Gazier, B. (1999). Employability: Concepts and Policies. *Berlin Institute for Applied Socioeconomics*.

Lassnigg, L., Burzlaff, H., & Rodrigues, M. A. D. (2007). *Lifelong Learning: Building Bridges Through Transitional Labour Markets*. Het Spuinhuis.

Leschke, J. (2006). Are unemployment insurance systems in Europe capable of covering the special risks that persons with non-standard employment contracts are facing?, Otto-Suhr-Institut für Politikwissenschaften Freien Universität Berlin, Berlin.

Nussbaum, M. & Sen. A. (1993). *The quality of life*. Oxford: Clarendon Press.

Sen, A. (1999). *Development As Freedom*: Oxford University Press.

Quintini, G. & Manfredi, T. (2009). Going Separate Ways? School-to-Work Transitions in the United States and Europe. *OECD Social, Employment and Migration Working Papers 90*

Schmid, G. & Gazier, B. (2002). *The Dynamics of Full Employment: Social Integration Through Transitional Labour Markets*: Edward Elgar.

Schmid, G. (2008). *Full Employment In Europe: Managing Labour Market Transitions and Risks*: Edward Elgar.

Singer, J. D. & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. Oxford University Press.

Studer, M., & Ritschard, G. (2015). What matters in differences between life trajectories: a comparative review of sequence dissimilarity measures. *Journal of the Royal Statistical Society: Series A (Statistics in Society*.

Discovering and Explaining Patterns of Work-Family Reconciliation in Luxembourg.

Analysis of Administrative Records.

Nevena Zhelyazkova · Gilbert Ritschard

1 Introduction

This abstract presents an analysis of administrative records from Luxembourg whereby the work-family trajectories of parents are re-constructed and analyzed using tools from state sequence analysis (see Abbott 1990; Abbott and Tsay 2000; Aisenbrey and Fasang 2010; Gauthier et al. 2014) and a clustering algorithm is applied to search for typical patterns. In the final step available co-variates are linked to cluster membership via a multinomial logit model.

From a theoretical point of view, this analysis draws on the life-course perspective (Elder and Giele 2009). In short, the life-course perspective recognizes that life courses are: longitudinal; emerging as a result of continuous interaction between individuals and the social structure; influenced by the historical and geographical context; affected by the timing of events; interconnected with the life courses of others. As outlined in McDaniel and Bernard (2011) and Bernard (2007) the life course perspective can guide academic research and the dialogue between policy makers and academics and provide new insights for policy making. Recognizing that human lives are more than the consecutive events that take place in each life trajectory, the life course perspective can guide the design of comprehensive provisions to build and preserve social, health and human capital through all life stages.

Administrative records for the analysis have been provided by the *Inspection générale de la sécurité sociale (IGSS)* in an anonymous form. The data set covers over 10 000 persons and the substantial time span of eight years. The administrative records describe actual behaviour of individuals, which is of particular

N.Zhelyazkova
Paris Dauphine University & UNU-MERIT
E-mail: nevena.zhelyazkova@dauphine.fr

G. Ritschard
NCCR LIVES, University of Geneva
E-mail: Gilbert.Ritschard@unige.ch

relevance for the analysis of leave policies as survey respondents have been shown to not always accurately report their parental leave status (Chan et al. 2012).

Luxembourg is a small country located in the heart of Western Europe, characterized by a very high GDP per capita, high employment rate and a high share of international labour migrants. Although findings from Luxembourg could not directly generalize to larger European countries, the size and the composition of the workforce make it comparable to many European regions, especially around large cities (Brosius and Ray 2012). The high proportion of foreign in-commuters particularly resembles that of regions with a dynamic economy located close to international borders, such as, for example, Geneva, Ticino and Basel in Switzerland (OECD 2011, see).

At the descriptive level, the collective analysis of the career trajectories of working parents shows that women tend to reduce their labour market participation after having a child. Taking parental leave full-time could result in both a transition back to employment or into withdrawing from the labour market, while part-time parental leave seems to be part of an overall strategy to reduce working hours after having a child. Interestingly the trajectories of mothers who took leave and who did not take it, did not seem to differ systematically before the event of the birth of the child. However, the trajectories of fathers who take parental leave and who did not were different even before the child is born. This suggests that use of parental leave of mothers can be seen more as reaction to the event of birth, while for fathers it may be related to a longer-term lifestyle preference.

To move the analysis to the analytical level, first a distance matrix has been computed. Next, a clustering algorithm has been applied to search for typical patterns. Finally, co-variates have been linked to cluster membership using a multinomial logit model.

The distance matrix has been computed using the Longest Common Subsequence (LCS). Elzinga (2007) (in Gabadinho et al. 2011, p.28) shows that using LCS is equivalent to using OM with constant indel costs (1) and substitution costs (2). LCS was chosen because it is more sensitive to duration than to sequencing. From a social security point of view it is most important to identify long-term trends, such as long-term transition to part-time hours of work or long spells outside the labour market. It is important to note that using the LCS method is based on a constant substitution and indel cost. In other words, the cost of substitutions between any two states is the same. Similarly, inserting a state has the same cost regardless of what states it is inserted next to. This implies an assumption that all states are equally different from each other.

For this study it was considered appropriate to use a hierarchical clustering procedure and not a partitional one because there was not enough prior knowledge or guidelines, which could have been used to pre-suppose the number of clusters. Another reason to choose a hierarchical procedure was that it has become somewhat of a standard in the sequence analysis literature. More specifically, agglomerative hierarchical clustering with Ward's method has been very frequently applied in previous literature, although it is acknowledged that this choice is mostly a result of convention Martin et al. (2008, p.186).

In order to choose the appropriate number of clusters, the cluster diagnostic tools proposed by Studer (2013) have been applied. He suggests using a number of measures of cluster quality available in the literature, which together can provide

some indications as to where to stop the clustering. These measures could also be used to compare the results obtained through different clustering algorithms[1].

The cluster analysis was initially performed under the assumption that the trajectories of mothers and fathers are entirely different from each other. Accordingly, the clustering was performed separately for both samples. A qualitative comparison of the obtained clusters, however, revealed that there are actually more similarities than differences in the trajectory types. The distribution of number of parents in each type is different (women, for example, were overrepresented in clusters dominated by part-time work). The types of clusters however, were not as different as expected. Therefore the analysis was performed again on the pooled sample of both types of trajectories. This has made it possible to compare directly the numbers of men and women in the same cluster.

The results of clustering algorithm applied in this analysis identified nine distinct groups of parents, summarized in Table 1. The gender difference became apparent in the varying male to female ratios in each cluster. The majority of the male trajectories were classified in the career types described by continuous full-time or overtime employment both before and after the birth. In comparison only about one third of the women's trajectories were classified in such clusters. A large fraction (about one third) of the mothers were classified in the trajectory types where the event of the birth marks a clear turn in the trajectory compared to the pre-birth period marked either by a reduction in the working hours or a withdrawal from the labour market. In addition, a large proportion (over 20 per cent) of the female trajectories were classified in clusters characterized by continuous part time employment.

The results from the analysis of the sample of mothers lend partial evidence to an economic justification of women's career decisions. Lower opportunity cost (in terms of foregone salary earnings) does seem to be associated with higher probability to leave the workforce. More children and the presence of a spouse in the household also make it more likely that a woman will leave her pre-birth employment than maintain a continuous full-time career track. However, these factors do not seem to be associated with a higher probability for women to reduce employment hours from full time to part time after having a child.

The results of the same analysis on the male sample provide further insight into the gendered nature of the relationship between employment and family. For both men and women higher levels of earnings and positive salary growth in the previous months are associated with higher odds of being in the cluster characterized by continuous full-time employment. Similarly to married women, married men are more likely to be in employment types characterized by less-than-full-time employment relative to having continuous full-time trajectories. However, in contrast to the results for women, the results for men suggest that presence of other children in the household does not seem to be associated with higher odds of being classified into one of the career types with reduced labour market participation.

---

[1] The diagnostic tools are available in the *WeightedCluster* (Studer 2013) package, which runs under the free statistical and programming environment R (R Core Team 2015).

4   Nevena Zhelyazkova, Gilbert Ritschard

	Women		Men	
	Count	Per Cent	Count	Per Cent
(1) Continuous part-time careers (20 - 39 hrs/week)				
	838	19 %	185	3 %
(2) Continuous self-employment				
	142	3 %	302	5 %
(3) Continuous full-time careers				
	675	15 %	1924	33 %
(4) Continuous overtime hours intermittent with part-time hours				
	477	11 %	2408	41 %
(5) Transition to labour market inactivity after birth in 2003				
	523	12 %	22	< 1 %
(6) Reduction of working hours after birth in 2003				
	682	15 %	39	1 %
(7) Continuous part-time careers (<=20 hrs/week)				
	307	7 %	63	1 %
(8) Continuous full-time careers with some irregular hours				
	439	10 %	662	11 %
(9) (Possibly) leaving Luxembourg after birth in 2003				
	398	9 %	222	4 %
Total				
	4481	100 %	5827	100 %

Table 1: Cluster sizes for men and women.
Notes: Counts and percentage correspond to the results obtained through clustering of the pooled data set.

References

Abbott, Andrew. 1990. A Primer on Sequence Methods. *Organization Science* 1 (4): 375–392.

Abbott, Andrew, and Angela Tsay. 2000. Sequence Analysis and Optimal Matching Methods in Sociology. Review and Prospect. *Sociological Methods & Research* 29 (1): 3–33.

Aisenbrey, Silke, and A. E. Fasang. 2010. New Life for Old Ideas: The "Second Wave" of Sequence Analysis Bringing the "Course" Back into the Life Course. *Sociological Methods & Research* 38 (3): 420–462. doi:10.1177/0049124109357532.

Bernard, Paul. 2007. The Interconnected Dynamics of Population Change and Life-course Processes. *Horizons* 9 (4): 13–16.

Brosius, Jacques, and Jean-Claude Ray. 2012. Wage Differentials Between Natives and Cross-border Workers Within and Across Establishments (Paper Presented

5

at 26th Annual Conference of The European Association for Population Economics). Retrieved from `https://espe.conference-services.net/resources/321/2907/pdf/ESPE2012_0353_paper.pdf`, Accessed April 24, 2013.

Chan, Chaowen, Dana Hamplová, and Céline Le Bourdais. 2012. Are Parental Leaves Considered as Work Interruptions by Survey Respondents? A Methodological Note. *Canadian Studies in Population* 1 (2): 31–44.

Elder, Glen H. Jr., and J Z Giele. 2009. *The Craft of Life Course Research*. New York: Guilford Publications.

Elzinga, Cees H. 2007. Sequence Analysis: Metric Representations of Categorical Time Series (Manuscript, Dept of Social Science Research Methods, Vrije Universiteit Amsterdam) Amsterdam, The Netherlands: Department of Social Science Research Methods, Vrije Universiteit Amsterdam.

Gabadinho, Alexis, Gilbert Ritschard, Nicolas Müller, and Matthias Studer. 2011. Analyzing and Visualizing State Sequences in R with TraMineR. *Journal Of Statistical Software* 40 (4): 1–37.

Gauthier, J A, F Bühlmann, and Philippe Blanchard. 2014. *Introduction: Sequence Analysis in 2014. Life course research and social policies*. New York: Springer. Chap. 1.

Martin, Peter, Ingrid Schoon, and Andy Ross. 2008. Beyond Transitions: Applying Optimal Matching Analysis to Life Course Research. *International Journal of Social Research Methodology* 11 (3): 179–199. doi:10.1080/13645570701622025.

McDaniel, Susan, and Paul Bernard. 2011. Life Course as a Policy Lens: Challenges and Opportunities. *Canadian Public Policy* 37 (S1): 1–13.

OECD. 2011. *OECD Territorial Reviews: Switzerland 2011*. Paris: OECD Publishing.

R Core Team. 2015. *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing, `http://www.r-project.org/`.

Studer, Matthias. 2013. Weighted Cluster Library Manual: A Practical Guide to Creating Typologies of Trajectories in the Social Sciences with R (LIVES Working Paper 2013/24), Lausanne, Switzerland. Retrieved from `http://lives-nccr.ch/sites/default/files/pdf/publication/24_lives_wp_studer_weightedcluster.pdf`.

Session 4B: Markov I

Using dynamic microsimulation to understand professional trajectories of the active Swiss population

Pauline Adamopoulou[*,1], Gilbert Ritschard[*,1], and André Berchtold[*,2]

[*] IP214 Measuring Vulnerability, NCCR-LIVES
[1] Institute of Demography and Socioeconomics, University of Geneva
[2] Faculty of Social and Political Sciences, University of Lausanne

January 2016

1 Introduction

Within the social and economic sciences and of particular interest to demographers are life course events. Looking at life sequences we can better understand which states, or life events, precede or are precursors to vulnerability (Marshall, 2011). A tool that has been used for policy evaluation and recently has been gaining ground in life course sequence simulation is dynamic microsimulation. Within this context dynamic microsimulation consists in generating entire life courses from the observation of portions of the trajectories of individuals of different ages.

A more general, and perhaps technical, definition would be that microsimulation uses micro-data collected on micro-units and, based on a set of rules (models) which are estimated, simulates the outcomes of these micro-units within a given system (Orcutt, 1957). To better illustrate this concept, lets take an example. An example of micro-data could be longitudinal data collected on individuals (the "micro-units") within Switzerland (the "system"). Assume there are longitudinal labour market data on these individuals and that the goal is, within the Swiss context, to project the labour market outcome for these individuals over time. The first step is to determine if an individual employed at the current time period would be more likely to remain employed in the following time period. In other words, what is the probability that this individual remains employed? What is the probability that this

individual from being employed transitions to unemployment? Conversely, what is the probability that an unemployed individual remains unemployed? Or that from unemployment she or he transitions into employment? These questions concern transition probabilities, and such transition probabilities would need to be estimated from the data using probability models. So the "rules" in this case would be transition probabilities, which would be estimated using probability models. The estimated transition probabilities could then be applied to simulate the future labour market outcomes of these individuals.

In this work, we aim to use dynamic microsimulation in order to analyse individual professional trajectories with a focus on vulnerability. The primary goal of this analysis is to deepen upon current literature by providing insight from a longitudinal perspective on the signs of work instability and the process of precarity.

An interesting feature of microsimulation is that it extends the range of possible uses of existing data. This is particularly important because within data collection funding is often an issue. As a consequence sample size as well as the time frame of data-collection are often limited. Additionally, the issue with data-collection is that oftentimes the data are used for a specific purpose by a research team and then are of limited use to other researchers. These reasons motivate the secondary goal of this work which is to show how, by using microsimulation, data collected for one purpose can be analysed under a different scope and used in a meaningful way.

The data to be used in this analysis are longitudinal and were collected by NCCR-LIVES IP207 under the supervision of Prof. Christian Maggiori and Dr. Grégoire Bollmann. Individuals aged 25 to 55 residing in the German-speaking and French-speaking regions of Switzerland were followed annually for four years. At the initial time period there were 2469 participants and the sample was roughly representative of the active Swiss population with women and the unemployed slightly over represented (Maggiori et al., 2014). These individuals were questioned regarding, *inter alia*, their personal, professional and overall well-being. At the end of the fourth and final wave, there were 1131 individuals who had participated in all four waves. The sample remained representative of the Swiss population with women and the unemployed slightly over represented.

Using the information collected from these surveys, we use simulation to construct various longitudinal data modules where each data module represents a specific life domain. The demographic module for instance would consist of: age, marriage, children, etc., whereas the education module would consist of: primary education, secondary education, tertiary education. We postulate the relationship between these modules and layout a framework of estimation. Within certain data modules a set of equations are created to model the process therein. For every dynamic (time-variant) data module, such as the labour-market module,

2

the transition probabilities between states (ex. labour market status) are estimated using a Markov model and then the possible outcomes are simulated.

The benefit of using dynamic microsimulation is that longitudinal sample observations instead of stylised profiles are used to model population dynamics. This is one of the main reasons large-scale dynamic microsimulation models are employed by many developed nations (such as the USA, Canada, Australia, the UK, France, the Netherlands, Sweden, Japan, etc.); thorough reviews of such models can be found in O'Donoghue (2001), Zaidi and Rake (2001), Cassells et al. (2006), Li and O'Donoghue (2013). There has been limited use, however, of such approaches with Swiss data. This work contributes to the analysis of professional trajectories of the active Swiss population by utilising dynamic microsimulation methods.

Acknowledgements

References

Cassells, R., Harding, A., and Kelly, S. (2006). Problems and Prospects for Dynamic Microsimulation: A review and lessons for APPSIM. NATSEM Working Paper Series 63, University of Canberra, National Centre for Social and Economic Modelling.

Li, J. and O'Donoghue, C. (2013). A survey of dynamic microsimulation models: uses, model structure and methodology. *International Journal of Microsimulation*, 6(2):3–55.

Maggiori, C., Rossier, J., Krings, F., Johnston, C., and Massoudi, K. (2014). Career pathways and professional transitions: preliminary results from the 1st wave of a 7-year longitudinal study. Book chapter to be published.

Marshall, V. W. (2011). Risk, vulnerability, and the life course. Inaugural Presentation. LIVES Working Paper 2011/1.

O'Donoghue, C. (2001). Dynamic Microsimulation: A Methodological Survey. *Brazilian Electronic Journal of Economics*, 4(2).

Orcutt, G. H. (1957). A new type of socio-economic system. *Review of Economics and Statistics*, 39(2):116–123.

Zaidi, A. and Rake, K. (2001). Dynamic microsimulation models: a review and some lessons for SAGE. Technical report, The London School of Economics, Simulating Social Policy in an Ageing Society (SAGE), Discussion Paper.

Mechanisms of the transition to adulthood: an application of Hidden Markov Models

Y. (Sapphire) Han, A. C. (Aart) Liefbroer and C. H. (Cees) Elzinga

Abstract An increasing number of studies focuses on understanding the processes underlying the transition to adulthood. However, the transition to adulthood is a complex process of a series of events that are often interlinked. Even though life courses are greatly varying sequences of roughly the same life course events, the complexity is caused by the fact that these sequences consist of correlated events and spells and these correlations depend on gender, social class, cohort and cohort-related macro events. Our previous work demonstrated that the application of stochastic models like the Latent-Class model helps to describe the variation in life courses and its correlation with gender and social class. But the Latent-Class model cannot account for correlated events within life courses nor can it account for switches between latent types during the life course. We argue that (Hidden) Markov models, as a simple generalization of the Latent-Class model, has the ability to account for correlations between events and spells and also allows for switches between latent types or model life courses. Therefore, this study will use (Hidden) Markov models to produce a typology of trajectories of the transition to adulthood. Furthermore, we will test hypotheses on social class- and gender differences in observed life courses and latent types or model-life courses, using data from the Gender and Generation Programme (GGP), which provides full monthly life course sequence data between age 15 to 40.

Y. (Sapphire) Han
Netherlands Interdisciplinary Demographic Institute, The Hague, University of Groningen, e-mail: han@nidi.nl

A. C. (Aart) Liefbroer
Netherlands Interdisciplinary Demographic Institute, The Hague, University Medical Centre Groningen, Vrije Universiteit Amsterdam, the Netherlands, e-mail: liefbroer@nidi.nl

C.H. (Cees) Elzinga
Netherlands Interdisciplinary Demographic Institute, The Hague, Vrije Universiteit Amsterdam, the Netherlands e-mail: c.h.elzinga@vu.nl

1 Introduction

The ultimate goal of life course research is to understand how life courses come about and what variables affect their shape. Essentially, this is a holistic question: to answer it requires the postulation of a mechanism of the generation of the complete life course. Holistic life course models must satisfy a few simple properties. First, these models must have a memory, a sense of the past, as it is generally assumed that events in the early stages of the life course may affect stages or outcomes later on [Mayer, 2009]. A second requirement is that the process that generates the life course, is affected by variables that that are supposed to influence the life course: gender, religion, parental education, etc. Finally, the model should be formulated in terms of a process that is not directly observable: since the life course is generated through mental, not directly observable, processes that are conscious or unconscious, and decisions that are voluntary or involuntary. Of course, such models should be testable and amenable to causal analysis.

Over the past decades, life course research has been dominated by two different paradigms: Event History analysis (EH) [Blossfeld et al., 2007] and Sequence Analysis (SA) [Cornwell, 2015]. EH-models are not holistic: they try to explain the waiting times for certain life course events to occur and the SA-approach leads to finding most frequent patterns in the wide variety of observed life courses, but it does not account for this variety. Recently, we have seen other methods and models being applied as well, for example Latent Class analysis [Barban and Billari, 2012] and Structural Equation Models [Pakpahan et al., 2015] but neither of these methods satisfies all of the three requirements as formulated above.

However, there is a broad class of models that does satisfy the above requirements: the class of so called Hidden Markov Models. These models have a memory in the sense as intended, they allow for time-constant and time-varying covariates and are formulated on the basis of a latent, hidden, random process over a finite set of states, a Markov chain. The models are testable in the sense that their parameters can be estimated [Bartolucci et al., 2012, Rabiner, 1989] and easily allow for causal analysis once formulated as a log-linear regression model [Paas et al., 2007]. Hidden Markov Models belong to a larger family of latent structure models that has been amply described by [Langeheine and Van de Pol, 1990, Vermunt, 1997].

This paper aims to model the life course, confined to the relatively turbulent transition to adulthood, through using Hidden Markov Models. The transition to adulthood is usually described by a collection of events [Elder Jr, 1985] from which two correlated processes can be distilled [Buchmann and Kriesi, 2011]: the school-to-work transition and the process of family formation. Here, we provide an example that solely focusses on the transition into adulthood in the family domain, as there is a large body of literature on the processes involved. More specifically, we example the family-life trajectories of French men and women born between 1956 and 1965, using data from the French Generations and Gender Survey (GGS).

The paper is structured as follows: in this lengthy introduction we discuss the main concepts of Hidden Markov Models and make some general remarks on their application to life course research, Section 2 discusses our data and methods used,

Mechanisms of the transition to adulthood: an application of Hidden Markov Models   3

Section 3 discusses our results and Section 4 summarizes, concludes and suggests further research.

1.1 Hidden Markov Models

Hidden Markov Models generalize the much simpler idea of a Markov chain. A Markov-model or Markov-chain is a random process over a set of states such that the probability of being in a particular state at the next observation only depends on the state-history of the process. If the relevant state history just consists of the present state, such a chain is called "first-order". Figure 1 shows a graphical representation of a first-order 2-state Markov-chain and its matrix of transition probabilities.

Fig. 1 A graph showing a first-order, 2-state Markov chain and its transition probability matrix **A**. The states are labeled as "0" and "1" and the arrows represent the transition probabilities.

Let us denote the k distinct states of a Markov chain as $Q = \{q_1, \ldots, q_k\}$ and let S_t denote the state that the system is in at time t, i.e. S_t could have any of the "values" or labels from the set Q. Then we say that a random process over Q is a first-order Markov-chain, precisely when

$$Prob(S_t = q_j | S_0 \ldots S_{t-1}) = Prob(S_t = q_j | S_{t-1} = q_i) = a_{ij}, \tag{1}$$

and we denote this probability by a_{ij}. If we now define the initial state-probabilities as $Prob(S_0 = q_i) = \pi_i$, the Markov-chain λ is fully defined by the k-vector $\pi = (\pi_1, \ldots, \pi_k)$ of initial state probabilities and the $k \times k$-matrix of transition probabilities $\mathbf{A} = \{a_{ij}\}$: $\lambda = (\pi, \mathbf{A})$.

In an ordinary Markov chain, transition probabilities depend on the present state only (see Equation (1)). However, it is easy to extend this model to account for the effect of covariates. Let $\mathbf{v}$ and $\mathbf{w}_t$ denote vectors of time-constant and time-varying covariates. Then a direct extension of the Markov-chain model is formulated as

$$Prob(S_0 = q_i | \mathbf{v}, \mathbf{w}_t) = \pi_i(\mathbf{v}, \mathbf{w}_t), \tag{2}$$

$$Prob(S_t = q_j | S_0 \ldots S_{t-1}, \mathbf{v}, \mathbf{w}_t) = Prob(S_t = q_j | S_{t-1}, \mathbf{v}, \mathbf{w}_t) = q_i) = a_{ij|\mathbf{v}, \mathbf{w}_t}. \tag{3}$$

Clearly, this formulation implies a separate Markov-chain for each point in the $(\mathbf{v}, \mathbf{w}_t)$-space.

A Markov chain could be used to model a set of observed life course sequences by simply identifying each of the observed states as a model state and estimating the transition probabilities from the relative transition frequencies of the observed sequences. The result of that would be a more or less accurate summary of the

observed transition frequencies. However, it would not lead to a credible model for the way these sequences were generated. Therefore we now turn our attention to an extension of the Markov chain: the Hidden Markov Model.

In a Hidden Markov Model (HMM), the Markov chain is defined over a set of latent, unobservable states. So, the stochastic process as such is not observable. Furthermore, it is supposed that, at each state, the process 'emits' an observable (an observable can be univariate or multivariate) according to a state-specific probability distribution over the full set of observables, in the present context the observable states of a life course. Thus, in a k-state HMM with a set of observables $Y = \{y_1, \ldots, y_n\}$, there must be a set B of k state-specific probability distributions $\mathbf{b}_j = (b_{j1}, \ldots, b_{jn})$, each satisfying $\sum_i b_{ji} = 1$:

$$b_{ji} = Prob(o_t = y_i | s_t = q_j). \tag{4}$$

This allows us to represent the set B as a $(k \times n)$-matrix

$$\mathbf{B} = \begin{pmatrix} b_{11} & \ldots & b_{n1} \\ \vdots & \ddots & \vdots \\ b_{1k} & \ldots & b_{nk} \end{pmatrix} = \begin{pmatrix} \mathbf{b}_1 \\ \vdots \\ \mathbf{b}_k \end{pmatrix} \tag{5}$$

whereof each row is a distinct probability distribution over the observables and the complete HMM $\lambda = (\pi, \mathbf{A}, \mathbf{B})$ is specified by the initial state distribution π, the $(k \times k)$-matrix $\mathbf{A}$ of transition probabilities and the $(n \times k)$-matrix $\mathbf{B}$ of emission probabilities.

In Fig. 2, we show a graph of the HMM-generated events in a time-window $(t-1, t+1)$: at $t-1$, the system arrives in state S_{t-1} and emits observable o_{t-1} (governed by $\mathbf{B}$) and then switches to state S_t (governed by $\mathbf{A}$) and again emits an observable, etc.. The reader should be aware that the system may, depending on the

Fig. 2 A graph showing the time-window $(t-1, t+1)$ of a Hidden Markov process. At each time t, the system is at some latent state S_t and emits an observable. Note that the hidden state S_t is not necessarily different from S_{t+1}. The observable is a random sample from the set of observables, according to a probability distribution that is specific for each state q_i, $i = 1, \ldots, k$.
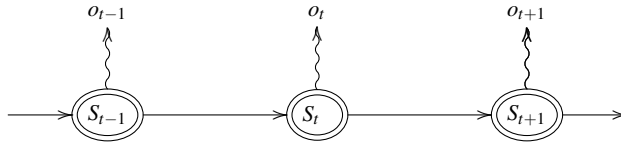
probability a_{jj}, actually stay in the same state j for quite a while and during that time emit various different observables. Similarly, the observables may remain the same for quite a while, at the same time but "below the surface", the system actually switches state several times. In practice, if we observe that people stay in the same

observable state for many years, it is to be expected that the diagonal elements of **A** are relatively big, i.e. close to 1.

1.2 Modelling with HMM's: Some practical considerations

Let $O_i = o_{i1} \ldots o_{iT}$ denote an observed sequence from a set $O = \{O_1, \ldots, O_N\}$ of such sequences and let $Prob(O_i|\lambda)$ denote the likelihood of that sequence, given the model. Furthermore, let $Q_i = q_{i1} \ldots q_{iT}$ denote the path along the latent states that maximizes $Prob(Q_i|O_i, \lambda)$, i.e. the latent sequence that "best accounts" for the observations, given the model.

Being able to calculate the likelihood of the observations given the model is a precondition for EM-estimation of the parameters of the model and calculating Q_i, the most probable latent sequence, is a precondition for a substantive interpretation of the model. Both problems, evaluating $Prob(O|\lambda)$ [Baum et al., 1970], and calculating Q_i [Viterbi, 1967] were already solved in the sixties of the previous century and have been amply described in many sources [Rabiner, 1989, Zucchini and MacDonald, 2009, Bartolucci et al., 2012]. Here, we will not deal with the intricacies of these methods. Instead, we will discuss some practical issues that are related to these methods and their output.

First, one should be aware that evaluating a HMM involves the estimation of quite some parameters: with k postulated latent states, we have to estimate $k-1$ parameters $\hat{\pi}_i$; $k-1$ since we must have that $\sum_i^k \hat{\pi}_i = 1$. Likewise, we have to estimate $k(k-1)$ parameters to obtain $\hat{\mathbf{A}}$ and $k(n-1)$ parameters to get $\hat{\mathbf{B}}$. So, the surface of the likelihood function $Prob(O|\lambda)$ is quite irregular and therefore, attempts to find its maximum, be it through EM [Dempster et al., 1977] or through any other method like simulated annealing [Andrieu and Doucet, 2000], will most often converge to a local instead of the global maximum. Extending the HMM to incorporate covariates will only aggravate this problem. Therefore, the estimation of a HMM should be repeated quite some times to find a configuration $(\hat{\pi}, \hat{\mathbf{A}}, \hat{\mathbf{B}})$ that (probably) comes close to the maximum sought for. For example, peeking around the corner of our modelling life courses with HMM's, we display the density of the BIC-values as obtained over 1000 repetitions of estimating a 4-state model. Clearly, these BIC-values are quite different, as are the underlying configurations $(\hat{\pi}, \hat{\mathbf{A}}, \hat{\mathbf{B}})$. Obtaining this curve took almost three hours of computation time and quite some memory. Increasing the number of states and the number of trials soon requires unfeasible computation times and memory for this exercise.

However, we do not consider this to be a serious problem for applying HMM's to model life courses. Normally, the size of the observation alphabet will be small and the number of postulated states k will be rather small too. The latter number should reflect the number of stages or states in which the subjects will take demographic decisions and these decisions are small in number; they pertain to leaving the parental home, partnering, reproducing and, eventually, returning to the parental home or breaking up a partnership. Therefore, in practice, k should be small. Con-

6   Y. (Sapphire) Han, A. C. (Aart) Liefbroer and C. H. (Cees) Elzinga

Fig. 3 BIC-density plot as obtained from repeating the estimation of a 4-state HMM 1000 times with random initial values.
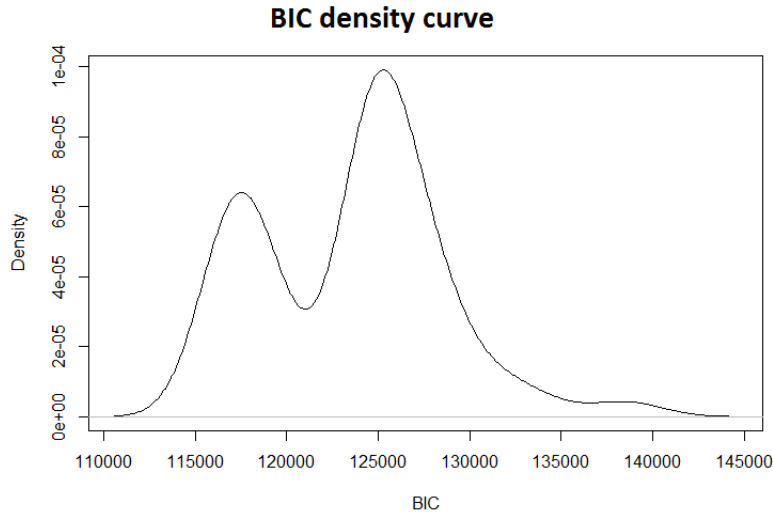
sidering the big data sets that social demographers use today, it is to be expected that the optimal value of the information criteria that we use to evaluate the fit of a HMM - variants of minimum-χ^2, AIC or BIC - cannot be good indicators of the optimal number of latent states as these indicators will drive us to accept large numbers of latent states while substantive interpretation is problematic. Rather, the size of k is to be fixed on a theoretical basis: can we assign a credible interpretation to these states in view of the latent trajectories and the way these trajectories and the emission distributions, i.e. the probability of picking particular behavioral, demographic alternatives, change as a result of covariates that pertain to societal pressure on rather than social capital of the subjects studied.

Then how do we start interpreting the latent states? A first clue to this interpretation is provided by studying the latent state sequences over time and evaluating the probabilities $Prob(S_t = q_j | t, \lambda)$, $t = 1, \ldots, T$, $j = 1, \ldots, k$, i.e. the relative frequencies of state occupancy, aggregated over the sample studied. Again peeking around the corner of our analysis yet to be presented, we show a plot of these relative frequencies for a 4-state HMM in Figure 4: This plot shows something that is not evident from the estimated transition probabilities: most subjects start in the latent state labeled as LS1 so it should be associated with a decision about leaving the parental home and this should be reflected in the emission probability distribution over the observables: it should be characterised by a relatively high probability of emitting the event "leaving the parental home". So, the marginal state occupancies

Mechanisms of the transition to adulthood: an application of Hidden Markov Models 7

Fig. 4 Plot of the marginal probabilities over time of latent state-occupancy in a 4-state HMM.
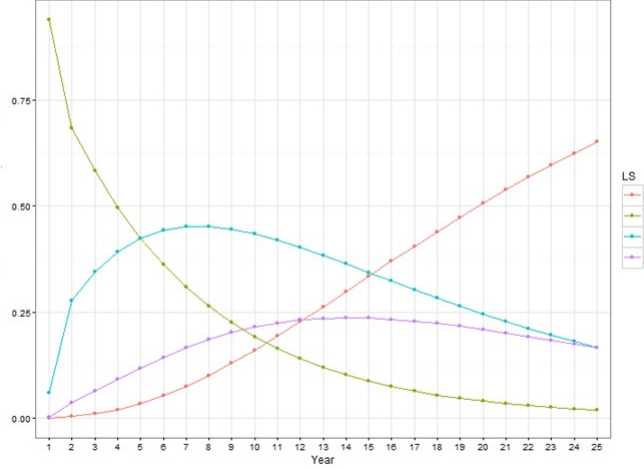
over time and the emission distributions will help us to interpret the latent states. However, these considerations do not suffice for a credible interpretation.

A credible interpretation can only arise in the light of the way covariates affect the parameters of the model: do the estimated effects of covariates corroborate, or at least are not at variance with, the knowledge that we already have about the effects of these covariates on the occurrence and timing of life course events. For example, we may expect that low-educated will enter parenthood earlier than high-educated and this should be reflected in the differences between transition probabilities and/or the emission probabilities of lower- and higher-educated. Therefore, it is not enough to only evaluate a HMM as such: we need to enrich the model with relevant covariates in order to decide on the credibility of the interpreted model.

How to incorporate covariates into a HMM? Thereto, we consider the likelihood $Prob(O_i|\lambda,\mathbf{v})$ of a particular sequence O_i from a set $O = \{O_1,\ldots,O_N\}$ of such sequences, wherein $\mathbf{v}$ denotes a vector of covariates. Since, according to the model, the observed sequences result from the latent sequences Q_i, we can decompose this likelihood as follows:

$$Prob(O_i|\lambda,\mathbf{v}) = Prob(O_i,Q_i|\lambda,\mathbf{v}) \tag{6}$$

$$= Prob(O_i|Q_i,\lambda,\mathbf{v})Prob(Q_i|\lambda,\mathbf{v}), \tag{7}$$

and thus, the likelihood of our data given the model can be decomposed as

$$Prob(O|\lambda,\mathbf{v}) = \underbrace{\prod_i Prob(O_i|Q_i,\lambda,\mathbf{v})}_{\text{"operational"}} \underbrace{\prod_i Prob(Q_i|\lambda,\mathbf{v})}_{\text{"structural"}}, \tag{8}$$

the multiplications arising from the assumption that the sequences have arisen independently from each other. From the latter Equation 8, we observe that the total likelihood of our data given the model can be decomposed into two separate parts. The second part is called the "structural" part of the model since it pertains to the unobservable structure (the transition probability matrix) of the stochastic process over the latent states, the components of the postulated model. The first part of the multiplicative structure (additive if we consider log-likelihood) we called "operational" for two reasons. The first reason is that we may consider the emission probabilities as choice-options given the latent state the person is in. For example, if the subject is in a state where partnering is the key issue, it may chose between different options to operationalize the positive choice for partnering: marriage, cohabitation or a relational agreement without living together. The second reason for calling this part "operational" is the fact that, given our belief in the validity of the number and structure of the latent states, this part of the model is affected by our way of operationalizing the life course in observational labels: choosing a different alphabet will affect the model fit without altering the structural validity[1]. Unfortunately, we cannot separate these parts of the model when assessing the model's fit as the observations are our only entry to the latent, structural part of the model. Therefore, it is not wise to have the same covariates play a role both in the structural and in the operational part of the model as it would severely hinder the substantive interpretation of the model. Here, we prefer to assume that covariates do not have a role in the operational part of the model, i.e. we assume that

$$Prob(O_i|\lambda, \mathbf{v}) = Prob(O_i|Q_i, \lambda)Prob(Q_i|\lambda, \mathbf{v}), \tag{9}$$

or, equivalently, that only the initial and state transition probabilities are affected by covariates. The reason for this preference is that we know that most life courses in developed countries only differ in the timing and duration of the various stages on the rout to adulthood. This implies that, in most countries, the behavioral alternatives and the order in which they are expressed are roughly the same for most people.

1.3 Applications of HMM

HMM's have been successfully applied in several fields [Bartolucci et al., 2012], including but not limited to psychological and educational measurement [Vermunt et al., 1999], medicine and health [Cook et al., 2000], criminology [Bijleveld and Mooijaart, 2003], marketing and related fields [Paas et al., 2007], interactions during survey-interviews [Elzinga et al., 2007] and labor market research [Richardson et al., 2011]. However, we have not yet seen the application of HMM's to life course research. One reason

---

[1] [Bartolucci et al., 2012] call this second part of the model the 'measurement model'. This is adequate when the observables contain a measurement error as is often the case in responses to psychological test or survey items; in life course encoding such errors are rare (but see [Manzoni et al., 2010]).

Mechanisms of the transition to adulthood: an application of Hidden Markov Models   9

for that could be that only recently, big demographical data sets, cheap software and fast PC's have become widely available.

2 Data and Method

2.1 Data

The Gender and Generation Programme (GGP) is a Longitudinal Survey of 18-79 year olds in 19 countries that that examines the relationships between generations and genders, by collecting nationally representative data in all participating countries. [Fokkema et al., 2016] provide extensive information on design and representativeness of the GGS. In this study, we select respondents (males and females, in total 1900) in the France GGP data of a cohort between birth year 1956 and 1965. In the selected dataset, full annual fertility, partnership and leaving parental home information between age 15 and 40 are available and background information such as gender, education level, parental education, parental divorce.

To demonstrate the application of HMM, we construct a multi-channel sequence dataset of respondents' fertility history (4 categories: no child, 1 child, 2 children, 3 and more children), partnership history (3 categories: single, cohabitation, marriage), and leaving parental home (2 categories: yes or no). To investigate the link between background variables, we include gender (2 categories: female and male) and education level (2 categories: high and low). To visualize the multi-channel sequence dataset, four sequence index plots separated by gender and education level are shown in Figure 5 (a) and (b). Take Figure 5 (a) left panel for example, it contains the fertility (4 categories), partnership (3 categories) and leaving home annual information throughout young adulthood (age 15 - 40) of high educated male respondents. The complexity of the dataset is obvious: it contains 24 categories ($4 \times 3 \times 2$) of 25 repeated annual measures.

Insert Figure 5

2.2 Method

During the analysis, the selected multi-channel sequence dataset was fitted to hidden Markov model (HMM) with hidden states equals 3-6, each of a time-homogeneous model with 1000 random starting values. This paper selectively shows the result of HMM 4 state solution as proof of concept of the application of HMM in the transition to adulthood research. Two types of covariates, i.e., gender (female vs. male) and education (low vs. high) were introduced in the latent model of HMM 4 state solution (also a time-homogeneous model with 1000 random starting values). The

reason of using 1000 random starting values for HMMs is to reduce the influence of local maximum. The estimation procedure of HMM relies on Baum-Welch or EM (Expectation-modification) algorithm [Rabiner, 1989]. With the increment of hidden states, the number of parameters to be fitted also increase drastically. During the analysis, it is found out that for the given dataset, HMM 6 state solution is unstable. It took more than 1 GB in the RAM, and 1000 random starting value repentances were not enough to generate stable solution. It might be possible to perform HMM with high number of hidden states on HPC (high performance computing) environment. HMMs with hidden states from 3 to 5 generate stable solution. Choosing the HMM with 4 hidden states is due to the fact that it largely reduce the data complexity at the same time providing substantively interesting interpretation.

All analyses were performed in R environment for statistical computing and graphics in a 64 bit PC with 32 GB RAM. R packages LMest [Bartolucci et al., 2015] and markovchain [Spedicato et al.,] were utilized for Hidden Markov models. Sequence visualization and related techniques were performed by R package TraMineR [Gabadinho et al., 2011].

3 Result

In this section, results of fitting HMM 4 state solution (without and with covariates in the latent model) to the multi-channel France GGP sequence data are presented. In each model, the time cost, the model fit parameter (BIC), the output parameters (initial probability distribution π, transition probability distribution a_{ij} and emission probability distribution b_j), the visualization and interpretation of these output parameters and the mechanisms revealed by HMM are presented.

3.1 Hidden Markov model 4 hidden state solution

The HMM with 4 hidden states were performed with 1000 random stating values to reduce the influence of local maximum. It took 2.6 hours and achieved a minimum BIC of 115961. The fitted HMM with the lowest BIC was chosen as the HMM 4 solution. The interpretation of the HMM is based on its output parameters, i.e., initial probability distribution, transition probability distribution and emission probability distribution. As described in Introduction Section, initial probability distribution reveals the proportion of hidden states that respondents occupy in the beginning of their life courses; transition probability distribution reveals the transition rate (per year in this study) to other hidden states once respondents arrive at a certain hidden state; emission probability distribution links the hidden states to the observed life course.

The output parameters are shown in Table 1 (the hidden states are ordered as 'A', 'B', 'C' and 'D'). It is difficult to interpret these number without graphic illustra-

Mechanisms of the transition to adulthood: an application of Hidden Markov Models   11

tion. Latent probability distribution graph (shown in Figure 6) is useful as a first step to interpret Table 1. This graph is based on the initial probability distribution and transition probability distribution shown in Table 1. There are four curves representing the dynamics of the four hidden states during respondents' 25 year young adult life course. Curve A is the state where almost every respondent begins with, and the proportion of respondents in this state has been dropping ever since. Curve B shows that, between age 20 and 30, overall majority of respondents take this hidden state. Proportion of respondents in state B starts increasing since the age 15 until the age 22. Curve C also shows a 'first increase then decrease' pattern as state B, however, proportion of respondents in this state is always lower than state B and the timing of decreasing (age 30) is later than that of state B. Curve D indicates that the proportion of respondents in state D keeps increasing throughout the whole young adulthood and becomes the majority after age 30.

<center>Insert Table 1</center>
<center>Insert Figure 6</center>

To understand the mechanisms behind the dynamic transition between these latent states, one can plot the state transition graph. As shown in Figure 6, from starting state A, one is 14 times more probable to transit to state B than to state C, given the transition probabilities of 0.14 (A to B) and 0.01 (A to C) in Table 1. Combined with emission probability distribution, state A is featured as being single (probability = 0.97), no child (probability = 0.99) and living with parents. State B is featured as being single (probability = 0.53), cohabiting (probability = 0.27) or married (probability = 0.20), no child and left parental home, whereas state C is featured as being single (probability = 0.17), cohabiting (probability = 0.25) or married (probability = 0.59), 1 child and left parental home. State C can be reached also from state B, which is 9 times more probable from State A. From state C, one can transit to state D. State D is an absorbing state, which means once one arrives at this state, transition to other states is not possible any more. State D is featured as low probability of being married (probability = 0.76), having 2 (probability = 0.68) or more (probability = 0.33) children. Summarizing the information given by Table 1, Figure 6 and Figure 7, one can interpret these four hidden states as inclinations of transition, reflecting the respondents' tendency to act during the stay of a certain state. Respondents in state A as the beginning of young adulthood: they are single, living with their parents, and having no child. They are also probably in school or training for future employment, which are not observable in the current dataset. Their life course activities are preparing them to leave parental home and start an independent life. Therefore, state A can be interpreted as inclining 'Leaving home'. Respondents are mainly in state B between age 20 and 30: they have different partnership status (mainly being single probability = 0.53), left parental home, and having no child. In this state, respondents' behaviors in are preparing themselves in 'Family formation'. In state C, the probability of partnership status shows high proportion of being married (0.59) and respondents already have one child. This state can be interpreted as 'Family extension'. State D, the absorbing state, is the 'Family completion' state. The young adulthood life course end at age 40, where the observed life course stops

in this study. Note that, the above-mentioned transition pattern applies to the whole sample, but respondents from different background (gender or education) may have different transition rate between states. The hidden Markov model with covariates are useful in studying the differences between social classes.

Insert Figure 7

After interpreting the hidden states as inclinations leading to young adult demographic transitions, it is necessary to visualize the hidden states paths throughout the whole young adulthood of respondents. The necessity comes on the one hand from need to check whether the hidden states paths fit substantive expectation and on the other hand from the three basic problems in any HMM application (describe in Introduction Section). The hidden states paths throughout the whole young adulthood of respondents (sort from end) are shown as sequence index plot in Figure 8 (a). Sequence index plots of longitudinal data use stacked bars or line segments to show how individuals move between a set of conditions or states over time [?]. Compared with the multi-channel sequence life course shown in Figure 5, the complicated partnership, fertility and leaving home trajectories are reduced into four category of inclinations. To better understand the heterogeneity in the transition into adulthood among these inclinations, sequence analysis with OM [?] was performed on the hidden state paths. Four typologies of the hidden states (chosen by cluster quality statistics [?]) were presented in Figure 8 (b) and Figure 8 (c). Figure 8 (b) are the sequence index plots of each typology and Figure 8 (c) is the sequence medoid plot [?], which is the most representing existing sequence in each typology. With the help of Figure 8 (b) and (c), four types of transition into adulthood have been identified, namely, 1: Late fertility or no fertility, 2: Traditional pathway (Leaving home at age 21, Family formation at age 26, Family extension at age 29, and followed by Family completion), 3: Small family (remaining in Family extension state until end of observation age 40), 4: Early transition (Leaving home at age 18, Family formation at age 21, Family extension at age 24, and followed by Family completion).

Insert Figure 8

3.2 HMM 4 with covariates in its latent model

As discussed in Introduction Section, HMM can allow covariates in its latent model to explain the heterogeneity in the population. One of the most intuitive way to include covariates in latent model is to allow for different transition probability distribution for different groups of respondents. It take 7.2 hours to perform HMM with 4 hidden states, 1000 random starting values, and including 2 variables, namely, education (high vs. low) and gender (female vs. males). The lowest BIC among these 1000 repentances is 113005, which is lower than the BIC of HMM 4 without covariates. The output model parameters are initial probability distribution, emission probability distribution (shown in Table 2) and 4 different transition probability

(shown in Table 3) distributions. Compared with the initial probability distribution, transition probability distribution and emission probability distribution of HMM 4 without covariate (shown in Table 1), there are some unnoticeable change in some probabilities, and the interpretation of these 4 hidden states remains the same.

Insert Table 2

Insert Table 3

The information of Table 3 reveals the difference in the transition to adulthood in gender and education level. From state A: 'Leaving home', high educated females are 0.01 faster (transition probability 0.16 vs. 0.15) to state B: 'Family formation' than the low educated males and similar pattern can be found in high educated males against low educated males. Compared with males, females are faster in transition from 'Leaving home' to 'Family formation'. From state B: 'Family formation', high educated females are 0.05 slower (transition probability 0.08 vs. 0.13) to state C: 'Family extension' than low educated females, and similar pattern can be found in high educated males against low educated males. Besides, males moves slower from 'Family formation' to 'Family extension' than females. Transition from state 'Family extension' to state D: 'Family completion' shows different pattern than from state 'Family formation' to state 'Family extension'. From State 'Family extension' to state 'Family extension', high educated females are 0.05 (0.19 vs. 0.14)than low educated females, and similar pattern can be found in high educated males against low educated males. For this transition, males are faster than females. To summarize, (1) high educated move out of parental home faster than low educated, females faster than males; (2) high educated start having child and change their partnership status slower than low educated, females faster than males; (3) high educated are faster having more child once they have one child, males faster than females.

4 Conclusion and Discussion

Most peoples life courses are made up of a multitude of changes in multiple life domains. A key challenge of life course research is to make sense of this complexity by searching for fundamental processes that drive these observable transitions and by examining which factors influence them. In this paper, we claim that Hidden Markov modeling (HMM) holds great promise in unraveling these processes, and we provide a relatively simple example of its potential by applying it to the family transition into adulthood among French men and women born between 1956 and 1965.

From a substantive point of view, what the HMM results reveal is that two fundamental viewpoints on the transition to adulthood can be distinguished. The HMM solution with four hidden states views the family transition to adulthood as a process that leads to the intergenerational reproduction of family life. The first challenge that young adults face is about leaving the parental home and finding a suitable partner relationship. The next steps in this intergenerational reproductive process are about

the initiation of a family (entry into parenthood), followed by successive phases of family expansion and family completion. Thus, the 4-HMM solution suggests a model of the full family cycle starting as a child in a family of origin and ending up as an adult in a next generation family. Our analysis also reveals clear differences in the speed and likelihood of making this transition between men and women and between the higher and lower educated. For instance, higher educated women make the first fundamental transition (out of the parental home) at earlier ages than lower educated women, but postpone the establishment of a family of their own. However, once they decide to establish a family, higher educated women are faster in making the family expansion step. Thus, our analysis shows that the pace and rhythm of this fundamental family succession model differs strongly between low and high educated women.

The 5-HMM solution (not presented) provides another interesting view on the family transition into adulthood. Rather than viewing this transition as a unilinear trajectory where young adults only differ in the likelihood and speed of moving to successive stages as is central to the 4-HMM solution, the 5-HMM solution distinguishes between two alternative family pathways into adulthood. As in the 4-HMM solution, the first challenge every young adult faces is when to leave the parental home. One pathway strongly resembles the traditional pathway where young adults first establish a traditional family, characterized by marriage and possibly a child, followed by a subsequent stage of family expansion. However, a second pathway is distinguished as well, where young adults opt for a more autonomous lifestyle, characterized by single living and/or unmarried cohabitation. After this stage, these young adults are confronted by another fundamental choice, either to continue this alternative lifestyle track and opt for children outside marriage, or to align themselves into the traditional pattern by moving back into the traditional family pathway. As with the 4-HMM solution, linking covariates to the 5-HMM structural model offers interesting insights. For instance, highly educated women are more likely to start off on the alternative track than low educated women, but once they enter this track, they are also more likely to revert to the traditional pattern than low educated women who start off on the alternative track.

Whether one interprets the data on the basis of the 4-HMM or 5-HMM solution at least partly depends on ones theoretical interests. The 4-HMM solution offers a succinct interpretation of the traditional family life pattern, pointing at three major decisions to be taken in the course of the family-life cycle [Glick, 1955]. The 5-HMM solution incorporates more heterogeneity into this family life cycle [Glick, 1989], and offers interesting opportunities to study the process of family change that is often captured under the heading of the Second Demographic Transition [Lesthaeghe, 1995].

A major advantage of both of these models is that they greatly limit the complexity of the process of transition into adulthood, by reducing the large number of transitions between observable states to a small number of transitions between unobservable, latent states. This property could be even more useful if the number of potential states and transitions becomes even larger, for instance if one wants to study both family transitions and career-related transitions in one model.

Mechanisms of the transition to adulthood: an application of Hidden Markov Models   15

The models introduced in this paper have clear merit for life course research. Several extensions of the Hidden Markov model could be envisaged, for example, constrained HMM. Constrained HMM is useful when one has a clear idea about the structure of the transition pattern, and want to test the hypothesized transition probability distribution. This paper did not elaborate on this type of topic yet, but it can be of great interest for future research. Generally, in applying these models to life course data, researchers have to be aware of both theoretical and practical restrictions on the analyses. Models should not become too complex in order for them to be mathematically feasible to estimate and to be theoretically interpretable. Our paper suggests a number of guidelines in this respect that may prove useful to future users.

Acknowledgement

References

[Andrieu and Doucet, 2000] Andrieu, C. and Doucet, A. (2000). Simulated annealing for maximum a posteriori parameter estimation of Hidden Markov Models. *IEEE Transactions on Information Theory*, 46(3):994–1004.

[Barban and Billari, 2012] Barban, N. and Billari, F. C. (2012). Classifying life course trajectories: a comparison of Latent Class and Sequence Analysis. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 61(5):765–784.

[Bartolucci et al., 2015] Bartolucci, F., Farcomeni, A., Pandolfi, S., and Pennoni, F. (2015). LMest: an R package for latent Markov models for categorical longitudinal data. *arXiv preprint:1501.04448*.

[Bartolucci et al., 2012] Bartolucci, F., Farcomeni, A., and Pennoni, F. (2012). *Latent Markov models for longitudinal data*. CRC Press.

[Baum et al., 1970] Baum, L. E., Petrie, T., Soules, G., and Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Annals of Mathematical Statistics 41:164-71*, 41(1):164–171.

[Bijleveld and Mooijaart, 2003] Bijleveld, C. C. and Mooijaart, A. (2003). Latent Markov modelling of recidivism data. *Statistica Neerlandica*, 57(3):305–320.

[Blossfeld et al., 2007] Blossfeld, H.-P., Golsch, K., and Rohwer, G. (2007). *Event History Analysis with STATA*. Lawrence Erlbaum.

[Buchmann and Kriesi, 2011] Buchmann, M. C. and Kriesi, I. (2011). Transition to adulthood in Europe. *Annual Review of Sociology*, 37:481–503.

[Cook et al., 2000] Cook, R. J., Ng, E., and Meade, M. O. (2000). Estimation of operating characteristics for dependent diagnostic tests based on latent markov models. *Biometrics*, 56(4):1109–1117.

[Cornwell, 2015] Cornwell, B. (2015). *Social Sequence Analysis: Methods and Applications*. Structural Analysis in the Social Sciences. Cambridge University Press, New York.

[Dempster et al., 1977] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likely-hood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B*, 39:1–38.

[Elder Jr, 1985] Elder Jr, G. H. (1985). Perspectives on the life course.

[Elzinga et al., 2007] Elzinga, C. H., Hoogendoorn, A. W., and Dijkstra, W. (2007). Linked Markov sources: Modeling outcome-dependent social processess. *Sociological Methods & Research*, 36:26–47.

[Fokkema et al., 2016] Fokkema, T., Kveder, A., Hiekel, N., Emery, T., and Liefbroer, A. C. (2016). Generations and gender programme wave 1 data collection: An overview and assessment of sampling and fieldwork methods, weighting procedures, and cross-sectional representativeness. *Demographic Research*, 34:499.

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

[Glick, 1955] Glick, P. C. (1955). The life cycle of the family. *Marriage and Family Living*, 17(1):3–9.

[Glick, 1989] Glick, P. C. (1989). The family life cycle and social change. *Family relations*, pages 123–129.

[Langeheine and Van de Pol, 1990] Langeheine, R. and Van de Pol, F. (1990). A unifying framework for Markov modeling in discrete space and discrete time. *Sociological Methods & Research*, 18(4):416–441.

[Lesthaeghe, 1995] Lesthaeghe, R. (1995). The Second Demographic Transition in Western countries: An interpretation. *Gender and family change in industrialized countries*, pages 17–62.

[Manzoni et al., 2010] Manzoni, A., Vermunt, J. K., Luijkx, R., and Muffels, R. (2010). Memory bias in retrospective collected employment careers: a model-based approach to correct for measurement error. *Sociological Methodology*, 40(1):39–73.

[Mayer, 2009] Mayer, K. U. (2009). New directions in life course research. *Annual Review of Sociology*, 35:413–433.

[Paas et al., 2007] Paas, L. J., Vermunt, J. K., and Bijmolt, T. H. (2007). Discrete time, discrete state latent Markov modelling for assessing and predicting household acquisitions of financial products. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 170(4):955–974.

[Pakpahan et al., 2015] Pakpahan, E., Hoffmann, R., and Kröger, H. (2015). Statistical methods for causal analysis in life course research: an illustration of a cross-lagged structural equation model, a latent growth model, and an autoregressive latent trajectories model. *International Journal of Social Research Methodology*.

[Rabiner, 1989] Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286.

[Richardson et al., 2011] Richardson, K., Harte, D., and Carter, K. (2011). *Understanding health and labour force transitions: Applying Markov models to SoFIE longitudinal data*. Statistics New Zealand.

[Spedicato et al.,] Spedicato, G. A., Kang, T. S., and Yalamanchi, S. B. The markovchain package: A package for easily handling discrete Markov chains in R.

[Vermunt, 1997] Vermunt, J. K. (1997). *Log-Linear Models for Event Histories*. Advanced quantitatiev Techniques in the Social Sciences 8. Sage, Thousand Oaks.

[Vermunt et al., 1999] Vermunt, J. K., Langeheine, R., and Bockenholt, U. (1999). Discrete-time discrete-state latent Markov models with time-constant and time-varying covariates. *Journal of Educational and Behavioral Statistics*, 24(2):179–207.

[Viterbi, 1967] Viterbi, A. J. (1967). Error bounds for convolutional codes and an asymptotically optimal decoding algorithm. *IEEE Transactions of Information Theory*, IT-13:260–269.

[Zucchini and MacDonald, 2009] Zucchini, W. and MacDonald, I. L. (2009). *Hidden Markov models for time series: an introduction using R*. CRC press.

Table 1: HMM model with 4 hidden states output parameter: Transition probability distribution, initial probability distribution and emission probability distribuion (ordered).

Transition probability distribution

State	State			
	A	B	C	D
A	.85	.14	.01	0
B	0	.91	.09	0
C	0	0	.85	.15
D	0	0	0	1

Initial probability distribution

State	A	B	C	D
	.94	.06	0	0

Emission probability distribution

Fertility category	State			
	A	B	C	D
0	.99	1	0	0
1	.01	0	1	0
2	0	0	0	.68
3+	0	0	0	.33

Partnership category	State			
	A	B	C	D
S	.97	.53	.17	.11
U	.02	.27	.25	.14
M	.01	.20	.59	.76

Leaving home category	State			
	A	B	C	D
No	1	0	.02	.01
Yes	0	1	.98	.99

Table 2: HMM model 4 with covariates output parameter: initial probability distribution and emission probability distribuion (ordered).

Initial probability distribution

State	A	B	C	D
	.94	.06	0	0

Emission probability distribution

Fertility category	State A	B	C	D
0	.99	1	0	0
1	.01	0	1	0
2	0	0	0	.67
3+	0	0	0	.33

Partnership category	State A	B	C	D
S	.97	.50	.17	.11
U	.02	.29	.25	.14
M	.01	.21	.58	.75

Leaving home category	State A	B	C	D
No	1	0	.02	.01
Yes	0	1	.98	.99

Table 3: HMM model 4 with covariates output parameter: Transition probability distributions of low educated males, high educated males, low educated females and high educated females (ordered).

Transition probability distribution of low educated males

State	State			
	A	B	C	D
A	.86	.13	.01	0
B	0	.91	.09	0
C	0	0	.86	.14
D	0	0	0	1

Transition probability distribution of high educated males

State	State			
	A	B	C	D
A	.85	.14	0	0
B	0	.94	.06	0
C	0	0	.81	.19
D	0	0	0	1

Transition probability distribution of low educated females

State	State			
	A	B	C	D
A	.84	.15	.01	0
B	0	.87	.13	0
C	0	0	.86	.14
D	0	0	0	1

Transition probability distribution of high educated females

State	State			
	A	B	C	D
A	.83	.16	.01	0
B	0	.91	.08	0
C	0	0	.82	.19
D	0	0	0	1
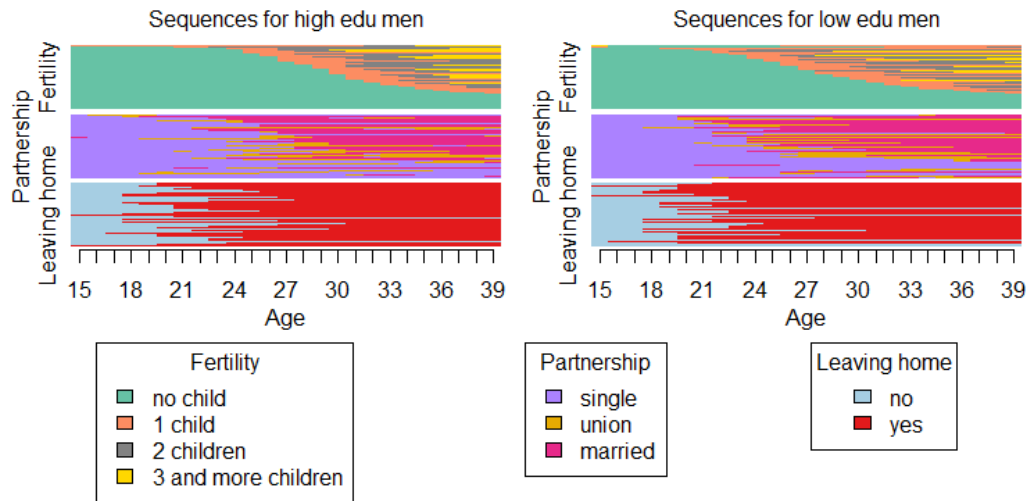
(a)
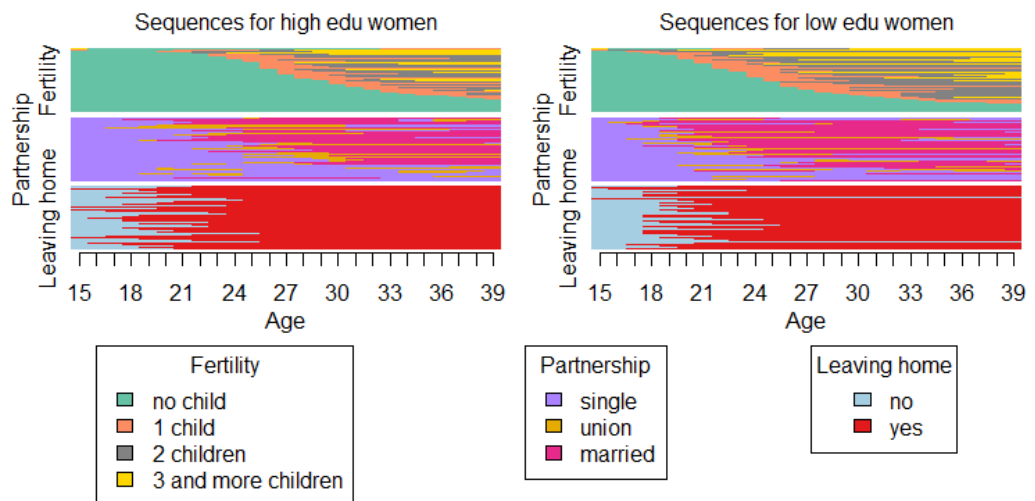
(b)

Figure 5: Multi-channel sequence presentation of the fertility, partnership and leaving home annual data of respondents between age 15 and 40 in France GGP birth cohort 1956-1965 (a) are sequences for high educated and low educated males. (b) are sequences for high educated and low educated females.
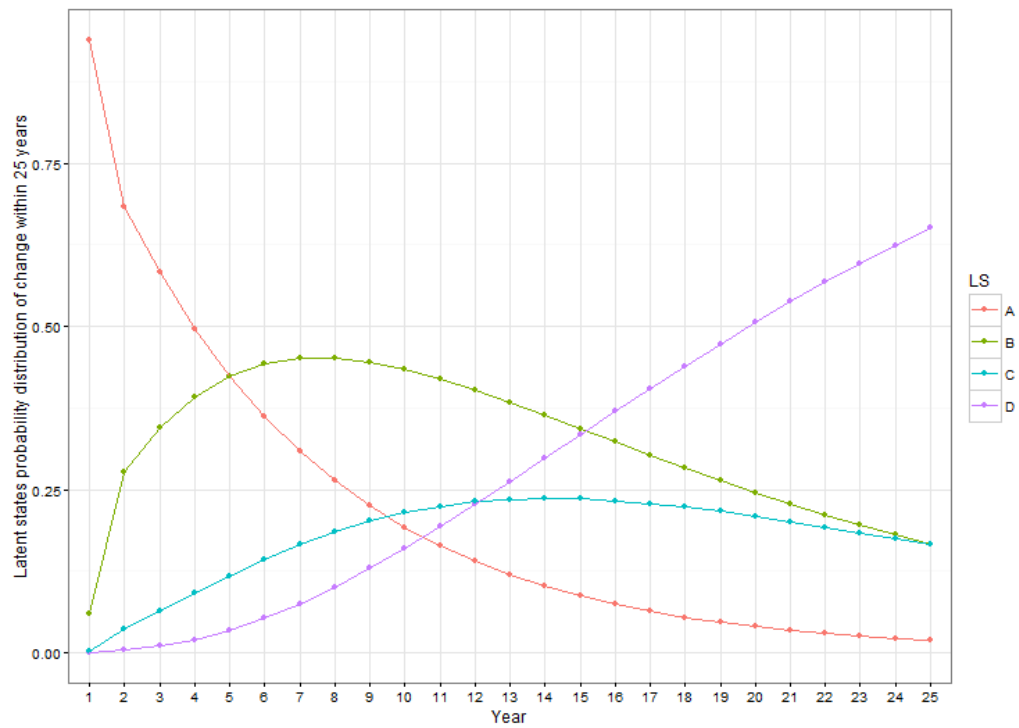
Figure 6: Latent probability distribution of change in the 25 years of HMM 4 based on the initial probability and transition proability distribution of HMM 4. State ordering corresponding to Table 1.
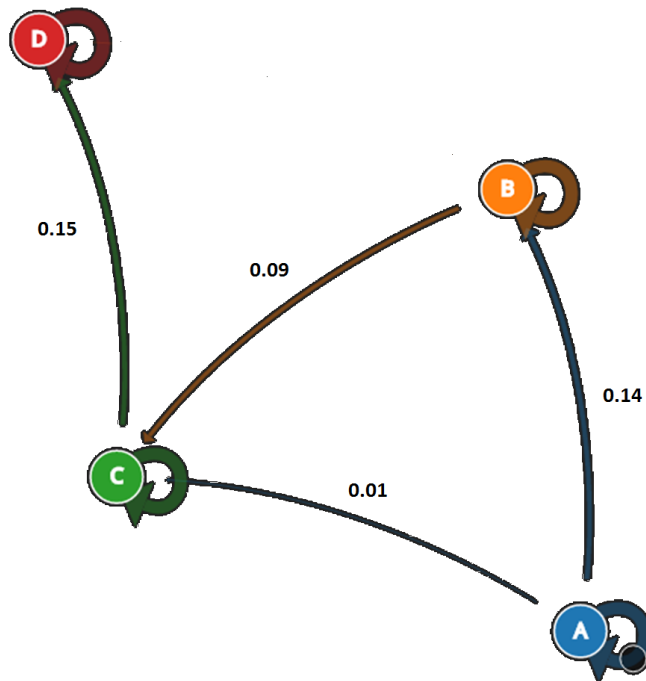
Figure 7: State transition graph based on the transition probability distribution of HMM 4. Thickness of the arrows reflects the transition probabilities (the transition probability not to itself are shown next to the arrow). State ordering corresponding to Table 1.

(a)

(b)

(c)
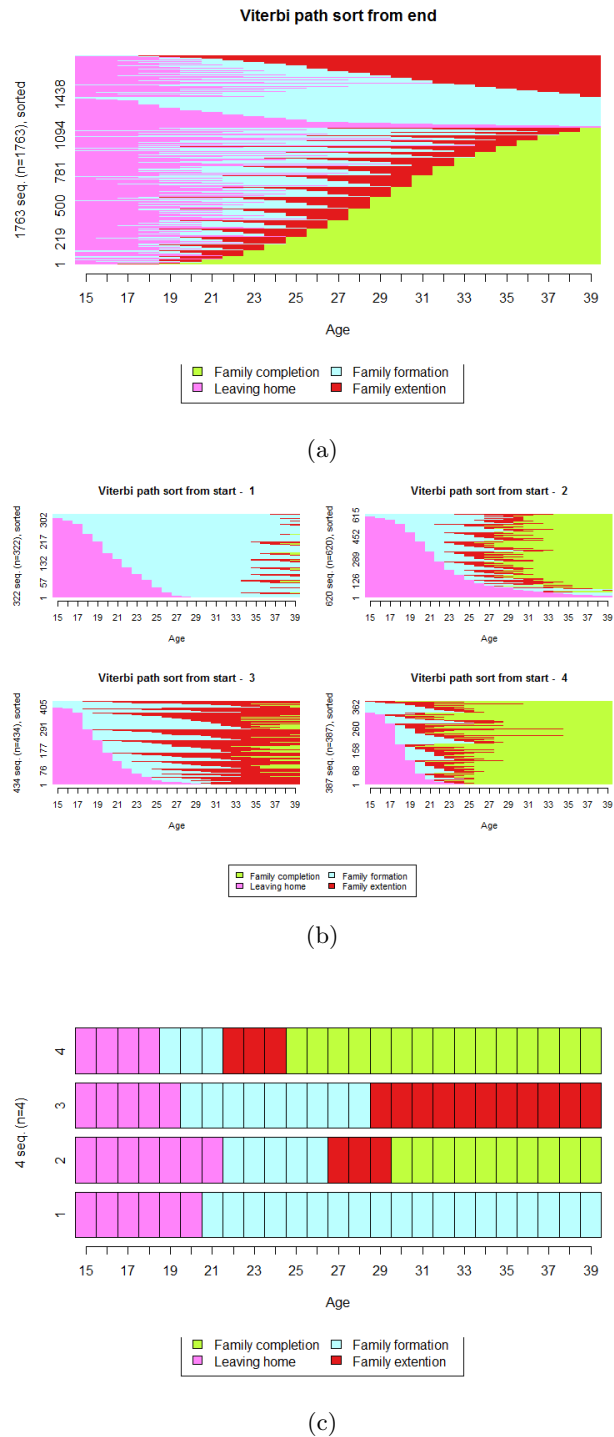
Figure 8: a: Sequence index plot of Viterbi path of all unique multi-channel sequence in the selected French GGP dataset. b: Sequence index plot of classified (Sequence Analysis) viterbi path. c: Sequence medoid plot of classified (Sequence analysis) viterbi path.

Markovian-Based Clustering of Internet Addiction Trajectories

Zhivko Taushanov and André Berchtold

Abstract A hidden Markov clustering procedure is applied to a sample of n=185 longitudinal Internet Addiction Test trajectories collected in Switzerland. The best solution has 4 groups. This solution is related to the level of emotional wellbeing of the subjects, but no relation is observed with age, gender and BMI.

1 Introduction

Excessive Internet use is an emerging health issue in the medical literature. The most current tool to quantify the degree of addiction to Internet is the Internet Addiction Test (IAT) developed by Young [7]. However, this scale based on 20 items is quite long and it was not validated for agreement between successive measurements performed on the same subjects, so it is difficult to distinguish between real changes of the subjects and changes due to the IAT itself.

We consider data taken from the *ado@internet.ch* study [6], a longitudinal study about the use of Internet among youth in the Swiss canton of Vaud. Data were collected at 5 occasions every 6 months from Spring 2012 (T0) to Spring 2014 (T4). A convenience sample of n=185 adolescents having answered to all 5 waves is used in the present study (67% females, mean age at T0: 14.1 years). Our goals are 1) to classify the trajectories of Internet addiction into meaningful categories, 2) to test whether this classification is related to other characteristics such as age, gender, wellbeing (measured by the WHO-5 index) and Body Mass Index (BMI). We hypothesize that if significant relationships exist with other characteristics, then the IAT could be considered as a reliable measure of the evolution of Internet addiction.

Zhivko Taushanov
University of Lausanne, ISS & NCCR LIVES, e-mail: zhivko.taushanov@unil.ch

André Berchtold
University of Lausanne, ISS & NCCR LIVES, e-mail: andre.berchtold@unil.ch

2 Model

The clustering of longitudinal continuous data is still an open question. In this paper, we use a specific class of Markovian Models called Hidden Mixture Transition Distribution model (HMTD) for that purpose. These models consist in a latent and an observed levels [4]. The visible level is a Mixture Transition Model (MTD) model introduced by Raftery in 1985 as a modelling of high order Markov chains [5] and developed later by Berchtold [1],[2] and Berchtold and Raftery [3]. Here we use a Gaussian version of the MTD model where the mean of the Gaussian distribution is a function of past observations. The latent level of the model is a homogeneous Markov chain. Each state of the chain is associated to a different component (or model) at the visible level, and the transition matrix is used to determine which component best represents the current observation.

To use the HMTD model as a clustering tool, we fix the hidden transition matrix to the identity matrix. Each sequence of successive observations is then associated to only one component of the model. The dependence order for the Gaussian distributions is fixed to one, and the number of components varies from 2 to 5. We use the Bayesian Information Criterion (BIC) to select the best model.

3 Results

The best model identified by the BIC is the 4 components model. Figure 1 shows the IAT trajectories associated to each group. We clearly differentiate between one group with average volatility and IAT level (group 1); one group with relatively low scores and low variance (2); one group with very low variability and low and constantly diminishing IAT score (3); one erratic group with high variability (4). When comparing the classification with covariates, no significant relationship does appear with age, gender and BMI. However a significant relationship appears with the WHO-5 measured on each wave.

4 Discussion

The HMTD model is able to classify sequences of continuous longitudinal data into as many groups as required. In our example, the four resulting groups differ in terms of average value and of variability. The relationship observed between the IAT and emotional wellbeing suggests that both concepts are linked and that a higher risk of Internet addiction is related to a poorer wellbeing.

Markovian-Based Clustering of Internet Addiction Trajectories
3
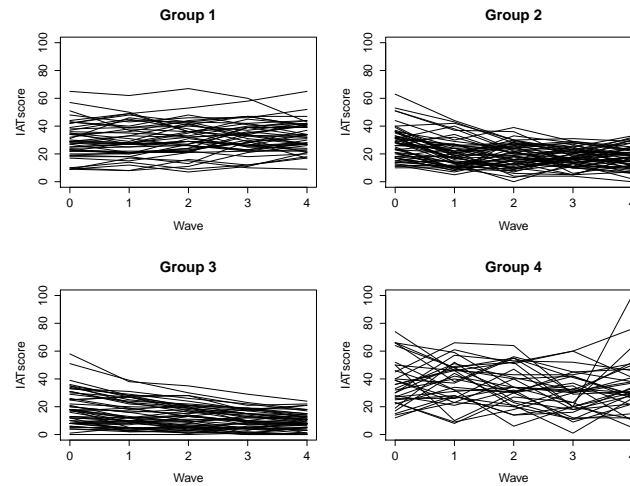
Fig. 1 IAT sequences associated with each cluster in the 4-groups solution.

References

1. Berchtold A. (2001) Estimation in the Mixture Transition Distribution Model. *Journal of Time Series Analysis* 22:379-397
2. Berchtold, A.(2003) Mixture transition distribution (MTD) modeling of heteroscedastic time series. *Computational statistics and data analysis* 41:399-411
3. Berchtold, A. and Raftery, A. (2002) The mixture transition distribution model for high-order Markov chains and non-Gaussian time series. *Statistical Science* 17:328-356.
4. Bolano, D., Berchtold, A. (2016) General framework and model building in the class of Hidden Mixture Transition Distribution models. *Computational Statistics and Data Analysis* 93:131-145
5. Raftery, A. (1985) A Model for High-Order Markov Chains. *Journal of the Royal Statistical Society* 47:528-539
6. Suris, J.-C., Akré, C., Berchtold, A., Fleury-Schubert, A., Michaud, P.-A. and Zimmermann, G. (2012) Ado@Internet.ch: Usage d'internet chez les adolescents vaudois. Lausanne: Institut universitaire de médecine sociale et préventive. (Raisons de santé, 208).
7. Young, K.S. (1998) Internet Addiction: The Emergence of a New Clinical Disorder. *CyberPsychology & Behavior* 1:237-244

Session 5A: Health

Application of 'pseudo panels' to investigate causal link between HIV and fertility in sub-Saharan Africa

M.A. Magadi, School of Social Sciences, University of Hull.

Abstract

Panel data are valuable for answering questions about change over time, but remain relatively scarce in most developing countries, especially sub-Saharan Africa (SSA). Where there exists a series of repeated cross-sectional data, 'pseudo-panels' provide a promising alternative. The use of 'pseudo panels' has received considerable attention in econometrics, but application in Demography remains rare. This paper explores the potential for using 'pseudo panels' to investigate causal link between HIV and fertility in SSA. The relationship between HIV and fertility is a complex one, partly because causality can run in either direction. We focus primarily on fertility as the outcomes of interest and HIV as a contributing factor. Repeated cross-sectional Demographic and Health survey (DHS) data from 20 countries in SSA are used to construct "pseudo panels" based on birth cohorts by country. The pseudo panels allow an investigation of possible causal link between HIV in an earlier survey and fertility behaviour of similar cohorts in a subsequent survey. Measures of HIV are based on risk perception and HIV status, while fertility is based on births within the last five years and future fertility intention. A total of 140 cohorts (7 age groups * 20 countries) were constructed, with 120 cohorts having data for at least two time periods. The analysis used two alternative approaches: (i) 'Conditional' models of HIV in an earlier survey and fertility behaviour in subsequent surveys; and (ii) Repeated measures multilevel analysis, with cohort as Level-2, and measurement occasion as level-1. An evaluation /assessment of the analysis involved a comparison of findings from 'pseudo cohort' with individual-level analysis, and a multi-level estimation of intra-cohort correlation coefficients to assess the degree of similarity of individuals in the same cohort. A multivariate analysis based on fixed effects models was used to determine the extent to which observed patterns may be attributable to key demographic/ socio-economic differences or infer possible causal links. Preliminary analysis shows promising results on application of 'pseudo panels' in investigation of demographic causal links in settings with limited panel data such as sub-Saharan Africa. However, further analysis is necessary for conclusive results. In particular, advanced modelling using Multiprocess modelling or Structural equation modelling will be used to address possible endogeneity in the relationships observed.

Background context

Panel data are valuable for answering questions about change over time, but remain relatively scarce in most developing countries, especially sub-Saharan Africa (SSA). Where there exists a series of repeated cross-sectional data, 'pseudo-panels' provide a promising alternative. In the 'pseudo panel' approach "similar" individuals are grouped in cohorts, and aggregate measures within these cohorts are treated as observations in a synthetic panel (Deaton, 1985). The use of 'pseudo panels' has received considerable attention in econometrics (Browning, Deaton and Irish, 1985;Verbeek and Nijman 1992; Moffitt, 1993; Collado, 1997; Propper, Rees and Green, 2001; McKenzie, 2004; Verbeek, 2008), but application in Demography remains rare. This paper explores the potential for using 'pseudo panels' to investigate causal link between HIV and fertility in SSA. The relationship between HIV and fertility is a complex one, partly because causality can run in either direction. We focus primarily on fertility as the outcomes of interest and HIV as a contributing factor.

Data and Methods

Repeated cross-sectional Demographic and Health survey (DHS) data from 20 countries in SSA are used to construct "pseudo panels" based on birth cohorts by country. The pseudo panels allow an investigation of possible causal link between HIV in an earlier survey and fertility behaviour of similar cohorts in a subsequent survey. The 20 countries included in the analysis represent all countries in sub-Saharan Africa where the DHS has included HIV testing on nationally representative samples of women of reproductive age in at least two consecutive surveys (see Table 1). Measures of HIV are based on risk perception and HIV status, while fertility is based on births within the last five years and future fertility intention.

A total of 140 cohorts (7 age groups * 20 countries) were constructed, with 120 cohorts having data for at least two time periods. The analysis used two alternative approaches: (i) 'Conditional' models of HIV in an earlier survey and fertility behaviour in subsequent surveys; and (ii) Repeated measures multilevel analysis, with cohort as Level-2, and measurement occasion as level-1. An evaluation /assessment of the analysis involved a comparison of findings from 'pseudo cohort' with individual-level analysis, and a multi-level estimation of intra-cohort correlation coefficients to assess the degree of similarity of individuals in the same cohort. A multivariate analysis based on fixed effects models was used to determine the extent to which observed patterns may be attributable to key demographic/ socio-economic differences or infer possible causal links.

Table 1: Summary of DHS (and AIS[1]) data in the study

	Country /Year of survey	Women tested for HIV	% HIV+	Number of cases HIV+
1	*Burkina Faso (2003, **2010**)	8384	1.2	101
2	*Cameroon (2004, **2011**)	7221	5.6	404
3	*DR Congo (2007, **2013/14**)	9264	1.6	148
4	*Cote d'Ivoire (2005**, 2012**)	4509	4.6	207
5	*Ethiopia (2005, **2011**)	5942	1.9	113
6	*Guinea (2005, **2012**)	4622	2.1	97
7	*Kenya (2003, **2008/9;** AIS: 2007, 2012)	3641	8.0	291
8	*Liberia (2007, **2013**)	4397	2.4	106
9	*Lesotho (2004, **2009**)	3800	27.0	1026
10	*Malawi (2004, **2010**, 2015-16)	7091	12.9	915
11	*Mali (2006, **2013**)	4806	1.3	62
12	*Mozambique (**2009**)	6000	13.1	786
13	*Niger (2006, **2012**)	5000	0.4	20
14	*Rwanda (2005, **2010**)	6917	3.7	256
15	*Senegal (2005, **2011**)	5300	0.8	42
16	*Siera Leone (2008, **2013**)	7695	1.7	131
17	*Tanzania (**2010**,AIS:2003, 2007, 2011/2)	8179	6.6	540
18	*Uganda (2004/05, **2011**)	11950	8.3	992
19	*Zambia (2007, 2**013**)	14719	15.1	2223
20	*Zimbabwe (2005/6, **2010/11**, 2015)	7313	18.0	1316
	Overall sample	179747		13071

*Data with HIV test results for these countries are available for multiple rounds of DHS surveys, enabling construction of 'pseudo-panels' for causal analysis. Data shown (marked in **bold**) are for the latest survey, unless the latest survey is still on-going and data not yet available**.**

1 – AIDS Indicator Survey

Preliminary findings

Preliminary findings suggest that perceived risk has a stronger correlation with fertility behavior and intentions, than HIV status. Observed patterns suggest that in settings where a large proportion of individuals do not know their HIV status, perceived risk of HIV is a stronger determinant of fertility behavior than the actual HIV status. Stronger associations were observed between HIV and fertility intentions than actual fertility, consistent with patterns observed at individual level. It is interesting to note that for HIV status, the difference in fertility intention between HIV-positive and HIV-negative women is notably larger in later than in earlier survey, presumably since a higher proportion of individuals in the later surveys knew their HIV status. For perceived risk, the differences between none/low and moderate/high risk is more or less the same at different time periods.

Discussion and further analysis

The preliminary analysis shows promising results on application of 'pseudo panels' in investigation of demographic causal links in settings with limited panel data such as sub-Saharan Africa. Indeed, 'pseudo panels' do have some merits over true panels since such data are immune to attrition bias or other threats to internal validity such as testing bias or design contamination. However, they are prone to instrumentation bias if the study variables were measured differently in subsequent surveys. Furthermore, an important consideration in this application is whether the number of DHS rounds considered here (typically two) is adequate. Nevertheless, it has been noted that 'In longitudinal repeated measures design, we usually have a large number of level 2 units with rather few Level-1 units (Rasbash et al, 2015: 195), and an advantage of multilevel modelling of repeated data is the ability to handle unequal measurement intervals (Rasbash et al, 2015).

Although overall preliminary results seem plausible, further analysis is necessary for conclusive results. First, it is important to establish whether it is appropriate to lag or not to lag the dependent variable. More importantly, advanced modelling is necessary to address possible endogeneity in the relationships observed. This will be undertaken using Multiprocess modelling or Structural equation modelling.

References

Browning M, Deaton A, Irish M (1985). A profitable approach to labor supply and commodity demands over the life cycle. *Econometrica* 53:503-543

Collado, M.D. (1997), Estimating Dynamic Models from Time Series of Independent Cross-Sections, *Journal of Econometrics*, 82, 37–62.

Deaton, A (1985). Panel data from time series of cross-sections. *Journal of Econometrics* vol. 30(1-2): 109-126.

McKenzie, D.J. (2004), Asymptotic Theory for Heterogeneous Dynamic Pseudo-Panels, *Journal of Econometrics,* 120, 235–262.

Moffitt, R. (1993), Identification and Estimation of Dynamic Models with a Time Series of Repeated Cross-Sections, *Journal of Econometrics*, 59, 99–123.

Propper, C., H. Rees and K. Green (2001), The Demand for Private Medical Insurance in the UK: A Cohort Analysis, *The Economic Journal*, 111, C180–C200.

Rasbash, J., Steele, F., Browne, W.J. and Goldstein, H. (2015) *A User's Guide to MLwiN, v2.33.* Centre for Multilevel Modelling, University of Bristol.

Verbeek, M (2008). Pseudo-panels and repeated cross-sections in Matyas, L. and Sevestre, P (Eds). *The Econometrics of Panel Data: Fundamentals and recent developments in theory and practice.* Advanced Studies in theoretical and Applied econometrics, Third Edition, Volume 46: 369-383 (Chapter 11), Springer.

Verbeek M. and Nijman T (1992). Can Cohort Data be Treated as Genuine Panel Data? *Empirical Economics* 17:9-23

Early Maternal Employment Sequences and Child Body Weight at Age Six
Evidence from the German Socio-Economic Panel

Michael Kühhirt

Abstract Using data from the German Socio-Economic Panel on children born between 2002 and 2006, this study investigates the relationship between mothers' monthly employment sequences in the first five years after birth and their children's body weight around age six. Employment history is measured by different variables derived from sequence analysis that each capture specific aspects such as typical employment patterns (based on additional cluster analysis), complexity, and turbulence. The association between these measures and children's body weight is estimated by inverse probability of treatment weighting of marginal structural models, a method developed in epidemiology to deal with the challenges of estimating the effects of time-varying exposures such as maternal employment. The results indicate that children who experienced very different maternal employment sequences but are similar with regard to background characteristics such as maternal education, household income, and family structure show no substantively or statistically significant disparities on body weight around age six. The study goes beyond the literature by focusing on maternal employment history beyond simple duration measures and by explicitly accounting for time-varying confounders of the relationship between maternal employment and child development.

Michael Kühhirt
Institute of Sociology and Social Psychology, University of Cologne, Germany
e-mail: michael.kuehhirt@uni-koeln.de

Care pathways of patients affected with multiple sclerosis in France from 2007 to 2013 using administrative databases and state sequence analysis

Jonathan Roux, Nolwenn Le Meur, Olivier Grimaud and Emmanuelle Leray

Abstract In France there is a lack of accurate and up-to-date data on care pathways and care consumption of patients affected with multiple sclerosis (MS). The aim of this study was to describe care consumption of MS people in France and build a typology of their care pathways over the 2007-2013 period. To answer this issue, sequence analysis and agglomerative hierarchical clustering were used on a random sample of 1,000 patients, issued from French health-care databases, split according to the data available. The typologies were then described according to individual characteristics and care consumption. Two similar partitions, using indel costs fixed to 0.9 and transition-based substitution costs, were obtained : a five-cluster and a six-cluster typology for respectively complete and incomplete care pathways (i.e. less than seven years). Low care consumption was associated with older and less treated patients. Patients having a medium-low care consumption were younger and treated, whereas those having a high consumption seems to be older and have more comorbidities than the others. This pioneer study, using an innovative method in health field, gives a first overview of the care consumption of MS-affected people in France using objective and quantitative information. To go further, same work will be performed on the whole French population affected with MS and will include biological and medical imaging exams, and specialists' care to complete care pathways.

Keywords : State sequence analysis, Care pathways, Multiple sclerosis, Administrative databases

Jonathan Roux[1,2],Nolwenn Le Meur[1], Olivier Grimaud[1], Emmanuelle Leray[1,2]
[1] EHESP French School of Public Health, Sorbonne Paris Cité, Rennes, France
[2] CIC-P 1414, CHU de Rennes, Rennes, France

1 Introduction

The increase of life expectancy and medical progress are leading to a steady growth of the number of people affected with a chronic disease [Organisation Mondiale de la Santé (OMS), 2002]. In France, 15 million people are affected with at least one chronic disease, namely slightly more than 20% of the whole population [Ministère de la Santé, de la Jeunesse, des Sports et de la Vie associative, 2007]. They require large amounts of medical care, which lead to new constraints of long-term care for healthcare systems [Organisation Mondiale de la Santé (OMS), 2002]. In France, the Long Disease Duration (LDD) status permits people to have a 100% coverage of their healthcare. Thirty illnesses are apart of the list of LDD, including multiple sclerosis (MS).

MS is a chronic neurological disease affecting 2.3 million people around the world with a women:men sex-ratio varying from 2:1 to 2.5:1 [Browne et al., 2014, Trojano et al., 2012, MS International Federation, 2015]. In France, the estimated prevalence is 151.2 cases per 100 000 inhabitants on 31st December 2012, namely about 100 000 people [Foulon et al., 2015]. MS mostly starts at young adulthood, between 20 and 40 years old, evolves for several decades and frequently leads to a disability, the median duration before the need for a walking aid being from 20 to 25 years [Confavreux et al., 2000, Confavreux et al., 2003]. Moreover, because of a life expectancy reduced of about 7 years [Leray et al., 2015] and availability of more and more DMTs (Disease Modifying Therapies), MS cost becomes bigger and bigger over time for each patient and thus for the society and was equal to 1 billion euros in France for the year 2013 [Direction de la stratégie, des études et des statistiques (DSES), 2015, Lefeuvre et al., 2016].

Care consumption of MS patients is expected to be high due to disease length, disease activity and disability progression. However, there is a lack of accurate and up-to-date data on this topic in France. Indeed, French studies on care consumption mainly focus on MS costs from the point of view of the French national Health Insurance system [Fromont et al., 2014, Lefeuvre et al., 2016] and not on the type, amount and chronology of care. In addition, there are no clearly definite care's recommendations for patients affected with MS at the moment in France.

In order to fill this gap and have an overview of the care consumption of people with MS in France, the first step is to measure the consecutive consultations and treatments constituting the care pathway of each patient. Therefore, this study aimed at describing the overall care consumption of people affected with MS, analysing their care pathways and creating a typology over the 2007-2013 period in France.

2 Methods

2.1 Study population

Data was issued from a French permanent sample of health insurance care users, named *Echantillon Généraliste des Bénéficiaires (EGB)*. It is a random sample of $1/97^{th}$ of the French national Health Insurance system, gathering the out-hospital reimbursed care consumption (consultations and home visits to general practitioners (GP), specialists, drugs deliveries, nurses, physiotherapists, ...). This sample is dynamic, which means that new people can enter the sample anytime. Moreover, we used data from the French Hospital Discharge Database (*Programme de Médicalisation des Systèmes d'Information (PMSI)*), gathering in-hospital care consumption (admissions including day hospital, but except outpatient visits).

Patients affected with MS were identified if at least one of the following criteria was fulfilled in the period stretching from 1st January 2007 to 31st December 2012 : either MS LDD status, either at least one DMT specific of MS dispensed (interferon-β, glatiramer acetate, fingolimod or natalizumab), or at least one hospital admission with an ICD-10 (International Classification of Diseases, 10th version) diagnosis code "G35" [World Health Organization, 2015]. Out-hospital care consumption and hospital admissions for these patients were then extracted from 2007 up to end 2013. As a whole, 1,003 patients were identified, of whom 3 were excluded due to a total lack of care consumption (MS-related and not) during the study period.

2.2 Data

Based on data of care consumption, monthly number of consultations or home visits with a GP, a private neurologist and a PM&R (Physical Medicine and Rehabilitation) physician were estimated for each of the 1,000 patients. Only MS-related hospital admissions were considered (main or related diagnosis equal to "G35", except DMT injection) and a parameter was created to count the monthly corresponding number for each individual. Nine DMT used for MS delivered out of the hospital (four interferon-β, glatiramer acetate, fingolimod, mycophenolate mofetil, methotrexate, azathioprine) and one in-hospital treatment (natalizumab) were also explored if applicable. The total duration of each treatment was estimated as the product of the number of boxes and the number of corresponding days. None of the two databases permitted to access consultations with public neurologists. Therefore we assumed that patients receiving a DMT, compulsorily prescribed by a neurologist, without having a visit with a private one, has one visit with a public neurologist (outpatient visit) at the treatment initiation and one every year as long as the treatment was ongoing.

A monthly then annual composite variable characterizing the individual care consumption was created by adding monthly (respectively annual) numbers of consultations with GPs, private and public neurologists, PM&R physicians and MS-related hospital admissions (except monthly DMT injections). Each yearly variable ranged widely and followed a Poisson distribution. Therefore, to have an overview of the care consumption, it was categorized in five groups according to the quartile distribution : no consumption ("0"), less than the first quartile ("]0; $Q1$]"), between the first quartile and the median included ("]$Q1$; $Q2$]"), between the median and the third quartile included ("]$Q2$; $Q3$]") and strictly superior to the third one (">$Q3$"). The sequence of the seven annual categorized parameters constituted the care pathway of each patient during the 7-year period. As inclusion criteria offered the opportunity to identify prevalent cases of MS, patients were not at the same stage of the disease during the 2007-2013 period and thus the sequences were not left-aligned for people entering the study after 2007.

For patients who died during follow-up, the "death" status was attributed the year following the death's year, since only annual states were considered. Therefore only deaths over the 2007-2012 period were considered. Death was coded as NA state, creating care pathways of different lengths. This solution permitted to take into account the death of people without the creation of a specific group, which may have happened with a devoted state.

2.3 Statistical analysis

Two different typologies of care pathways were created using Optimal Matching (OM) analysis and then agglomerative hierarchical clustering using Ward's criterion. The first one was obtained using patients with complete care pathways (n=648), i.e. sequences beginning in 2007, whereas the second focused on incomplete sequences (n=352), corresponding to people entering the cohort later than 2007 (due to dynamic cohort). This separation was chosen to avoid the creation of a cluster devoted to patients entering the cohort with delay appearing when analysing the 1,000 sequences together.

No theoretical knowledge about costs was available, therefore substitution costs were estimated empirically using observed transition rates. To ensure Needleman-Wunsch algorithm could be applied to compute OM distance, metric properties were verified for substitution costs, especially triangle inequality [Studer and Ritschard, 2015]. Costs are one major parameter permitting to have an optimal matching and thus influence clustering results [Lesnard, 2010]. For our purpose, the aim was to compare care pathways by keeping the temporality of events and studying series of sub-sequences and thus long-term care consumption rather than succession of one-state events. Moreover, we did not want to focus on outstanding events but rather on the whole sequence. Therefore, we chose to fix indel costs to 0.9 in order to

approach the Levenshtein II distance and thus compare sequences according to the longest common subsequence [Lesnard, 2010].

An agglomerative hierarchical clustering analysis using Ward's criterion was then performed on the dissimilarity matrix to create homogeneous groups. Several parameters were used to assess the quality of the partition, namely weighted Average Silhouette Width (ASWw) [Kaufman and Rousseeuw, 1990, Studer, 2013], Hubert's C (HC) [Hubert and Levin, 1976], Hubert's Gamma (HG) [Hubert and Arabie, 1985] and Point Biserial Correlation (PBC) [Hennig and Liao, 2010, Milligan and Cooper, 1985]. The maximisation of ASWw, HG and PBC and the minimisation of HC permitted to set the optimal number of clusters. Once the clustering obtained, each group was characterized according to individual characteristics and care consumption.

All computational and statistical analyses were performed using R (Version 3.2.3) [R Core Team, 2015]. Sequence analysis and clustering were performed using the *TraMineR* library (Version 1.8-11), the *WeightedCluster* library (Version 1.2) and the *cluster* library (Version 2.0.3) [Gabadinho et al., 2011, Studer, 2013, Maechler et al., 2015].

3 Results

3.1 Population characteristics

According to the selection criteria, 1,000 care users were identified in the databases, whose characteristics are presented in Table 1. They were mostly women (71.1%) and had a median year of birth of 1963 (range 1914-1997), i.e. aged about 44 years at study start. Patients with a complete care pathway were significantly older than patients having a shorter one (p<0.001). As a whole, 577 patients (57.7%) were under the LDD status for MS at the beginning of follow-up, with a median LDD time of 6.8 years (range 0.0-29.2), and 261 (26.1%) acquired the status on the 2007-2013 period. 278 (27.8%) patients had at least another LDD different from MS. The median follow-up duration was 6.8 years (range 0.0-7.0) out of a maximum of 7 years. As a whole, 71 (7.1%) deaths were observed after a median follow-up duration of 3.9 years (range 0.0-6.7). About half the population (53.9%) received at least one DMT and was treated during 63.0% of the follow-up duration in median.

Table 1 Population characteristics according to the length of care pathways (N=1000)

	Complete pathway n=648	Incomplete pathway n=352	Total N=1000	p-value[a]
Number of women (%)	473 (73.0%)	238 (67.6%)	711 (71.1%)	0.086
Birth Year *	1961 (1914-1991)	1966 (1924-1997)	1963 (1914-1997)	<0.001
Number of LDD for MS [b] (%)	589 (90.9%)	249 (70.7%)	838 (83.8%)	<0.001
Duration since MS-LDD's beginning [b,c]* (year)	12.7 (0.0-34.6)	6.5 (0.2-33.8)	11.6 (0.0-34.6)	<0.001
At least another LDD [b,d] (%)	186 (28.7%)	92 (26.1%)	278 (27.8%)	0.429
Follow-up duration * (year)	6.9 (0.0-7.0)	3.5 (0.0-6.0)	6.8 (0.0-7.0)	<0.001
Number of deaths observed (%)	55 (8.5%)	16 (4.5%)	71 (7.1%)	0.029
At least one DMT prescription [e] (%)	368 (56.8%)	171 (48.6%)	539 (53.9%)	0.015
Part of the follow-up duration under treatment (%) [f]*	61.6 (3.3-100.0)	66.0 (7.6-100.0)	63.0 (3.3-100.0)	0.218
GPs [g]*	6.1 (0.0-50.0)	4.3 (0.0-48.0)	5.6 (0.0-50.0)	<0.001
Neurologists [h]*	6 (0-48)	2 (0-19)	4 (0-48)	<0.001
At least one MS-related hospital admission [i] (%)	307 (47.3%)	166 (47.2%)	473 (47.3%)	1.000

* Median (minimum-maximum). [a] P-value of the comparison of complete and incomplete pathways' groups using either Kruskal-Wallis test, Pearson's chi-squared test or Fisher's exact test if needed. [b] LDD : Long Disease Duration. [c] Calculated at the date of last information (31/12/2013 or date of death). [d] LDD other than MS. [e] Disease Modifying Therapy. [f] For patients having at least one DMT use. [g] Annualized number of consultations and home visits taken together per patient. [h] Total number of consultations, home and imputed outpatient visits taken together per patient over the 2007-2013 period. [i] Total MS-related hospital admissions per patient over the 2007-2013 period (except monthly DMT injections).

3.2 Complete care pathways' clustering

The partition of the 648 complete care pathways led to a typology of five clusters as presented in Figure 1, which conducted to the best quality partition's parameters with ASWw, HC, HG and PBC respectively equal to 0.238, 0.102, 0.765 and 0.551.

Among the five clusters, two are dominating, i.e. clusters 1.5 and 1.4, with respectively 188 (29.0%) and 180 (27.8%) patients. These two groups corresponded to people having an overall high care consumption according to the mean duration in each state (Table 2). The group 1.3 characterized by a medium consumption is formed of 128 (19.8%) patients. At the opposite, people categorized in group 1.2 (n=124 (19.1%)) had a low care consumption and the 28 (4.3%) in the cluster 1.1 had no consumption during almost half the follow-up period.
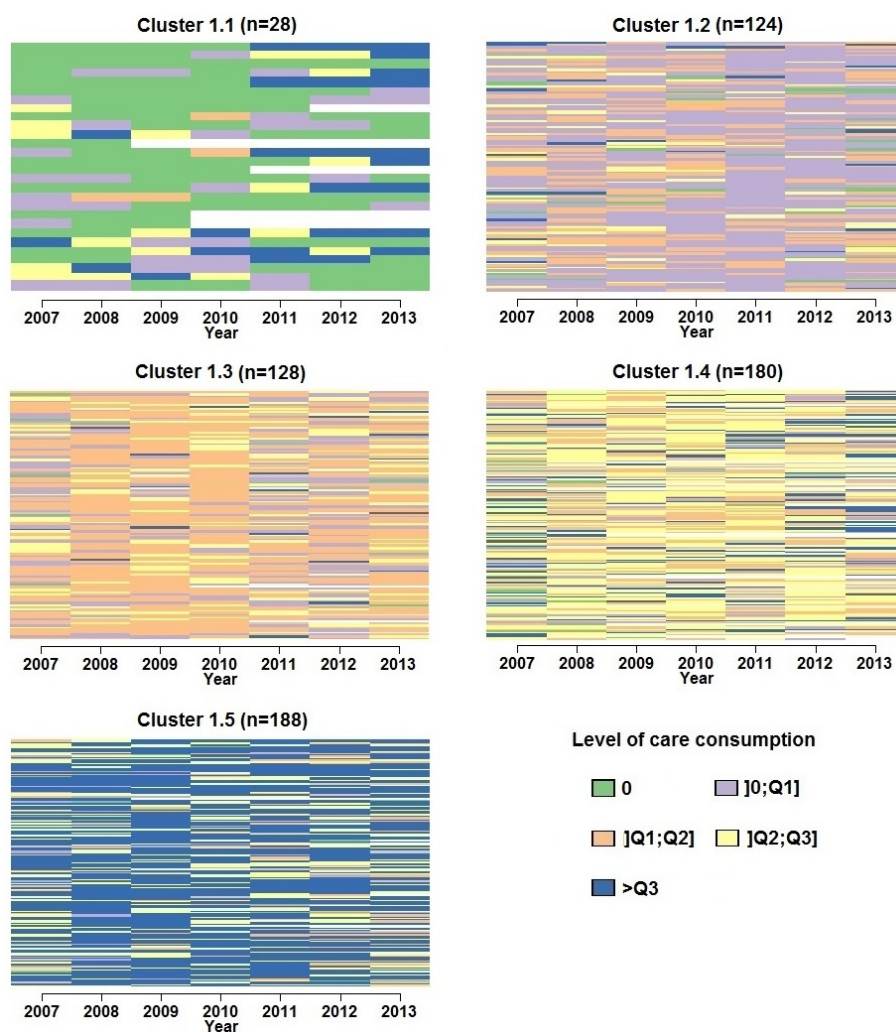
Fig. 1 Index plots of the groups obtained after the clustering procedure with indel costs fixed to 0.9 and transition-based substitution costs for complete care pathways (N=648)

Table 2 Mean duration in years (part of total duration) in each state according for complete care pathways' clustering (N=648)

State	"0"	"]0;Q1]"	"]Q1;Q2]"	"]Q2;Q3]"	">Q3"
Group 1.1	**3.46 (49.4%)**	1.18 (16.9%)	0.14 (2.0%)	0.64 (9.1%)	0.93 (13.3%)
Group 1.2	0.31 (4.4%)	**4.06 (58.0%)**	1.86 (26.6%)	0.55 (7.9%)	0.19 (2.7%)
Group 1.3	0.09 (1.3%)	1.39 (19.9%)	**4.07 (58.1%)**	1.20 (17.1%)	0.16 (2.3%)
Group 1.4	0.13 (1.9%)	0.57 (8.1%)	1.76 (25.1%)	**2.95 (42.1%)**	0.95 (13.6%)
Group 1.5	0.10 (1.4%)	0.24 (3.4%)	0.45 (6.4%)	1.37 (19.6%)	**4.71 (67.3%)**

The predominant state for each cluster is indicated in bold.

3.3 Incomplete care pathways' clustering

According to quality parameters, the 352 incomplete care pathways were partitioned into six clusters presented in Figure 2. Indeed, ASWw, HG and PBC were maximised respectively with 0.238, 0.731 and 0.487, and HC was equal to 0.139.

The different groups had similar consumption to those obtained with complete care pathways, plus a sixth cluster (group 2.6), representing 26.4% of incomplete sequences, concerning patients mostly entering the cohort after 2010 and having a quite low-medium consumption (Table 3). Group 2.5 was formed of 52 patients (14.9%) having an overall high consumption and cluster 2.4 composed of 55 patients (15.6%) having a medium-high consumption. The cluster 2.3 characterized by a medium consumption was formed of 68 patients (19.3%). At the opposite, people categorized in group 2.2 (n=49 (13.9%)) had a low care consumption and the 35 (9.9%) in the cluster 2.1 had no consumption during almost three quarters of the follow-up period.

Table 3 Mean duration in years (part of total duration including NA state) in each state according for incomplete care pathways' clustering (N=352)

State	"0"	"]0;Q1]"	"]Q1;Q2]"	"]Q2;Q3]"	">Q3"
Group 2.1	**3.34 (73.6%)**	1.03 (22.7%)	0.14 (3.1%)	0.00 (0.0%)	0.03 (0.7%)
Group 2.2	0.41 (8.4%)	**3.31 (67.6%)**	0.59 (12.0%)	0.43 (8.8%)	0.06 (1.2%)
Group 2.3	0.29 (6.0%)	1.09 (22.6%)	**2.34 (48.5%)**	0.94 (19.5%)	0.16 (3.3%)
Group 2.4	0.20 (4.3%)	0.44 (9.5%)	0.95 (20.5%)	**2.36 (50.9%)**	0.67 (14.4%)
Group 2.5	0.35 (7.5%)	0.37 (8.0%)	0.35 (7.5%)	0.77 (16.6%)	**2.77 (59.6%)**
Group 2.6	0.22 (8.6%)	**0.82 (32.0%)**	**0.85 (33.2%)**	0.28 (10.9%)	0.30 (11.7%)

The predominant state for each cluster is indicated in bold.

3.4 Characteristics of the final clustering

The different patients' characteristics and care consumption according to the two partitioning are presented in Tables 4 & 5.

As expected, the number of consultations and home visits with GPs and neurologists increased from cluster 1.1 to 1.5 (Table 4). In group 1.1, the follow-up duration was the shortest one, which was probably related to the higher proportion of deaths observed over the study period (p<0.001). Moreover, the number of patients having used at least one DMT was lower in this cluster (p=0.001). Patients belonging to group 1.2 were significantly younger (p<0.001) and were the most treated for a median treatment duration of about three-quarters of follow-up. Furthermore, this cluster was almost composed of half men and half women, compared
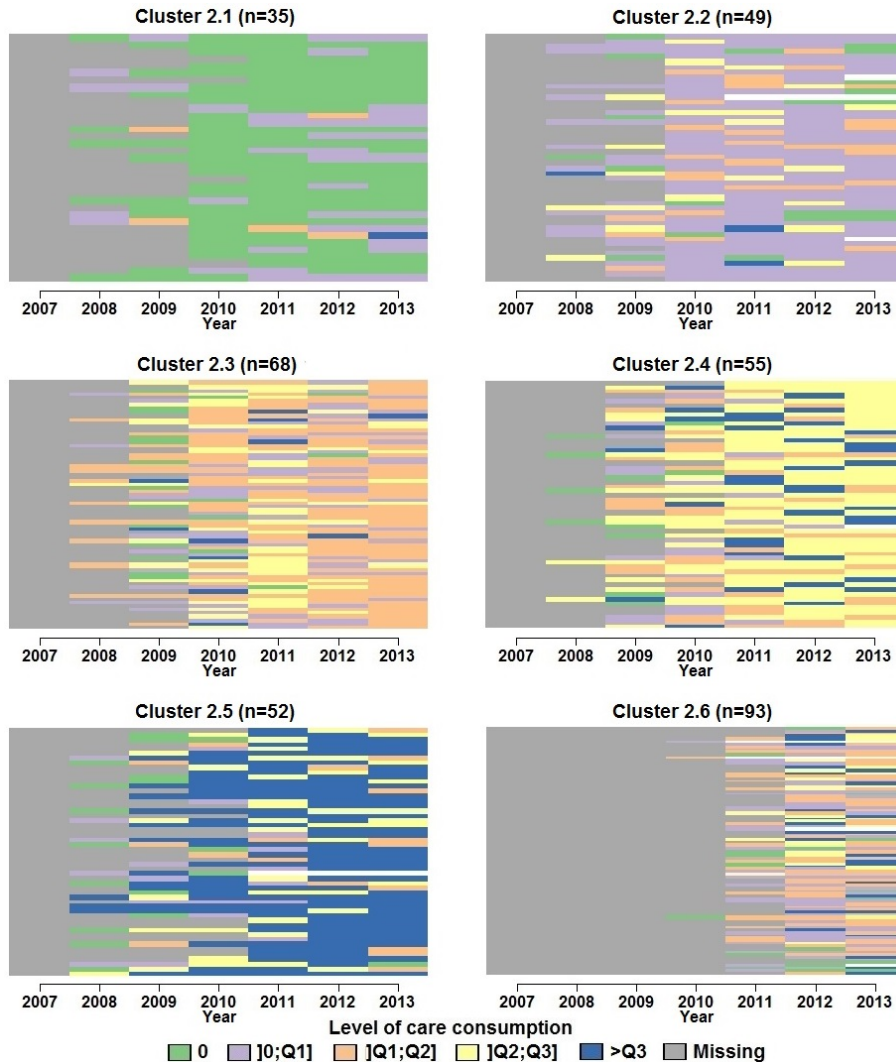
Fig. 2 Index plots of the groups obtained after the clustering procedure with indel costs fixed to 0.9 and transition-based substitution costs for incomplete care pathways (N=352)

to others mostly formed of women. The cluster 2.3 was also composed of patients using a DMT for the most part of follow-up duration. Group 1.4 was the oldest group, which could explain the percentage of deaths in this cluster. In group 1.5, i.e. having the highest consumption, patients seemed to be old with significantly more comorbidities, expressed as other LDD (p<0.001). These patients, who went more to hospital, were mostly treated.

As for the precedent typology, the care consumption with GPs and neurologists rose from cluster 2.1 to 2.5 in the second typology (Table 5). Patients in cluster 2.1 were the oldest and had the longest LDD duration compared to other groups, which could explain that they were significantly less treated (p<0.001). In groups 2.2 and 2.3, men and women were almost equally represented. Patients from cluster 2.2 were the youngest and those being the most treated. Like in the first typology, patients belonging to the highest consumption cluster 2.5 seemed to have more comorbidities than the others. Patients in cluster 2.6 were among the youngest ones with the most recent LDD status.

Table 4 Description of the complete care pathways' clustering ordered according to care consumption (N=648)

Cluster	1.1 n=28 (4.3%)	1.2 n=124 (19.1%)	1.3 n=128 (19.8%)	1.4 n=180 (27.8%)	1.5 n=188 (29.0%)	p-value[a]
Number of women (%)	23 (82.1%)	72 (58.1%)	92 (71.9%)	142 (78.9%)	144 (76.6%)	<0.001
Birth Year *	1958.5 (1930-1989)	1965.5 (1929-1991)	1960 (1923-1988)	1957.5 (1914-1986)	1959 (1916-1987)	<0.001
Number of LDD for MS [b] (%)	24 (85.7%)	110 (88.7%)	120 (93.8%)	167 (92.8%)	168 (89.4%)	0.373
Duration since MS-LDD's beginning [b,c]* (year)	12.4 (1.0-26.2)	12.6 (1.0-28.6)	13.0 (3.0-34.6)	12.1 (0.6-34.0)	12.8 (0.0-34.3)	0.629
At least another LDD [b,d] (%)	9 (32.1%)	19 (15.3%)	32 (25.0%)	48 (26.7%)	78 (41.5%)	<0.001
Follow-up duration * (year)	6.5 (0.0-7.0)	6.9 (0.0-7.0)	6.9 (0.0-7.0)	7.0 (0.0-7.0)	7.0 (0.0-7.0)	<0.001
Number of deaths observed (%)	6 (21.4%)	4 (3.2%)	4 (3.1%)	26 (14.4%)	15 (8.0%)	<0.001
At least one DMT prescription [e] (%)	6 (21.4%)	78 (62.9%)	74 (57.8%)	95 (52.8%)	115 (61.2%)	0.001
Part of the follow-up under treatment (%) [f]*	14.3 (3.6-59.0)	75.5 (3.3-100.0)	76.5 (4.4-100.0)	60.2 (3.3-100.0)	51.8 (4.4-100.0)	<0.001
GPs [g]*	2.7 (0.0-10.4)	2.6 (0.0-7.3)	4.6 (1.4-7.7)	6.9 (0.0-13.7)	12.7 (4.4-50.0)	<0.001
Neurologists [h]*	0 (0-8)	6 (0-23)	7 (0-35)	5 (0-38)	7 (0-48)	<0.001
At least one MS-related hospital admission [i] (%)	8 (28.6%)	61 (49.2%)	44 (34.3%)	84 (46.7%)	110 (58.5%)	<0.001

* Median (minimum-maximum). [a] P-value of the comparison of the five groups using either Kruskal-Wallis test, Pearson's chi-squared test or Fisher's exact test if needed. [b] LDD : Long Disease Duration. [c] Calculated at the date of last information (31/12/2013 or date of death). [d] LDD other than MS. [e] Disease Modifying Therapy. [f] For patients having at least one DMT use. [g] Annualized number of consultations and home visits taken together per patient. [h] Total number of consultations, home and imputed outpatient visits taken together per patient over the 2007-2013 period. [i] Total MS-related hospital admissions per patient over the 2007-2013 period (except monthly DMT injections).

Table 5 Description of the incomplete care pathways' clustering ordered according to care consumption (N=352)

Cluster	2.1 n=35 (10.0%)	2.2 n=49 (13.9%)	2.3 n=68 (19.3%)	2.4 n=55 (15.6%)	2.5 n=52 (14.8%)	2.6 n=93 (26.4%)	p-value[a]
Number of women (%)	23 (65.7%)	26 (53.1%)	39 (57.4%)	42 (76.4%)	38 (73.1%)	70 (75.3%)	0.022
Birth Year *	1960 (1930-1992)	1969 (1930-1997)	1966.5 (1931-1988)	1962 (1924-1989)	1964 (1924-1987)	1969 (1926-1994)	0.010
Number of LDD for MS[b] (%)	31 (88.6%)	32 (65.3%)	48 (70.6%)	39 (70.9%)	36 (69.2%)	63 (67.7%)	0.324
Duration since MS-LDD's beginning[b,c]* (year)	8.2 (0.2-33.0)	5.3 (0.3-14.7)	4.2 (1.5-33.2)	4.6 (1.6-30.8)	4.3 (2.0-27.0)	2.1 (1.0-33.8)	<0.001
At least another LDD[b,d] (%)	5 (14.3%)	10 (20.4%)	17 (25.0%)	17 (30.9%)	18 (34.6%)	25 (26.9%)	0.303
Follow-up duration * (year)	3.3 (0.0-5.6)	4.1 (0.0-5.9)	4.2 (0.0-5.9)	4.4 (0.0-5.9)	4.2 (0.0-6.0)	1.9 (0.0-3.0)	<0.001
Number of deaths observed (%)	0 (0.0%)	4 (8.2%)	2 (2.9%)	1 (1.8%)	3 (5.8%)	6 (6.5%)	0.399
At least one DMT prescription[e] (%)	4 (11.4%)	30 (61.2%)	35 (51.5%)	27 (49.1%)	27 (51.9%)	48 (51.6%)	<0.001
Part of the follow-up under treatment (%)[f]*	28.2 (15.4-44.4)	67.6 (15.8-100.0)	73.2 (18.4-98.3)	58.9 (10.2-97.8)	59.9 (7.6-98.1)	70.7 (11.2-95.6)	0.061
GPs[g]*	0.0 (0.0-4.5)	2.2 (0.0-6.3)	4.2 (1.2-7.5)	7.6 (2.4-12.7)	11.5 (2.8-48.0)	3.5 (0.0-24.0)	<0.001
Neurologists[h]*	0 (0-2)	4 (0-12)	4 (0-19)	4 (0-14)	4 (0-15)	2 (0-10)	<0.001
At least one MS-related hospital admission[i] (%)	10 (28.6%)	23 (46.9%)	33 (48.5%)	24 (43.6%)	34 (65.4%)	42 (45.2%)	0.031

* Median (minimum-maximum). [a] P-value of the comparison of the six groups using either Kruskal-Wallis test, Pearson's chi-squared test or Fisher's exact test if needed. [b] LDD : Long Disease Duration. [c] Calculated at the date of last information (31/12/2013 or date of death). [d] LDD other than MS. [e] Disease Modifying Therapy. [f] For patients having at least one DMT use. [g] Annualized number of consultations and home visits taken together per patient. [h] Total number of consultations, home and imputed outpatient visits taken together per patient over the 2007-2013 period. [i] Total MS-related hospital admissions per patient over the 2007-2013 period (except monthly DMT injections).

4 Discussion

The main aim of this pioneer study was to estimate the care consumption of the patients affected with MS over the 2007-2013 period in France. State sequence analysis using optimal matching was used to approach the concept of care pathway and build a typology of care consumption. Sequence analysis is usually employed in social sciences (transitions from education to work [Brzinsky-Fay, 2007, McVicar and Anyadike-Danes, 2002], career patterns [Anyadike-Danes and McVicar, 2005, Biemann and Datta, 2014], etc.) and genetics. Our research question on chronological events and type of data at our disposal offered the opportunity to apply this method in health field.

The study of care pathways reveals five groups of consumption for complete pathways and six for incomplete ones. In the two cases, clusters are mostly driven by a particular state dominating the sequences and show a gradation in care consumption. So, it seems that patients having a low care consumption are older, with a long LDD duration and less treated. Patients having a medium-low care consumption are younger and use DMTs, whereas those having a high consumption seems to be older and have more comorbidities than the others. It appears that the two partitions' clusters are similar on individual characteristics and care consumption. This result are observed despite the split realised on sequences' length and the last cluster of the second typology of people entering the cohort later than 2010, which could need to be studied separately. Therefore, it would be interesting to study more accurately the two partitions gathered together in order to have a five clusters' typology.

Our study has several advantages. Firstly, it is based on a representative random sample of the French population, reducing selection bias. Characteristics of our study population being close to those of MS patients in France [Foulon et al., 2015], it tends to consider that our population is quite representative of MS-affected people in France. Then, the two databases permit to access all reimbursed care consumption of patients, independently from self declaration and thus minimising memory bias. Moreover, it conducts to a care consumption as closely as possible to the true consumption of the study population (except concerning adherence to DMTs).

Concerning the statistical analysis, state sequence analysis is an innovative method in the field of care pathways to our knowledge. Furthermore, the choice to split and then analyse sequences according to their length avoids the creation of a cluster devoted to patients entering the cohort with delay. Similarly, the creation of a state devoted to the "death" status would have conducted to a cluster mainly formed of people dying during the follow-up.

However, some limitations can be mentioned. The first comes from the database itself, since people can get out from the EGB without any indication in the data if they leave one of the covered system, but this should concern only a very little number of people and not affect our results. Furthermore the database does not include outpatient visits with public neurologists, despite we know that MS expert

centres are mainly located in French university hospital. Therefore they were imputed, but the true consumption is certainly greater than the one assigned. Indeed, with the chosen scheme of imputation, we can't take into account adverse events at DMT initiation or follow-up care visits of untreated patients, but only the visits needed for yearly prescriptions. However these two limits can be overcome with the access to the French National Health Insurance Information System (*Système National d'Information Inter-Régimes de l'Assurance Maladie (SNIIRAM)*) from which our data were randomly selected.

Concerning statistics, only few individual characteristics were available to describe the partitions, since administrative databases do not contain clinical data. Although a discrepancy analysis and a regression model were considered, they were not presented since the pseudo-R^2 calculated were too small (lower than 0.05). We are currently developing algorithms to estimate relevant clinical parameters such as motor disability, based on the rentals and purchases of mobility aids and on spasticity's drugs deliveries, or adherence to treatments, using parameters described by Hess et al. [Hess et al., 2006]. Comorbidities will also be deeper explored, following works on Canadian databases and recommendations from Marrie et al. [Marrie et al., 2014, Marrie et al., 2016].

Another limit of this study comes from the weakness of the quality of the two partitions obtained, which can not permit to assess a strong structure in the data. However, this frailty can come from the fact that patients were not at the same stage of the disease during the 2007-2013 period and that care consumption may depend on several factors, such as MS-relapses or DMTs for example. Moreover, we worked on calendar years and had no information on MS-onset of patients, which do not permit to left-align sequences. To overcome this problem, we are trying to identify incident MS cases with an algorithm based on the work of Marrie et al. [Marrie et al., 2010] and thus study the impact of MS-onset on care consumption.

Finally, this study permits to describe for the first time the care consumption of MS-affected people in France over the 2007-2013 period with objective and quantitative information. The two typologies based on care pathways and obtained through sequence analysis are close to each other and reveal different ways of consumption among patients. This work, only constituting a preliminary draft, is going to be continued on the whole French population affected with MS, thanks to French National Health Insurance Information System. It will aim at examining more accurately care pathways, by analysing outpatient visits, biological exams, medical imaging exams, paramedical and other specialists care amongst others. In parallel, it will be designed to exploit administrative databases in order to have more relevant parameters explaining care consumption, such as level of disability or adherence to treatments.

References

[Anyadike-Danes and McVicar, 2005] Anyadike-Danes, M. and McVicar, D. (2005). You'll never walk alone: Childhood influences and male career path clusters. *Labour Economics*, 12(4):511 – 530. European Association of Labour Economists 16th Annual Conference, Universidade Nova de Lisboa, Lisbon, 9th – 11th Sepetember, 2004.

[Biemann and Datta, 2014] Biemann, T. and Datta, D. K. (2014). Analyzing sequence data: Optimal matching in management research. *Organizational Research Methods*, 17(1):51–76.

[Browne et al., 2014] Browne, P., Chandraratna, D., Angood, C., Tremlett, H., Baker, C., Taylor, B. V., and Thompson, A. J. (2014). Atlas of multiple sclerosis 2013: A growing global problem with widespread inequity. *Neurology*, 83(11):1022–1024.

[Brzinsky-Fay, 2007] Brzinsky-Fay, C. (2007). Lost in transition? labour market entry sequences of school leavers in europe. *European Sociological Review*, 23(4):409–422.

[Confavreux et al., 2003] Confavreux, C., Vukusic, S., and Adeleine, P. (2003). Early clinical predictors and progression of irreversible disability in multiple sclerosis: an amnesic process. *Brain*, 126(4):770–782.

[Confavreux et al., 2000] Confavreux, C., Vukusic, S., Moreau, T., and Adeleine, P. (2000). Relapses and progression of disability in multiple sclerosis. *N. Engl. J. Med.*, 343(20):1430–1438.

[Direction de la stratégie, des études et des statistiques (DSES), 2015] Direction de la stratégie, des études et des statistiques (DSES) (2015). Personnes prises en charge pour sclérose en plaques (sep) en 2013. Fiche pathologie, Caisse Nationale de l'Assurance Maladie des Travailleurs Salariés (CNAMTS). Update on July 22nd, 2015.

[Foulon et al., 2015] Foulon, S., Weill, A., Maura, G., Dalichampt, M., Debouverie, M., and Moreau, T. (2015). Prévalence de la sclérose en plaques en france en 2012 et mortalité associée en 2013 à partir des données du sniiram-pmsi. *Revue d'Épidémiologie et de Santé Publique*, 63, Supplement 1:S17 – S18. XXVIIIe Congrès national Émois, Nancy, 26 et 27 mars 2015.

[Fromont et al., 2014] Fromont, A., Lehanneur, M.-N., Rollot, F., Weill, A., Clerc, L., Bonithon-Kopp, C., Binquet, C., and Moreau, T. (2014). Coûts de la sclérose en plaques en france. *Revue Neurologique*, 170(6-7):432–439.

[Gabadinho et al., 2011] Gabadinho, A., Ritschard, G., Müller, N. S., and Studer, M. (2011). Analyzing and visualizing state sequences in r with traminer. *Journal of Statistical Software*, 40(4):1–37.

[Hennig and Liao, 2010] Hennig, C. and Liao, T. (2010). Comparing latent class and dissimilarity based clustering for mixed type variables with application to social stratification. Technical report, Department of Statistical Science, UCL, Department of Sociology, University of Illinois.

[Hess et al., 2006] Hess, L. M., Raebel, M. A., Conner, D. A., and Malone, D. C. (2006). Measurement of adherence in pharmacy administrative databases: a proposal for standard definitions and preferred measures. *The Annals of Pharmacotherapy*, 40(7-8):1280–88.

[Hubert and Arabie, 1985] Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of Classificationcation*, 2:193–218.

[Hubert and Levin, 1976] Hubert, L. and Levin, J. (1976). A general statistical framework for assesing categorical clustering in free recall. *Psychological Bulletin*, 83:1072–1080.

[Kaufman and Rousseeuw, 1990] Kaufman, L. and Rousseeuw, P. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley Series in Probability and Statistics. Wiley, New York.

[Lefeuvre et al., 2016] Lefeuvre, D., Rudant, J., Foulon, S., Weill, A., and Alla, F. (2016). Prise en charge des patients atteints de sclérose en plaques en france en 2013–consommations de soins et montants remboursés par l'assurance maladie. *Revue d'Épidémiologie et de Santé Publique*, 64, Supplement 1:S26 – S27.

[Leray et al., 2015] Leray, E., Vukusic, S., Debouverie, M., Clanet, M., Brochet, B., de Sèze, J., Zéphir, H., Defer, G., Lebrun-Frenay, C., Moreau, T., Clavelou, P., Pelletier, J., Berger, E., Cabre, P., Camdessanché, J. P., Kalson-Ray, S., Confavreux, C., and Edan, G. (2015). Excess mortality in patients with multiple sclerosis starts at 20 years from clinical onset: Data from a large-scale french observational study. *PLoS ONE*, 10(7):e0132033.

[Lesnard, 2010] Lesnard, L. (2010). Setting cost in optimal matching to uncover contemporaneous socio-temporal patterns. *Sociological Methods & Research*, 38(3):389–419.

[Maechler et al., 2015] Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., and Hornik, K. (2015). cluster: Cluster analysis basics and extensions. Technical report, R package version 2.0.3.

[Marrie et al., 2014] Marrie, R. A., Fisk, J. D., Stadnyk, K. J., Tremlett, H., Wolfson, C., Warren, S., Bhan, V., and Yu, B. N. (2014). Performance of administrative case definitions for comorbidity in multiple sclerosis in manitoba and nova scotia. *Chronic Diseases and Injuries in Canada*, 34(2-3):145–153.

[Marrie et al., 2016] Marrie, R. A., Miller, A., Sormani, M. P., Thompson, A., Waubant, E., Trojano, M., O'Connor, P., Fiest, K., Reider, N., Reingold, S., Cohen, J., and the attendees of the International Workshop on Comorbidity in Multiple Sclerosis, F. (2016). Recommendations for observational studies of comorbidity in multiple sclerosis. *Neurology*.

[Marrie et al., 2010] Marrie, R. A., Yu, N., Blanchard, J., Leung, S., and Elliott, L. (2010). The rising prevalence and changing age distribution of multiple sclerosis in manitoba. *Neurology*, 74(6):465–471.

[McVicar and Anyadike-Danes, 2002] McVicar, D. and Anyadike-Danes, M. (2002). Predicting successful and unsuccessful transitions from school to work by using sequence methods. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 165(2):317–334.

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

[Ministère de la Santé, de la Jeunesse, des Sports et de la Vie associative, 2007] Ministère de la Santé, de la Jeunesse, des Sports et de la Vie associative (2007). Plan 2007-2011 pour l'amélioration de la qualité de vie des personnes atteintes de maladies chroniques. Technical report, Ministère de la Santé, de la Jeunesse, des Sports et de la Vie associative.

[MS International Federation, 2015] MS International Federation (2015). Atlas of MS. http://www.msif.org/about-us/advocacy/atlas/. Last accessed on Jul 28, 2015.

[Organisation Mondiale de la Santé (OMS), 2002] Organisation Mondiale de la Santé (OMS) (2002). Des soins novateurs pour les affections chroniques : Eléments constitutifs : Rapport mondial. Technical report, OMS.

[R Core Team, 2015] R Core Team (2015). R: A language and environment for statistical computing. Technical report, R Core Team, Vienna, Austria.

[Studer, 2013] Studer, M. (2013). Weightedcluster library manual: A practical guide to creating typologies of trajectories in the social sciences with r. Technical report, LIVES Working Papers 24. DOI: 10.12682/lives.2296-1658.2013.24.

[Studer and Ritschard, 2015] Studer, M. and Ritschard, G. (2015). What matters in differences between life trajectories: a comparative review of sequence dissimilarity measures. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*.

[Trojano et al., 2012] Trojano, M., Lucchese, G., Graziano, G., Taylor, B. V., Simpson, Jr, S., Lepore, V., Grand'Maison, F., Duquette, P., Izquierdo, G., Grammond, P., Amato, M. P., Bergamaschi, R., Giuliani, G., Boz, C., Hupperts, R., Van Pesch, V., Lechner-Scott, J., Cristiano, E., Fiol, M., Oreja-Guevara, C., Saladino, M. L., Verheul, F., Slee, M., Paolicelli, D.and Tortorella, C., D'Onghia, M., Iaffaldano, P., Direnzo, V., Butzkueven, H., Group, M. S., and the New Zealand MS Prevalence Study Group (2012). Geographical variations in sex ratio trends over time in multiple sclerosis. *PLoS ONE*, 7(10):e48078.

[World Health Organization, 2015] World Health Organization (2015). ICD-10 Version:2016. Last accessed on Jun 22, 2015.

Session 5B: Markov II

Analysing Complex Life Sequence Data with Hidden Markov Modelling

Satu Helske, Jouni Helske, and Mervi Eerola

Abstract When analysing complex sequence data with multiple channels (dimensions) and long observation sequences, describing and visualizing the data can be a challenge. Hidden Markov models (HMMs) and their mixtures (MHMMs) offer a probabilistic model-based framework where the information in such data can be compressed into hidden states (general life stages) and clusters (general patterns in life courses).

We studied two different approaches to analysing clustered life sequence data with sequence analysis (SA) and hidden Markov modelling. In the first approach we used SA clusters as fixed and estimated HMMs separately for each group. In the second approach we treated SA clusters as suggestive and used them as a starting point for the estimation of MHMMs.

Even though the MHMM approach has advantages, we found it to be unfeasible in this type of complex setting. Instead, using separate HMMs for SA clusters was useful for finding and describing patterns in life courses.

1 Introduction

In social science applications, sequence analysis (SA) has gained more and more interest since its introduction in the mid-80s. It is now central to the life course perspective where it has been used to understand various trajectories and crucial transitions (Gauthier et al., 2014).

Satu Helske
University of Jyvaskyla, PO Box 35, FI-40014, University of Jyväskylä, Finland, e-mail: satu.helske@jyu.fi

Jouni Helske
University of Jyvaskyla, Finland

Mervi Eerola
University of Turku, Finland

1

Often the goal in SA is to find a typology of life sequences described as categorical time series data. Dissimilarities between each pair of sequences is determined using some criterion. Common choices have been optimal matching (McVicar and Anyadike-Danes, 2002) and Hamming distances (Hamming, 1950; Lesnard, 2010), but many modifications to these and also more fundamentally different methods have been developed (see, e.g., Aisenbrey and Fasang, 2010; Elzinga and Studer, 2014). Usually these dissimilarities are then grouped using cluster analysis such as Ward's agglomerative algorithm.

Life course data often consists of not only one sequence per subject, but multiple parallel sequences, one for each life domain of interest. We refer to *complex sequence data* for data which consist of multiple subjects and long multichannel (multidimensional) sequences.

One option for studying such data is to combine the sequences of each subject time point by time point by extending the state space of observations. This approach is simple if the number of possible combinations is moderate, but the combined state space grows rapidly as the number of domains and/or states grows. Multichannel sequence analysis (Gauthier et al., 2010) has been used for computing pairwise dissimilarities and finding clusters in complex sequence data (see, e.g., Eerola and Helske, 2016; Müller et al., 2012; Spallek et al., 2014). However, the dissimilarities are largely affected by the chosen dissimilarity metric and the cluster allocation may not be well suited to borderline cases. Also, describing, visualizing, and comparing such data is difficult. We use hidden Markov modelling for gaining a probabilistic descriptions of complex sequence data.

Hidden Markov models (HMMs) have been widely used in biological sequence analysis (Durbin et al., 1998) and speech recognition (Rabiner, 1989). Typically, the interest is in one long time series or another type of sequence. In social sciences this approach has been called latent Markov modelling. Typically, the data consists of a few measurements for multiple subjects.

Mixture hidden Markov model (MHMM) is a generalization of the HMM. There we assume that the data consists of latent subpopulations with different model structures. In the context of social sciences, the mixture hidden Markov model approach was formulated by van de Pol and Langeheine (1990) as the mixed Markov latent class model and later generalized to include time-constant and time-varying covariates by Vermunt et al. (2008) (who named the resulting model as the mixture latent Markov model, MLMM).

Multidimensional responses are included in the formulation of the MLMM but, to our knowledge, there are no empirical studies with complex life sequence data. Few studies use (M)HMMs for multichannel social science data. Helske and Helske (2016) have illustrated HMMs and MHMMs for multichannel data but do not conduct actual analyses with real data. Bartolucci et al. (2007) have studied criminal trajectories using HMMs with multiple binary sequences per subject. The data were large in the number of subjects (684 000 individuals), but sequences were short (6 age categories) and they had fixed groups (men and women) instead of latent clusters. Crayen et al. (2012) have used a hierarchical MLMM for two-channel categorical sequences to model dynamics of mood regulation of university students

during one week. The sequences were longer (56 time points) but the number of subjects is moderate (164) and they used only three states in both channels. In their hierarchical model there were two parallel latent structures; one between the days and the other within the days.

We study two approaches to analysing complex sequence data. The first is to use sequence analysis and cluster analysis for finding a few sets of clusters and then, separately for each cluster, to estimate an HMM. In this approach, hidden Markov modelling is used to compress and describe life course information within the clusters and to help choosing the number of clusters.

The second approach is to estimate a mixture model. Now the clustering is not fixed but we get a probability of each individual belonging to each cluster. For large data, estimating the MHMM with the maximum likelihood can be a complex and time-consuming task unless the set of candidate models is restricted. We study the option of using SA clusters and simple HMMs as a starting point for mixture modelling.

2 Interpretation of hidden Markov models for life sequences

One rationale behind using the HMM approach for life sequence analysis was the attempt to identify similar life course patterns based on similar hidden state trajectories. The similarity of hidden state sequences can be attributed to both external factors, which are common to groups of populations, or to internal behavioural similarities between individuals with similar features. Finding hidden dynamics is thus important for analysing and grouping life courses and also for understanding relationships between factors that are measured. The significance of hidden states in life sequence data is dependent on the chosen structure of the model. The goals of our analysis were two-folded:

1. to group individuals with similar life course patterns (clusters) and
2. to compress information in observed states across life domains to capture patterns and dynamics within a group (hidden states)

The aim was to find hidden states that compress the information across several life domains into more general life stages. These life stages could be either stable episodes between two transitions (e.g., employed and married without children) or characterized by transitions in some of the life domains (e.g., moving between unemployment and short-term jobs). We restricted to left-to-right models where transitions back to previous hidden states are not possible. Such representation makes it easier to comprehend the overall dynamics within a group and is also natural from the life course perspective: even though individuals may be in similar states at different times, the second time has a different history compared to the first time. E.g., there could be a group where, at some points of their lives, individuals are married with children, then divorced for a while, and later again married with children (but with the history of having experienced a divorce).

3 Data

We illustrate the analysis of complex life sequence data using a subsample of the German National Educational Panel Survey (NEPS) (Blossfeld et al., 2011).

We restricted to life courses of an age cohort born in 1955–1959. Only individuals who were born in Germany or moved there before age 14 were included.

The data consisted of monthly life statuses of 1731 individuals in three life domains (career, partnerships, and parenthood) from age 15 to age 50. For each individual, there were three parallel sequences of length 434, which made altogether 2,253,762 data points. Using the monthly time scale allowed for detecting also smaller fluctuations in life courses, e.g. recurrent transitions between unemployment and employment.

3.1 Sequences

The sequences in three life domains were constructed as follows:

Career with 4 states:

- Studying (in school, vocational training, or vocational preparation)
- Employed (full-time or part-time)
- Unemployed
- Else (parental leave, military or non-military service, voluntary work, or other gap in employment history)

Partnerships with 4 states:

- Single (never lived with a partner)
- Cohabiting
- Married/in a registered partnership
- Divorced/separated/widowed

Parenthood with 2 states:

- No children
- Has (had) children (biological, adopted, or foster children)

The coding for parenthood was very simple. A practical reason was that this record was available for most individuals, whereas more detailed information was often missing. On the other hand, we can argue that specifically the experience of becoming a parent is relevant as one step in the developmental process into adulthood.

For the latter two life domains, the status of each month was usually determined from the latest event. An exception was made for the rare partnerships that lasted for less than a month; there separation was coded from the following month onward. In a case of multiple records per month in the career domain, the final status was

Analysing Complex Life Sequence Data with Hidden Markov Modelling   5

given according to assumed importance: school and vocational training came before employment, which in turn dominated over vocational preparation, unemployment, and other non-employment statuses.

Altogether 306 individuals (17.7%) had some missing information in one or two life domains. Thus, at each time point we have at least some information from each individual.

4 Hidden Markov models

In the context of hidden Markov models, observed states are determined via a Markov process of hidden states. These hidden states cannot be observed directly, but only through the sequence(s) of observations, since hidden states generate ("emit") observations on varying probabilities.

Assume we have multichannel sequence data for N individuals with C parallel sequences of length T. Naturally, the following applies for single-channel data (subjects with one sequence only) by setting $C = 1$. Let us denote the observation in channel c, $c = 1, \ldots, C$, of individual i, $i = 1, \ldots, N$, at time t, $t = 1, \ldots, T$, with y_{itc} and the corresponding hidden state with z_{it}. A discrete first order hidden Markov model $\mathcal{M}$ is characterized by the following parameters:

- Initial probability of hidden state s:

$$\pi_s = P(z_{i1} = s); \; s \in \{1, \ldots, S\}, \text{for all } i = 1, \ldots, N.$$

- Transition probability from hidden state s to hidden state r:

$$a_{sr} = P(z_{it} = r | z_{i(t-1)} = s); \; s, r \in \{1, \ldots, S\}, \text{for all } i = 1, \ldots, N.$$

- Emission probability of observed state m_c in channel c given the hidden state s:

$$b_s(m_c) = P(y_{itc} = m_c | z_{it} = s); \; s \in \{1, \ldots, S\}, \; m_c \in \{1, \ldots, M_c\},$$
$$\text{for all } i = 1, \ldots, N. \quad (1)$$

The (first order) Markov assumption states that the hidden state transition probability at time t only depends on the hidden state at the previous time point $t - 1$:

$$P(z_{it} | z_{i(t-1)}, \ldots, z_{i1}) = P(z_{it} | z_{i(t-1)}). \quad (2)$$

Also, the observed states at time t are independent of all other observations and hidden states given the hidden state at t. For multichannel sequence data, we assume the same latent structure applies for all channels, i.e., the hidden state at time t for individual i generates the observed state y_{itc} in all channels c. Observations $y_{it1}, \ldots, y_{itC}$ are assumed independent of each other given the hidden state z_{it}, i.e.,

$P(\mathbf{y}_{it}|z_{it}) = P(y_{it1}|z_{it})\cdots P(y_{itC}|z_{it})$. Fig. 1 illustrates an HMM with a hidden state sequence and two channels.
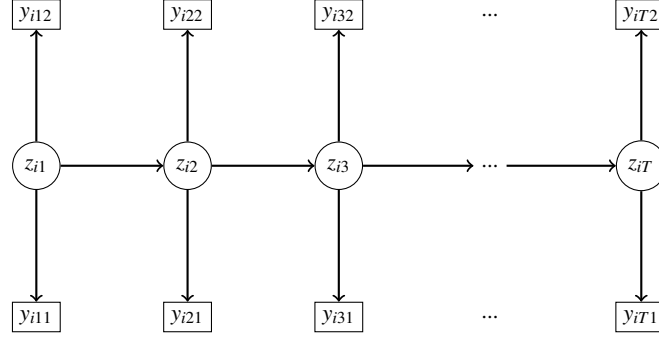
Fig. 1 Illustration of the hidden Markov model structure for two-channel sequence data for individual i with hidden states $z_{i1}\ldots,z_{iT}$ and observed states $y_{i1c}\ldots,y_{iTc}, c = 1,2$.

The log-likelihood for the HMM is written as

$$\log L = \sum_{i=1}^{N} \log P(Y_i|\mathcal{M}),\tag{3}$$

where Y_i are the observed sequences in channels $1,\ldots,C$ for subject i and $\mathcal{M}$ describes the model and its parameters $\{\pi, A, B_1, \ldots, B_C\}$, where $A = \{a_{sr}\}$ is a matrix of transition probabilities and $B_c = \{b_s(m_c)\}$ is a matrix of emission probabilities for channel c. The probability of observation sequences for subject i given the model is

$$
\begin{aligned}
P(Y_i|\mathcal{M}) &= \sum_{\text{all } z} P(Y_i|z,\mathcal{M}) P(z|\mathcal{M}) \\
&= \sum_{\text{all } z} P(z_1|\mathcal{M})P(\mathbf{y}_{i1}|z_1,\mathcal{M})\prod_{t=2}^{T} P(z_t|z_{t-1},\mathcal{M})P(\mathbf{y}_{it}|z_t,\mathcal{M}) \\
&= \sum_{\text{all } z} \pi_{z_1} b_{z_1}(y_{i11})\cdots b_{z_1}(y_{i1C})\prod_{t=2}^{T}\left[a_{z_{t-1}z_t} b_{z_t}(y_{it1})\cdots b_{z_t}(y_{itC})\right],
\end{aligned}\tag{4}
$$

where the hidden state sequences $z = (z_1,\ldots,z_T)$ take all possible combinations of values in the hidden state space $\{1,\ldots,S\}$ and where $\mathbf{y}_{it}$ are the observations of subject i at t in channels $1,\ldots,C$; π_{z_1} is the initial probability of the hidden state at time $t = 1$ in sequence z; $a_{z_{t-1}z_t}$ is the transition probability from the hidden state at time $t-1$ to the hidden state at t; and $b_{z_t}(y_{itc})$ is the probability that the hidden state of subject i at time t emits the observed state at t in channel c.

Analysing Complex Life Sequence Data with Hidden Markov Modelling   7

4.1 Mixture hidden Markov model

The mixture hidden Markov model is, by definition, a mixture of simple hidden Markov models. We assume that the population consists of subpopulations of individuals (latent classes or clusters) with different life patterns. Respectively, the mixture model consists of varying submodels that characterize the clusters. Transitions from one cluster to another are not allowed.

Assume that we have a set of HMMs $\mathscr{M} = \{\mathscr{M}^1, \ldots, \mathscr{M}^K\}$, where $\mathscr{M}^k = \{\pi^k, A^k, B_1^k, \ldots, B_C^k\}$ for clusters $k = 1, \ldots, K$. We denote $P(\mathscr{M}^k) = w_k$ as the prior probability that an arbitrary observation sequence is generated by the submodel $\mathscr{M}^k$ such that $\sum_{k=1}^{K} w_k = 1$.

The log-likelihood of the MHMM is of the form

$$
\begin{aligned}
\log L &= \sum_{i=1}^{N} \log P(Y_i|\mathscr{M}) \\
&= \sum_{i=1}^{N} \log \left[\sum_{k=1}^{K} P(\mathscr{M}^k) \sum_{\text{all } z} P\left(Y_i|z, \mathscr{M}^k\right) P\left(z|\mathscr{M}^k\right) \right] \\
&= \sum_{i=1}^{N} \log \left[\sum_{k=1}^{K} w_k \sum_{\text{all } z} \pi_{z_1}^k b_{z_1}^k(y_{i11}) \cdots b_{z_1}^k(y_{i1C}) \prod_{t=2}^{T} \left[a_{z_{t-1}z_t}^k b_{z_t}^k(y_{it1}) \cdots b_{z_t}^k(y_{itC}) \right] \right].
\end{aligned}
$$
(5)

For more detailed description of MHMMs, see Helske and Helske (2016) or Vermunt et al. (2008).

4.2 Model estimation

The log-likelihoods of (4) and (5) are efficiently calculated with the *forward–backward algorithm* (Baum and Petrie, 1966; Rabiner, 1989). A common maximum likelihood estimation method is the Baum–Welch algorithm, i.e., the expectation–maximization (EM) algorithm in the HMM context.

The Baum–Welch algorithm requires starting values for model parameters. In order to reduce the risk of being trapped in a poor local optimum, a large number of initial values should be tested. Simpler models with few parameters are fast to estimate; therefore, it is possible to fit the model numerous times with varying random starting values for finding the model with the best likelihood. When the model is large, estimation is more time-consuming and good starting values for model parameters are useful or even essential.

The most probable path of hidden states for each subject given their observations and the model can be computed using the *Viterbi algorithm* (see, e.g., Rabiner, 1989). This path maximizes the probability of $P(z|Y_i, \mathscr{M})$.

The forward–backward algorithm can also be used for computing posterior cluster probabilities (the probability that subject i belongs to a certain cluster) for MHMMs. These can be used for classifying subjects into different groups.

4.3 Model comparison

Models with the same number of parameters can be compared with the value of the log-likelihood function. For choosing between models with a different number of hidden states, we need to take account of the number of parameters.

Bayesian information criterion (BIC) is the usual criterion for comparing (M)HMMs. We define it as

$$BIC = -2\log(L) + p\log\left(\sum_{i=1}^{N}\sum_{t=1}^{T}\frac{1}{C}\sum_{c=1}^{C} \mathrm{I}(y_{itc} \text{ observed})\right), \qquad (6)$$

where L is given in equation 3, p is the number of estimated parameters, I is the indicator function, and the summation in the logarithm is the size of the data. If data are completely observed, the summation is simplified to $N \times T$. The smaller the BIC, the better the model.

When computing the log-likelihood for the combined model with fixed SA clusters we simply sum the log-likelihoods of the cluster-wise HMMs. BIC of the combined model is determined as

$$BIC = -2 \times \sum_{k=1}^{K}\log(L_k) + \sum_{k=1}^{K} p_k\log\left(\sum_{i=1}^{N}\sum_{t=1}^{T}\frac{1}{C}\sum_{c=1}^{C} \mathrm{I}(y_{itc} \text{ observed})\right), \qquad (7)$$

where L_k is the likelihood of the HMM of cluster k, p_k is the number of estimated parameters in the HMM for cluster k, and the summation in the logarithm is the size of the full data set.

5 Visualizing sequence data and models

Visualization is an important tool throughout the analysis process from the first glimpses into the data to presenting the results. As an example, we consider the data and the HMM for one of the preliminary clusters described "Long education and later family" (from the ten-cluster solution).

Fig. 2 illustrates a five-state HMM with the following life stages:

1. Single and (mostly) studying
2. Cohabiting, separated, or divorced; studying or employed
3. Married, studying or employed
4. Married with children, non-employed

Analysing Complex Life Sequence Data with Hidden Markov Modelling   9
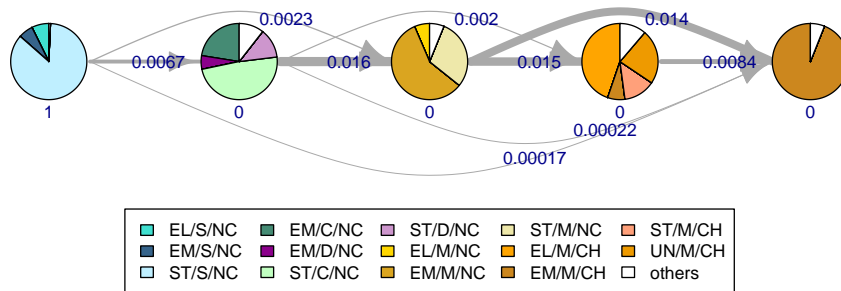
Fig. 2 Illustrating the hidden Markov model for the cluster of individuals with long education and later family. Pies present five hidden states, with slices showing the emission probabilities of combinations of observed states. States with emission probability less than 0.05 are combined into one slice for easier interpretation. The edges show the transtion probabilities – the thicker the edge, the higher the probability. Initial probabilities of the hidden states are given below the pies. The descriptions of the combined states show career/partnership/parenthood statuses: ST=studying, EM=employed, UN=unemployed, EL=else; S=single, C=cohabiting, M=married, D=divorced/separated; NC=no children, CH=has child(ren).

5. Married with children, employed

The hidden states are described by the most probable emitted observations, but there are also less probable states that are omitted from the plot for readability. E.g., the second state also emits marriages with a small probability—from the most probable hidden state paths in Fig. 3 we can see that these are marriages which end in divorce relatively fast. We could interpret that the second hidden state describes a life stage of searching for a partner before forming a long-lasting marriage.

All subjects start from the first state at age 15. At the start of the follow-up they are all single and mostly studying. The most common transition is to the second state, but the third state is quite probable also. Due to the monthly data, the transition probabilities are small—individuals usually spend years in each state.

Most individuals move to the third hidden state which describes childless marriage. It is the hidden state where individuals spend the least time on average. Transitions to the fourth and the fifth hidden state are almost as common. These both describe parenthood; some move out of workforce for a while or until the end of the follow-up, while some continue working.

6 Analysis

Estimating a large MHMM for complex sequence data can be difficult and time-consuming unless the structure of the model is fixed or known, even approxim-
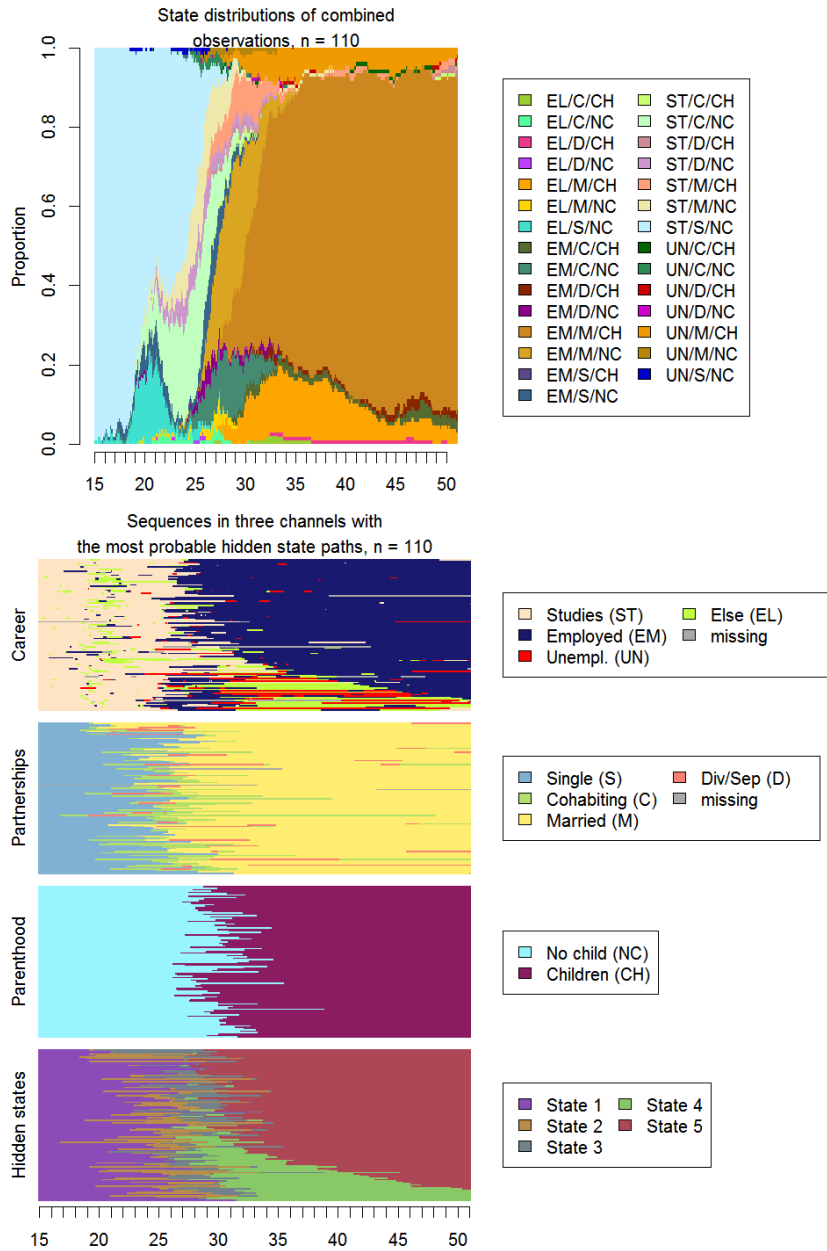
Fig. 3 State distributions of combined observations (top) and sequences of observations in each channel as well as the most probable paths of hidden states (bottom). Sequences are ordered by multidimensional scaling scores. States 1–5 correspond to the hidden states presented in Fig. 1.

Analysing Complex Life Sequence Data with Hidden Markov Modelling   11

ately. In other cases, the set of candidate models must be somehow restricted. In this case we had little prior knowledge on the structure of the model; hence, how many clusters to choose and how many hidden states to include in each cluster? As transitions were frequent in some of the trajectories and infrequent in others, it was clear that some of the clusters should contain more hidden states than others, leading to an unfeasible large number of possible model structures.

We compared two different approaches for the analysis of complex sequence data, of which both were conducted in a stepwise manner. The first two steps applied for both approaches, whereas step 3 was different (denoted as 3a and 3b). More detailed descriptions of the analysis process are given in the following sections.

1. **Sequence analysis**. Computing the dissimilarities between the subjects with the Hamming distance. Using Ward's hierarchical method for clustering individuals with similar life courses. Choosing a set of reasonable clustering solutions for preliminary analysis.
2. **Hidden Markov models.** Separately for each SA cluster, fitting simple HMMs with a different number of hidden states. Choosing the best model for each preliminary cluster.
3a. **Combined HMMs.** Constructing a combined model from separate HMMs (from step 2), keeping parameters fixed. Computing the likelihood and BIC for combined models with 7–12 clusters for determining the number of clusters. Computing the most probable path of hidden states for each individual.
3b. **Mixture hidden Markov models.** For each clustering solution (7–12 clusters), estimating an MHMM by using parameters of the corresponding HMMs (from step 2) as starting values. Computing the likelihood and BIC of the MHMMs for determining the number of clusters. Computing the most probable path of hidden states for each individual.

6.1 Step 1: Sequence analysis and preliminary clustering

We started by applying multichannel sequence analysis and computed the dissimilarities between the sequences. These were then used in cluster analysis.

6.1.1 Sequence dissimilarities

We compared a few dissimilarity metrics that are suitable for multichannel data: optimal matching (OM), generalized Hamming distance (HAM), and dynamic Hamming distance (DHD) (Lesnard, 2010). We chose the generalized Hamming distance with theory-driven substitution costs (see Table 1). The metric compares observed states time point by time point and gives a cost for mismatches. It generally works relatively well in a problem where timing is important and also here resulted in meaningful clusters with high goodness-of-fit (see Sect. 6.1.2).

Table 1 Substitution costs for Hamming distances.

Career status →	ST →	EM →	UN →	EL →	*
Studying (S) →	0	3	2	1	0
Employed (EM) →	3	0	2	2	0
Unemployed (UN) →	2	2	0	1	0
Else (EL) →	1	2	1	0	0
Missing (*) →	0	0	0	0	0

Partnership status →	S →	C →	M →	D →	*
Single (S) →	0	2	2	3	0
Cohabiting (C) →	2	0	1	2	0
Married (M) →	2	1	0	2	0
Divorced/sep. (D) →	3	2	2	0	0
Missing (*) →	0	0	0	0	0

Parenthood status →	NC →	CH →	*
No children (NC) →	0	3	0
Has children (CH) →	3	0	0
Missing (*) →	0	0	0

6.1.2 Cluster analysis

Ward's method was chosen for clustering since it typically produces usable and relatively even-sized clusters compared to most of the other clustering methods (Aassve et al., 2007; Helske et al., 2015). We chose six clustering solutions with 7–12 clusters for further examination. The choice was based on the dendrogram and interpretability of the clusters. Ward's method is agglomerative, so when two smaller clusters are merged, all other clusters remain the same. This means that within the six sets of clustering results there were only $7 + 2 + 2 + 2 + 2 + 2 = 17$ distinct clusters (see Fig. 4 for an illustration).

Table 2 shows the goodness-of-fit statistics for different clustering results and dissimilarity metrics, as measured by the proportion of the variation explained by the clusters (pseudo coefficient of determination (R^2); see Studer et al., 2011). Here, generalized Hamming distances resulted in meaningful clusters with a relatively high goodness-of-fit. OM resulted in clusters with as high goodness-of-fit while DHD resulted in somewhat lower values of R^2 (though not by much). OM clusters were similar to HAM clusters in many ways but had more variation in the timings of first transitions into employment, partnerships, and parenthood.

Analysing Complex Life Sequence Data with Hidden Markov Modelling 13
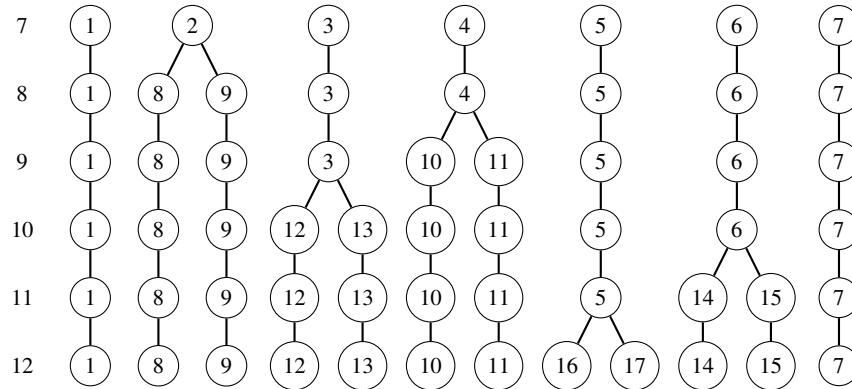
Clusters

Fig. 4 Clustering structure for Ward's agglomerative method shown for six sets of clustering results with 7–12 clusters.

Table 2 Proportion of variation covered by 7–12 clusters. Clustering was based on different dissimilarity metrics; generalized Hamming distance (HAM), optimal matching (OM), and dynamic Hamming distance (DHD)).

Clusters	HAM	OM	DHD
7	0.38	0.38	0.35
8	0.40	0.40	0.37
9	0.42	0.42	0.38
10	0.43	0.43	0.40
11	0.44	0.44	0.41
12	0.44	0.45	0.42

6.2 Step 2: Simple hidden Markov models for clusters

At the next step, we estimated five HMMs with 4–9 hidden states separately for each of the 16 clusters—fewer hidden states for simpler clusters, more for more complex ones. Since the goal was to find life stages between adolescence and middle age, having too few or too many hidden states was not plausible nor interpretational.

6.2.1 Model estimation

We set starting values for parameters by determining candidate hidden states from observed data and re-estimated the model numerous times by altering these values as follows. At first, we estimated the model 10,000 times with a large variation in starting values. For each re-estimation step we added noise from the $N(0,0.3^2)$ distribution to the the original starting values (with proper scaling and correction of

signs). The aim of this estimation was to broadly explore the parameter space and to get closer to the global maximum.

To make sure that we were at or near the global optimum, we re-estimated the model by using the model with the highest likelihood as a stating point, now adding noise from the $N(0, 0.15^2)$ distribution. If the model with the highest likelihood was found only a few times, similar estimation was repeated (again using the best model as the new starting point) in order to be fairly certain to have found the global optimum. For clusters with fewer members and models with fewer hidden states, the first estimation step was often enough for finding the (assumed) global maximum.

6.2.2 Model comparison

For each cluster, the HMMs with a different number of hidden states were compared to find the best model to use in the mixture models. BIC and other information criteria are common choices for comparison of HMMs with different numbers of hidden states. Another common option for model selection is cross-validation.

We chose to use BIC as it generally selects parsimonious models. BIC has been proven consistent for ergodic stationary HMMs (Whiting and Pickett, 1988), but not to left-to-right HMMs. Here, also BIC consistently chose models with more hidden states and clusters than is interpretational or plausible.

A likely reason for poor performance of information criteria in this problem was that we were comparing models which all were considerably simple compared to the complexity of real life. The goal was to simplify and describe the overall patterns and dynamics in life trajectories, not to find data-generating models.

However, we did use BIC as one source of information for choosing the number of hidden states by looking for turning points in BIC after which additional hidden states were not as profitable. In addition to BIC, the choice of the number of hidden states was based on interpretability of the model and the prevalence of an additional hidden state in the most probable hidden state paths—if a hidden state was "visited" only rarely it was regarded as unnecessary.

6.3 Step 3 a: Combined HMMs

At this step we used the separate cluster-specific HMMs to construct combined models with 7–12 clusters. For each combined model, we computed the likelihood and BIC to determine the best number of clusters.

The combined model with the smallest BIC was used for determining the best number of clusters. Given the best clustering, we computed the most probable paths of hidden states for each individual.

Analysing Complex Life Sequence Data with Hidden Markov Modelling   15

6.4 Step 3 b: Mixture hidden Markov models

At this step we constructed six MHMMs with 7–12 clusters. We used the estimated parameters of respective cluster-wise HMMs as starting values for mixture models. To avoid non-structural zeros in starting values, we added a small amount of 0.001 to each starting value (with proper scaling). We estimated models in a similar manner to the previous step, by using randomized starting values—first with a larger noise and, after getting closer to the optimum, again with a smaller noise.

6.5 Software

Analyses were conducted with the R software (R Core Team, 2015) by using packages TraMineR (Gabadinho et al., 2011) for sequence analysis, cluster (Maechler et al., 2015) for cluster analysis, and seqHMM (Helske and Helske, 2016) for hidden Markov modelling.

7 Results

The number of hidden states per cluster varied between six and eight. We applied both the combined model and the mixture model approach for describing data and determining the best number of clusters.

7.1 Combined model approach

Table 3 shows the BICs for models with 7–12 clusters. The model with eight clusters resulted in smallest BIC (even the highest likelihood) and was chosen as the best model. The model with seven clusters was almost as good; the only difference was that the two childless clusters (see Figures xx and xxx) were combined into one.

Table 3 Number of parameters, log-likelihood, and BIC for combined models with 7–12 clusters. The smallest value of BIC is shown in bold.

Clusters	Parameters	Log-likelihood	BIC
7	533	−369075.7	745059.4
8	**595**	**−364825.9**	**743368.2**
9	643	−370746.2	755208.7
10	705	−368985.0	751686.5
11	767	−368977.5	751671.5
12	800	−373550.3	760817.0

Fig. 5 and Fig. 6 illustrate the HMM structure for each of the eight clusters. More detailed visualizations with observed sequences and most probable hidden state paths are shown in the Appendix.

The clusters were well separated from each other by the timing and occurrence of career and family states. The two largest clusters were characterized by (mostly) short education and family. They differed in the timing of partnership and parenthood transitions which occurred either earlier in life (cluster A with 461 members of which 59% were females) or later (cluster B, 403 members, 54% males) The third largest cluster (cluster C, 266 members, 68% males) mostly consisted of individuals with long education and later family. Another cluster with early family transitions (cluster D, 159 members, 96% females) was characterized with a long career break for mostly taking care of children.

Two clusters were characterized by no or very late parenthood. They differed in timing of the partnerships; the larger cluster (cluster E, 177 members, 51% males) had earlier first partnerships while in the smaller cluster (cluster F, 116 members, 59% males) partnerships were delayed or omitted altogether.

The two smallest clusters consisted of single parents (cluster G, 47 individuals, 72% females) or parents living divorced or separated (cluster H, 102 individuals, 61% females).

7.2 Mixture model approach

The estimation of ordinary HMMs can be challenging due to multiple local optima in likelihood surfaces, since typical parameter estimation algorithms often only find these suboptimal solutions. Therefore, multiple starting values for the estimation are needed to ensure that the global optimum is found. The same problem is even more prevalent in complex MHMM settings with a large amount of parameters and mixture components. In addition, when the structure of the model (the number of mixture components and/or hidden states) is unknown, the amount of required computing resources naturally multiplies.

Therefore, even after using allegedly reasonable starting values (from simple HMMs), parallel computation, and extensive computing resources, we were not able reach satisfactory results. With different starting values the estimation always resulted in a different solution, so finding the global optimum would have required an unfeasible amount of computing time and/or resources.

Even though we were not able to find optimal MHMMs, we did study some of the suboptimal solutions. To study the differences of SA and MHMM clusters, we estimated a mixture model by keeping the initial, transition, and emission parameters of the submodels fixed (i.e., estimating only prior cluster probabilities, later referred to as the "non-estimated MHMM"). This approach was similar to the combined model approach, but instead of keeping the cluster memberships fixed we allowed individuals to switch clusters. Each individual was assigned to the cluster with the highest posterior cluster probability given their observed sequences.

Analysing Complex Life Sequence Data with Hidden Markov Modelling 17

Short education & early family, n = 461

State 1 | State 2 | State 3 | State 4 | State 5 | State 6 | State 7

☐ ST/S/NC	■ EM/S/NC	☐ ST/C/NC	■ EM/C/CH	■ ST/M/CH
■ EL/S/NC	■ EM/C/NC	☐ ST/M/NC	■ EM/D/CH	■ UN/M/CH
■ EM/S/CH	■ EM/M/NC	■ EM/M/CH	■ EL/M/CH	☐ others

Short education & later family, n = 403

State 1 | State 2 | State 3 | State 4 | State 5 | State 6 | State 7

☐ ST/S/NC	■ EM/C/NC	■ EM/M/NC	■ EM/M/CH	☐ others
■ EL/S/NC	■ EM/D/NC	■ EM/C/CH	■ UN/M/CH	
■ EM/S/NC	☐ ST/C/NC	■ EM/D/CH	■ EL/M/CH	

Long education & later family, n = 266

State 1 | State 2 | State 3 | State 4 | State 5 | State 6 | State 7

■ EL/S/NC	■ EM/C/NC	■ ST/D/NC	■ EL/M/CH	■ UN/M/CH	☐ others
☐ ST/S/NC	■ EM/D/NC	■ EM/M/NC	■ EM/M/CH	■ EM/C/CH	
■ EM/S/NC	☐ ST/C/NC	■ ST/M/NC	■ ST/M/CH	■ EM/D/CH	

Long career break & early family, n = 159

State 1 | State 2 | State 3 | State 4 | State 5 | State 6 | State 7 | State 8

☐ ST/S/NC	■ EM/C/NC	■ EL/C/CH	■ EM/M/CH	■ EM/D/CH
■ EM/S/NC	■ EM/M/NC	■ EL/M/CH	■ ST/M/CH	■ ST/D/CH
■ EL/M/NC	☐ ST/M/NC	■ EM/C/CH	■ UN/M/CH	☐ others

Fig. 5 HMM graphs for the eight cluster solution (clusters A–D). State abbreviations show career/partnership/parenthood statuses: ST=studying, EM=employed, UN=unemployed, EL=else; S=single, C=cohabiting, M=married, D=divorced/separated; NC=no children, CH=has child(ren).

Partnership(s) & no children, n = 177

No or late family, n = 116

Divorced parents, n = 102

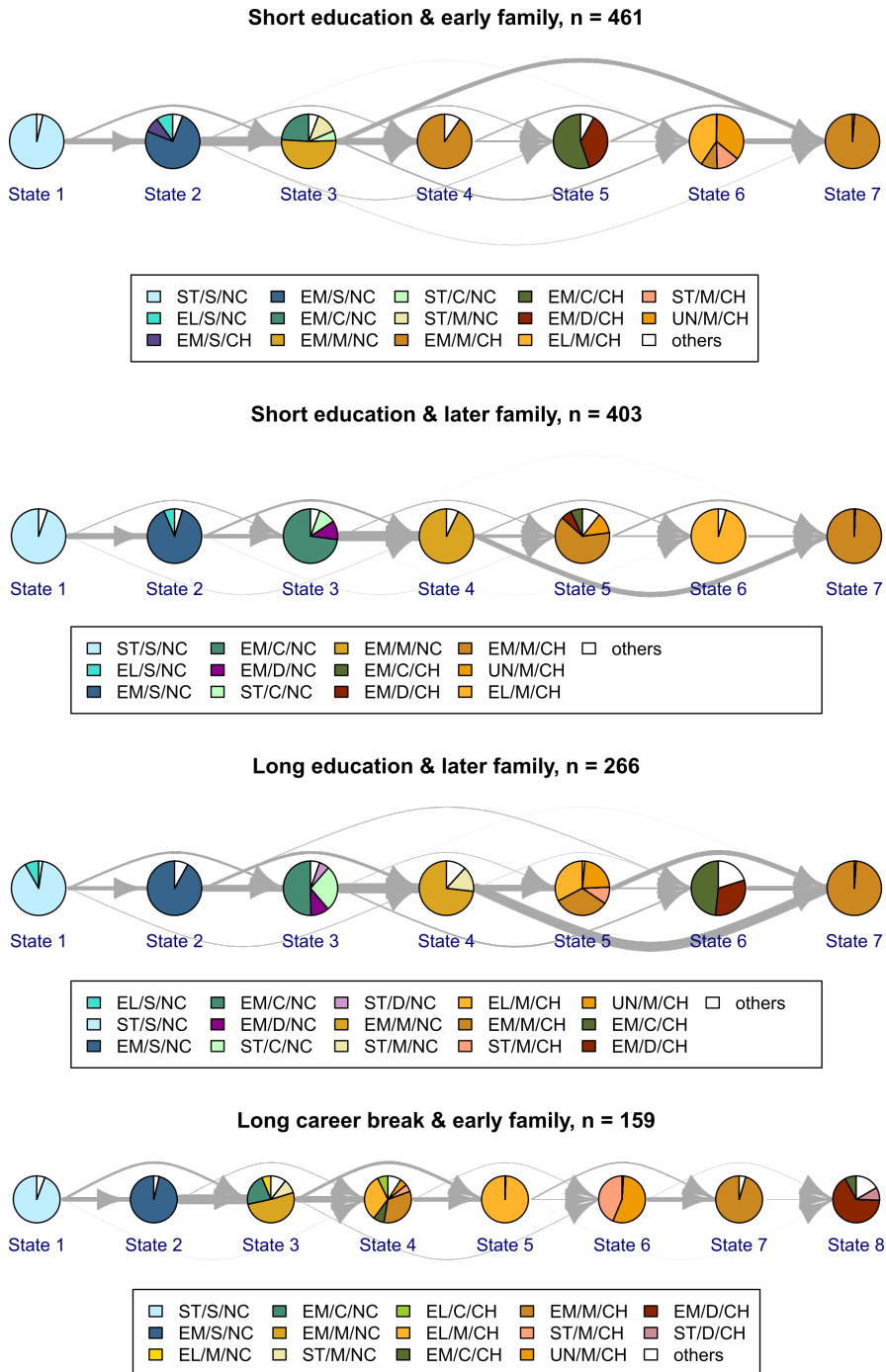

Single parents, n = 47

Fig. 6 HMM graphs for the eight cluster solution (clusters E–H). State abbreviations show career/partnership/parenthood statuses: ST=studying, EM=employed, UN=unemployed, EL=else; S=single, C=cohabiting, M=married, D=divorced/separated; NC=no children, CH=has child(ren).
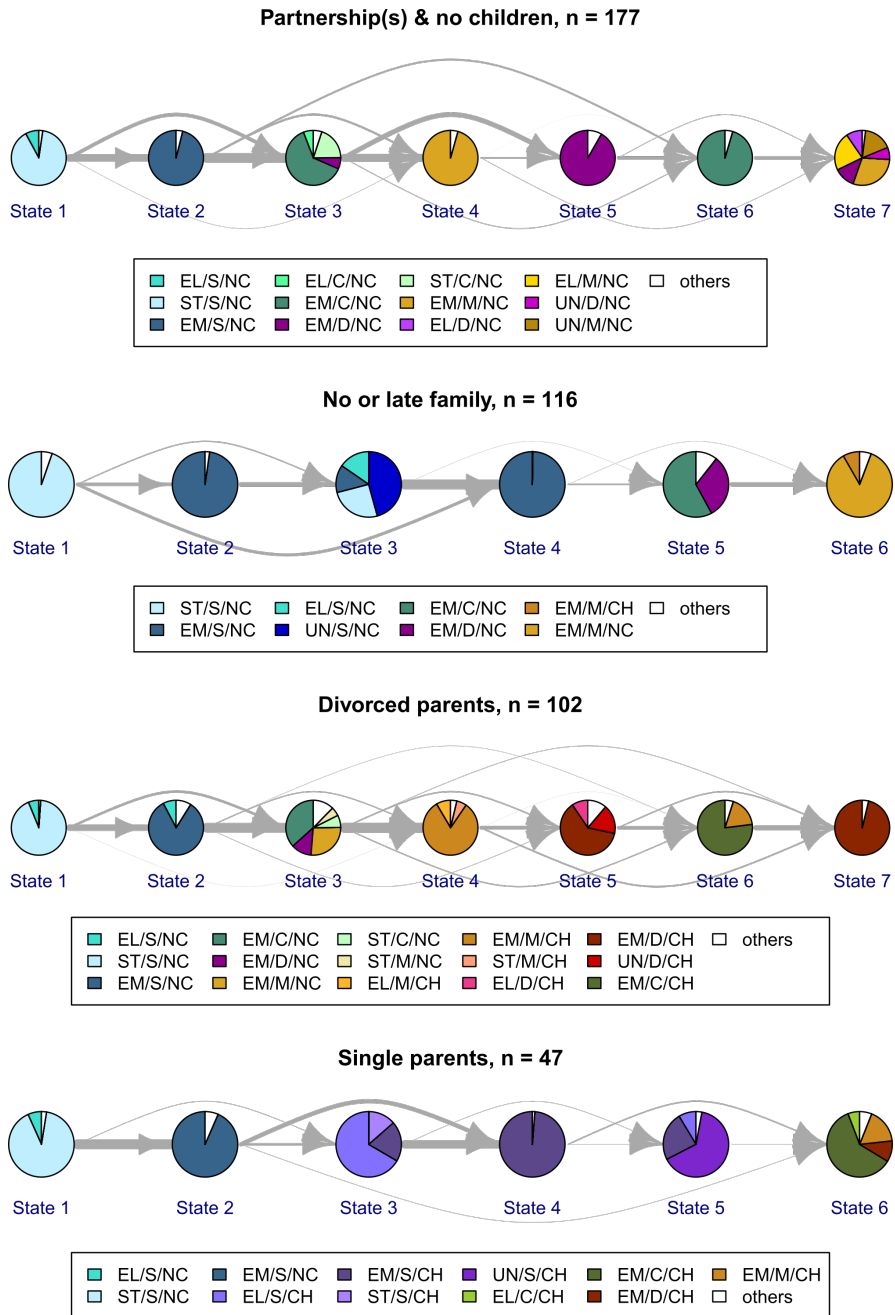
Analysing Complex Life Sequence Data with Hidden Markov Modelling 19

Many individuals switched clusters compared to the SA solution (see Table 4). Some cluster were more stable; close to 90% of the members of the SA clusters "Single parents" and "Partners and no children" stayed in the same cluster in the MHMM solution. Others had many switchers; less than half of the members of SA clusters "Short education and early family" and "Long education and later family" stayed in their original clusters in the MHMM solution.

Table 4 Comparison of SA cluster memberships (left) to most probable cluster memberships from the non-estimated MHMM (top). Probabilities of staying in the same cluster are shown in bold.

SA clusters	MHMM clusters								
	A	B	C	D	E	F	G	H	Members
Short educ. & early fam. (A)	**0.32**	0.35	0.15	0.11	0.00	0.00	0.06	0.01	461
Short educ. & later fam. (B)	0.09	**0.64**	0.16	0.09	0.00	0.00	0.03	0.00	403
Long educ. & later fam, (C)	0.06	0.32	**0.43**	0.13	0.00	0.00	0.07	0.00	266
Career break & early family (D)	0.04	0.39	0.03	**0.54**	0.00	0.00	0.00	0.00	159
Partnership(s) & no child (E)	0.00	0.05	0.03	0.00	**0.87**	0.05	0.00	0.01	177
No or late family (F)	0.00	0.03	0.01	0.03	0.32	**0.60**	0.00	0.01	116
Divorced parents (G)	0.04	0.00	0.03	0.16	0.00	0.00	**0.77**	0.00	102
Single parents (H)	0.00	0.00	0.02	0.00	0.00	0.00	0.04	**0.94**	47
Number of cluster members	207	577	260	228	191	79	138	51	1731

If the MHMM parameters were estimated jointly, the differences compared to the SA clusters were even larger (we do not report the findings as we were not able to find the globally optimal model). In both MHMM approaches, the order and occurrence of states were generally more determining for the cluster memberships than the timing and duration of states. Fig. 7 illustrates this difference seen in the cluster "Short education and early family", showing the observed and hidden state sequences of members of the SA cluster and the cluster from the non-estimated MHMM. One can easily see that the variation in the timing of transitions between states (both observed and hidden) is much larger in the MHMM cluster compared to the SA cluster.

8 Discussion

When analysing complex sequence data with multiple channels, describing and visualizing the data can be a challenge. Hidden Markov models and their mixtures offer a probabilistic model-based framework where the information in data can be compressed into hidden states (different life stages) and clusters (general patterns in life courses). Hidden states can capture general life stages that include not only rather stable episodes (as the fifth hidden state of work, marriage, and children in Fig. 2) but also life stages characterized by change (as the second hidden state of searching for a partner in Fig. 2).
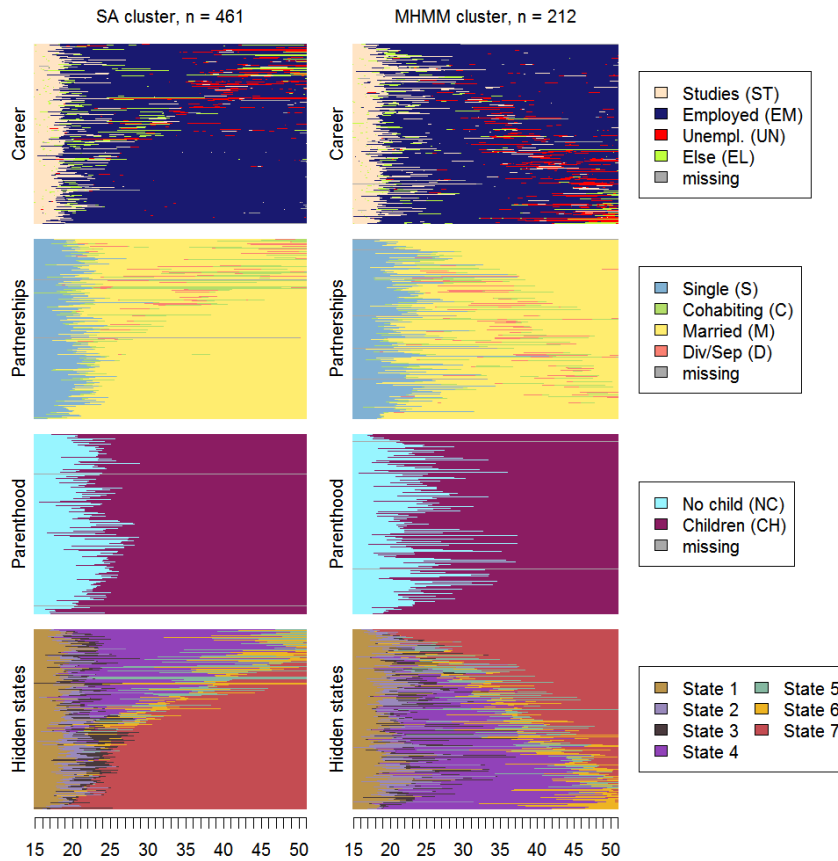
Fig. 7 Comparison of a cluster (Short education and early family) given by SA and the non-estimated MHMM.

Mixture hidden Markov modelling has several advantages. With posterior cluster probabilities we get information on certainty of the clustering for each individual and a measure for the goodness of the classification. We can also extend the model by adding covariates for explaining cluster memberships or transitions between hidden states. The MHMM approach has been used successfully in simpler settings, e.g., for accounting for measurement error and for finding clusters of "movers" and "stayers" between two hidden states.

The downsides of MHMM analysis are related to computational issues. Maximum likelihood estimation of parameters of a complex MHMM is computationally heavy. Due to multimodality of the likelihood surface we need to estimate the model numerous times with different starting values. Also, often the structure of the model (in terms of the number of hidden states and/or clusters) is not known and in general

Analysing Complex Life Sequence Data with Hidden Markov Modelling   21

selecting the best structure is a nontrivial task. Thus, finding the globally optimal MHMM can become unfeasible without constraining the problem.

Using sequence analysis and cluster analysis as a starting point might be useful by providing preliminary classification and by limiting the set of candidate models for a complex MHMM setting. In our study we were not able to reach satisfactory results. Our data was much more complex than in a typical MHMM analysis where sequences often come from panel data with a moderate number of measurement points. The multichannel structure, long sequences, and the relative large number of individuals in our data was a challenging combination for parameter estimation. Also, typically the number of candidate models is rather limited; when HMMs are used for accounting for measurement error, the number of hidden states is known in advance and usually the state space is very limited (e.g., poor/nonpoor or drug user/nonuser). In our study the model structure was unknown and we expected to find several clusters, each with an unknown number of hidden states.

Instead of using mixture models, we treated the SA clusters as fixed and estimated HMMs separately for each cluster (the combined model approach). With SA we found clusters that were adequately well separated by the timing and duration of life states. Hidden Markov models were used for choosing the number of clusters and for describing the overall dynamics within clusters.

Clusters found using SA and the MHMM were different in several ways. When defining sequence dissimilarities, we considered the timing of the events very important and used Hamming distances. In the MHMM analysis many individuals switched clusters; the order of states was generally more determining than their timing and duration. Further research is needed in order to determine distance metrics that result in SA clusters which capture similar features as HMMs. Metrics that weight the order of states instead of their timing such as the number of matching subsequences or the subsequence vectorial representation metric (Studer and Ritschard, 2016), might produce clustering results that are better suited for the starting point of MHMM estimation. Unfortunately, using these metrics with multichannel data is not a straightforward task.

Another topic for further research is model selection of left-to-right HMMs and MHMMs. In our study, BIC performed poorly. Further theoretical and empirical studies are needed for detecting the reasons for its failure and for discovering selection criteria that are better suited for finding parsimonious HMMs.

The aim of our study was to describe complex life sequence data. For that goal, SA and the combined HMM approach gave satisfactory results in a reasonable time. We were able to find meaningful clusters and to visualize their complex life course information by using stacked sequence plots, combined state distributions, and HMM graphs.

9 Acknowledgements

Appendix

Detailed visualizations of the eight SA clusters and the respective HMMs. Figures show state distributions of combined observations at each time point (top), observed sequences in three life domains and the most probable hidden state paths given the HMM (middle), as well as HMM graphs with initial and transition probabilities (bottom). See Sect. 5 for more information on how to interpret the visualizations.

Analysing Complex Life Sequence Data with Hidden Markov Modelling   23

24   Satu Helske, Jouni Helske, and Mervi Eerola

Short education & later family, n = 403

Analysing Complex Life Sequence Data with Hidden Markov Modelling 25

Long education & later family, n = 266

26   Satu Helske, Jouni Helske, and Mervi Eerola

Long career break & early family, n = 159

Analysing Complex Life Sequence Data with Hidden Markov Modelling   27

Partnership(s), no children, n = 177

No or late family, n = 116

Analysing Complex Life Sequence Data with Hidden Markov Modelling   29

Single parents, n = 47

References

Aassve, A., Billari, F. C., and Piccarreta, R. (2007). Strings of adulthood: A sequence analysis of young British women's work-family trajectories. *European Journal of Population/Revue européenne de Démographie*, 23(3-4):369–388.

Aisenbrey, S. and Fasang, A. (2010). New life for old ideas: The "second wave" of sequence analysis – bringing the "course" back into the life course. *Sociological Methods & Research*, 38(3):420–462.

Bartolucci, F., Pennoni, F., and Francis, B. (2007). A latent Markov model for detecting patterns of criminal activity. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 170(1):115–132.

Baum, L. E. and Petrie, T. (1966). Statistical inference for probabilistic functions of finite state Markov chains. *The Annals of Mathematical Statistics*, 67(6):1554–1563.

Blossfeld, H.-P., Roßbach, H.-G., von Maurice, J., Schneider, T., Kiesl, S. K., Schönberger, B., Müller-Kuller, A., Rohwer, G., Rässler, S., Prenzel, M. S., et al. (2011). Education as a lifelong process-the German National Educational Panel Study (NEPS). *Age*, 74(73):72.

Crayen, C., Eid, M., Lischetzke, T., Courvoisier, D. S., and Vermunt, J. K. (2012). Exploring dynamics in mood regulation—mixture latent Markov modeling of ambulatory assessment data. *Psychosomatic Medicine*, 74(4):366–376.

Durbin, R., Eddy, S., Krogh, A., and Mitchison, G. (1998). *Biological sequence analysis: Probabilistic models of proteins and nucleic acids*. Cambridge, UK: Cambridge University Press.

Eerola, M. and Helske, S. (2016). Statistical analysis of life history calendar data. *Statistical Methods in Medical Research*, 25(2):571–597.

Elzinga, C. H. and Studer, M. (2014). Spell sequences, state proximities, and distance metrics. *Sociological Methods & Research*, pages 3–47.

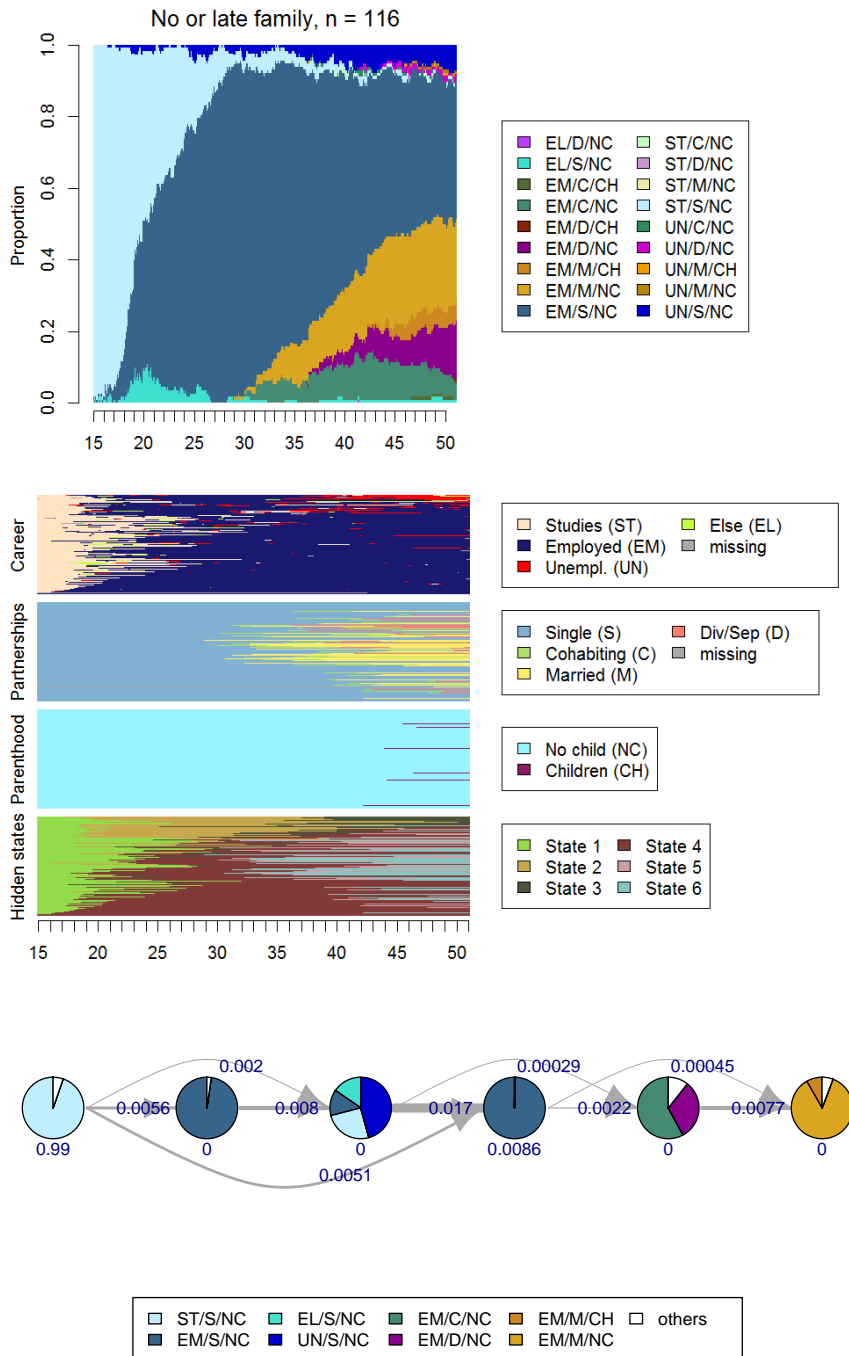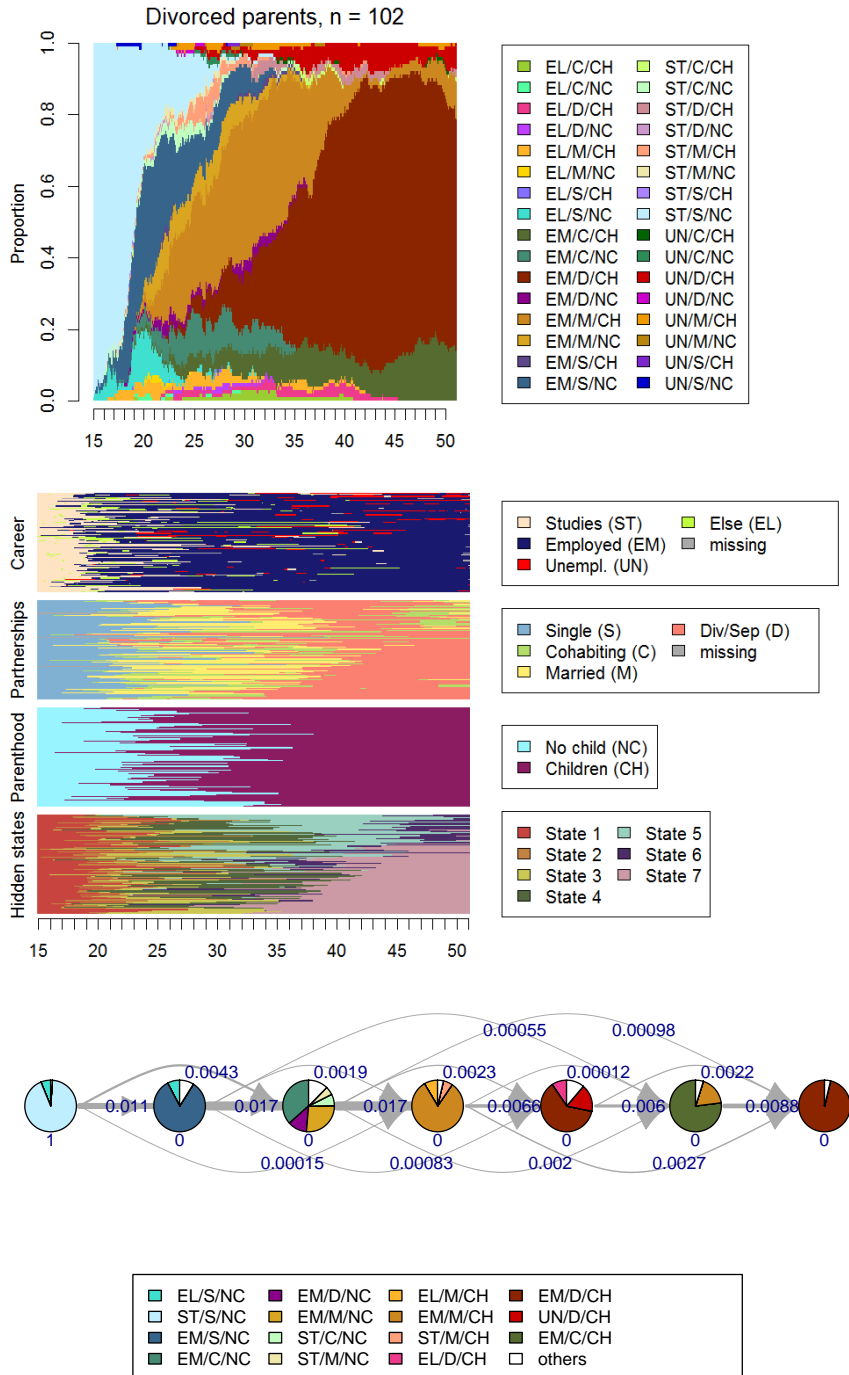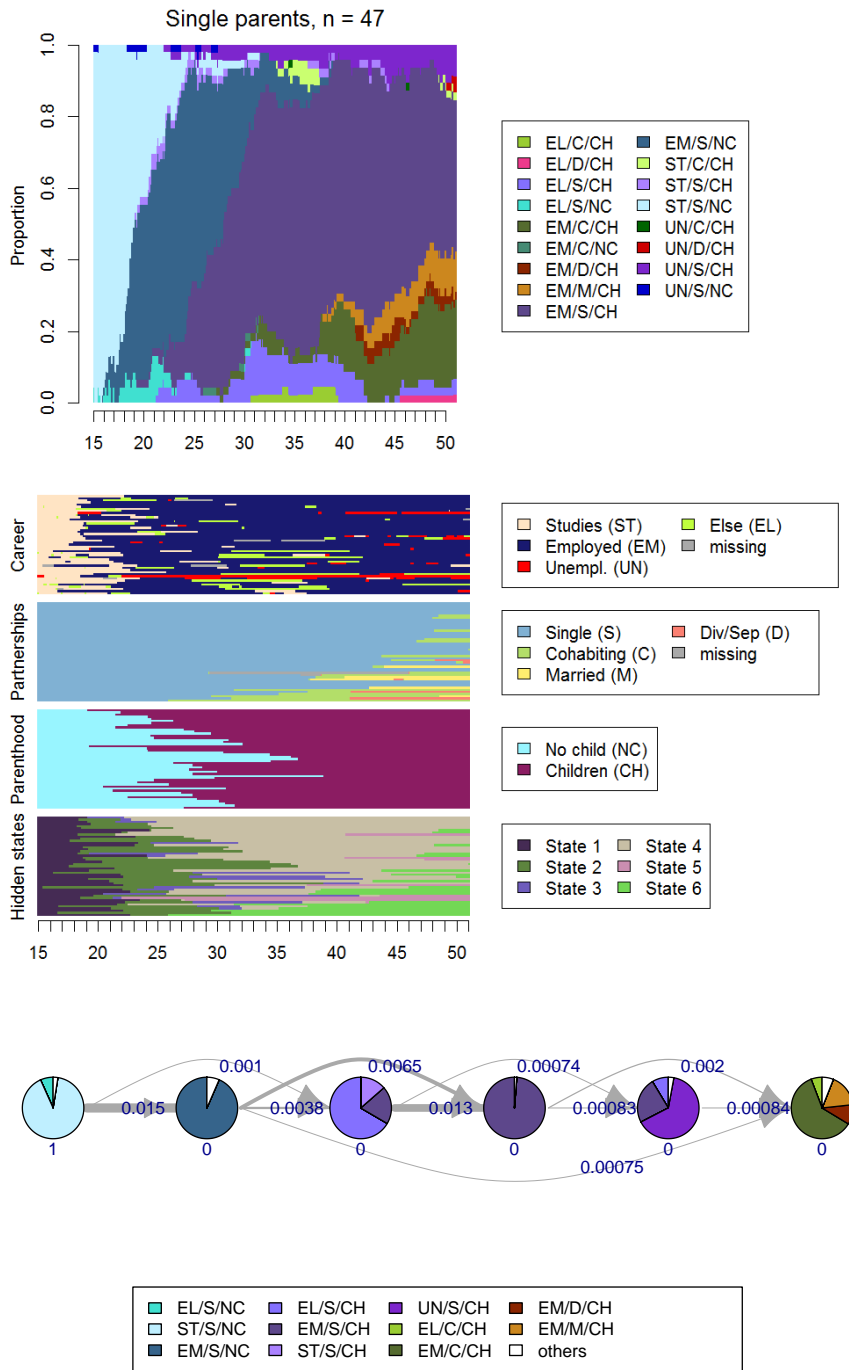Gabadinho, A., Ritschard, G., Müller, N. S., and Studer, M. (2011). Analyzing and visualizing state sequences in R with TraMineR. *Journal of Statistical Software*, 40(4):1–37.

Gauthier, J.-A., Bühlmann, F., and Blanchard, P. (2014). Introduction: Sequence analysis in 2014. In *Advances in Sequence Analysis: Theory, Method, Applications*, pages 1–17. Springer.

Gauthier, J.-A., Widmer, E. D., Bucher, P., and Notredame, C. (2010). Multichannel sequence analysis applied to social science data. *Sociological Methodology*, 40(1):1–38.

Hamming, R. W. (1950). Error detecting and error correcting codes. *Bell System Technical Journal*, 29(2):147–160.

Helske, S. and Helske, J. (2016). Mixture hidden Markov models for sequence data: the seqHMM package in R. *Submitted*.

Helske, S., Steele, F., Kokko, K., Räikkönen, E., and Eerola, M. (2015). Partnership formation and dissolution over the life course: applying sequence analysis and event history analysis in the study of recurrent events. *Longitudinal and Life*

Course Studies, 6(1):1–25.

Lesnard, L. (2010). Setting cost in optimal matching to uncover contemporaneous socio-temporal patterns. *Sociological Methods & Research*, 38(3):389–419.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., and Hornik, K. (2015). *cluster: Cluster Analysis Basics and Extensions*. R package version 2.0.3.

McVicar, D. and Anyadike-Danes, M. (2002). Predicting successful and unsuccessful transitions from school to work by using sequence methods. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 165(2):317–334.

Müller, N. S., Sapin, M., Gauthier, J.-A., Orita, A., and Widmer, E. D. (2012). Pluralized life courses? an exploration of the life trajectories of individuals with psychiatric disorders. *International Journal of Social Psychiatry*, 58(3):266–277.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rabiner, L. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286.

Spallek, M., Haynes, M., and Jones, A. (2014). Holistic housing pathways for Australian families through the childbearing years. *Longitudinal and Life Course Studies*, 5(2):205–226.

Studer, M. and Ritschard, G. (2016). What matters in differences between life trajectories: a comparative review of sequence dissimilarity measures. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 179(2):481–511.

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

van de Pol, F. and Langeheine, R. (1990). Mixed Markov latent class models. *Sociological Methodology*, 20:213–247.

Vermunt, J. K., Tran, B., and Magidson, J. (2008). *Latent Class Models in Longitudinal Research*, pages 373–385. Handbook of Longitudinal Research: Design, Measurement, and Analysis. Elsevier, Burlington, MA.

Whiting, R. and Pickett, E. (1988). On model order estimation for partially observed Markov chains. *Automatica*, 24(4):569–572.

A discussion on Hidden Markov Models for Life Course Data

Danilo Bolano, André Berchtold, Gilbert Ritschard

Abstract This is an introduction on discrete-time Hidden Markov models (HMM) for longitudinal data analysis in population and life course studies. In the Markovian perspective, life trajectories are considered as the result of a stochastic process in which the probability of occurrence of a particular state or event depends on the sequence of states observed so far. Markovian models are used to analyze the transition process between successive states. Starting from the traditional formulation of a first-order discrete-time Markov chain where each state is liked to the next one, we present the hidden Markov models where the current response is driven by a latent variable that follows a Markov process. The paper presents also a simple way of handling categorical covariates to capture the effect of external factors on the transition probabilities and existing software are briefly overviewed. Empirical illustrations using data on self reported health demonstrate the relevance of the different extensions for life course analysis.

Key words: : Sequence Analysis, Life course approach, Hidden Markov Model.

———————————————

Danilo Bolano

Institute for Social Science Research, Life Course Center, The University of Queensland, e-mail: D.Bolano@uq.edu.au

André Berchtold

Institute for Social Sciences, NCCR LIVES, University of Lausanne e-mail: Andre.Berchtold@unil.ch

Gilbert Ritschard

Institute of Demography and Socioeconomics, NCCR LIVES University of Geneva e-mail: Gilbert.Ritschard@unige.ch

Danilo Bolano, André Berchtold, Gilbert Ritschard

1 Introduction

Markovian models are stochastic models dedicated to the analysis of the transitions between successive states in sequences. More specifically, a Markovian model aims to describe how the current distribution of the possible values of a characteristic of interest depends on the previously observed values, and possibly how this dependence may be moderated by observable covariates or some unobservable latent factor.

Markov models have been extensively applied in many research areas such as speech recognition, behavior analysis, climatology, or finance. In demographic and population studies, they have been used for multistate analysis (e.g. Rogers, 1975; Land and Rogers, 1982; Willekens, 2014), for modelling population processes and life cycles (e.g., Feichtinger, 1973; Caswell, 2009), for analysing social mobility (since e.g., Hodge, 1966; McFarland, 1970), and so on. However, most of these studies consider only a simple Markov process, a first-order homogeneous Markov chain. Extensions such as higher order Markov chains, latent-based and non-homogeneous Markov models are instead rarely used. The aim of this paper is to underline the potential of latent-based Markov models, the so-called Hidden Markov Models, for life course studies.

In the Markovian perspective, life trajectories are considered as the result of a stochastic process in which the probability of occurrence of a particular state or event depends on the sequence of states observed so far. In other words, considering life trajectories as sequences of mutually exclusive states—e.g., sequences of employment statuses or of health conditions—a Markovian process focuses on successive transitions and attempts to depict the life history of an individual looking at the probabilities to switch to the different states of interest given the state history lived so far.

Markovian models form an extremely flexible class of models. The most immediate application is the study of transition mechanisms between successive states (e.g., studies on working career, social mobility, evolution of health conditions). A basic Markov chain models directly the transitions between visible states, while hidden Markov models prove useful to study how the succession of observed states may be governed by an underlying latent process. The latter approach is particularly interesting in life course studies where many not- or hardly observable aspects such as motivations, beliefs, levels of frailty, may influence the observed behavior of an individual. The salient aspect of hidden Markov models is that, unlike other latent models, it allows for a time-varying latent characteristic. Such models prove useful, for example, to study the transitions between unobserved vulnerability statuses revealed through some observable variable such as the health condition or the general satisfaction level.

The paper focuses on several important aspects of using Markov based models for life course studies addressing the analysis of transition mechanisms (Section 3), the modelling of latent processes (Section 4), probabilistic clustering, and how to evaluate the impact of covariates on the transition process (Section 6). Finally, existing software are briefly commented in Section 7 and the concluding discussion

A discussion on Hidden Markov Models for Life Course Data

in Section 8 recaps the scope and applications of Markov modelling for life course analysis. The concepts presented all over the article are illustrated with longitudinal data on self-rated health represented in Section 2.

2 Longitudinal data on self-rated health condition

For illustration we use data from 14 waves of the Swiss Household Panel (Voorpostel et al., 2013). It is a yearly panel study started in 1999. We focus here on an unbalanced sub-sample of 1,331 individuals aged 50 years and more at the first interview and with at least three measurement occasions.

We intend to study the change over time in self-rated health conditions (SRH). We will analyse the transitions between SRH conditions (Section 3) and, by the means of a hidden Markov model, we will test whether there is some underlying hidden process that drives the observed changes (Section 4). Finally we will investigate the effects of the educational level on the process (Section 6).

The SRH condition is defined from the question "How do you feel right now?". Five possible answers were proposed: "not well at all", "not very well", "so-so", "well", "very well" that we shall denote respectively as P (poor), B (bad), M (medium), W (well) and E (excellent) health condition. The distribution on the overall dataset shows a general condition of good health. Near 80% of the respondents feel well (W) or very well (E) and only 2% bad (B) or very bad (P).

3 Markov chains

A discrete-time Markov chain is a stochastic process that describes how individuals transit between a finite number of pre-defined categorical states.

A Markov process models the probabilities of being in a state given the states visited in the past. In its basic formulation, the next value of the variable of interest depends only on the current state that is assumed to summarize the whole history of the individual. This is known as the *Markov property* and it defines a first-order Markov chain (Figure 1). For instance, with the two state alphabet corresponding to being in good (G) and bad health (B), we would consider, for people in a bad health condition, the probability to stay in the same condition the next period, $p(B|B)$, versus the probability to improve their health condition, $p(G|B)$. And, for those who are in a good health, the probability to have a deterioration in health condition, $p(B|G)$, versus the probability to stay in good health, $p(G|G)$.

The probability of switching from a given state to another is often assumed to remain unchanged over time. This defines a *time-homogeneous* Markov process. This assumption is tenable in fields as machine learning, biology but is often violated in applications in the social sciences. For example the probability of recovering from a bad health condition is likely to change over time with the age of the respondent.

Danilo Bolano, André Berchtold, Gilbert Ritschard
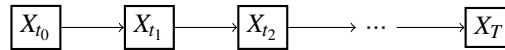
Fig. 1: A graphical representation of a first-order Markov chain.

For simplicity, this paper focuses on time-homogenous Markov processes but extensions to relax the homogeneity assumption have been proposed in the literature (see for instance the Double Chain Markov models proposed by Berchtold, 1999).

Table 1: First order Markov chain for SRH trajectories. The transition matrix.

$$
\mathbf{A} = \begin{array}{c} \\ X_{t-1} \\ P \\ B \\ M \\ W \\ E \end{array} \begin{array}{ccccc} & & X_t & & \\ P & B & M & W & E \\ \left(0.261 \right. & 0.217 & 0.348 & 0.13 & \left. 0.043 \right) \\ 0.024 & 0.159 & 0.53 & 0.271 & 0.018 \\ 0.006 & 0.049 & 0.473 & 0.435 & 0.038 \\ 0.002 & 0.009 & 0.141 & 0.726 & 0.123 \\ 0.000 & 0.003 & 0.043 & 0.521 & 0.433 \end{array}
$$

The transition probabilities are generally represented in matrix form in the so-called *transition matrix* (see for example Table 1). It is a square matrix of order m, with m the number of states.

Considering our illustration on self-rated health conditions, Table 1 reports the transition probabilities between health conditions estimated for a first order Markov chain. The probabilities to stay in the current state, reported on the main diagonal, are, with one exception, smaller than 50% meaning that there are frequent changes within the health trajectories. The probability to change the health condition is particularly high for people in bad condition. The probabilities of changing from a poor and a bad condition are respectively 73.9% and 84.1%.

4 Hidden Markov models

4.1 *Including a latent process in life course data*

Instead of modelling the stochastic process of the variable of interest—the health status in our illustration—it is often more realistic to assume that the successive values of this variable are governed by the underlying process of a latent variable such as motivation, belief, or vulnerability. Assuming that such categorical latent variable can change over time (i.e., is time-varying) following a Markov process,

A discussion on Hidden Markov Models for Life Course Data

we get the so-called hidden Markov model (HMM) (see, e.g., Rabiner, 1989). The modalities assumed by the latent variable are called hidden states.

HMMs are widely used in biosciences and genetics (e.g., Le Strat and Carrat, 1999; Shirley et al., 2010) to study sequences of DNA and protein. An extensive literature exists in speech recognition since Baum and Petrie (1966). HMMs are also used in behavioral and criminal studies (Bijleveld and Mooijaart, 2003; Bartolucci et al., 2007), psychology (e.g., Visser et al., 2002) and in economics and finance where they are known as regime switching models (e.g., Elliott et al., 1998; Hayashi, 2004; Netzer et al., 2008).

There are several alternative ways of interpreting the latent variable in HMM providing multiple potential usages of this approach in life course studies. First, the HMM latent variable can be seen as a related unobserved characteristic of interest. For instance, assuming the reported health condition depends on the unobserved frailty level of the individual, an HMM would allow to study the stochastic process of this unobserved frailty.

Second, the hidden states may serve to capture the process heterogeneity or more specifically the 'person-position' heterogeneity, i.e., differences in the individual outcome probabiility distribution at the successive positions (e.g., McLachlan and Peel, 2000; Zucchini and MacDonald, 2009). In that case, the levels of the latent variable do not receive any specific interpretation but are just supposed to render the diversity of the person-period behaviors.

Third, HMM can be used for probabilistic clustering (see Section 5). This is similar to the capture of the process heterogeneity except that here a higher focus is put on each level of the latent variable being interpreted as a distinct latent class.

4.2 The HMM framework

When modelling life course data with an HMM, the sequence of observed events/states are supposed to be stochastically generated from a hidden Markov process. For each hidden state we have a different distribution of the visible state. So, it is the hidden process that selects at each position the distribution of the visible state.[1] Figure 2 shows a path diagram of a first-order hidden Markov process.

While a basic discrete first order Markov model is characterized by a response variable $X(t)$ with m modalities and a $m \times m$ matrix of transition probabilities $\mathbf{A}$, a first-order discrete HMM consists of five elements: i) a response variable $X(t)$ with m modalities;[2] ii) a categorical latent variable $S(t)$ with k modalities; iii) a $k \times k$ matrix $\mathbf{Q}$ of transition probabilities between two successive hidden states; iv) the emission—or outcome—probabilities, i.e., the probabilities $p_i(x_t)$, of observing

---

[1] For the sake of simplicity, in the rest of the paper, we shall omit the adjective "hidden" or "visible" when the nature of the state is unambiguous from the context.

[2] Even though the outcome variable $X(t)$ could also be numeric, we consider here only the case of a categorical response variable for simplicity.

Danilo Bolano, André Berchtold, Gilbert Ritschard
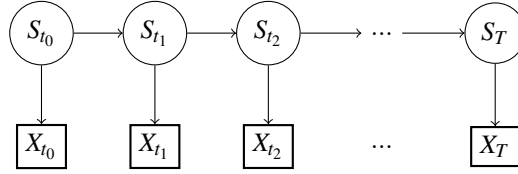
Fig. 2: A graphical representation of a first-order hidden Markov model. S_t is the latent variable at time t, X_t the observed response variable.

$X_t = x_t$ when being in the hidden state i; v) the initial probability distribution π of the k hidden states

The simplest HMM—a homogeneous hidden Markov model of order one—can be summarized using the following equations:

$$q_{ij} = p(S_t = i | S_{t-1} = j) \quad t = 1, ..., T \tag{1a}$$

$$\pi_i = p(S_{t_0} = i) \quad i = 1, ..., k \tag{1b}$$

$$p_i(x_t) = p(X_t = x_t | S_t = i) \quad i = 1, ..., k \tag{1c}$$

The first two equations represent the unobservable part of the model. Equation (1a) states that the latent variable S_t follows a first-order Markov process. So the current hidden state depends only on the previous one. As for visible Markov chains, a higher order dependence can be introduced. Equation (1b) gives the initial probability of the hidden states, i.e., the probability at the first time point $t = 1$.

The third equation (Eq 1c) refers to the measurement part of the model. It states that the visible state is determined by a hidden-state-dependent process—the emission probabilities (vertical arrows in Figure 2). Such emission probabilities are also known in the literature as response probabilities. The probability distribution of X_t depends only on the current hidden state and does not depend on previous observations nor on previous hidden states. In other words, Equation (1c) assumes that the observations are conditionally independent given the latent process. This is known as the *local independence* assumption.

4.3 Model comparison and selection of the number of hidden states

An important aspect with HMM is the choice of the number of hidden states. In some settings, the relevant number of states can be determined on theoretical grounds. This would be the case, for instance, when the latent variable explicitly stands for an a priori defined unobservable characteristic such as frail versus non-frail. Alternatively, we may want to let the number of states be determined from the data, i.e., choose the number of states on statistical grounds.

A discussion on Hidden Markov Models for Life Course Data

Since the HMM is an extension of mixture models, the issue of finding the number of hidden states is similar to find the number of mixture components (McLachlan and Peel, 2000). Information criteria such as the AIC and the BIC can be used. The model with the lowest BIC (or AIC) is chosen. Although other model selection criteria exists (e.g. cross-validated likelihood, Celeux and Durand, 2008), the BIC is the most commonly used.

For instance, in order to select the optimal number of hidden states in our empirical example, we compare several models in terms of likelihood and BIC increasing the number of hidden states up to 5 (Table 2). The lowest BIC (17971.8) is observed

Table 2: SRH trajectories. The choice of the number of hidden states.

Model	No. of hidden states	Free parameters	Log-Likelihood	BIC
HMM	2	11	-10532.6	21167.2
HMM	3	20	-8893.17	17971.82
HMM	4	31	-8887.315	18062.1
HMM	5	44	-8782.427	17972.87

Note: the number of parameters and the BIC do not include the null transition probabilities.

for the model with three hidden states.

4.3.1 A 3-state HMM for SRH trajectories

As for other latent-based models, the order of the hidden states is not meaningful. The relationship between the data (i.e., the outcome variable) and the hidden states have to be analyzed using the emission probabilities (Eq. 1c, here reported in Table 3a) in order to give a "name" (labeling) each state. An alternative is to estimate for each individual the most likely sequence of hidden states (by the means of the Viterbi algorithm, Viterbi, 1967) and then to provide a cross tabulation between observations and the predicted hidden states (Table 4).

The first hidden state refers to individuals with perfect health condition with high chances to be in excellent (56.1%) or well (42.1%) condition. We will refer to this hidden state as a situation of "very good" condition (VG). Hidden State 2 is instead mainly associated with state M (65%) or with a worse health conditions (10% of probability of feeling B "not very well" or P "not well at all"). We will then label this state as "frail" health condition (F). This is the only state with individuals potentially at risk of frailty since in hidden state 1 and 3 we have almost no chance to be in a poor or very bad health condition. Finally, Hidden State 3 is an intermediate situation mainly associated with W (84%) or M (almost 10%). We will refer to State

Danilo Bolano, André Berchtold, Gilbert Ritschard

Table 3: Three-state HMM

(a) Emission probability distribution by hidden states (columns).

SRH	Hidden State 1 F	Hidden State 2 G	Hidden State 3 VG
P	0.013	0.000	0.000
B	0.082	0.002	0.002
M	0.649	0.098	0.016
W	0.245	0.841	0.421
E	0.011	0.059	0.561

(b) Initial hidden state distribution π

F	G	VG
0.199	0.528	0.273

(c) Transition matrix between hidden states.

$$\mathbf{Q}= \begin{array}{c} \\ S_{t-1} \\ F \\ G \\ VG \end{array} \begin{array}{c} S_t \\ \begin{array}{ccc} F & G & VG \end{array} \\ \left(\begin{array}{ccc} 0.943 & 0.057 & 0.000 \\ 0.034 & 0.957 & 0.009 \\ 0.000 & 0.084 & 0.916 \end{array} \right) \end{array}$$

Table 4: Cross tabulation between observed states (by rows) and predicted hidden states (1,331 individual sequences.)

SRH	F	G	VG
P	29	0	0
B	181	10	3
M	1383	655	29
W	426	5471	798
E	22	414	1228

3 as a state of "good" health (G)[3]. In the rest of the paper we will then (re)label and (re)order the hidden states as F, G and VG.

The transition probabilities between hidden states can be represented in matrix form (Table 3c) or since we have only three hidden states (i.e. the latent variable has three categories) as a path diagram (Table5). In the diagram, the arrows correspond-

---

[3] Here, the labels of the hidden states will be printed in italics since such states are not observed but inferred from the data.

A discussion on Hidden Markov Models for Life Course Data

ing to probabilities estimated as zero are not shown and for readability purposes transition probabilities have been rounded to two decimals.

Table 5: Transition probabilities as path diagram.
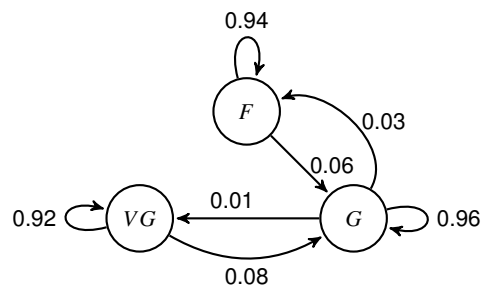
Despite the overall healthy aging of the Swiss population, we identify a relevant risk of vulnerability since the first observation. The initial distribution of hidden state (π in Table 3b) even if dominated by the hidden state G (52.7%), reports a 19.9% of chance to start the trajectory in a potentially at risk situation of frailty (hidden state F).

According to the transition probabilities (Table 3c), the states are very persistent meaning that a stability in health patterns is observed. There is more than 90% of probability to stay in the same state for two consecutive periods and three transitions, $(F - VG)$, $(G - VG)$, $(VG - F)$, are extremely rare or impossible. The transition probabilities for individuals with a good health condition, hidden state G, are particularly interesting. It is the most common hidden state and estimated to be used 59% of the time. Apart from those who stay in the same hidden state, they have more chance to fall down in the at frail condition rather than to improve their situation (3.4% vs 0.9%) confirming that a risk of a slight deterioration in health condition in the following period exists despite a general tendency of stability over time.

Once estimated the hidden states via the Viterbi algorithm (Viterbi, 1967) we can graphically represent their cross-sectional distributions as in the chronogram reported in Figure 3.

These results confirm the one reported for a visible Markov chain with a Swiss population in good health and with a pattern of stability or slightly deterioration over time. However, unlike before, a hidden Markov model allows to distinguish between three different conditions (the three hidden states) instead of analyzing the transition probabilities among the five observed health states.

Danilo Bolano, André Berchtold, Gilbert Ritschard
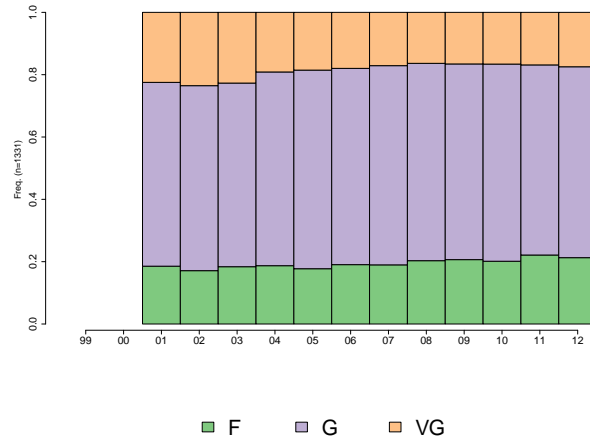
Fig. 3: Sequence of cross-sectional distributions of hidden states.

5 Using HMM for probabilistic clustering

As already mentioned, hidden states can be interpreted as latent classes and, therefore, HMM can serve for clustering. HMM belong to the class of mixture models so clustering with HMM is a model-based probabilistic clustering method. The class membership is given by the different emission probability distributions assigned to the hidden states.

HMM can perform two main types of clustering depending on whether the latent class is allowed to vary or not over time. We cluster person-position (person-period) states when the hidden state can vary over time, i.e., when using an unconstrained HMM, and trajectories when the latent variable is constrained to remain fixed over time. Other alternatives have been proposed in the literature as (e.g. Bicego et al., 2003) deriving pairwise dissimilarities between estimated sequences of hidden states and then to proceed with a dissimilarity-based clustering from those dissimilarities. Here, we focus on the first two alternatives.

5.1 Person-position state clustering

In an unconstrained conventional hidden Markov model, the latent variable is time-varying. Therefore, each individual may move from one cluster to another over time. In that case, we do not get clusters of individuals but clusters of states, actually of person-position states. Considering, for example, the three-hidden-state HMM fitted

A discussion on Hidden Markov Models for Life Course Data

Table 6: Emission probability distributions (columns). Constrained model.

SRH	$\overline{F}$	$\overline{G}$	$\overline{VG}$
P	0.011	0.001	0.000
B	0.069	0.003	0.003
M	0.55	0.114	0.021
W	0.346	0.801	0.458
E	0.024	0.081	0.518

before, we would have a cluster of frail states (F), one of good (G) and one of very good (VG) states of health conditions.

The membership probability at each time point is obtained multiplying the initial distribution π by the matrix of transition probabilities $\mathbf{Q}$ of the hidden process. For example, in our empirical example, the overall cluster membership probability distribution would be at position 3 given by $\pi * \mathbf{Q} * \mathbf{Q}$

$$\pi\mathbf{Q}^2 = \begin{matrix} F & G & VG \\ \begin{pmatrix} 0.212 & 0.549 & 0.239 \end{pmatrix} \end{matrix}, $$

5.2 Trajectory clustering

To cluster the trajectories—the entire individual sequences—we have to constrain the transition matrix of the hidden process to be the diagonal identity matrix. This makes the latent variable time-invariant and individuals will belong to one and only one cluster during the whole period of observation. As in the period-position state case, we have to resort to the Viterbi algorithm to get cluster memberships.

5.2.1 HMM trajectory clustering of SRH conditions

We illustrate the clustering of trajectories with the SRH data by constraining the transition matrix $\mathbf{Q}$ of the hidden process to be the identity matrix $\mathbf{I}$. Re-estimating the model with this constraint, we get the emission probability distributions reported in Table 6. We use the notation $\overline{S}$ to designate the hidden states estimated using such a constrained transition matrix.

According to the initial probabilities,

$$\pi = \begin{matrix} \overline{F} & \overline{G} & \overline{VG} \\ (0.245 & 0.554 & 0.201) \end{matrix}$$

Danilo Bolano, André Berchtold, Gilbert Ritschard
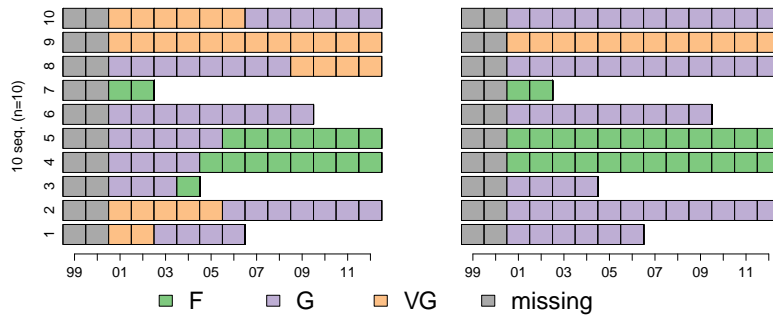
Fig. 4: Sequences of predicted hidden states. Unconstrained HMM (left-hand side) and with identity transition matrix (right-hand side).

we get a probability of 20.1% to belong to the hidden state $\overline{VG}$, 55.4% of being in hidden state $\overline{G}$ and 24.5% to be in "at risk" group $\overline{F}$

If we compare Table 6 with the emission probabilities for the unconstrained HMM (Table 3a), we observe only slight differences. Therefore, the interpretation of the hidden states remains the same and this justifies using a similar labelling of the clusters. Using the constrained transition matrix, the frail hidden state $\overline{F}$—actually a frail trajectory—seems to include a few more individuals with good health conditions than the frail state in the unconstrained case. People classified in this frail 'trajectory' have 37% chances to declare a well or excellent condition at one time point versus 26% for those in the frail state of the unconstraint model.

Figure 4 shows the differences in ten sequences of hidden states estimated using an unconstrained HMM (clusters of person-period states, left panel) and with the constrained transition matrix (clusters of trajectories, right panel). In the right panel where we cluster trajectories, each case remains in the same estimated hidden state for all periods where he/she responded. Looking at the first sequence for example, the individual is, according to the unconstrained HMM, in the hidden states VG for two periods and then, because of a worsening in its health status, he/she moves to hidden states G for the next four periods. According to the constrained model, he/she is estimated to have a good health $\overline{G}$ trajectory.

The appropriate approach should be chosen according to the research question. If the goal is to cluster individuals according to the entire trajectories and to study variation between groups, then a constrained matrix should be used. If we want to focus on the evolution of the situations lived by the individuals such as short periods at risk and recovery from positive or negative shocks, a person-period state clustering, i.e., an unconstraint HMM, is better suited.

A discussion on Hidden Markov Models for Life Course Data

6 Covariates

So far we have seen how Markov models can describe the probability to be in one or the other state of a response variable—the SRH in our illustrative example—in terms of the previously observed states of this same variable or of a latent Markov process. However, it is natural to make this relationship with previous values also depend on external factors so that we can test whether the fitted Markov process remains the same for different values of covariates. For example, we may want to know whether the evolution of the SRH is the same for men and women, or for people with different education levels. Covariates are a concern for both visible Markov models and HMMs.

In the literature, several methods have been considered to account for covariates. Berchtold and Raftery (1999) distinguish between two main approaches: By making the transition probabilities depend on the covariates, e.g., by means of a multinomial regression model, or through the interaction between the previous states and the values taken by the covariates. The first alternative is flexible. It can be used with multiple categorical as well as continuous covariates and also in case of multiple response variables (Bartolucci et al., 2015). However, parametrizing the transition probabilities dramatically increases the complexity of the model, which in turn involves many numerical computation difficulties. For a detail discussion on this approach refer to Bartolucci et al. (2012).

Here, we focus on the second alternative. The advantage of this second approach is its simplicity. It can be applied straightforwardly without modifying the estimation procedure both for Markov chains and Hidden Markov Models. However, this way of doing implies an increase of the number of probabilities to be estimated and works only with categorical covariates.

There are two main alternatives to make the current state depend on interaction between the modalities of categorical covariates and the previous states. We can either estimate directly one, possibly very large, transition matrix or approximate the matrix with a mixture model similar to the Mixture Transition Distribution model model proposed by Berchtold and Raftery (2002) for high-order Markov chains. We focus here on the first alternative where we estimate a single transition matrix with a row for each combination of the values taken by the covariates and the lag of the variable. The estimation of this matrix requires to simply count the number of observed transitions for each combination of modalities of the covariates. This approach is easy to implement however, increases the number of parameters to estimate. For the SRH data with two covariates—e.g. age and education—with each three modalities, the transition matrix would have $5 \times 3 \times 3 = 45$ rows and this number would increase exponentially with the number of variables and proportionally with the number of modalities of each variable. Below we include the covariates in the the three-state hidden model.

Danilo Bolano, André Berchtold, Gilbert Ritschard

6.1 Effect of education level on the SRH process

We illustrate how we can account for covariates by studying the effect of educational level on the SRH trajectories More specifically, we include the covariate in the HMM with three hidden states considered before.

The level of education has been coded into three categories: *Low*, lower secondary level (22.29% of the 13,976 data points); *Medium*, secondary level and vocational school (44%); *High*, high educational level combining high level vocational school, maturity and university degrees (33.71%).[4]

The result of the direct estimation of the HMM with the education covariate is given in Table 7. The emission probabilities are very close from those of the model without covariate (Table 3). Therefore, we can maintain the same interpretation and labels for the hidden states.

From the transition matrix **D**, the probability of falling in the frail hidden state (F) decreases with the level of education. Moreover, less educated people have a probability to already be in a frail situation at the beginning of the sequence (Table 7b) twice bigger than the most educated ones (0.311 versus 0.139). The level of education has a slight positive impact also on chances to recover from a frail condition. For instance, people with high level of education have 2% more chances to move from a frail (F) to a good (G) situation (transition F-G) than those with a lower level of education (7% against 5%). Similarly, the probability of a worsening in the health condition (transition G-F) decreases with the educational attainment. All this indicates that education affects positively the evolution of the health condition.

7 Available software

According to some scholars (Ritschard and Oris, 2005; Scott et al., 2005), the limited use of Markovian models in life course and medical studies is also due to the limited software offer. Even the conventional homogeneous first-order Markov chains are not directly available in standard packages such as SPSS, Stata or SAS, and few dedicated software have been implemented. The March software[5] for categorical variables, for instance, was introduced in 2000 but, even though it offers advanced tools, March is not a free software, it runs only under the Microsoft Windows environment and is no longer maintained.

Fortunately, there are now several R packages that offer functions to model Markovian models. Most of these packages are designed for specific aims (e.g. for univariate time series or with specific type of processes) and not all propose the possibility to include the effect of external covariates. The more interesting R packages for fitting Markov models are `msm` (Jackson, 2011) and `Biograph` (Willekens,

---

[4] The percentages reported reflect the overall distribution of the educational level over the 14 waves considered.

[5] Software available at http://www.andreberchtold.com/march.html

A discussion on Hidden Markov Models for Life Course Data

Table 7: Direct estimation of the HMM with education as covariate.

(a) Emission probability distributions

	F	G	VG
P	0.013	0.000	0.000
B	0.082	0.002	0.002
M	0.647	0.097	0.016
G	0.247	0.844	0.423
V	0.011	0.057	0.559
H CI-Width	0.026	0.015	0.026

(b) Initial hidden state distributions (π)

	F	G	VG
Low	0.311	0.500	0.189
Medium	0.187	0.502	0.311
High	0.139	0.563	0.298

(c) Hidden transition distributions (**D**)

$t-1$	Education	F	G	VG
F	Low	0.950	0.050	0.000
F	Medium	0.949	0.051	0.000
F	High	0.922	0.074	0.005
G	Low	0.049	0.937	0.014
G	Medium	0.035	0.954	0.011
G	High	0.026	0.968	0.006
VG	Low	0.000	0.137	0.864
VG	Medium	0.000	0.091	0.909
VG	High	0.000	0.065	0.935

2014) for multi-state models in continuous time; `markovchain` (Spedicato and Signorelli, 2014) for discrete-time Markov chains; `HiddenMarkov` (Harte, 2010), `seqHMM` (Helske and Helske, 2016), and in particular `depmixS4` (Visser, 2010), `LMest` (Bartolucci et al., 2014) and `march` (the R port of the above mentioned March software, Berchtold, 2014) for discrete-time hidden Markov models. One of the main differences between the last three packages is the type of dependent variable used. In `depmixS4` the models can be fitted on data with distributions from the generalized linear model (glm) family, the (logistic) multinomial, or the multivariate normal distribution (i.e., continuous and discrete outcomes). The packages `LMest` and `march` are explicitly designed for discrete variables. These three packages differ also in the way of including covariates. `depmixS4` and `LMest` use a

Danilo Bolano, André Berchtold, Gilbert Ritschard

parametrization approach. The current version (version 1.0) of `march` does not support covariates, but will propose the method of transition probabilities by level of covariate values as presented in Section 6 in a next release. The empirical examples we provide in the paper have been computed using the Windows version of `march`.

8 Conclusion

The article illustrates the basic aspects and the flexibility of Markovian models with an illustration in life course studies to propose a general discussion on the relevance and possible applications of this class of models.

Starting from the traditional formulation of an homogeneous first order Markov chain, we presented the Hidden Markov Model and an intuitive and easy to understand way of including categorical covariates. Instead of using a parametrization of the transition probabilities, we consider directly the interaction between the states (observed or hidden) and the modalities assumed by the covariates. One of the main features that make Markovian models an interesting approach for life course studies is the specific role of time. The serial dependence between repeated measures is directly taken into account. In this transitional setting, the current measurement is described as a function of previous outcomes (Molenberghs and Verbeke, 2005) so that the distribution of the states depends on the own past of the subject. This aspect is particularly relevant for time-structured data such as individual life trajectories. For example, we have shown the significant relationship between previous self-reported health conditions and the current one.

In a life course perspective, modeling individual sequences at two levels, a visible and a latent one, as in the HMM, proves particularly interesting since many aspects of a life trajectory are not fully observable. For example, in the empirical example we have demonstrated that SRH trajectories are related to a latent variable representing frailty regimes. Moreover, unlike conventional latent variable approaches such as latent class or mixed effect models, HMM is a time-varying model. In many applications, the interest is not only to analyze the inter-individual differences in the response variable, but also the way in which individuals change their responses over time. In a HMM approach, the unobservable characteristic that drives the observed behavior has its own dynamics following a Markov process. The HMM then explore the dynamics in unobserved aspects which are measured by one or more response variables. The HMM can also be used as a clustering tool. HMM is a generalization of the mixture model (e.g., McLachlan and Peel, 2000) where each component is associated to one of the hidden states of the model. In particular, HMM can perform a probabilistic clustering in two ways. In its conventional formulation with a time-varying latent variable individuals can move among latent classes and we have a clustering of individual state observations. With a time-invariant latent variable (i.e., a HMM with each hidden state being fully absorbing), HMM performs a static probabilistic clustering of the individuals according to their entire observed sequences that can be seen as a clustering of the trajectories.

A discussion on Hidden Markov Models for Life Course Data

Although we focused here on the case of a single categorical outcome variable (the self-rated health condition), hidden Markov models can be applied to multivariate data (Bartolucci et al., 2012) and to numeric outcome variables. Bolano and Berchtold (2016) for example considered a double chain Markov model for numeric outcomes.

Interesting extensions of the HMM framework we did not discuss in the paper are the Double Chain Markov Models (see Berchtold, 1999)) that allow to relax the homogeneity assumption keeping the model parsimonious in terms of free parameters and the introduction of mixed effects (the so called Mixed Hidden Markov Model, Altman, 2007; Maruotti, 2011). By including (individual specific) random effects, the mixed HMM relaxes the assumption of conditional independence of the observations (Eq 1c). The resulting model is more flexible than a conventional HMM and it allows to distinguish between two sources of heterogeneity. The random effects capture the between-subject variations and the hidden states capture the heterogeneity in the individual trajectories.

Some aspects limit the diffusion of Markovian models in life course studies and related fields. For example, in social sciences and medical studies, a key aspect is to analyze the effect of external factors on the dynamics of the dependent variable. In the framework of hidden Markov models, although several alternatives to include covariates have been considered in the literature, an easily usable framework for estimating the effect of the covariates is still laking. Another limitation concerns the readability of the results in particular in presence of multiple states and/or multiple covariates. Comprehensive data visualization tools for representing empirical results are crucially missing too. However, given the numerous potential applications of Markovian models, and given the facilities offered by recently released software, we can expect Markovian modelling to overcome the above mentioned issues and this way to gain popularity also in life course research.

Acknowledgments

References

Altman, R. M. (2007). Mixed hidden markov models: An extension of the hidden markov model to the longitudinal data settings. *Journal of the American Statistical Association*, 102(477):201–210.

Danilo Bolano, André Berchtold, Gilbert Ritschard

Bartolucci, F., Farcomeni, A., and Pennoni, F. (2012). *Latent Markov Models for Longitudinal Data*. Chapman and Hall/CRC Press. Statistics in the Social and Behavioral Sciences, New York, NY.

Bartolucci, F., Farcomeni, A., and Pennoni, F. (2014). The r package "lmest". comprehensive r archive network cran http://cran.r-project.org/web/packages/lmest/.

Bartolucci, F., Montanari, G. E., and Pandolfi, S. (2015). Three-step estimation of latent markov models with covariates. *Computational Statistics and Data Analysis*, 170(1):115–132.

Bartolucci, F., Pennoni, F., and Francis, B. (2007). A latent markov model for detecting patterns of criminal activity. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 170(1):115–132.

Baum, L. E. and Petrie (1966). Statistical inference for probabilistic functions of finite state markov chains. *The Annals of Mathematical Statistics*, 37 (6):1554–1563.

Berchtold, A. (1999). The double chain markov model. *Communications in Statistics: Theory and Methods*, 28(11):2569–2589.

Berchtold, A. (2014). The r package "march". comprehensive r archive network cran http://cran.r-project.org/web/packages/lmest/.

Berchtold, A. and Raftery, A. (1999). The mixture transition distribution (mtd) model for high-order markov chains and non-gaussian time series. Technical Report 360, Department of Statistics - University of Washington.

Berchtold, A. and Raftery, A. (2002). The mixture transition distribution model for high-order markov chains and non-gaussian time series. *Statistical Science*, 17(3):328–359.

Bicego, M., Murino, V., and Figueiredo, A. T. (2003). Similarity-based clustering of sequences using hidden markov models. *Machine Learning and Data Mining in Pattern Recognition. Lecture Notes in Computer Science*, 2734:pp 86–95.

Bijleveld, C. C. and Mooijaart, A. (2003). Latent markov modelling of recidivism data. *Statistica Neerlandica*, 57(3):305–320.

Bolano, D. and Berchtold, A. (2016). General framework and model building in the class of hidden mixture transition distribution models. *Computational Statistics & Data Analysis.*, 93:http://dx.doi.org/10.1016/j.csda.2014.09.011.

Caswell, H. (2009). Stage, age and individual stochasticity in demography. *Oikos*, 118(12):1763–1782.

Celeux, G. and Durand, J.-B. (2008). Selecting hidden markov model state number with cross-validated likelihood. *Computational Statistics*, 23(4):541–564.

Craig, B. A. and Sendi, P. P. (2002). Estimation of the transition matrix of a discrete-timemarkov chain. *Health economics*, 11(1):33–43.

Elliott, R. J., Hunterb, W. C., and Jamieson, B. M. (1998). Drift and volatility estimation in discrete time. *Journal of Economics Dynamics and Control*, 22:209–218.

Feichtinger, G. (1973). Markov models for some demographic processess. *Statistische Hefte*, 14(4).

Harte, D. S. (2010). R package "hiddenmarkov". comprehensive r archive network cran http://cran.r-project.org/web/packages/hiddenmarkov/.

A discussion on Hidden Markov Models for Life Course Data

Hayashi, T. (2004). A discrete-time model of high-frequency stock returns. *Quantitative finance*, 4:140–150.

Helske, S. and Helske, J. (2016). Mixture hidden markov models for sequence data:the seqhmm package in r.

Hodge, R. W. (1966). Occupational mobility as as probability process. *Demography*, 3(1):19–34.

Jackson, C. H. (2011). Multi-state models for panel data: The msm package for r. *Journal of Statistical Software*, 38(8):1–28.

Land, K. C. and Rogers, A. (1982). *Introduction to multiregional mathematical demography*. Academic Press, Waltham, MA.

Le Strat, Y. and Carrat, F. (1999). Monitoring epidemiologic surveillance data using hidden markov models. *Statistics in medicine*, 18(24):3463–78.

Maruotti, A. (2011). Mixed hidden markov models for longitudinal data: An overview. *International Statistical Review*, 79(3):427–454.

McFarland, D. D. (1970). Intragenerational social mobility as a markov process: Including a time-stationary markovian model that explains observed declines in mobility rates over time. *American Sociological Review*, 35(4):465–43.

McLachlan, G. J. and Peel, D. (2000). *Finite Mixture Models*. Wiley Series in Probability and Statistics, New York, NY.

Molenberghs, G. and Verbeke, G. (2005). *Models for Discrete Longitudinal Data*. Springer-Verlag. Series in Statistics, Berlin.

Netzer, O., Lattin, J. M., and Srinivasan, V. (2008). A hidden markov model of customer relationship dynamics. *Marketing Science*, 27(2):185–204.

Rabiner, L. I. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77:257–286.

Ritschard, G. and Oris, M. (2005). Life course data in demography and social sciences: Statistical and data-mining approaches. In Levy, R., Ghisletta, P., Le Goff, J.-M., Spini, D., and Widmer, E., editors, *Towards an Interdisciplinary Perspective on the Life Course. Advances in Life Course Research*, volume 10, pages 283–314.

Rogers, A. (1975). *Introduction to multiregional mathematical demography*. Wiley, New York, NY.

Scott, S. L., James, G. M., and Sugar, C. A. (2005). Hidden markov models for longitudinal comparisons. *Journal of the American Statistical Association*, 100(470):359–369.

Shirley, K. E., Small, D. S., Lynch, K. G., Maisto, S. A., and Oslin, D. W. (2010). Hidden markov models for alcoholism treatment trial data. *The Annals of Applied Statistics*, 4(1):366–395.

Spedicato, G. A. and Signorelli, M. (2014). The r package "markovchain": Easily handling discrete markov chains. comprehensive r archive network cran http://cran.r-project.org/web/packages/markovchain/.

Visser, I. (2010). depmixs4 : An r package for hidden markov models. *Journal of Statistical Software*, 36(7):1–21.

Visser, I., Raijmakers, M. E. J., and Molenaar, P. C. M. (2002). Fitting hidden markov models to psychological data. *Scientific Programming*, 10:185–199.

Danilo Bolano, André Berchtold, Gilbert Ritschard

Viterbi, A. J. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 16(2):260–269.

Voorpostel, M., Tillmann, R., Lebert, F., Kuhn, U., Lipps, O., Ryser, V.-A., Schmid, F., Rothenbühler, M., and Boris, W. (2013). Swiss household panel user guide. *Swiss Household Panel Userguide (1999-2012), Wave 14, November 2013. Lausanne FORS*.

Wiggins, L. M. (1973). *Panel Analysis: Latent Probability Models for Attitude and Behavior Processes*. Elsevier Scientific Pub, Amsterdam, Netherlands.

Willekens, F. (2014). *Multistate Analysis of Life Histories with R*. Springer-Verlag. Use R!, Berlin.

Zucchini, W. and MacDonald, I. L. (2009). *Hidden Markov models for time series. An introduction Using R*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability, New York, NY.

Session 6A: Transition to adulthood

Russian Generations: Sequencing the Transition to Adulthood

Ekaterina S. Mitrofanova

Institute of Demography, National Research University Higher School of Economics

Abstract This paper demonstrates how Russian men and women from different generations undertake their first demographic and socio-economic events. For the first attempt at understanding how these events are ordered, we used the descriptive techniques of Sequence Analysis, including chronograms, parallel coordinate plots, the durations of statuses, and the frequencies of subsequences on tables. The study was performed on a panel of the Russian part of the "Generations and Gender Survey" (GGS: 2004, 2007, and 2011). The subsample consists of 4,595 respondents of 1935-1984 years of birth.

Our analysis reveals the changes between sexes and generations. Men devote a significant part of their youth to achieving socio-economic events, while women much earlier and more actively initiate their demographic careers. Nevertheless, by the age of 35, there are more respondents among men who have children and relationships than among women. Young people, compared to older generations, much more actively enter into cohabitations and have children in them, but they delay the onset of all these events, especially childbearing, to later ages.

1 Introduction

The transition to adulthood (or adolescence, or early adulthood) is a stage of life course when person experience biological, emotional, cognitive, and societal maturation. As sociologists and demographers, we are interested in the last component of this complicated process.

Transition to adulthood is one of the important status passages in a life course, when many new roles and statuses come with occurring events in different spheres of life: leaving the parental house, finishing education, entrance into the job market, family formation. The sequence and timing of onset of these events are changing from generation to generation. It is conditioned by the demographic, cultural, economic, political transformations of societies (Buchmann 1989; Kaa 1987; Lesthaeghe 1995; Liefbroer 1999).

The behavior of Russians in socio-economic and demographic spheres has undergone many changes over the past decade (Frejka и Zakharov 2012; Mitrofanova 2013; Mitrofanova и Artamonova 2014; Zakharov 2008). It is visible especially in

terms of the transformation of life course starting events, as these changes appear in the biographies of young people most quickly and help to define the difference between generations before the completion of reproduction, marriage and other "careers".

In this paper, the biographies of Russians were studied through sequence analysis. We also promote the author's approach to data visualization using Lexis grids.

2 Database and methods

The study was performed on a panel of the Russian part of the "Generations and Gender Survey" (GGS: 2004, 2007, 2011). The subsample consists of 4,595 respondents (32% men and 68% women)[1]. It has been taken across five 10-year generations of 1935-1984 years of birth. Based on empirical data and existing classifications, we define the cohorts of 1935-1974 years of birth as "Soviet generations" (those who socialized in the Soviet Union), and cohorts of 1975-1984 years of birth as "modern generations".

The life course of an individual consists of a set of different events, which may happen sequentially or immediately. It is difficult to explore this multiplicity of "clocks"[2] (marriage, reproductive, labor, educational, etc.) by conventional methods, whereby the events are analyzed either individually or in small groups. The transition to the study of event chains makes it possible to achieve a new level of understanding of the structure of individuals' lives. An advanced method known as Sequence Analysis (SA) helps demographers and sociologists to achieve this aim (Abbott и Tsay 2000; Aisenbrey и Fasang 2007, 2010; F. C. Billari 2001; F. Billari и Piccarreta 2005). We used several descriptive techniques of SA: chronograms, parallel coordinate plots, the durations of statuses, and the frequencies of subsequences on tables.

3 Results

We analyzed starting life course events, which we grouped according to three dimensions (corresponding statuses are in parentheses)[3]: the presence of children (no children, first child), marital status (single, first cohabitation, first marriage),

---

[1] The gender imbalance is due to the "rash" of the sample and the inability to correct the panel data by weights.

[2] "Multiple clocks" is a term of the Life Course concept which includes the notion of a life as a set of different spheres, each of which has its own timing (Mills 2000).

[3] We used the R package TraMineR (Gabadinho et al. 2011).

3

socio-economic status (no events, first separation from parents, first job, and completed education of the highest level). The number of combinations of statuses is very high, so in order to reduce their amount, we focused only on the first event in pairs of socio-economic events, and on the last one in the triple events. The list of determined statuses are specified and their codes are shown in Figure 1. Grey color indicates censored events which have not yet occurred at the time of the survey. The censoring is possible because the representatives of the youngest generation are 27-36 years old, so only two cohorts of ten (from this generation) can have events at the age of 35.

Socio-economic events	Demographic events					
	no children			1st child		
	single	1st cohabitation	1st marriage	single	1st cohabitation	1st marriage
no events or one event	SC00 (no events) / SC0L (separation from parents) / SC0J (job) / SC0E (education)	P1C01	M1C01	SC11	P1C11	M1C11
separation from parents > some event	SC0L+	P1C0L+	M1C0L+	SC1L+	P1C1L+	M1C1L+
job > some event	SC0J+	P1C0J+	M1C0J+	SC1J+	P1C1J+	M1C1J+
education > some event	SC0E+	P1C0E+	M1C0E+	SC1E+	P1C1E+	M1C1E+
2 events concurrently	SC02	P1C02	M1C02	SC12	P1C12	M1C12
2 events > separation from parents	SC0++L	P1C0++L	M1C0++L	SC1++L	P1C1++L	M1C1++L
2 events > job	SC0++J	P1C0++J	M1C0++J	SC1++J	P1C1++J	M1C1++J
2 events > education	SC0++E	P1C0++E	M1C0++E	SC1++E	P1C1++E	M1C1++E
3 events concurrently	SC03	P1C03	M1C03	SC13	P1C13	M1C13
censoring						

Fig. 1: Groups of statuses.

Using the information on the data occurrences, we reconstructed the segments of the respondents' biographies. We created statuses for each month from age 15 to age 35 for each respondent. We chose the age of 15 as the margin of childhood and capped the observation period at the age of 35 to equalize the chances of different generations in terms of the events occurring, and to exclude marginal cases (since the first events most likely occur in the first half of life).

We obtained the frequency distribution of occurrences of different statuses at any given time for each generation; this distribution became the framework for building chronograms representing these frequencies for men (Fig. 2) and for women (Fig. 3). Presented chronograms were placed on Lexis grids, thus allowing for comparison across the three time dimensions: the X-axis – the calendar date, the Y-axis – age, and the diagonal – generation. The X-axis represents the proportion of delayed

4

status at each particular time inside the corridors of each generation; the Y-axis depicts ages from 15 to 35 years.
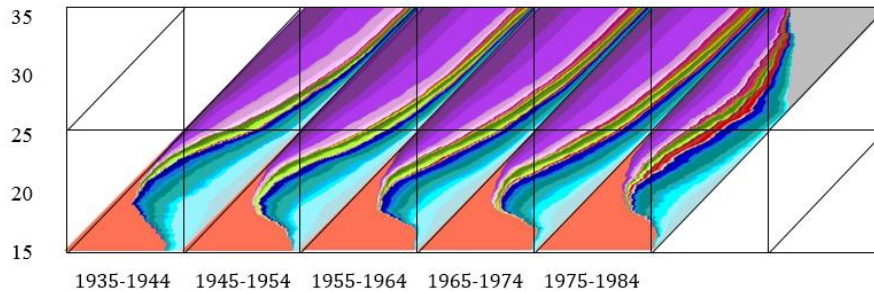
Fig. 2: Chronograms for men of 1935-1984 years of birth.
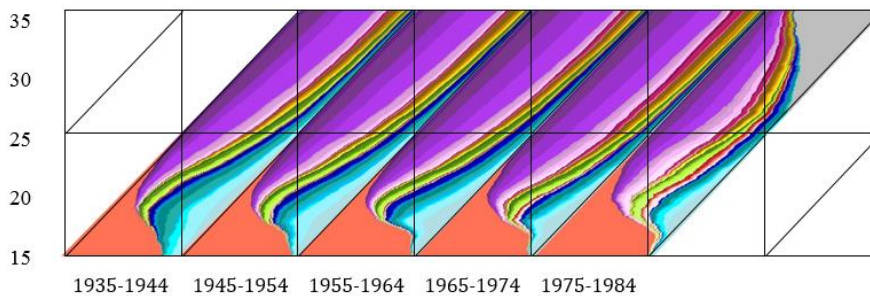
Fig. 3: Chronograms for women of 1935-1984 years of birth.

We can make the following observations according to the data depicted on the chronograms. The form of the start of biographies (the "neck", from which the colors appear) indicates that, at the age of 15, the older generations had a much larger number of events than the younger generations; for older generations, such events were mainly socio-economic, while young generations face more demographic ones. Almost 90% of men of the older generations experience exclusively socio-economic events (blue palette) by the age of 20. By about the same age, men of other generations begin to acquire demographic events, but only 70-80% of them have socio-economic ones. Women begin demographic careers two years earlier and have approximately 35-65% of socio-economic events by this age.

At the age of 35, 70-95% of men born in 1935-1974 are married and have at least one child (purple palette), while among women from similar generations there are less than 80% with such statuses, and 8% are in cohabitation with a child (pink palette). There are 10-20% of women who have a child and are not married or cohabiting (yellow colors). In contrast, a man with a child is almost always a man in a relationship. Among the men and women of the younger generations, there are only 10% of those who are married and have a child, but these representatives have had very few other events, because only a small portion of the respondents have reached the age of 35. For both sexes, when we move to younger generations, there

is tendency towards reduction in the share of those who are married and an increase in the share of those who are in cohabitation.

All generations include respondents who do not have any demographic event by the age of 35, but their share does not exceed 5%. The most popular final socio-demographic event men achieve at the age of 35 is the first separation from their parents; a bit less popular is education. The first job started to compete with education only for the generation of 1965-1974 years of birth. For Soviet women, there is an identical structure to the final socio-economic events, but, starting from the generation of 1955-1964 years of birth, the shares of each event will turn out to be equal. To clarify these statements, we placed the chronograms of the individual biographies of respondents, sorted "from the end", in Appendix 1.

There are the mean durations of states in Table 1. We included in this table only the events which lasted three or more months. The demographic events are indicated by bold print.

Table 1. Mean durations of states for both sexes, men, women.

#	Both sexes		Men		Women	
	Status	**Duration, months**	**Status**	**Duration, months**	**Status**	**Duration, months**
1	SC00	34.0	SC00	36.2	**M1C1++L**	33.6
2	**M1C1++L**	33.7	**M1C1++L**	34.1	SC00	33.0
3	**M1C1++J**	23.5	**M1C1++E**	19.5	**M1C1++J**	26.0
4	**M1C1++E**	20.5	**M1C1++J**	18.1	**M1C1++E**	21.0
5	SC0J	10.0	SC0J	12.1	**M1C1J+**	9.5
6	**M1C1J+**	9.3	SC0J+	11.7	SC0J	9.0
7	SC0L	8.8	SC0E+	10.4	SC0L	8.5
8	SC0J+	8.6	SC0L	9.5	SC0J+	7.1
9	SC0E+	8.1	**M1C1J+**	9.0	SC0E+	7.0
10	**M1C13**	5.9	SC0++J	6.7	**M1C13**	6.8
11	**M1C1L+**	5.8	SC0++L	6.1	**M1C1L+**	6.2
12	SC0++J	5.6	SC0E	5.5	SC0++J	5.0
13	**M1C1E+**	4.4	SC0L+	4.9	**M1C1E+**	4.8
14	**M1C0++L**	4.1	**M1C1L+**	4.8	SC1++L	4.7
15	SC0L+	4.0	**M1C0++L**	4.4	**M1C0++L**	4.0
16	SC0++L	3.9	SC0++E	4.0	SC0L+	3.5
17	SC0E	3.8	**M1C13**	3.9	SC1++E	3.5
18	SC1++L	3.7	**M1C1E+**	3.7	**M1C11**	3.5
19	SC0++E	3.5	**M1C0++J**	3.4	SC0++E	3.3
20	**M1C11**	3.3	SC02	3.2	SC1++J	3.2
21	**M1C0++J**	3.0			SC0E	3.0

The mean duration of an average event for both sexes is 5.7 months: for men it is 7.6 months, and for women it is 4.8 months. The ranking shows that the longest state for men (more than 3 years) is an absence of events of all types. The second one (less than 3 years) is first marriage, first child and all the three events, the third of which is leaving the parental home. For women, we have the opposite situation, but the difference between the duration of the events is less than a month.

6

The next two events are also the same for men and women, but they appear in opposite succession. Men have all the events, with the last one – education – for more than a year and a half and the last one – the first job – for a period of a month shorter. Women are staying longer in the same statuses, but the longest one ends with the first job (more than 2 years), the next one – with education (less than 2 years). Other statuses are incomplete (respondents do not have the events of all the types) and the duration of events is less than a year. Nevertheless, we revealed that women stay in demographic statuses longer than men.

In Figure 4, there are so called "parallel coordinates" which indicate the transitions of respondents from one status to another. The X-axis represents the number of transitions, the Y-axis depicts each possible status from SC00 (no events) to M1C13 (1st marriage, 1st child, all socio-economic events). The horizontal lines are dividing different groups of statuses from each other.

7

8
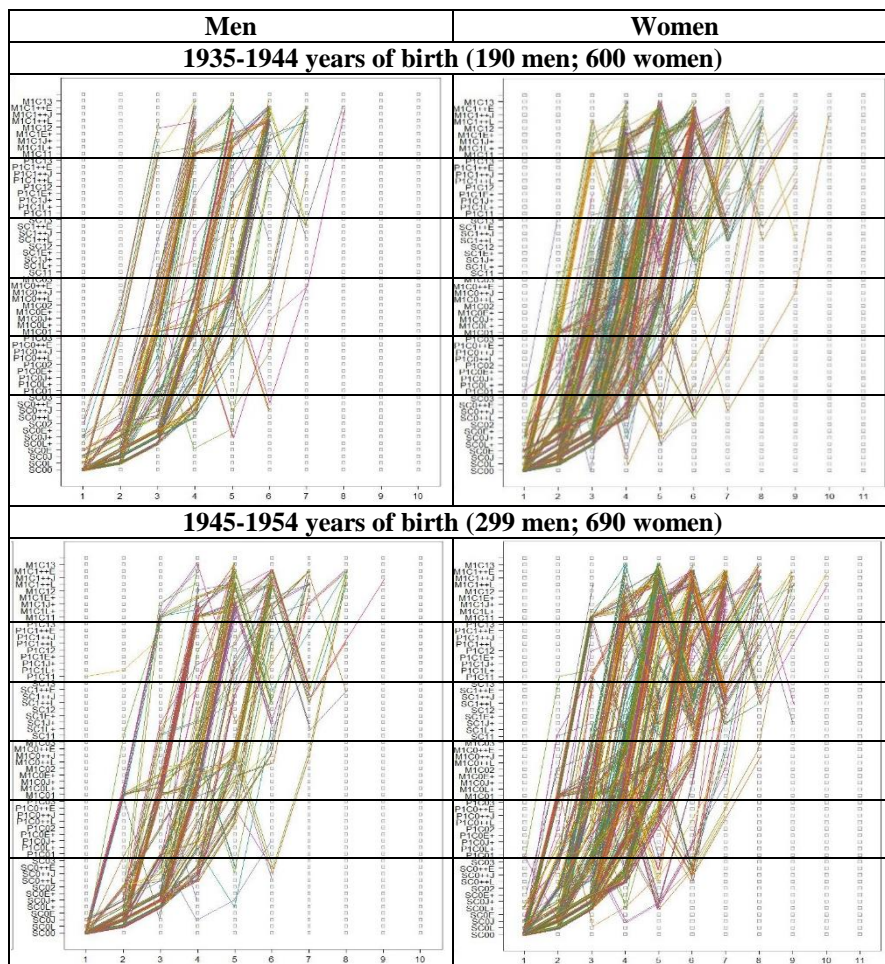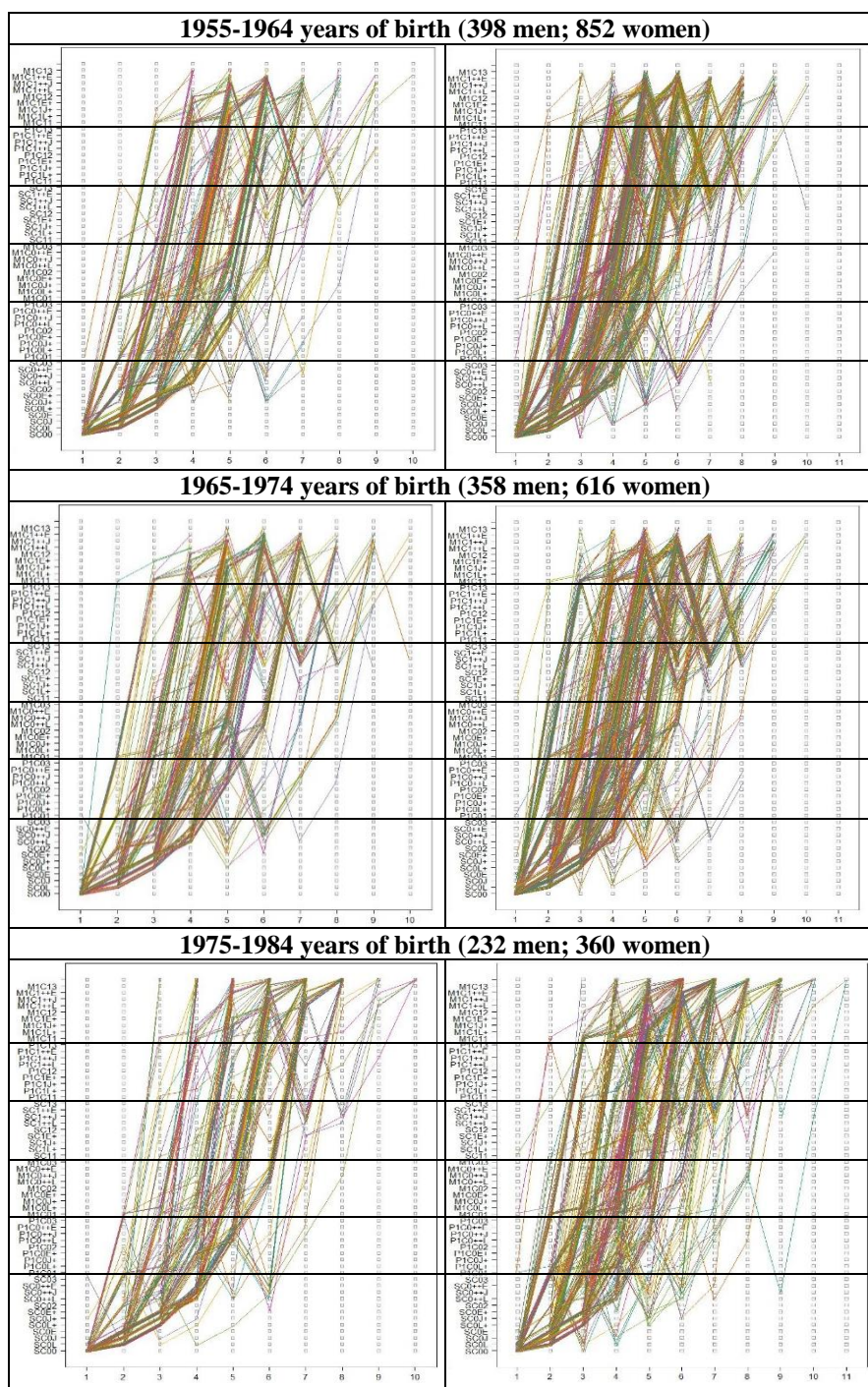
Fig. 4: "Parallel coordinates" for men and women of each generation.

Due to the imbalance between the number of men and women, there is much more information for women than for men, though this does not affect the results: the typical transitions are seen in both cases. The first transition (from status one to status two) is often from the position with no statuses to some socio-economic positions. Sometimes we can see individuals gaining a marriage or a partnership. The trend of all the graphs is a sharp rise, so people are achieving demographic statuses relatively rapidly. The "speed" is slowing in modern generations, especially among men: they are entering their marriage "careers" only after third step.

The most frequent subsequences are listed in Table 2. We have information for both sexes, men and women. We excluded the subsequences with a support (the amount of people of a group, who have the subsequent) of less than 10%.

Table 2. Frequencies of appearances of subsequences for both sexes, men and women.

#	Both sexes			Men			Women		
	Subsequence	Support	Count	Subsequence	Support	Count	Subsequence	Support	Count
1	(SC00)	91%	4438	(SC00)	93%	1441	(SC00)	90%	2997
2	(SC00>SC0J)	27%	1306	(SC00>SC0J)	29%	453	(SC00>SC0J)	26%	853
3	(SC00>SC0L)	25%	1208	(SC00>SC0E)	27%	415	(SC00>SC0L)	25%	828
4	(SC00>SC0E)	23%	1120	(SC00>SC0L)	25%	380	(SC00>SC0E)	21%	705
5	(SC0E>SC0E+)	20%	956	(SC00)-(SC0E>SC0E+)	23%	357	(SC0L>SC0L+)	19%	632
6	(SC0L>SC0L+)	19%	946	(SC0J>SC0J+)	23%	354	(SC0E>SC0E+)	18%	590
7	(SC0J>SC0J+)	19%	929	(SC0L>SC0L+)	20%	314	(SC0J>SC0J+)	17%	575
8	(SC00)-(SC0E>SC0E+)	18%	886	(SC00)-(SC0J>SC0J+)	20%	313	(SC00)-(SC0L>SC0L+)	17%	548
9	(SC00)-(SC0L>SC0L+)	17%	835	**(M1C0++L>M1C1++L)**	19%	295	**(M1C0++L>M1C1++L)**	16%	532
10	**(M1C0++L>M1C1++L)**	17%	827	(SC00)-(SC0L>SC0L+)	19%	287	(SC00)-(SC0E>SC0E+)	16%	529
11	(SC00)-(SC0J>SC0J+)	17%	826	**(SC00)-(M1C0++L>M1C1++L)**	17%	269	(SC00)-(SC0J>SC0J+)	15%	513
12	**(SC00)-(M1C0++L>M1C1++L)**	15%	726	**(SC00)-(M1C0++J>M1C1++J)**	10%	152	**(SC00)-(M1C0++L>M1C1++L)**	14%	457
13	**(SC00)-(M1C0++J>M1C1++J)**	11%	511				**(SC00)-(M1C0++J>M1C1++J)**	11%	359
14	**(M1C0J+>M1C1J+)**	10%	464				(SC0L>SC0L+)-(SC0L+>SC0++J)	10%	323
15							**(M1C01>M1C11)**	10%	318

The table shows that women have more events with a support of more than 10%, but the percentage of people in socio-economic statuses is more common among men than women. The majority of respondents (90%) have the subsequence when there are no events. This means that 10% of people have some events at the age of

9

15. The next popular subsequences (more than 20% each) are transitions from no-events status to the first job, education and leaving parents. 19% of men and 16% of women have a subsequence, which includes demographic events: it is a transition from "the first marriage without children, but with all the socio-economic events (last is leaving parents)" to "the first marriage with the first child, and with all the socio-economic events (last is leaving parents)".

4 Conclusion

This paper illuminates how men and women from different generations gaining their first demographic and socio-economic events. We represented the individual biographies on chronograms, the status transitions on parallel coordinate plots and the durations of statuses and the frequencies of subsequences on tables.

Our analysis reveals the changes between sexes and generations. Men devote a significant part of their youth to achieving socio-economic events, while women much earlier and more actively initiate their demographic careers. Nevertheless, by the age of 35, there are more respondents among men who have children and relationships than men who do not face such events. Young people, compared to older generations, much more actively enter into cohabitations and have children in them, but they delay the onset of all these events, especially childbearing, to later ages.

Acknowledgements

10

References

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

Aisenbrey, S., & Fasang, A. (2007). Beyond optimal matching: the 'second wave' of sequence analysis. *New Haven: Center for Research on Inequalities and the Life Course, Yale University*.

Aisenbrey, S., & Fasang, A. E. (2010). New Life for Old Ideas: The «Second Wave» of Sequence Analysis Bringing the «Course» Back Into the Life Course. *Sociological Methods & Research*, *38*(3), 420–462.

Billari, F. C. (2001). Sequence Analysis in Demographic Research. *Canadian Studies in Population*, *28*(2), 439–458.

Billari, F., & Piccarreta, R. (2005). Analyzing Demographic Life Courses through Sequence Analysis. *Mathematical Population Studies*, *12*(2), 81–106.

Buchmann, M. (1989). *The script of life in modern society: entry into adulthood in a changing world*. Chicago: University of Chicago Press.

Frejka, T., & Zakharov, S. (2012). *Comprehensive analyses of fertility trends in the Russian Federation during the past half century*. Max Planck Institute for Demographic Research, Rostock, Germany.

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

Kaa, D. J. van de. (1987). Europe's Second Demographic Transition. *Population Bulletin*, *42*, 3–57.

Lesthaeghe, R. (1995). The second demographic transition in Western countries: An interpretation. B *Gender and family change in industrialized countries* (cc. 17–62). Oxford: Clarendon Press.

Liefbroer, A. C. (1999). From Youth to Adulthood: Understanding Changing Patterns of Family Formation From a Life Course Perspective. B *Van Wissen L J G, Dykstra P A (eds.) Population Issues, an Interdisciplinary Focus* (cc. 53–85). The Netherlands: Kluwer Academic.

Mills, M. (2000). Providing space for time - The impact of temporality on life course research. *TIME & SOCIETY*, *9*(1), 91–127.

Mitrofanova, E. S. (2013, Spring). Demographic behaviour of Russians: family and fertility patterns across generations. *The Macrotheme Review*, *2*(3), 71–80.

Mitrofanova, E. S., & Artamonova, A. V. (2014). Is cohabitation an alternative to marriage in Russia? *Social Science Research Network*.

Zakharov, S. (2008). Russian Federation: From the first to second demographic transition. *Demographic Research*, *19*, 907–972.

11

Appendix

Appendix 1. Individual biographies of the respondents sorted "from the end".

The biographies of all the respondents, who were sorted "from the end" (i.e. from the last event in the segment of the biography), are displayed below. We chose this medium of presentation to fix the sets of the first events, which people of different sexes and generations experienced at the age of 35. In addition, we can trace the individual tracks to the final events.

12

1955-1964 years of birth (398 men; 852 women)

1965-1974 years of birth (358 men; 616 women)

1975-1984 years of birth (232 men; 360 women)
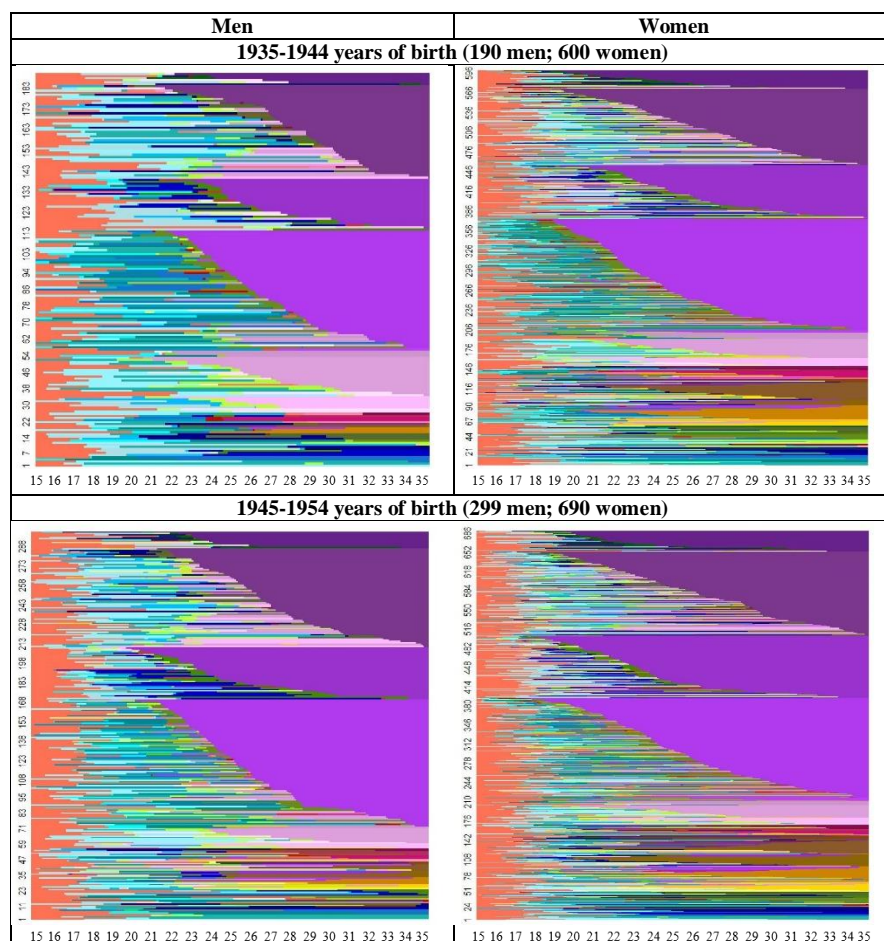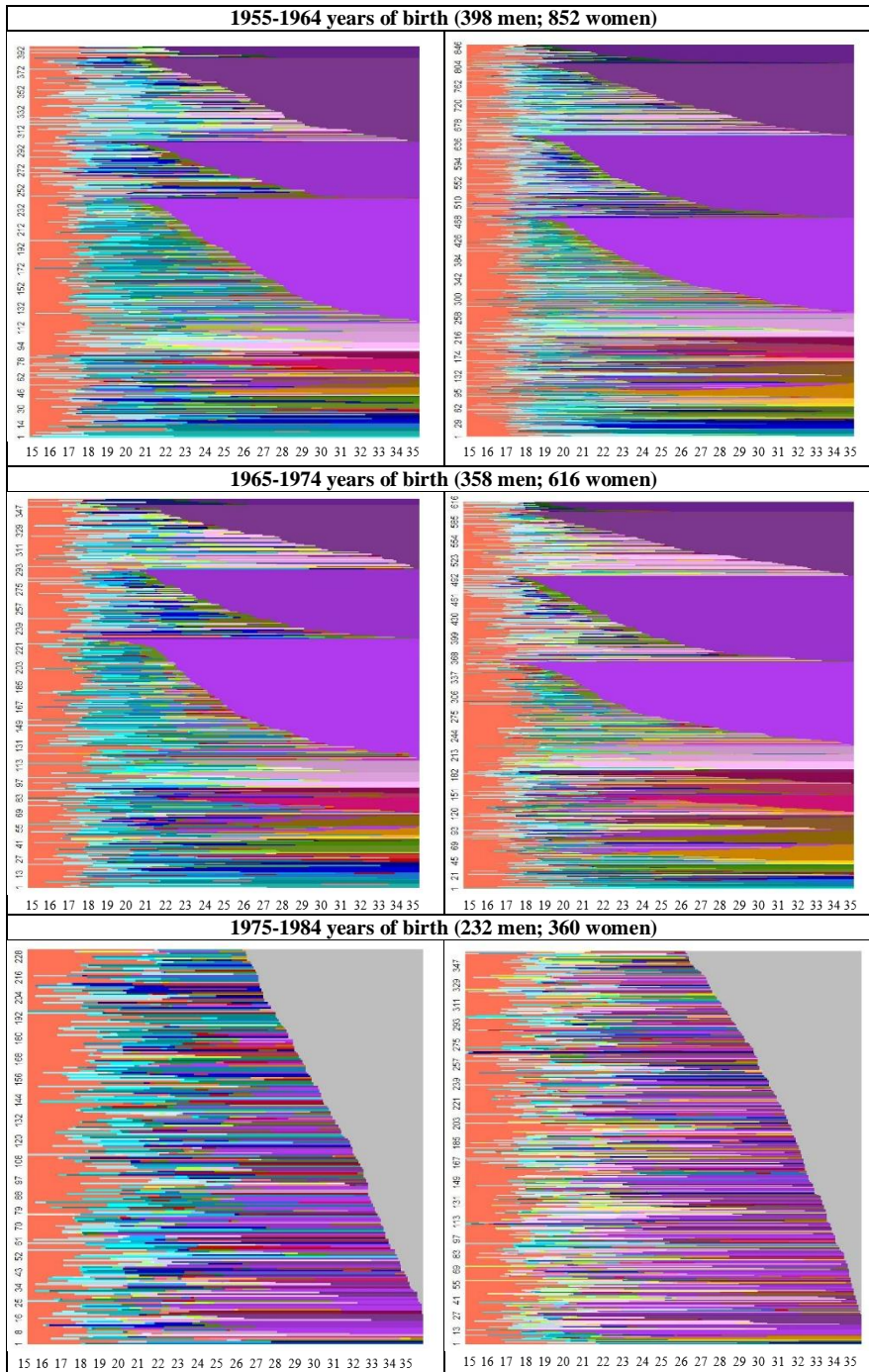
On the road to success? The intergenerational transmission of disadvantage through the transition to adulthood

Jarl Mooyaart[a,b]

Aart C. Liefbroer[a,c,d]

Francesco C. Billari[e]

[a]*Netherlands Interdisciplinary Demographic Institute, the Hague*

[b]*University of Groningen*

[c]*Department of Epidemiology, University Medical Centre Groningen (UMCG) / University of Groningen*

[d]*Department of Sociology, VU University Amsterdam*

[e]*University of Oxford*

[Draft prepared for LaCOSA II conferene, June 2016]

[Please do not cite or distribute without permission of authors]

Abstract

How does the intergenerational transmission of disadvantage come about? This study aims to broaden our understanding by examining the extent to which income trajectories in later stages of young adulthood are influenced by the work- and family-related pathways young people take into adulthood. The transition to adulthood is a demographically dense period, in which individuals make important decisions regarding their future career and family life, which in turn are likely to have a large impact on their future earnings. This study assesses to what extent the influence of family background, in terms of parental income, education,

family structure and race, is mediated by the career and demographic pathways that youths choose during the transition to adulthood. It is examined to what extent incomes diverge between those opting for different pathways to adulthood and whether within groups choosing for the same pathway to adulthood, family background remains to have an influence on these income trajectories. This study uses panel data from the National Longitudinal Survey of Youths of 1997 (N=4966). Sequence analysis is used to define different career (based on education and employment) and demographic pathways (based on household, relationship and parenthood status) between age 17 and 25, separately for men and women. The family background variables and the different clusters are included as categorical variables in a growth curve model, with annual income between age 25 and 32 as the dependent variable. Results indicate that the effects of family background variables mostly disappear once the career and demographic clusters are included. Career pathways appear to be more important in explaining differences in income trajectories in early adulthood than demographic pathways. Incomes diverge for individuals who are in career clusters with longer college enrollment compared to those who are in clusters that have little college education.

Introduction

Is America still the land of opportunity? There is an ongoing debate in society and social sciences on whether children of all social backgrounds have the opportunity to have a decent life . An important indicator of whether someone is successful in life is income. Do the opportunities for youths of disadvantaged background to escape from poverty increase or decrease? Research on intergenerational income mobility does not find that the United States has become a more open society in the last decades. Aaronson & Mazumder (2008) even find

a decline in income mobility, although others do not find a clear trend (Hauser, 2010; Hertz, 2007; Lee & Solon, 2009). Chetty et al. (2014) claim that while rank-based mobility has remained stable, the differences between the ranks have increased, from which they infer that the social class that children are raised in has become more important. Whether intergenerational mobility has increased or not, in comparison to other countries the United States shows a strong intergenerational gradient (Corak, Lindquist, & Mazumder, 2014; Ermisch et al, 2012).

While classic sociological research on intergenerational transmission of socio-economic status has mainly focused on the role of parental investment in education (Becker & Tomes, 1979; Blau & Duncan, 1967; Breen & Goldthorpe, 1997), more recently, increased attention is paid to the influence of family structure on intergenerational mobility (Amato, Booth, McHale, & Van Hook, 2015; Putnam, 2015). McLanahan (2004) claims that destinies of children with high and low educated mothers are diverging. She shows how children with low maternal education have increasingly fewer resources at their disposal as they are more likely to be raised by their mother alone, therefore missing out on resources that are provided by the father. Those who are raised by single low educated mothers do not only receive less financial investment (Kornrich & Furstenberg, 2013), but also less childcare (Kalil, Ryan, & Corey, 2012). Part of the intergenerational transmission of disadvantage may also be the result of the children of these fragile families making the same choices regarding family formation.

This study contributes to the literature on intergenerational transmission of income by examining how differences in the pathways into adulthood can explain divergence in income inequality between contemporary youths from advantaged and disadvantaged backgrounds, in terms of both socio-economic status and family structure. Although there have been studies that have aimed to identify different work-life pathways to adulthood (Garrett & Eccles,

2009; P. Martin, Schoon, & Ross, 2008; Oesterle, David Hawkins, Hill, & Bailey, 2010; Osgood, Ruth, Eccles, Jacobs, & Barber, 2005; Salmela-Aro, Kiuru, Nurmi, & Eerola, 2011), no previous research has yet linked these transitions to (early) adult life outcomes. Naturally, schooling and employment decisions in young adulthood can have important implications for one's earnings later in adulthood. However, demographic decisions regarding the timing and ordering of events, such as leaving the parental home, relationship formation (marriage or cohabitation) and parenthood as suggested by the diverging destinies literature may also have an impact on one's future income. By examining these pathways simultaneously we can have a better understanding to what extent intergenerational inequality is reproduced by career and demographic pathways in young adulthood.

In this study the following research questions are addressed: 1) To what extent are young adults' social backgrounds related to their income trajectories during the later stages of young adulthood? 2) To what extent is this relationship mediated by their career and demographic pathways during the early stages of young adulthood? The transition to adulthood may be the life-phase in which youths of advantaged background realize an advantaged position for themselves. On the other hand, youths from disadvantaged background could benefit from following a career and demographic pathway that is associated with better income trajectories. This study examines income trajectories in young adulthood in order to assess whether the destinies of young adults are diverging. Furthermore, it is investigated whether there is a cumulative advantage for those who come from an advantaged family background and follow the "right" pathways during their transition to adulthood.

The transition to adulthood is an important life phase, in which youths make major decisions regarding career and family that shape their adult life-course (Arnett, 1998; Rindfuss, 1991). The transition to adulthood is a process that has become more complex as

over the last decades pathways have become more diverse, less standardized and prolonged (Shanahan, 2000). In examining the transition to adulthood of contemporary youths, it is therefore important to take into account the wide variety in which children of today become adults. This study provides a holistic approach to the transition to adulthood. Rather than examining the effects of single career or demographic events, sequences of events are studied. By examining sequences one can assess not only the effect of certain events, but also the effect of timing and ordering of these events (Billari, 2001). Finally, a contribution of this study is that career and demographic pathways are investigated separately for men and women.

This study uses panel data of the National Longitudinal Survey of Youth from 1997 (NLSY97), to examine to what extent current youths are diverging in income. In order to map out different pathways to adulthood we use sequence analysis. Clusters of career and demographic pathways from age 17 to 25 are created separately using Optimal Matching Analysis (OMA) (Abbott, 1983). After defining a distinct set of clusters, we will examine to what extent people belonging to certain clusters will have higher incomes compared to those of other clusters and to what extent there is divergence in incomes examining the yearly incomes from age 25 to 32.

Theory

Family background

An extensive body of research has demonstrated that higher socio-economic status of the parents is related to better economic outcomes of their children. More wealthy and better educated parents are more likely to spend more resources on the development of their children, particularly in their education (Becker & Tomes, 1979; Breen & Goldthorpe, 1997; Putnam, 2015). Theory on explaining the intergenerational transmission of social class

5

mostly focusses on differences in resource availability and socialization between children from advantaged and disadvantaged backgrounds. High status parents already invest in children´s education early on as they are more likely to send their children to pre-school and childcare and are more likely to send their kids to private schools (Kornrich & Furstenberg, 2013; Putnam, 2015; Temple & Reynolds, 2007). Since housing prices are higher in neighborhoods with better schools (Black, 1999; Haurin & Brasington, 1996), wealthy parents are more able to move to neighborhoods with better schools. Not only do they spend more money on their children, higher educated parents also spend more time with their children in their first years compared to lower educated mothers (Altintaş, 2015; Kalil et al., 2012). However, it is not only about higher educated parents spending more time with their children, but also *how* they spend their time with their children. According to Bourdieu children of high status parents are socialized in a way that they adopt certain attitudes, preferences and behaviors, which constitute cultural capital, which helps them with their educational and occupational careers (Bourdieu & Passeron, 1990). In an ethnographic study Lareau describes how upper-middle class parents adopt a strategy of "concerted cultivation" in raising their children, whereas parents from lower class families are more likely to adopt a "accomplishment of natural growth" strategy (Lareau, 2011). Parents using concerted cultivation make sure that their children spend more time in structured activities such as sports, music classes, art, clubs etc., have extensive conversations with their children speaking in rich vocabulary and long sentences, and learn their children how to negotiate institutions, for instance with school. On the other hand, parents using "accomplishment of natural growth", provide less structure for their children (they hang out with them or their other kin), have little discussion and rarely allow questioning, and show a sense of powerlessness when it comes to negotiating with institutions. There is indeed evidence that concerted cultivation mediates the effect of family background on educational achievement

6

(Bodovski & Farkas, 2008; J. E. Cheadle & Amato, 2011; Cheadle, 2009; Martin, 2012). In determining to what extent parental resources or parental socialization, it is important to the distinguish different dimensions of parental SES, by examining both the effect of parental income and education (Amato et al., 2015).

Family structure is another important aspect of family background that has been linked to the future income prospects of children. Being raised by a single parent or having experienced a parental divorce have been associated with lower income in later life (McLanahan, 2004; McLanahan, 2009). Children raised in non-intact households are more likely to have fewer resources at their disposal than children raised in intact households (in which children raised in marriage rather than cohabitation are found to have most resources). In these broken families children are less likely to receive as much resources from both their parents as children from intact families. Not only do these children have less financial resources, but also resources such as parental care and social capital (Putnam, 2015) are often lacking. Lack of parental care has been associated with poorer cognitive development and behavioral problems (Ermisch et al., 2012), which may decrease the likelihood for them to attain higher education (or not be able to attain higher secondary education before that) and a high earning job. Parental divorce may not only have an impact through a reduction of resources during childhood. Experiencing a parental divorce is often considered as an adverse life-event, which may not only have short-term, but also more long-term consequences (Putnam, 2015). Parental conflict may cause stress for children and adolescents or make children dissociate with their parents, most often with the father, which can result in long-term behavioral problems (Amato & Gilbreth, 1999). Parental divorce may therefore be an important risk factor for low socio-economic attainment (Amato, 2000).

The diverging destinies literature shows that parental divorce is increasingly concentrated among parents with low SES (McLanahan & Jacobsen, 2015; Mclanahan, 2004;

McLanahan, 2009). Therefore it is often a combination of parental divorce and poverty that strikes children of disadvantaged background making it more likely to start the transition to adulthood with a disadvantage, for instance because they dropped out of high school. In the next section we will discuss the transition to adulthood and its relation with family background.

Transition to adulthood

The transition to adulthood has been described as a demographically dense period (Rindfuss, 1991). During this life stage individuals usually experience multiple transitions. It is a stage in which careers are started, either by immediately entering the labor market or by enrolling in higher education. Furthermore, it is a stage in which individuals leave the parental home either to live on their own or to enter a union. A traditional final marker of reaching adulthood has been the entry of parenthood. However, that is not to say that this transition always occurs last. The order and timing of each of these events can have important implications for the future life-course.

Although the transition to adulthood is a stage in which they learn to be independent of their parents, social origin is still an important predictor of the timing, occurrence and sequencing of these transitions. As mentioned earlier, children from advantaged backgrounds are likely to spend more time in education, i.e. they are more likely to finish high school and enter college. On the other hand, children from disadvantaged backgrounds are more likely to enter the labor market after high school, whether they drop out or complete high school. Even with a high school diploma, youths of disadvantaged background may forego on going to college as they and their parents are more likely to view entering college as a risk, because if no degree is obtained then costs are more likely to be covered by the youth itself, whereas youths of advantaged background may be more inclined to go to college knowing that their

parents usually prefer them to go to college and will financially support them even if the drop out (Breen & Goldthorpe, 1997). Furthermore, those with high status parents are more likely to return to school, whether this is high school (Raymond, 2008) or college (Baum, Ma, & Payea, 2013). Moreover, children from advantaged backgrounds are more likely to choose for 4-year instead of 2-year college programs (Baum et al., 2013).

However, there are many young adults who combine education and employment (Kalenkoski & Pabilonia, 2010). It may especially for those from disadvantaged backgrounds be necessary to have a job to cover the costs of college education (Bozick, 2007). Overall, research indicates that working at a high intensity decreases college achievement and increases the likelihood of drop-out (Bozick, 2007; Kalenkoski & Pabilonia, 2010; Staff & Mortimer, 2007; Triventi, 2014). There are two explanations for this relationship. First, from a time-use perspective, individuals that spend much time on employment have less hours available for studying. Second, there may be selection, as those who perform poorly at college and receive more satisfaction from employment are likely to work more (Bozick, 2007). However, there are some studies that indicate that working up to 20 hours can enhance academic performance (Bozick, 2007; Triventi, 2014). Thus, children with little parental resources may complete their education if they are able to find a good balance between education and work, but they are still disadvantaged compared to those with many parental resources who will require less work hours to make ends meet.

Disadvantaged youths who enter the labor market without a college degree and even more so those without a high school diploma also have a higher risk of being unemployed during young adulthood (Taylor et al., 2011). Long spells of unemployment during young adulthood may lower one's socio-economic status not only in the short, but also in the long term. Mroz and Savage (2006) find that unemployment continues to negatively affects earnings up to ten years later. Thus, the choice to forego college does not only increase the

9

risk of unemployment, but also decreases future earnings compared to those with a college degree. Indeed, income studies have demonstrated that those with higher educational attainment have on average higher earnings and that the gaps between those with and without a college degree are expanding (Taylor et al., 2011).

Whether someone becomes a parent early in the adult life course has a great impact on adult life outcomes, including income, not only for women but also for men (Dariotis, Pleck, Astone, & Sonenstein, 2011). Raising a child requires resources, and children from disadvantaged backgrounds will not be able to rely on parental financial resources, but rather have to provide these resources on their own. This means that they will have to enter the labor market and forego higher education in order to provide for the child(ren). Those from disadvantaged background are more likely to have children early in life. They may view it as a legitimate way to enter adulthood as they have low career aspirations for themselves (Smith & Roberts, 2011). Early childbearing is more likely to be less deviant behavior among those of disadvantaged backgrounds as the mothers are likely to have experienced teenage childbearing themselves (Jennifer S. Barber, 2001). Parenthood at a young age is often unplanned. Higher educated parents may be more able to inform their children about the risk of unprotected sex (Miller, 2002).

Another important event in the life of young adults, which most experience, is leaving the parental home. Parental background also plays a crucial role in this decision. Young adults can have different reasons to leave the parental home. Traditionally, young adults left the parental home in order to marry. Nowadays, there are multiple ways to leave the parental home. Young adults may have to leave the parental home, because of the large geographical distance between the parental home and the college they wish to attend (Mulder & Clark, 2002). More affluent parents are more likely to provide the necessary means in order for their children to live on campus. Bozick (2007) finds that students from low-income families are

more likely to stay in the parental home. In general, wealthy parents are more able to help their children to set up their own household, whether this is to live independently or to enter a union (Avery, Goldscheider, & Speare, 1992; Sassler, 2004; Spitze & Waite, 1981). On the other hand, children with high status parents may be less willing to leave as their parental home is likely to provide them many resources that they would not have if they were to live on their own or with a partner (Avery et al., 1992; Easterlin, 1980; Goldscheider & Goldscheider, 1998)

Parental background also influences the timing and choice of union formation. Children from advantaged backgrounds are found to postpone their union formation compared to those from disadvantaged backgrounds (e.g. Axinn & Thornton, 1992; South, 2001; Wiik, 2009). An important reason is that, as mentioned above, children with high status parents are more likely to be enrolled in education. Research has indicated that the educational system works as a moratorium in which union formation is postponed (e.g. Blossfeld & Huinink, 1991; Liefbroer & Corijn, 1999; Raymore, Barber, & Eccles, 2001; Thornton, Axinn, & Teachman, 1995). Another reason why young adults with high status parents are more likely postpone relationship formation and parenthood is because they want a spouse of similar social status as their parents (Oppenheimer, 1988; Wiik, 2009). Since acquiring a high status job usually requires extensive education, children of advantaged backgrounds may postpone marriage until after the potential spouse has reached his/her full potential. Thus, children from advantaged backgrounds may be more risk averse in settling for a partner than children of disadvantaged background, therefore postponing their union formation. On the other hand, children from disadvantaged background are more likely to enter their unions early, because the home environment does not provide any comfort (Easterlin, 1980; Gierveld, Liefbroer, & Beekink, 1991). Furthermore, they may be more risk taking in their partner choice and leading to higher probabilities that they will divorce.

Indeed, Berrington & Diamond (1999) find that those who enter a union early are more likely to divorce. In turn, experiencing a divorce has been associated with higher unemployment (Covizzi, 2008).

Regarding union formation it is not only the question when young adults enter it, but also whether they opt for married or unmarried cohabitation. Marriage has been associated with better adult life-outcomes including income (Ahituv & Lerman, 2007; Waite & Gallagher, 2002). Children with high SES parents are more likely to marry than those with low SES parents (Bumpass & Lu, 2000; Kennedy & Bumpass, 2008; Lichter, Qian, & Mellott, 2006; Manning & Cohen, 2015; Seltzer, 2004). However, the negative effects of cohabitation may be especially visible among those who have children. Young adults with high status parents may not marry, but rather cohabit if they are not sure about the partner and only have children in marriage. Cohen and Manning (2010) find that young adults with highly educated mothers are more likely to serial cohabit. This may mean that cohabitation for advantaged youths may serve as a weeding process (Klijzing, 1992), in which one leaves the partners that are not fulfilling their potential and marries the one that does. On the other hand, disadvantaged youths may have less resources to cover the costs of marriage or married life and therefore remain in a cohabiting relationship (Clarkberg, 1999). Cohabitation than serves as a poor man's marriage (Hiekel, Liefbroer, & Poortman, 2014).

The domains of career and family are linked. Those who enter a union and/or parenthood early are less likely to enter education and vice versa (Blossfeld & Huinink, 1991; Raymore et al., 2001). Some may have to leave the parental home in order to attend education, which has also been found to enhance study performance. (Bozick 2007). In general, it could be argued that for one's career it better to postpone major demographic events. However, there is some indication that those attending education usually not the ones that remain in the parental home until there mid-twenties (Amato et al., 2015). Married

individuals (with children) may take up more work, because they feel more responsible for their family situation, which may influence their current earning, but possible also future earnings (Ahituv & Lerman, 2007; Amato et al., 2015). Thus experiencing some demographic transitions may also enhance career performance.

Income trajectories

During the later stages of early adulthood, young adults have usually finished their education and are on the labor market. The education and work experience that young adults have obtained during the transition to adulthood is not only crucial for their income in young adulthood, but also for their potential future income. It may therefore be that for those who followed a career pathway with education and (some) employment may increasingly diverge in their income from those who have little education and work experience. A process that can be described as cumulative advantage or "Matthew effect" (Merton, 1968). Indeed research has indicates that between educational level groups there is cumulative advantage in wages (DiPrete & Eirich, 2006; Elman & O'Rand, 2004). Furthermore, those who start with a higher income may also be more likely to have a higher income rise during their career (Cheng, 2015).

   If disadvantaged youths are able to follow a 'successful' pathway during the transition to adulthood differences with their peers with an advantaged background are expected to diminish. However, young adults of high status background may still hold an advantage over those from low status background in terms of income accumulation even when they follow the same career and demographic pathways. High status parents facilitate their children with the transition from school to work by providing their children information and contacts that may help them to obtain jobs (Ermisch et al., 2012). On the other hand, youths of disadvantaged background with a college degree, lacking this parental social capital, may

13

have more difficulty to find a job that matches their educational credentials. Thus, there may still be divergence on the bases of parental background even within groups who choose a similar path to the transition to adulthood.

Earning a higher income may also prevent relational instability. Ahituv and Lerman (2007)indicate that marriage and high earnings may reinforce one another. On the other hand, disadvantaged youths who experienced a parental divorce may be less able to have steady relationships, especially if they also do not have a stable job (Oppenheimer, 2003).

Gender and race considerations

Thus far we have not distinguished between gender and race. However, there is ample evidence that suggests that pathways may be different or that the distribution of different kind along these dimensions (Oesterle et al., 2010). Women have overtaken men in both college enrollment and graduation (Buchmann & DiPrete, 2006; Dwyer, Hodson, & McCloud, 2012). Regarding demographic transitions, women enter unions and parenthood earlier (Uecker & Stokes, 2008; Winkler-Dworak & Toulemon, 2007). Furthermore, when they enter parenthood women are often expected to be the main responsible caregiver (Barber, 2000; Wiik, 2009)). This may mean that for women childbearing increases the difficulty to work on a career, whereas men may retract themselves from parental responsibilities . However, there also research that indicates that parenthood for men may have long-term consequences in terms of lower earnings(Dariotis et al., 2011).

Regarding racial differences being black has been associated with disadvantage. Compared to whites they are likely to face more difficulty in obtaining a job, because of discrimination and cultural differences may provide them with less cultural capital, which may make it more difficult for them to enter college and obtain a high status occupation (Black & Sufi, 2002; Hardaway & McLoyd, 2009). Blacks are also found to marry less and

more often have children outside of wed-lock compared to other racial groups (Loomis & Landale, 1994; Manning & Smock, 1995; Schoen & Cheng, 2006) and as mentioned above marriage is associated with many positive outcomes.

In the analyses we will therefore will construct different clusters for men and women and control for racial differences.

Data & Methods

Data

This study uses the National Longitudinal Survey of Youth from 1997 (NLSY97), a panel study conducted by the U.S. Bureau of Labor Statistics. Respondents were selected in 1997 at the ages 12 to 17 (born in 1980 to 1984), using a multi-stage stratified random sampling design and have been interviewed annually until 2011 and a last wave was conducted in 2013. The NLSY97 contains an oversample of respondents of Afro-American and Latino decent. However, when weighted the NLSY97 provides a nationally representative sample of youths. The total sample consists of 8984 respondents. However, we only select respondents who have participated in all waves and who have at least some information on personal income, between the ages 25 to 32, leading to a selection of N=4966 cases of which 2301 are male and 2665 are female. There are a number of reasons why the NLSY97 is a good dataset to answer our research questions. First, it contains a high level of detail when it comes to demographic and career characteristics at all waves. Second, income is measured at all waves making it possible to assess income over the early adult life-course. Third, the NLSY97 contains information on those who have recently become adults, therefore answering the question how contemporary youths may diverge in their income trajectories.

15

Sequencing

In the NLSY97 youths record at which year and month a specific event related to the transition to adulthood occurred. In terms of education, youths were asked in each round to report whether they had entered or exited an educational institution the year before. Respondents were also asked to report the level of education they enrolled in, i.e. secondary school, 2-year college, 4-year college. Regarding employment, youths were asked to provide the start and end dates of each job they had the last year. There is also information on the type of job (also reporting if youth joined the military) and the number work hours it provided[1]. With respect to demographic characteristics, respondents were asked whether they had started or ended a marriage or cohabiting relationship in the previous year. Youths also had to report the birth year and month of each of their children. Each wave youths indicated the household composition, in which they reported the people that were living in their household. Furthermore, respondents were asked the month and year in which they first left and returned to the parental home (if they did)[2].

The information available in the data is used to construct a sequence dataset for both career and demographic pathways. In order to create such a dataset one has to define the different states that individuals can be in each month. For the career sequence states can differ in two dimensions: education and employment. In terms of education, youths can be enrolled in high school, college or not be enrolled. To limit the number of states we opted not to distinguish between attending a 2-year and 4-year college. However, the sequence does capture how long an individual is enrolled in college. Regarding employment, individuals either have employment over 20 hours per week, employment less than 20 hours per week or no employment (which includes people who are unemployed, but also for instance stay at

---

[1] The NLSY97 reports weekly job status. We recoded this to monthly statuses using the by NLS recommended conversion. If someone is employed for at least one week during that period, this person is considered employed.

[2] These questions were included in 2003 onwards, but in the 2003 question respondents also indicated the month and year of home return if they occurred in any of the years before.

home mothers). The cut-off of 20 hours is chosen as working 20 hours or more has been defined as moderate to high levels of work intensity for those enrolled in college (Roksa & Velez, 2012). This leads to a number of 9 (3 x 3) possible different career states, which individuals can be in.

The demographic pathways are also defined along two dimensions: residential status and parenthood. The first dimension captures with whom one lives in a household, in which we define four options: living with parents, living alone/independent, living with partner (cohabiting), living with spouse (marriage). The second dimension is parenthood, indicating whether someone has become a parent at some point or not. Entering parenthood is considered irreversible, as once one becomes a parent they stay a parent for the rest of the sequence. This leads to a total of 8 (4 x 2) possible demographic states.

Each sequence contains 96 spells as youths pathways are recorded monthly between the age of 17 and 25. This particular age range is chosen for a couple of reasons. First, it covers the range proposed by Arnett (2000) in describing a life-phase called emerging adulthood. The sequence starts from age 17 as in this year most people are still in high school and the transition to college (for those who go to college) still has to take place. Second, Schulenberg and Schoon (2012) state that differences in pathways become most visible during ones mid-twenties. In order to establish how different sequences are from one another (referred to as distance) Optimal Matching (OMA) is used (Abbott 1983). This method establishes how many indels, i.e. substitutions, deletions or insertions, are required to transform one sequence into another. The more operations are required, the more distant sequences are from another. However, some transitions may be occurring more often than others. For instance, people who recorded that they live with their parents may be less likely to become parents in the next month compared those who reported being married (also known as data-driven approach). Therefore, we assign costs of indels to be based on the

transition rates between different states. Thus, some operations are costly than others, meaning that the increase in distance as a result of an indel is higher if the operation has to change a state for which the transition rate with the other state is low (difficult sentence).

In order to create cluster we use the TraMineR package in R. Based on the distance defined by the OMA procedure different clusters can be defined. A weighted (using NLSY97 weights) hierarchical clustering procedure using the Ward method is chosen to produce clusters. This procedure is executed separately for men and women as pathways for both groups are likely to be different. For instance, women on average enter relationships earlier than men (Winkler-Dworak & Toulemon, 2007). For both men and women a number of clusters has to be chosen for both career and demographic pathways. A higher number of clusters can provide more detail in how individuals vary in their sequences, but also increases complexity and lower group sizes. One therefore has to establish whether each extra cluster one introduces represents a specific group of people or whether this group can really be categorized as part of another group.

Family background variables

The first round also contains a parent questionnaire from which family background characteristics, such as parental income, education and family structure are derived. *Parental education* is coded as the highest education of mother or father in three categories: more than high school, high school or lower and missing if the education for both the father and mother was missing. *Parental income* is the income reported by one of the parents when the youth was 12 to 16 years old and is coded in quartiles, including also a missing category. The *family structure* has four categories: 1) Both biological parents – 2) 1 biological, 1 stepparent – 3) 1 biological parent – 4) other (no biological parents). Finally, *race* is coded as: 1) white (non-Hispanic), 2) black (non-Hispanic), 3) Hispanic, other(mixed).

18

[Include descriptives table]

Growth-curve modeling

In order to estimate the effects of family background and the transition to adulthood on income trajectories growth curve models are used. Each wave respondents are asked to report all income they received from salaries, wages and commissions in the previous year. Income trajectories are mapped out for the ages 25 until 32 (using log income). Not all individuals have reached the age of 32. Thus, higher ages of a lower number of observations. In determining the income trajectory we allow both the intercept and the slope to vary within individuals. The demographic clusters and family background variables are included in the model to explain the differences in these intercepts and slopes. The first model only contains the family background variables and in the second model the career and demographic pathways are included.

Results

Descriptive results

Figure 1.1 shows the medoid of the five cluster solution for women's career pathway. The medoid is the sequence of which the distance to all other sequence in that cluster is lowest. It therefore represents a typical pathway of someone within that cluster. The first cluster shows a pathway that involves much enrollment in education, in which there is a gradual increase in hours spend in employment. It is a pathway of a college attendee who focusses on education first and then more permanently enters the labor market. The second cluster involves women who have stable employment starting already in high school, but who do not attend any post-secondary education. The third cluster also contains no postsecondary education, but women

19

in this cluster have less stable employment. The fourth cluster contains those who both have employment, but also are enrolled in post-secondary education. Finally, the fifth cluster contains those who are mostly inactive after high school.

Figure 1.1

Medoid career pathways women
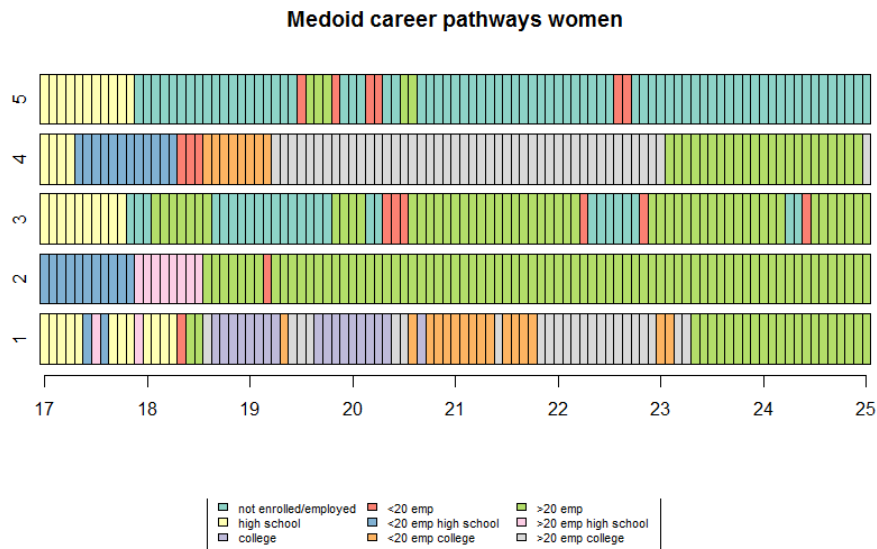
In Figure 1.2 the different medoid career sequences for men are presented. The first medoid sequence consists of unstable unemployment and no postsecondary education. Men in this cluster are most inactive compared to those in other clusters. The second cluster contains much enrollment in postsecondary education and towards the end more employment and finally only employment (around age 23). The third cluster shows a combination of college and work throughout the sequence only turning to only employment also around age 23. Finally, men in the fourth cluster are those who have steady employment after high school, but do not attend postsecondary education. For men we opt for the 4 cluster solution as a fifth cluster for men only contains 90 respondents. Furthermore, this 5th cluster contains stable employment for 20 hours or less, which fits with the relatively inactive status of those in

cluster 1 of which in the 4 cluster solution, they are part of. The different kind of career pathways appear to be quite similar. Both men and women have a *college with little employment* cluster, a *college with employment* cluster, an *employment no college* cluster and a *unstable employment no college* cluster. The only difference is that among women there appear to be those who are *mostly inactive* after high school and those who have unstable employment, but no college education.

Figure 1.2

Medoid career pathways men
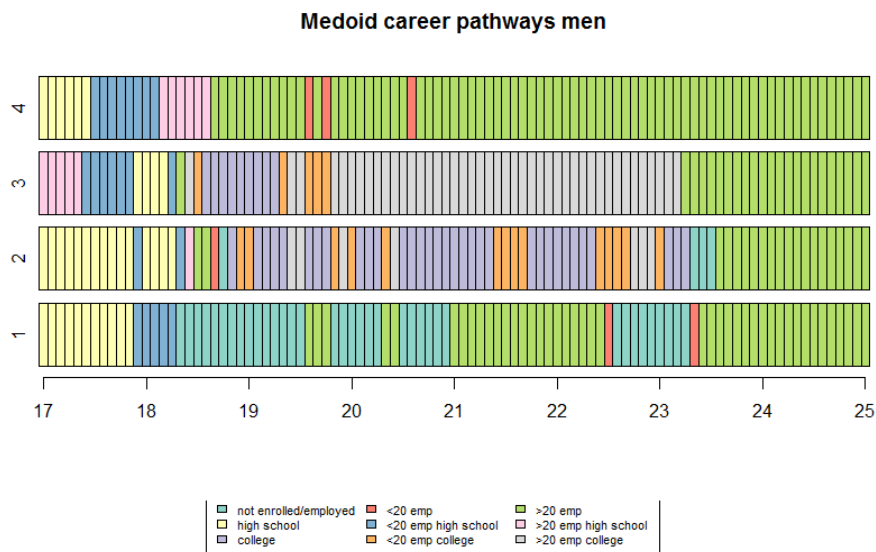
Figure 2.1 shows the medoid sequences of the 5 cluster solution on demographic pathways for women. The first cluster, represents those who stay in the parental home until age 25. The second cluster, involves parenthood at a young age and no stable unions, as the sequence contains only cohabitation for some period, but no marriage. In the third cluster women postpone leaving the parental home, but do at the end leave and start a union. Compared to the third cluster, women leave the parental home earlier in the fourth. Finally, women in the

fifth cluster appear to follow the most traditional pattern as they leave the parental home relatively early, but also shortly after get married and have children.

Figure 2.1

Medoid demographic pathways women
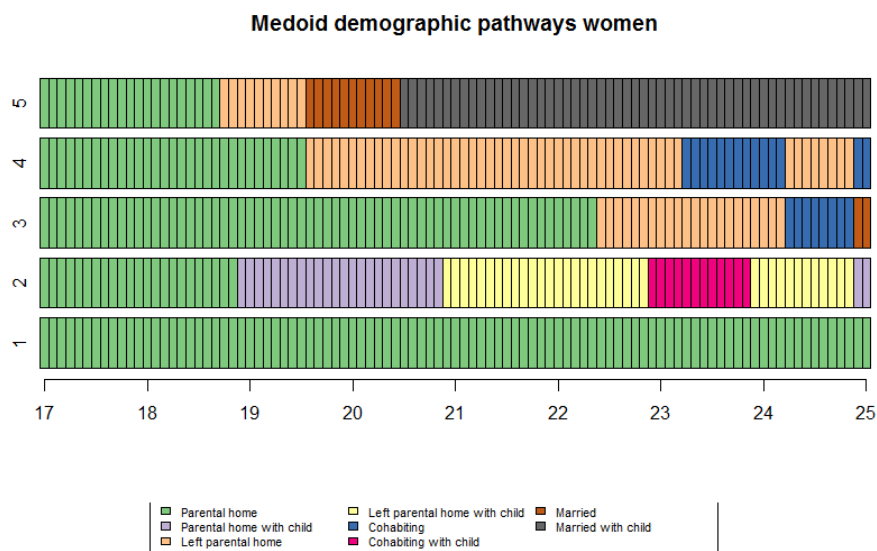
In Figure 2.2 the medoid sequences of the 5 cluster solution for men are presented. In the first cluster men leave the parental home, but only at the end (around age 23) in which they do not enter a union or parenthood. The medoid sequence of the second cluster shows a more traditional pattern, involving union formation and parenthood just after leaving the parental home. However, they first appear to cohabit before they marry and have children. The third cluster contains those who stay in the parental home. Men in the fourth cluster enter parenthood before entering a union. When they enter a union it is a cohabiting union and not a marriage. Finally, the fifth cluster contains those who leave the parental home at a relatively young age (around age 20), but do not enter a union or parenthood before age 25. Again there are similarities between the men's and women's clusters. Both men and women have a

cluster of those *staying in the parental home*. Furthermore, both have cluster that involves early parenthood but in which they do not enter marriage and appear to have little stability in their relationships (*parenthood and unstable union*). Also, men and women have a cluster in which they leave the parental home, but only around age 23 (*postponing parental home leaving*). The main difference between the clusters of men and women are that men appear to postpone relationship formation compared to women. However, both have a cluster that involves union formation followed by parenthood within marriage (for men: *union formation*, for women: *married with children*) and a cluster in which they remain relatively independent (for women: *independent and union*, for men: *independent living*.

Figure 2.2

Medoid demographic pathways men
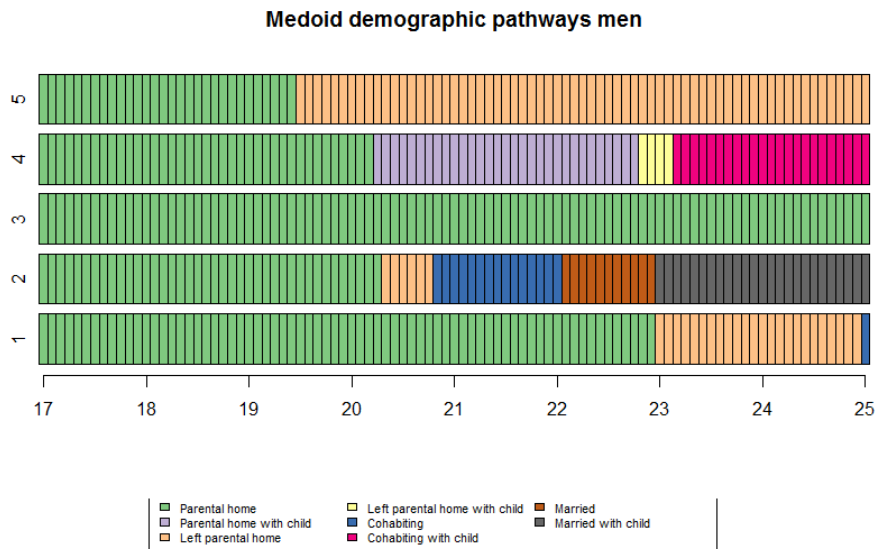
Figures 3.1 and 3.2 show the distribution of the states at each time point for the different career clusters for respectively women and men. These show that within a cluster some states are more common at a certain time point than at another. For instance in the college and work cluster, working for more than 20 hours while being in college occurs mostly in the middle of

the sequence (age 21/22). Furthermore, in some clusters certain states are more prevalent over the other states than in other clusters. In the *only employment* cluster, the state of being unemployed is much prevalent, whereas in the college with little employment there is more variation in the distribution of states at each time point.

[figures 3.1, 3.2, 4.1 and 4.2 about here]

Figures 4.1 and 4.2 show the distribution of states within the demographic clusters for respectively women and men. Again there is variance in the extent to which a certain state is dominant within a cluster. For both men and women the *staying in the parental home* cluster almost exclusively contains this state, whereas for instance in the *parenthood and unstable union* cluster there is no dominant state at the end of the sequence. These distribution figures show that within clusters there is quite some variation in sequences that individuals have. However, the sequences in each of the clusters do show similarities in the type of states that they experience and the timing at which they occur.

[tables 1.1 and 1.2 about here]

Table 1.1 shows a cross tabulation of the demographic and career clusters of women. In general demographic clusters that involve early childbearing (clusters *parenthood and unstable union* and *married with children*) have relatively few individuals that follow a career pathway that involves college education and vice versa. Furthermore, those who are in cluster with early parenthood are mostly in the cluster that involve more unemployment/no employment. Those who do not attend college, but have steady employment are mostly in the *independent and union* cluster. However, most people in this cluster attend college. For men,

presented in table 1.2, the distribution is mostly similar. A difference is that whereas for women staying at the parental home mostly attend college, the men in this cluster are more evenly distributed over all the career cluster, where in fact they are mostly present in the *unstable employment no college* cluster. Although some combinations of cluster have many whereas others have little, for both men and women, all cells contain at least some individuals.

[tables 2.1 through 3.2 about here]

Tables 2.1 and 2.2 contain a cross tabulation of respectively women's and men's career cluster with family background. In general, youths with higher parental income and education tend to be more in clusters containing college education, whereas youths with low parental income and education are more present in the clusters with little employment. Furthermore, youths from broken families are less likely to be in a cluster containing college and relatively more likely to be part of a cluster containing less activity in terms of employment and education. Finally, whites are more likely than other races to be in a cluster containing college education, although these differences appear to be higher among men. Tables 3.1 and 3.2 show the cross tabulation with demographic clusters and family background for respectively women and men. Those in the *parenthood and unstable union* cluster more often come from low income households with little parental education. Furthermore, the ones is this cluster are more likely to come from single-headed and black households. Those postponing the parental home leave and who are either independent or in a relationship are more often have high earning and educated parents. Furthermore, they are more often white and come from intact families. For women those staying in the parental home tend to come from high income, educated household, whereas for men the distribution is more even.

Although the distributions are different (all Chi2 tests are significant) all cells do contain cases (except one in mixed race, but this is because there are few individuals in the data that have a mixed race background).

[figures 5.1 and 5.2 about here]

In figure 5.1 the mean income trajectories of women for each career cluster are presented. Those who have a cluster with college education have the highest income, which also appears to increase the most compared to the other clusters. However, those in the *employment no college* cluster start to converge more with the cluster containing college education after age 29. Between the *college with little employment* and the *college and employment* there appears to be little difference. The mostly *inactive cluster* has the lowest income trajectory followed by the *unstable employment no college* cluster. These clusters have both the lowest slope and intercept. The same can be observed for men (figure 5.2), although for men the divergence between those with college education and those without (especially compared to the *unstable employment no college* cluster) appears to be stronger than for women.

[figures 6.1 and 6.2 about here]

In figure 6.1 the mean income trajectories of women per demographic cluster are presented. Women who have a child after age 25 appear to have the highest income at age 25 until 32. The most successful women appear to be those who leave the parental home between the ages 17 and 25 either to live on their own or to enter a union. However, women who remained in the parental home do catch up with those who had left the parental home earlier. The *postponing parental home leaving* cluster shows the highest mean at age 32, although the

26

confidence intervals show that they are not significantly distinct from the *independent and union* and the *staying in parental home* cluster. The cluster with the lowest intercept and slope is the *parenthood and unstable union* cluster. At all ages, except 25, women in this cluster have a lower mean income compared to all the other clusters. For men (figure 6.2) also those who postpone parenthood have a better income trajectory compared to those who do. A difference is that the income trajectory for men in the *staying at parental home cluster* is relatively lower compared to women. In fact, except for the *parenthood and unstable union* cluster, men in all other clusters have an higher income at all ages. Finally, there appears to be a little more divergence between the bottom two clusters and the other three clusters for men than for women.

Growth curve model results

Results of growth curve modeling for women are presented in table 4.1. Model 1 contains all the random slopes and intercepts for all the background variables. Parental income is significant at the intercept (each higher quartile providing a higher income), but does not explain variation in the random slope. That is, there are differences between youths, but these differences do not change between the ages 25 and 32. Those who have a parent with more than high school education also have a higher income at the intercept. Regarding family structure, there are no effects at the intercept, but women who had a stepparent in the household have a less steep slope compared to women who were raised by both biological parents. In model 2 the career and demographic pathways are included. At the intercept all career pathways have a higher income compared to the *unstable employment no college* cluster, except for the *mostly inactive* for which the income is substantially lower. The increase in income is higher for the *college with little education* , the *college and employment* and *mostly inactive* cluster. This indicates that the clusters with college education diverge

27

from the other clusters and that the *mostly inactive* cluster converges slightly with the cluster with more employment. Regarding the demographic pathways there is only one significant effect at the intercept, which is that women in the *independent and union* cluster have a higher income compared to those in the *staying in parental home cluster*. The family background variables are no longer significant at either the intercept or random slope.

[table 4.1 and 4.2 about here]

Table 4.2 shows the results of the growth curve model for men. Parental socio-economic status measures are significant, whereas family structure is not at the intercept. Men with a parental income in the highest quartile diverge in slope compared to those in the lowest quartile. The effects of parental income at the intercept remain mostly significant in model 2. For both the career and demographic pathways there are significant differences. Compared to men in the *unstable employment no college cluster* all other have a higher income and those in the *college with little employment* cluster diverge even more from this group with age. Regarding the demographic clusters, men in cluster that involve independent living and unions, but no parenthood have a higher income compared to those staying in the parental home. However, these differences do not increase with age. Finally, for men there is an intercept difference of race, indicating that black men have a lower income, but this difference does not increase as the interaction with age (random slope) is not significant.

Summary and Discussion

In this study we have linked the transition to adulthood to income trajectories in young adulthood. By examining the transition to adulthood holistically we were able to identity how patterns of life-courses rather than single events influence ones income in young adulthood.

Furthermore, it was investigated whether there is a visible divergence in the income between young adults of today and to what extent these differences are related to different transitions to adulthood and family background. The novelty of this research lies in incorporating the transition to adulthood in models of intergenerational transmission of (dis)advantage.

First it was examined whether family background had an effect on income trajectories during young adulthood. In line with the literature on diverging destinies (Amato et al., 2015; Mclanahan, 2004; McLanahan, 2009) family background had a significant impact on income during young adulthood, although it was mainly parental income that had significant effects on the income trajectories of young adults. However, little divergence in slopes of the income trajectories was found between those with high and low parental income. Parental education and family structure appeared to have little (additional) impact on income differences. However, if parental income and education were not included in the model, the effects of being raised by a one biological parent and a step-parent, being raised by a single parent or not being raised by either biological parent, had significant negative effects on income. This suggests that experiencing a parental divorce or having no parent in itself may not have a large influence on one's income during early adulthood, but rather the lack of financial resources that is associated with being raised by a single parent (Cohen, 2015).

Sequence analysis provided a number of distinct career and demographic clusters, mostly similar for men and women. Individuals who were in a cluster containing college education had higher incomes compared to those only in employment or those relatively inactive. Furthermore, the slopes of the income trajectories diverged for those who were enrolled in college education (particularly for those who worked relatively little beside education) compared to those who did not enroll in education and had little employment. These results appear to indicate a strong college premium. The effect of the demographic pathways appear to be less strong. Individuals in clusters containing independent living and

cohabitation or marriage towards the end (near age 25) fared better in terms of income compared to those who had early childbearing or those who remained in the parental home. However, these differences did not increase with age as with some of the career clusters. An interesting finding is that not only early childbearing, but also remaining in the parental home until age of 25 is associated with lower income in early adulthood. This supports the idea that being a slow starter may also have negative impact on your career as an adult (Amato et al., 2015).

Family background and the transition to adulthood were not unrelated. Indeed, children from intact families with higher parental education and income were more likely to be part of a career and demographic cluster that was associated with a higher income in early adulthood. However, there were individuals of all social backgrounds present in each of the different clusters, meaning that we were able assess the effects of the transition to adulthood in addition to the effects of family background. Including the career and demographic pathways strongly decreased the family background effects, in which almost all family background effects become insignificant, indicating that the career and demographic pathways mediate the relationship between family background and income in early adulthood. Thus, it appears that although advantaged youths may be more likely to attend college and avoid early parenthood, there is little cumulative advantage compared to disadvantaged youths beyond the transition to adulthood.

There were some gender and racial differences. For women we opted for an additional career cluster, containing those who were completely inactive, most probably housewives. Furthermore, in the demographic clusters there was a separate married with children cluster, whereas for men these were included in a cluster containing also married or cohabiting men without children. For men some direct effect of parental income remained whereas for women the effect of parental income was completely mediated by the transition to adulthood.

For women those who were raised by a biological and a step-parent, there was a lower slope in income compared to women from intact families. It is somewhat surprising that this was not found for women raised by single parents, although findings of Wilcox (2014) also indicate somewhat lower earnings for those raised by a biological parent plus stepparent compared to those raised by a single parent . However, this effect disappeared as the career and demographic clusters were included. Regarding racial differences, only for men there was a significant negative effect of being black on the intercept, but this effect remained significant also when the career and demographic clusters were included. Racial differences for men may partly be because of the relatively high incarceration rate among black men (Pettit & Western, 2004).

There are some limitations of this research. First, adults could only be followed until age 32, whereas income differences are likely to be more pronounced around age 40. On the other hand, study does provide an indication on how differences of the young adults of today arise. Since the NLSY97 is continuing data collection it would be interesting to conduct a follow-up study to examine whether these differences indeed become more pronounced. Second, only individuals that participated in all waves were included, meaning that many cases were excluded from the analysis. Although sample weights constructed for those who participated in all waves were incorporated in the analysis, there may still be a selection-bias. Third, there were many missings on both income of the respondent and parental income. The missings of parental income appeared to be random as the coefficient for unknown parental income was mostly between the coefficient of the 2nd and 3rd quartile. The missings on respondents income plus the fact that some of the younger respondents had not reached the higher ages (29-32) at the time of the last survey, may mean that slope differences of income trajectories are underestimated.

Although career tracks and parental SES appeared to be more important for income than demographic tracks and family structure this does not mean that this is the same for other important outcomes in early adulthood. For instance, family structure and demographic pathways may have a relatively stronger impact on health. Regarding health, there may also be differences in physical health and mental well-being. Future research could also assess whether the transmission of disadvantage through the transition to adulthood varies depending on the national context. Finally, future research could examine whether these relationships change over time or whether specific period effects, such as the latest economic crisis, change these relationships.

The lives of young adults in the United States are clearly stratified. Young adults with high status parents are more likely to attend college and usually avoid entering parenthood before age 25. It appears that the best way to help young adults from disadvantaged backgrounds to obtain a better income is to provide better access to college education. Higher college tuition fees are likely to discourage disadvantaged youth, but at the same time it becomes increasingly difficult to earn a decent wage without a college degree. Providing more scholarships or lowering tuition may help young adults from disadvantaged background not only to enter college, but more importantly leave college with a degree, as this may be most important in helping disadvantaged youths.

Literature

Aaronson, D., & Mazumder, B. (2008). Intergenerational Economic Mobility in the United States, 1940 to 2000. *Journal of Human Resources*, *43*, 139–172. http://doi.org/10.1353/jhr.2008.0008

Abbott, A. (1983). Sequences of Social Events: Concepts and Methods for the Analysis of Order in Social Processes. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, *16*(4), 129–147. http://doi.org/10.1080/01615440.1983.10594107

Ahituv, A., & Lerman, R. I. (2007). How do Marital Status, Work Effort, and Wage Rates Interact? *Demography*, *44*(3), 623–647. http://doi.org/10.1353/dem.2007.0021

Altintaş, E. (2015). The Widening Education Gap in Developmental Child Care Activities in the United States, 1965-2013. *Journal of Marriage and Family*. http://doi.org/10.1111/jomf.12254

Amato, P. R. (2000). The consequences of divorce for adults and children. *Journal of Marriage and Family*, *62*(4), 1269–1287. http://doi.org/10.1111/j.1741-3737.2000.01269.x

Amato, P. R., Booth, A., McHale, S. M., & Van Hook, J. (Eds.). (2015). *Families in an Era of Increasing Inequality* (Vol. 5). Cham: Springer International Publishing. http://doi.org/10.1007/978-3-319-08308-7

Amato, P. R., & Gilbreth, J. G. (1999). Nonresident fathers and children's well-being: A meta-analysis. *Journal of Marriage and the Family*, *61*(3), 557–573. http://doi.org/10.1086/250095

Arnett, J. J. (1998). Learning to Stand Alone: The Contemporary American Transition to Adulthood in Cultural and Historical Context. *Human Development*, *41*(5-6), 295–315. http://doi.org/10.1159/000022591

Avery, R., Goldscheider, F., & Speare, A. (1992). Feathered nest/gilded cage: parental income and leaving home in the transition to adulthood. *Demography*, *29*(3), 375–88. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/1426435

Axinn, W. G., & Thornton, A. (1992). The influence of parental resources on the timing of the transition to marriage. *Social Science Research*, *21*(3), 261–285. http://doi.org/10.1016/0049-089X(92)90008-5

Barber, J. S. (2000). Intergenerational Influences on the Entry into Parenthood: Mothers' Preferences for Family and Nonfamily Behavior. *Social Forces*, *79*(1), 319–348. http://doi.org/10.1093/sf/79.1.319

Barber, J. S. (2001). The Intergenerational Transmission of Age at First Birth among Married and Unmarried Men and Women. *Social Science Research*, *30*(2), 219–247. http://doi.org/10.1006/ssre.2000.0697

Baum, S., Ma, J., & Payea, K. (2013). Education pays 2013. *The College Board*.

Becker, G. S., & Tomes, N. (1979). An equilibrium theory of the distribution of income and intergenerational mobility. *The Journal of Political Economy*, 1153–1189.

Berrington, A., & Diamond, I. (1999). Marital dissolution among the 1958 British birth cohort: The role of cohabitation. *Population Studies*, *53*(1), 19–38.

Billari, F. C. (2001). Sequence analysis in demographic research. *Canadian Studies in Population*, *28*(2), 439–458.

Björklund, A., Jäntti, M., & Solon, G. (2007). Nature and nurture in the intergenerational transmission of socioeconomic status: Evidence from Swedish children and their biological and rearing parents. *The BE Journal of Economic Analysis & Policy*, *7*(2).

Black, S. E. (1999). Do Better Schools Matter? Parental Valuation of Elementary Education. *The Quarterly Journal of Economics*, *114*(2), 577–599. http://doi.org/10.1162/003355399556070

Black, S. E., & Sufi, A. (2002). Who Goes to College? Differential Enrollment by Race and Family Background. *California Center for Population Research*. Retrieved from

http://escholarship.org/uc/item/6dm686ft#page-1

Blau, P. M., & Duncan, O. D. (1967). The American occupational structure.

Blossfeld, H.-P., & Huinink, J. (1991). Human capital investments or norms of role transition? How women's schooling and career affect the process of family formation. *American Journal of Sociology*, 143–168.

Bodovski, K., & Farkas, G. (2008). "Concerted cultivation" and unequal achievement in elementary school. *Social Science Research*, *37*(3), 903–919. http://doi.org/10.1016/j.ssresearch.2008.02.007

Bourdieu, P., & Passeron, J.-C. (1990). *Reproduction in education, society and culture* (Vol. 4). Sage.

Bozick, R. (2007). Making It Through the First Year of College: The Role of Students' Economic Resources, Employment, and Living Arrangements. *Sociology of Education*, *80*(3), 261–285. http://doi.org/10.1177/003804070708000304

Breen, R., & Goldthorpe, J. H. (1997). EXPLAINING EDUCATIONAL DIFFERENTIALS: TOWARDS A FORMAL RATIONAL ACTION THEORY. *Rationality and Society*, *9*(3), 275–305. http://doi.org/10.1177/104346397009003002

Buchmann, C., & DiPrete, T. a. (2006). The Growing Female Advantage in College Completion: The Role of Family Background and Academic Achievement. *American Sociological Review*, *71*(4), 515–541. http://doi.org/10.1177/000312240607100401

Bumpass, L., & Lu, H.-H. (2000). Trends in cohabitation and implications for children s family contexts in the United States. *Population Studies*, *54*(1), 29–41. http://doi.org/10.1080/713779060

Cheadle, J. E. (2009). Parent educational investment and children's general knowledge development. *Social Science Research*, *38*(2), 477–491. http://doi.org/10.1016/j.ssresearch.2008.12.002

Cheadle, J. E., & Amato, P. R. (2011). A Quantitative Assessment of Lareau's Qualitative Conclusions About Class, Race, and Parenting. *Journal of Family Issues*, *32*(5), 679–706. http://doi.org/10.1177/0192513X10386305

Cheng, S. (2015). Unequal Origins, Unequal Trajectories: Social Stratification over the Life Course. Retrieved from http://deepblue.lib.umich.edu/handle/2027.42/113572

Chetty, R., Hendren, N., Kline, P., Saez, E., & Turner, N. (2014). Is the United States Still a Land of Opportunity? Recent Trends in Intergenerational Mobility.

Clarkberg, M. (1999). The price of partnering: The role of economic well-being in young adults' first union experiences. *Social Forces*, *77*(3), 945–968.

Cohen, J., & Manning, W. (2010). The relationship context of premarital serial cohabitation. *Social Science Research*, *39*(5), 766–776.

Cohen, P. (2015). Divergent Responses to Family Inequality. In P. R. Amato, A. Booth, S. M. McHale, & J. Van Hook (Eds.), *Families in an Era of Increasing Inequality SE - 2* (Vol. 5, pp. 25–33). Springer International Publishing. http://doi.org/10.1007/978-3-319-08308-7_2

Corak, M., Lindquist, M. J., & Mazumder, B. (2014). A comparison of upward and downward intergenerational mobility in Canada, Sweden and the United States. *Labour*

Economics, *30*, 185–200. http://doi.org/10.1016/j.labeco.2014.03.013

Covizzi, I. (2008). Does union dissolution lead to unemployment? A longitudinal study of health and risk of unemployment for women and men undergoing separation. *European Sociological Review*, *24*(3), 347–361.

Dariotis, J. K., Pleck, J. H., Astone, N. M., & Sonenstein, F. L. (2011). Pathways of Early Fatherhood, Marriage, and Employment: A Latent Class Growth Analysis. *Demography*, *48*(2), 593–623. http://doi.org/10.1007/s13524-011-0022-7

DiPrete, T. A., & Eirich, G. M. (2006). Cumulative advantage as a mechanism for inequality: A review of theoretical and empirical developments. *Annual Review of Sociology*, 271–297.

Dwyer, R. E., Hodson, R., & McCloud, L. (2012). Gender, Debt, and Dropping Out of College. *Gender & Society*, *27*(1), 30–55. http://doi.org/10.1177/0891243212464906

Easterlin, R. A. (1980). Birth and fortune: The effects of generation size on personal welfare. *New York*.

Elman, C., & O'Rand, A. M. (2004). The Race Is to the Swift: Socioeconomic Origins, Adult Education, and Wage Attainment. *American Journal of Sociology*, *110*(1), 123–160. http://doi.org/10.1086/386273

Ermisch, J., Jäntti, M., Smeeding, T., & Wilson, J. A. (2012). Advantage in Comparative Perspective. Russell Sage Foundation. Retrieved from http://www.diva-portal.org/smash/record.jsf?pid=diva2%3A527102&dswid=-6414

Garrett, J., & Eccles, I. (2009). Linking Late—Adolescent Lifestyles to Family and Work Status in the Mid-Twenties. In *... to work: Globalization, individualization, ...*. Retrieved from https://books.google.nl/books?hl=nl&lr=&id=AUghAwAAQBAJ&oi=fnd&pg=PA243&dq=Garrett,J.L.,&Eccles,J.S.(2009).Transitiontoadulthood:Linkinglateadolescentlifestylestofamilyandworkstatusinthemid-twenties.InI.+Schoon&R.Silbereisen(Eds.),+Transitions+from+sc

Giervveld, J. D. J., Liefbroer, A. C., & Beekink, E. (1991). The effect of parental resources on patterns of leaving home among young adults in the Netherlands. *European Sociological Review*, *7*(1), 55–71.

Goldscheider, F. K., & Goldscheider, C. (1998). The effects of childhood family structure on leaving and returning home. *Journal of Marriage and the Family*, 745–756.

Hardaway, C. R., & McLoyd, V. C. (2009). Escaping poverty and securing middle class status: How race and socioeconomic status shape mobility prospects for African Americans during the transition to adulthood. *Journal of Youth and Adolescence*, *38*(2), 242–256. http://doi.org/10.1007/s10964-008-9354-z

Haurin, D. R., & Brasington, D. (1996). School Quality and Real House Prices: Inter- and Intrametropolitan Effects. *Journal of Housing Economics*, *5*, 351–368. http://doi.org/10.1006/jhec.1996.0018

Hauser, R. M. (2010). Intergenerational Economic Mobility in the United States -- Measures, Differentials, and Trends. *EurAmerica*, *40*(3), 635–681. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=ahl&AN=54307533&site=ehost-live&scope=site

Hertz, T. (2007). Trends in the Intergenerational Elasticity of Family Income in the United States. *Industrial Relations: A Journal of Economy and Society*, *46*(1), 22–50. http://doi.org/10.1111/j.1468-232X.2007.00456.x

Hiekel, N., Liefbroer, A. C., & Poortman, A.-R. (2014). Understanding diversity in the meaning of cohabitation across Europe. *European Journal of Population*, *30*(4), 391–410.

Kalenkoski, C. M., & Pabilonia, S. W. (2010). Parental transfers, student achievement, and the labor supply of college students. *Journal of Population Economics*, *23*(2), 469–496.

Kalil, A., Ryan, R., & Corey, M. (2012). Diverging destinies: maternal education and the developmental gradient in time with children. *Demography*, *49*(4), 1361–83. http://doi.org/10.1007/s13524-012-0129-5

Kennedy, S., & Bumpass, L. (2008). Cohabitation and children's living arrangements: New estimates from the United States. *Demographic Research*, *19*, 1663.

Klijzing, E. (1992). "Weeding"in the Netherlands: First-union disruption among men and women born between 1928 and 1965. *European Sociological Review*, *8*(1), 53–70.

Kornrich, S., & Furstenberg, F. (2013). Investing in children: changes in parental spending on children, 1972-2007. *Demography*, *50*(1), 1–23. http://doi.org/10.1007/s13524-012-0146-4

Lareau, A. (2011). *Unequal childhoods: Class, race, and family life, with an update a decade later. Unequal Childhoods: Class, Race, and Family Life, With an Update a Decade Later*. Retrieved from http://www.scopus.com/inward/record.url?eid=2-s2.0-84887663310&partnerID=tZOtx3y1

Lee, C.-I., & Solon, G. (2009). Trends in Intergenerational Income Mobility. *Review of Economics and Statistics*, *91*(4), 766–772. http://doi.org/10.1162/rest.91.4.766

Lichter, D. T., Qian, Z., & Mellott, L. M. (2006). Marriage or Dissolution? Union Transitions Among Poor Cohabiting Women. *Demography*, *43*(2), 223–240. http://doi.org/10.1353/dem.2006.0016

Liefbroer, a C., & Corijn, M. (1999). Who, what, where, and when? Specifying the impact of educational attainment and labour force participation on family formation. *European Journal of Population = Revue Européenne de Démographie*, *15*(1), 45–75. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12159001

Loomis, L. S., & Landale, N. (1994). Nonmarital Cohabitation and Childbearing among Black and White American Women. *Journal of Marriage and Family*, *56*(4), 949–962. http://doi.org/10.2307/353605

Manning, W. D., & Cohen, J. A. (2015). Teenage cohabitation, marriage, and childbearing. *Population Research and Policy Review*, *34*(2), 161–177.

Manning, W. D., & Smock, P. J. (1995). Why marry? Race and the transition to marriage among cohabitors. *Demography*, *32*(4), 509–520. http://doi.org/10.2307/2061671

Martin, M. A. (2012). Family structure and the intergenerational transmission of educational advantage. *Social Science Research*, *41*(1), 33–47. http://doi.org/10.1016/j.ssresearch.2011.07.005

Martin, P., Schoon, I., & Ross, A. (2008). Beyond Transitions: Applying Optimal Matching

Analysis to Life Course Research. *International Journal of Social Research Methodology*, *11*(3), 179–199. http://doi.org/10.1080/13645570701622025

Mclanahan, S. (2004). Diverging Destinies : How Children Are Faring Under the Second Demographic Transition *, *41*(4), 607–627.

McLanahan, S. (2009). Are Faring Diverging Destinies : Transition. *Demography*, *41*(4), 607–627.

McLanahan, S., & Jacobsen, W. (2015). Diverging Destinies Revisited. In P. R. Amato, A. Booth, S. M. McHale, & J. Van Hook (Eds.), *Families in an Era of Increasing Inequality SE - 1* (Vol. 5, pp. 3–23). Springer International Publishing. http://doi.org/10.1007/978-3-319-08308-7_1

Merton, R. K. (1968). The Matthew effect in science. *Science*, *159*(3810), 56–63.

Miller, B. C. (2002). Family influences on adolescent sexual and contraceptive behavior. *Journal of Sex Research*, *39*(1), 22–26.

Mroz, T. A., & Savage, T. H. (2006). The long-term effects of youth unemployment. *Journal of Human Resources*, *41*(2), 259–293.

Mulder, C. H., & Clark, W. A. V. (2002). Leaving home for college and gaining independence. *Environment and Planning A*, *34*(6), 981–1000.

Oesterle, S., David Hawkins, J., Hill, K. G., & Bailey, J. A. (2010). Men's and women's pathways to adulthood and their adolescent precursors. *Journal of Marriage and Family*, *72*(5), 1436–1453. http://doi.org/10.1111/j.1741-3737.2010.00775.x

Oppenheimer, V. K. (1988). A theory of marriage timing. *American Journal of Sociology*, 563–591.

Oppenheimer, V. K. (2003). Cohabiting and marriage during young men's career-development process. *Demography*, *40*(1), 127–149. http://doi.org/10.2307/3180815

Osgood, D. W., Ruth, G., Eccles, J. S., Jacobs, J. E., & Barber, B. L. (2005). *Six paths to adulthood: Fast starters, parents without careers, educated partners, educated singles, working singles, and slow starters*. na.

Pettit, B., & Western, B. (2004). Mass Imprisonment and the Life Course: Race and Class Inequality in U.S. Incarceration. *American Sociological Review*, *69*(2), 151–169. http://doi.org/10.1177/000312240406900201

Putnam, R. D. (2015). *Our Kids: The American Dream in Crisis*. Retrieved from https://books.google.com/books?hl=nl&lr=&id=1UHqBgAAQBAJ&pgis=1

Raymond, M. (2008). *High school dropouts returning to school*. Statistics Canada, Culture, Tourism and the Centre for Education Statistics.

Raymore, L. A., Barber, B. L., & Eccles, J. S. (2001). Leaving Home, Attending College, Partnership and Parenthood: The Role of Life Transition Events in Leisure Pattern Stability From Adolescence to Young Adulthood. *Journal of Youth and Adolescence*, *30*(2), 197–223. http://doi.org/10.1023/A:1010345825065

Rindfuss, R. (1991). The Young Adult Years: Diversity, Structural Change, and Fertility. *Demography*, *28*(4), 493–512. http://doi.org/10.2307/2061419

Roksa, J., & Velez, M. (2012). A late start: Delayed entry, life course transitions and

bachelor's degree completion. *Social Forces*, *90*(3), 769–794.
http://doi.org/10.1093/sf/sor018

Salmela-Aro, K., Kiuru, N., Nurmi, J.-E., & Eerola, M. (2011). Mapping pathways to adulthood among Finnish university students: Sequences, patterns, variations in family- and work-related roles. *Advances in Life Course Research*, *16*(1), 25–41. http://doi.org/10.1016/j.alcr.2011.01.003

Sassler, S. (2004). Revisiting Jane Austen's Theory of Marriage Timing: Changes in Union Formation among American Men in the Late 20th Century. *Journal of Family Issues*, *25*(2), 139–166. http://doi.org/10.1177/0192513X03257708

Schoen, R., & Cheng, Y. H. A. (2006). Partner choice and the differential retreat from marriage. *Journal of Marriage and Family*, *68*(1), 1–10. http://doi.org/10.1111/j.1741-3737.2006.00229.x

Schulenberg, J., & Schoon, I. (2012). The transition to adulthood across time and space: overview of Special Section. *Longitudinal and Life Course Studies*, *3*(2), 164–172. http://doi.org/10.14301/llcs.v3i2.194

Seltzer, J. A. (2004). Cohabitation in the United States and Britain: Demography, kinship, and the future. *Journal of Marriage and Family*, *66*(4), 921–928.

Shanahan, M. J. (2000). Pathways to Adulthood in Changing Societies: Variability and Mechanisms in Life Course Perspective. *Annual Review of Sociology*, *26*(1), 667–692. http://doi.org/10.1146/annurev.soc.26.1.667

Smith, D. M., & Roberts, R. (2011). Social inequality and young pregnancy: The causal attributions of young parents in London, UK. *Health & Place*, *17*(5), 1054–1060.

South, S. J. (2001). The Variable Effects of Family Background on the Timing of First Marriage: United States, 1969–1993. *Social Science Research*, *30*(4), 606–626. http://doi.org/10.1006/ssre.2001.0714

Spitze, G. D., & Waite, L. J. (1981). Wives' employment: The role of husbands' perceived attitudes. *Journal of Marriage and the Family*, 117–124.

Staff, J., & Mortimer, J. T. (2007). Educational and work strategies from adolescence to early adulthood: Consequences for educational attainment. *Social Forces*, *85*(3), 1169–1194.

Taylor, P., Parker, K., Fry, R., Cohn, D., Wang, W., Velasco, G., & Dockterman, D. (2011). Is college worth it?: College presidents, public assess value, quality and mission of higher education.

Temple, J. A., & Reynolds, A. J. (2007). Benefits and costs of investments in preschool education: Evidence from the Child-Parent Centers and related programs. *Economics of Education Review*, *26*(1), 126–144. http://doi.org/10.1016/j.econedurev.2005.11.004

Thornton, A., Axinn, W. G., & Teachman, J. D. (1995). The Influence of School Enrollment and Accumulation on Cohabitation and Marriage in Early Adulthood. *American Sociological Review*, *60*(5), 762. http://doi.org/10.2307/2096321

Triventi, M. (2014). Does working during higher education affect students' academic progression? *Economics of Education Review*, *41*, 1–13. http://doi.org/10.1016/j.econedurev.2014.03.006

Uecker, J. E., & Stokes, C. E. (2008). Early marriage in the United States. *Journal of*

Marriage and Family, *70*(4), 835–846. http://doi.org/10.1111/j.1741-3737.2008.00530.x

Waite, L., & Gallagher, M. (2002). *The case for marriage: Why married people are happier, healthier and better off financially*. Broadway Books.

Wiik, K. A. (2009). "You"d Better Wait!'- Socio-economic background and timing of first marriage versus first cohabitation. *European Sociological Review*, *25*(2), 139–153. http://doi.org/10.1093/esr/jcn045

Wilcox, W. B., & Stokes, C. E. (2014). The Family Foundation: What Do Class and Family Structure Have to Do with the Transition to Adulthood?. Families in an Era of Increasing Inequality: Diverging Destinies, 5, 147

Winkler-Dworak, M., & Toulemon, L. (2007). Gender Differences in the Transition to Adulthood in France: Is There Convergence over the Recent Period? (Le passage vers l'âge adulte des hommes et des femmes en France: y-a-t-il convergence?). *European Journal of Population / Revue Européenne de Démographie*, *23*(3/4), pp. 273–314.

Figure 3.1 *Women's distribution in career pathways*
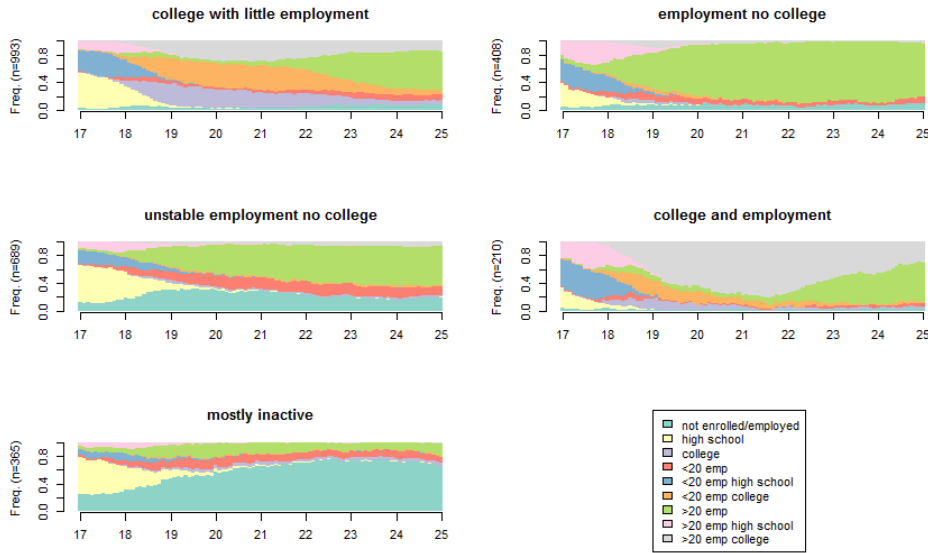
Figure 3.2 *Men's distribution in career pathways*

Figure 4.1 *Women's distribution in demographic pathways*
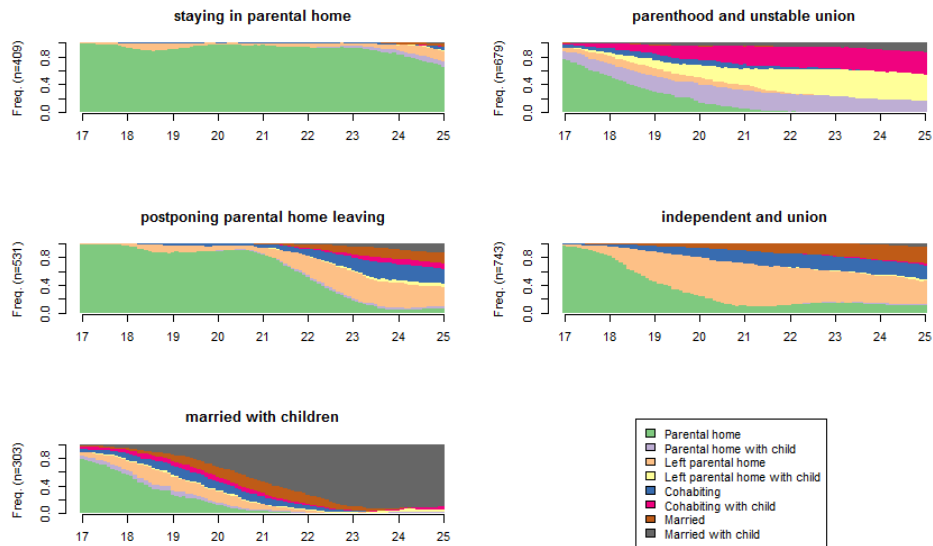
Figure 4.2 *Men's distribution in demographic pathways*
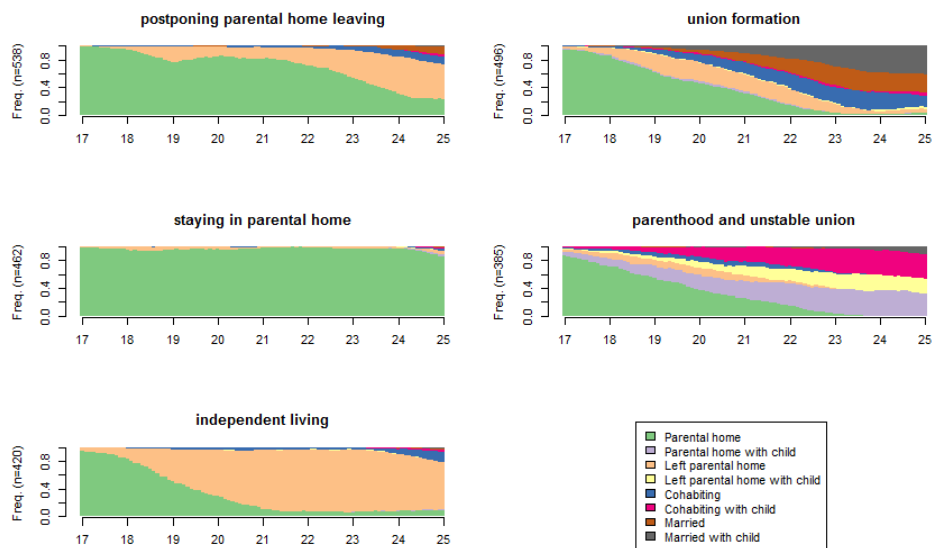
Table 1.1 Women's demographic and career cluster membership

<table>
<tr><td rowspan="18">Careeer pathways =></td><td colspan="6">Demographic pathways =></td></tr>
<tr><td></td><td>Staying in parental home</td><td>Parenthood and unstable union</td><td>Postponing parental home leaving</td><td>Independent and union</td><td>Married with children</td></tr>
<tr><td rowspan="3">College with little employment</td><td>210</td><td>90</td><td>281</td><td>368</td><td>44</td></tr>
<tr><td>152</td><td>253</td><td>198</td><td>277</td><td>113</td></tr>
<tr><td>0.079</td><td>0.034</td><td>0.105</td><td>0.138</td><td>0.017</td></tr>
<tr><td rowspan="3">Employment no college</td><td>53</td><td>96</td><td>76</td><td>124</td><td>59</td></tr>
<tr><td>63</td><td>104</td><td>81</td><td>114</td><td>46</td></tr>
<tr><td>0.020</td><td>0.036</td><td>0.029</td><td>0.047</td><td>0.022</td></tr>
<tr><td rowspan="3">Unstable employment no college</td><td>79</td><td>283</td><td>85</td><td>130</td><td>112</td></tr>
<tr><td>106</td><td>176</td><td>138</td><td>192</td><td>78</td></tr>
<tr><td>0.030</td><td>0.106</td><td>0.032</td><td>0.049</td><td>0.042</td></tr>
<tr><td rowspan="3">College and employment</td><td>28</td><td>24</td><td>53</td><td>94</td><td>11</td></tr>
<tr><td>32</td><td>54</td><td>42</td><td>59</td><td>24</td></tr>
<tr><td>0.011</td><td>0.009</td><td>0.020</td><td>0.035</td><td>0.004</td></tr>
<tr><td rowspan="3">Mostly Inactive</td><td>39</td><td>186</td><td>36</td><td>27</td><td>77</td></tr>
<tr><td>56</td><td>93</td><td>73</td><td>102</td><td>41</td></tr>
<tr><td>0.015</td><td>0.070</td><td>0.014</td><td>0.010</td><td>0.029</td></tr>
</table>

Pearson's Chi2 test=617.791, df=16, p<0.001. Each cell containing: actual count, expected count (under independence condition) and proportion of total.

Table 1.2 Men's demographic and career cluster membership

		Demographic pathways =>				
		Postponing parental home leaving	Union formation	Staying in parental home	Parenthood and unstable union	Independent living
	Unstable employment no college	85 141 0.037	129 130 0.056	142 121 0.062	177 101 0.077	72 110 0.031
	College with little employment	227 133 0.099	77 122 0.033	100 114 0.043	17 95 0.007	146 103 0.063
Careeer pathways =>	College and employment	104 93 0.045	79 85 0.034	104 80 0.045	23 66 0.010	86 72 0.037
	Employment no college	122 171 0.053	211 158 0.092	116 147 0.050	168 123 0.073	116 134 0.050

Pearson's Chi2 test=361.173, df=12, p<0.001. Each cell containing: actual count, expected count (under independence condition) and proportion of total.

43

Table 2.1 Women's career cluster and family background membership

	College with little employment	Employment no college	Unstable employment no college	College and employment	Mostly Inactive
Parental income	Pearson's Chi2 test=346.331, df=16, p<0.001				
Quartile 1	92	75	203	27	140
	200	82	139	42	74
	0.035	0.028	0.076	0.010	0.053
Quartile 2	150	96	134	36	75
	183	75	127	39	67
	0.056	0.036	0.050	0.014	0.028
Quartile 3	212	84	109	55	36
	185	76	128	39	68
	0.080	0.032	0.041	0.021	0.014
Quartile 4	299	65	61	51	20
	185	76	128	39	68
	0.112	0.024	0.023	0.068	0.015
missing	240	88	182	41	94
	240	99	167	51	88
	0.090	0.033	0.068	0.015	0.035
Parental education	Pearson's Chi2 test=328.137, df=8, p<0.001				
High school or less	381	241	386	97	189
	482	198	335	102	177
	0.143	0.090	0.145	0.036	0.071
Some college	492	96	112	81	53
	311	128	218	66	114
	0.185	0.036	0.042	0.030	0.020
missing	120	71	191	32	123
	200	82	138	42	74

	0.045	0.027	0.072	0.012	0.046
Family structure	Pearson's Chi2 test=195.453, df=12, p<0.001				
Both bio parents	627	175	257	114	116
	480	197	334	101	176
	0.236	0.066	0.097	0.043	0.044
1 bio, 1 stepparent	101	74	95	29	47
	129	53	89	27	47
	0.038	0.028	0.036	0.011	0.018
Single parent	229	145	282	56	166
	327	134	227	69	120
	0.086	0.055	0.106	0.021	0.062
other	34	13	54	10	35
	54	22	38	11	20
	0.013	0.005	0.020	0.004	0.013
Race	Pearson's Chi2 test=116.241, df=12, p<0.001				
White	584	212	265	113	127
	485	199	336	103	178
	0.219	0.080	0.099	0.042	0.048
Black	247	98	239	50	142
	289	119	201	61	106
	0.093	0.037	0.090	0.019	0.053
Non-white Hispanic	155	94	182	46	89
	211	87	146	45	78
	0.058	0.035	0.068	0.017	0.033
Mixed	7	4	3	1	7
	8	3	6	2	3
	0.003	0.002	0.001	0.000	0.003

Each cell containing: actual count, expected count (under independence condition) and proportion of total.

45

Table 2.2 Men's career cluster and family background membership

	Unstable employment no college	College with little employment	College and employment	Employment no college
Parental income	Pearson's Chi2 test=275.143, df=12, p<0.001			
Quartile 1	185 110 0.080	49 103 0.021	38 72 0.017	147 133 0.064
Quartile 2	136 116 0.059	55 109 0.024	80 76 0.035	171 141 0.074
Quartile 3	105 118 0.046	113 110 0.049	82 77 0.036	148 143 0.064
Quartile 4	53 121 0.023	204 131 0.089	106 79 0.046	97 147 0.042
missing	126 140 0.055	146 131 0.063	90 92 0.039	170 169 0.074
Parental education	Pearson's Chi2 test=345.13, df=6, p<0.001			
High school or less	336 293 0.146	161 274 0.070	166 192 0.072	450 355 0.196
Some college	109 201 0.047	339 188 0.147	176 131 0.076	139 243 0.060
missing	160 112	67 105	54 73	144 135

	0.070	0.029	0.023	0.063
Family structure	Pearson's Chi2 test=160.270, df=9, p<0.001			
Married parents	220	398	259	377
	328	308	217	400
	0.096	0.174	0.113	0.165
1 bio, 1 stepparent	87	40	39	97
	69	65	45	84
	0.038	0.017	0.017	0.042
Single parent	247	108	86	216
	172	162	114	210
	0.108	0.047	0.038	0.094
other	45	18	12	37
	30	28	20	40
	0.020	0.008	0.005	0.017
Race	Pearson's Chi2 test=166.80, df=9, p<0.001			
White	237	389	240	388
	330	309	216	399
	0.103	0.169	0.104	0.169
Black	249	94	66	166
	151	142	99	183
	0.108	0.041	0.029	0.072
Non-white Hispanic	115	72	86	170
	116	109	76	141
	0.050	0.031	0.037	0.074
mixed	4	12	4	9
	8	7	5	9
	0.002	0.005	0.002	0.004

Each cell containing: actual count, expected count (under independence condition) and proportion of total.

47

Table 3.1 Women's demographic cluster and family background membership

	Staying in parental home	Parenthood and unstable union	Postponing parental home leaving	Independent and union	Married with children
Parental income	Pearson's Chi2 test=219.460, df=16, p<0.001				
Quartile 1	73	218	88	87	71
	82	137	107	150	61
	0.027	0.082	0.033	0.033	0.027
Quartile 2	84	148	75	129	55
	75	125	98	137	56
	0.032	0.056	0.028	0.048	0.021
Quartile 3	69	92	113	166	56
	76	126	99	138	56
	0.026	0.035	0.042	0.062	0.021
Quartile 4	71	43	141	199	42
	76	126	99	138	56
	0.027	0.016	0.053	0.075	0.016
missing	112	178	114	162	79
	99	164	128	180	73
	0.042	0.067	0.043	0.061	0.030
Parental education	Pearson's Chi2 test=199.693, df=5, p<0.001				
High school or less	220	364	228	321	161
	199	330	258	361	147
	0.083	0.137	0.086	0.120	0.060
Some college	118	110	222	324	60
	128	212	166	233	95
	0.044	0.041	0.083	0.122	0.023
missing	71	205	81	98	82
	82	137	107	150	61

	0.027	0.077	0.030	0.037	0.031
Family structure	Pearson's Chi2 test=169.416, df=12, p<0.001				
Both bio parents	238	202	297	412	140
	196	329	257	360	147
	0.090	0.076	0.112	0.115	0.053
1 bio, 1 stepparent	31	93	69	109	44
	53	88	69	97	39
	0.012	0.035	0.026	0.041	0.017
Single parent	121	321	147	190	99
	134	224	175	245	100
	0.046	0.121	0.055	0.071	0.037
other	15	62	18	31	20
	22	37	29	29	17
	0.006	0.023	0.007	0.012	0.008
Race	Pearson's Chi2 test=275.758, df=12, p<0.001				
White	156	212	295	485	153
	200	331	259	363	148
	0.059	0.080	0.111	0.182	0.057
Black	144	320	135	130	47
	119	198	155	216	88
	0.054	0.120	0.051	0.049	0.018
Non-white Hispanic	109	140	96	122	99
	87	144	113	158	64
	0.041	0.053	0.036	0.046	0.037
Mixed	0	7	5	6	4
	3	6	4	6	3
	0.000	0.003	0.002	0.002	0.002

Each cell containing: actual count, expected count (under independence condition) and proportion of total.

49

Table 3.2 Men's demographic cluster and family background membership

	Postponing parental home leaving	Union formation	Staying in parental home	Parenthood and unstable union	Independent living
Parental income	Pearson's Chi2 test=171.756, df=16, p<0.001				
Quartile 1	70	91	78	124	56
	98	90	84	70	76
	0.030	0.040	0.034	0.054	0.024
Quartile 2	88	108	85	94	67
	103	95	89	74	81
	0.038	0.047	0.037	0.041	0.029
Quartile 3	91	119	90	59	89
	105	97	90	75	82
	0.040	0.052	0.039	0.026	0.039
Quartile 4	155	80	76	25	124
	108	99	92	77	84
	0.067	0.035	0.033	0.011	0.054
missing	134	98	133	83	84
	124	115	107	89	97
	0.058	0.043	0.058	0.036	0.037
Parental education	Pearson's Chi2 test=117.191, df=8, p<0.001				
High school or less	215	258	226	230	184
	260	240	224	186	203
	0.093	0.112	0.098	0.100	0.080
Some college	237	157	140	53	176
	178	164	153	128	139
	0.103	0.068	0.061	0.023	0.076
missing	86	81	96	102	60
	99	92	85	71	78

	0.037	0.035	0.042	0.044	0.026
Family structure	Pearson's Chi2 test=82.018, df=12, p<0.001				
Both bio parents	331	276	278	137	232
	293	271	251	210	228
	0.145	0.121	0.121	0.060	0.101
1 bio, 1 stepparent	53	63	37	62	48
	61	57	53	44	48
	0.023	0.028	0.016	0.027	0.021
Single parent	133	135	124	157	108
	154	142	132	110	120
	0.058	0.059	0.054	0.069	0.047
other	18	21	20	27	29
	27	25	23	19	21
	0.008	0.009	0.009	0.012	0.013
Race	Pearson's Chi2 test=265.802, df=12, p<0.001				
White	351	310	208	106	279
	293	270	252	210	229
	0.153	0.135	0.090	0.046	0.121
Black	99	65	131	196	84
	134	124	115	96	105
	0.043	0.028	0.057	0.085	0.037
Non-white Hispanic	78	114	117	80	54
	104	95	89	74	81
	0.034	0.050	0.051	0.035	0.023
Mixed	10	7	6	3	3
	7	6	6	5	5
	0.004	0.003	0.003	0.001	0.001

Each cell containing: actual count, expected count (under independence condition) and proportion of total.

51

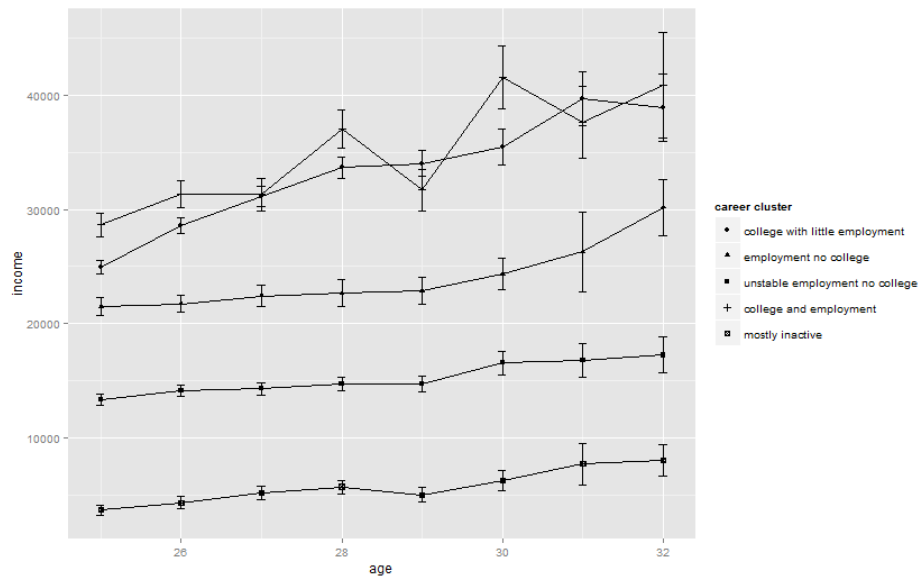Figure 5.1 *Mean income trajectories of women per career cluster*

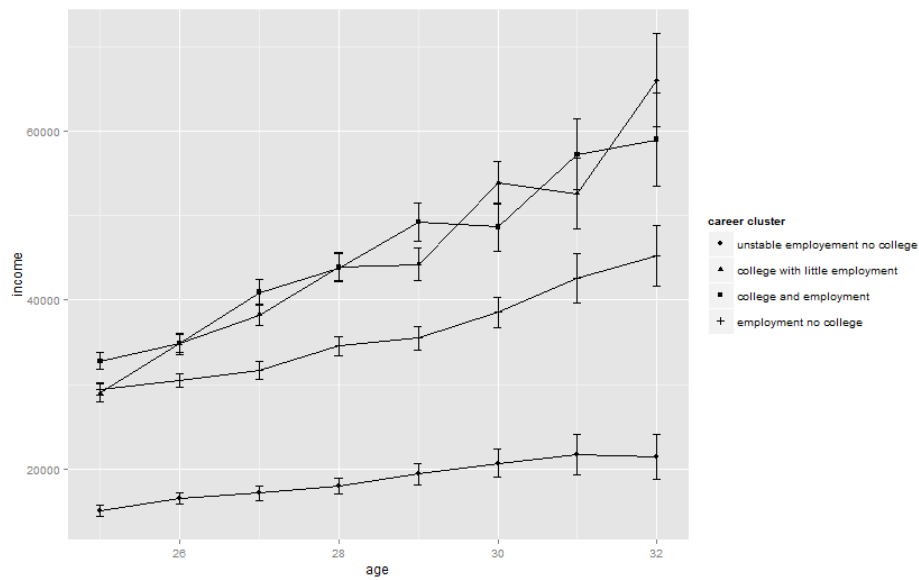Figure 5.2 *Mean income trajectories of men per career cluster*

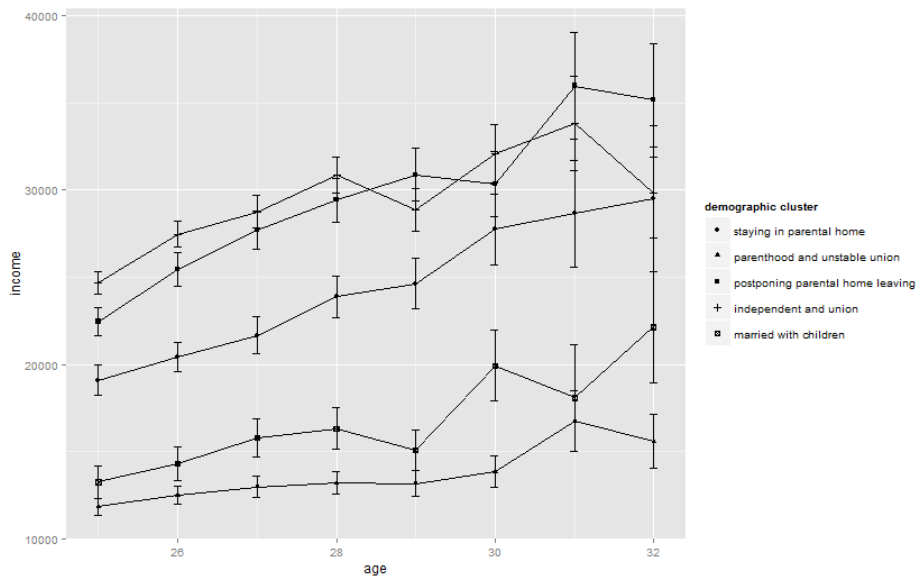Figure 6.1 *Mean income trajectories of women demographic cluster*

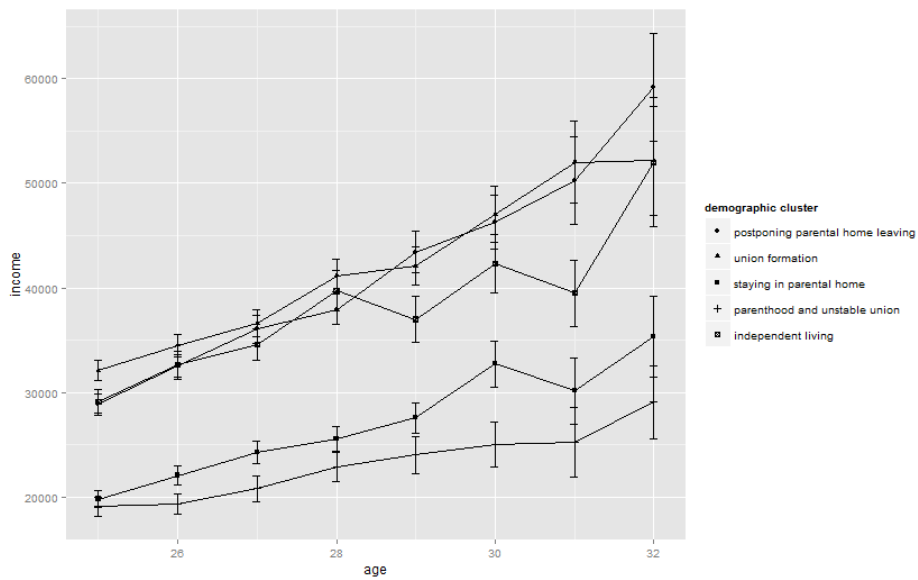Figure 6.2 *Mean income trajectories of men per demographic cluster*

Table 4.1 Growth curve model log income for women

	Model 1		Model 2	
	beta	Standard error	beta	Standard error
Intercepts				
Constant	6.779**	0.319	7.451**	0.327
Parental income (ref.=quartile 1)				
Quartile 2	0.673**	0.258	0.055	0.225
Quartile 3	1.252**	0.262	0.212	0.228
Quartile 4	1.739**	0.263	0.444†	0.234
Unknown	0.717**	0.240	0.020	0.209
Parental education (ref.=missing)				
<high school	0.369+	0.217	0.051	0.185
>high school	0.957**	0.237	0.219	0.205
Family structure (ref.=both bio parents)				
1 bio 1 step parent	-0.134	0.220	0.060	0.185
Single parent	-0.095	0.186	-0.026	0.164
No bio parent	-0.586	0.391	-0.231	0.320
Race				
Black	-0.211	0.180	-0.056	0.158
Hispanic	0.048	0.197	0.184	0.172
Mixed	-0.887	0.763	0.224	0.659
Career pathways (ref=unstable employ. no col.)				
college with little employment			1.111**	0.178
employment no college			1.231**	0.193
college and employment			1.648**	0.227
mostly inactive			-4.014**	0.256
Demographic pathways (ref=stay. par. home)				
parenthood and unstable union			-0.143	0.212
postponing parental home leaving			0.198	0.194
independent and union			0.554**	0.173
married with children			-0.522†	0.275
Random slope				
age	-0.037		-0.122	0.106
Parental income				

Quartile 2	-0.058	0.071	-0.056	0.072
Quartile 3	0.029	0.072	-0.041	0.074
Quartile 4	0.037	0.075	0.040	0.076
Unknown	0.122[†]	0.067	0.120[†]	0.068
Parental education (ref.=missing)				
<high school	0.008	0.061	0.025	0.061
>high school	-0.010	0.069	-0.009	0.069
Family structure (ref.=both bio parents)				
1 bio 1 step paremt	-0.125[*]	0.060	-0.111	0.059
Single parent	-0.061	0.053	-0.034	0.054
No bio parent	-0.073	0.114	-0.067	0.114
Race				
Black	0.035	0.049	0.044	0.050
Hispanic	-0.019	0.057	-0.010	0.057
Mixed	-0.073	0.215	-0.199	0.210
Career trajectories (ref=unstable employ. no col.)				
college with little employment			0.130*	0.058
employment no college			0.007	0.067
college and employment			0.146*	0.073
mostly inactive			0.180*	0.077
Demographic trajectories (ref=stay. par. home)				
parenthood and unstable union			-0.051	0.068
postponing parental home leaving			-0.022	0.059
independent and union			-0.068	0.058
married with children			0.116	0.084
Random effects				
sd (age)	0.540	0.029	0.537	0.029
sd (constant)	2.853	0.069	2.277	0.075
corr (age, constant)	-0.263	0.041	-0.291	0.049
sd (Residual)	2.473	0.044	2.468	0.044

Note ** p<0.01, * p<0.05, † p<0.10

Table 4.2 Growth curve model log income for men

	Model 1		Model 2	
	beta	Standard error	beta	Standard error
Intercepts				
Constant	8.306**	0.278	6.541**	0.335
Parental income (ref.=quartile 1)				
Quartile 2	0.471*	0.221	0.254	0.208
Quartile 3	0.693**	0.233	0.456*	0.218
Quartile 4	0.834**	0.244	0.539*	0.231
Unknown	0.627**	0.215	0.418*	0.206
Parental education (ref.=missing)				
<high school	0.463*	0.196	0.326†	0.195
>high school	0.472*	0.218	0.232	0.216
Family structure (ref.=both bio parents)				
1 bio 1 step parent	-0.277	0.206	-0.239	0.193
Single parent	-0.146	0.171	-0.108	0.161
No bio parent	-0.574	0.366	-0.538	0.348
Race				
Black	-1.105**	0.181	-0.604**	0.174
Hispanic	-0.014	0.164	0.049	0.162
Mixed	-0.474	0.693	-0.483	0.641
Career pathways (ref=unstable employ. no col.)				
college with little employment			1.297**	0.214
college and employment			2.157**	0.187
employment no college			1.888**	0.173
Demographic pathways (ref=stay. par. home)				
postponing parental home leaving			0.752**	0.193
union formation			1.296**	0.181
parenthood and unstable union			0.243	0.227
independent living			0.827**	0.204
Random slope				
age	-0.033	0.077	-0.091	0.099
Parental income (ref.=quartile 1)				
Quartile 2	0.085	0.064	0.080	0.064

Quartile 3	0.102	0.066	0.077	0.066
Quartile 4	0.162 *	0.071	0.103	0.070
Unknown	0.077	0.061	0.044	0.061
Parental education (ref.=missing)				
<high school	-0.042	0.054	-0.003	0.054
>high school	0.034	0.063	0.021	0.063
Family structure (ref.=both bio parents)				
1 bio 1 step parent	-0.029	0.054	0.004	0.056
Single parent	0.060	0.045	0.078 †	0.046
No bio parent	-0.154	0.098	-0.119	0.095
Race				
Black	-0.044	0.047	-0.040	0.049
Hispanic	0.012	0.047	0.023	0.047
Mixed	0.032	0.142	0.028	0.134
Career trajectories (ref=unstable employ. no col.)				
college with little employment			0.244 **	0.062
college and employment			0.071	0.059
employment no college			0.028	0.052
Demographic trajectories (ref=stay. par. home)				
postponing parental home leaving			-0.004	0.053
union formation			-0.057	0.053
parenthood and unstable union			-0.016	0.065
independent living			-0.048	0.059
Random effects				
sd (age)	0.494	0.034	0.476	0.033
sd (constant)	2.480	0.084	2.294	0.080
corr (age, constant)	-0.391	0.055	-0.426	0.053
sd (Residual)	2.115	0.046	2.119	0.046

Note ** p<0.01, * p<0.05, † p<0.10

Session 6B: Methods I

Surveys, Memories and Sequences: The Role of Recall Bias and Survey Mode

Dr. Christian Brzinsky-Fay, Wissenschaftszentrum Berlin für Sozialforschung (WZB)

Extended Abstract

Sequence analysis has relatively high demands regarding data properties: the information to be analysed needs to be longitudinal, is not allowed to have any gaps, and should contain categorical information for the statuses, which is exhaustive and mutually exclusive. Such kind of information is provided by either retrospective life course surveys, or administrative register data, or panel datasets. These kinds of datasets have different advantages and disadvantages regarding sequence data. The validity of sequence information in a particular dataset depends to a large extent on the recall bias and how the survey mode moderates it.

In retrospective data, people were surveyed on one time point and have to remember events of their life course, which might be decades ago. Since this circumstance, there is a large literature on recall error regarding employment careers (e.g. Dex & McCulloch, 1998; Horvath, 1982; Jacobs, 2002). However, whether there is an effect of recall bias on the properties of sequences generated from retrospective data is not researched so far. Administrative data (e.g. on employment histories) are continuously collected, even if the intervals are varying. In general, they are collected by institutions or third persons, so that recall bias doesn't play a role. Of course, administrative data have other disadvantages – such as limited number of variables available or distortion by the data generation process. However, a comparsion of administrative and retrospective data of the same individuals could help to find out, how recall bias influence sequence data.

Furthermore, panel data, where people are surveyed repeatedly, collect the longitudinal information on fixed time points, generally each year. Typical representatives are household panels, such as the German Socio-Economic Panel (SOEP) or the British Household Panel Study (BHPS). The recall bias in these datasets is comparatively low, because the distance between the event and the survey of the event remains small, i.e. people have to take information from their memory after 1.5 years on average. Although the expected recall bias should be the same between different yearly surveyed household panels, we find remarkable differences between the properties of sequences between the SOEP and the BHPS (see figure).

2

In order to assess this problem, I have created comparable samples of 16- to 18-year old individuals from both datasets and look at their employment statuses for five years using only very basic statuses, i.e. employment, unemployment, inactivity and education. First of all, there are structural differences as one would expect: In the UK, young people enter the labour market earlier than in Germany, where young people stay longer in education. This explains the higher share of employment (green) in the UK and the higher share of education (blue) in Germany. The second thing that is eye-catching, are the regular 'waves' in the SOEP data. These regularity doesn't have its origin in the real phenomenon of employment statuses, but in the survey mode applied: in the SOEP, employment careers are surveyed as calendar data, while in the BHPS there are surveyed in an episode format, where in each year there are questions regarding the beginning and the end date of an episode. In the former instance it seems that respondents simply make their crosses for every twelve months, whereas this 'wave' structure seems to be much flatter in the BHPS data. However, the survey mode seems to influence crucially the properties of the data.

The paper proposed aims at clarifying two issues: First, I want to estimate the recall bias of retrospective life course data compared to register data. Here, the data from the National Education Panel Study (NEPS, starting cohort 6) provides the unique possibility to combine the same individuals from the retrospective survey data with administrative data from the German employment service provided by the Institute for Employment Research (IAB). This allows assessing exactly the influence of recall bias on sequence characteristics. Second, with the comparison of the two datasets mentioned above (BHPS and SOEP), I would like to find out more about the effects of different survey modes on sequence characteristics, such as turbulence, episode number etc.

3
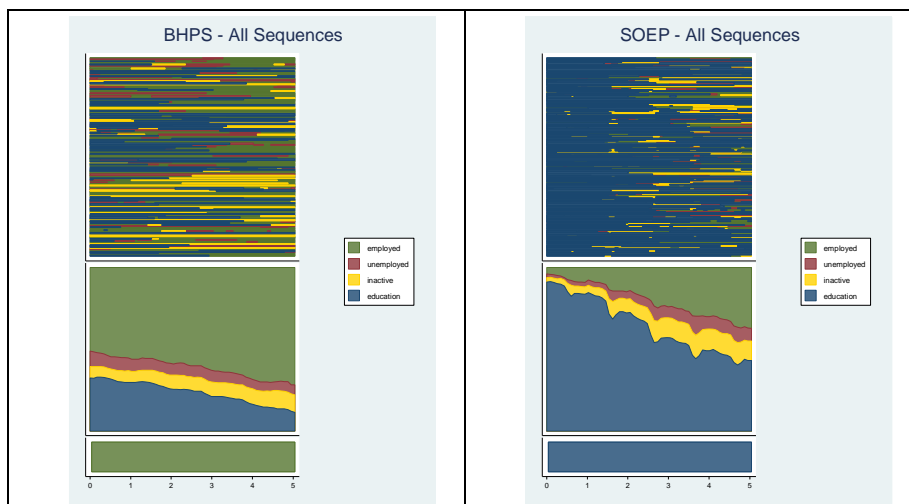
References

Dex, S., & McCulloch, A. (1998). The Reliability of Retrospective Unemployment History Data. Work, Employment and Society, 12(3), 497-509.

Horvath, F. W. (1982). Forgotten Unemployment: Recall Bias in Retrospective Data. Monthly Labor Review(March), 40-43.

Jacobs, S. C. (2002). Reliability and Recall of Unemployment Events Using Retrospective Data. Work, Employment and Society, 16(3), 537-548. doi: 10.1177/095001702762217489

A Complementary Study of Elite Fencing Tactics Using Lag Sequential, Polar Coordinate, and T-Pattern Analyses

Rafael Tarragó [(1)], Xavier Iglesias [(1)], Daniel Lapresa [(2)] and M. Teresa Anguera [(3)]

[(1)] Institut Nacional d'Educació Física de Catalunya (INEFC), Universitat de Barcelona, Spain

[(2)] Facultad de Letras y de la Educación, Universidad de La Rioja, Spain

[(3)] Facultad de Psicología, Universidad de Barcelona, Spain

Introduction

The aim of this study was to perform a diachronic analysis, using three complementary techniques, of behaviors in fencing, a sport in which the exchange of actions between two fencers is determined by a series of decisional processes (Iglesias, Gasset, González and Anguera, 2010; Tarragó, Iglesias, Michavila, Chaverri, Ruiz-Sanchis and Anguera, 2015). The study was performed within the context of systematic observation (Portell, Anguera, Chacón and Sanduvete, 2015; Sánchez-Algarra and Anguera, 2013).

The continuous interchange of actions and reactions during a fencing bout, the aim of which is to gain a touch, is known as a fencing phrase and can be analyzed sequentially from two perspectives. In the first case, the objective is to determine what happens within each fencing phrase, i.e., to identify actions-reactions triggered by the techniques employed by the two fencers; each phrase is characterized by an internal logic that determines the tactics used and in which the different actions executed influence the result (i.e. gain or loss of a touch). In the second case, the objective is to determine how these fencing phrases evolve throughout the bout, i.e., to analyze the diachronic relationships underlying the tactics employed (succession of actions in time), as this can shed light on strategic and tactical decisions that lead fencers to use or modify a certain behavior (technical action), giving rise to the repetition or diversification of sequences of actions, or fencing phrases, during the bout.

We employed an observational methodology design as the actions to be studied were perceivable, regular (i.e. performed repeatedly by professional fencers), and held in a setting that lends itself particularly well to observation. Systematic observation is also an ideal method for collecting data for subsequent analysis of whether behaviors that occur throughout episodes or periods of time have an internal sequential structure (Abbott, 1990, 1995; Abbott and Hrycak, 1990; Abbott and Tsay, 2000).

This internal structure can be analyzed using different data analysis techniques, each with its own algorithms and analytical rules. In our case, we chose lag sequential analysis, polar coordinate analysis, and T-patterns (temporal patterns) detection:

A) Lag sequential analysis (Bakeman and Quera, 2011) was used to detect communication patterns and investigate associated relationships between categories based on the calculation of observed and expected probabilities, and to compare them using a corrected binomial test. This method is applicable to datasets

of behaviors that occur in a certain order; the data can be type I or type II (where the parameter of interest is sequence) or type III or IV (where the parameter of interest is duration). Data types I and III are sequential (unidimensional), i.e., behaviors can never overlap, while data types II and IV are concurrent (multidimensional), i.e., behaviors from different dimensions may overlap (Bakeman, 1978). Sequential analysis can be used to analyze a single dimension within an observation instrument (type I or type III data) or several dimensions simultaneously (type II or type IV data).

We used here an observation instrument with six dimensions. The data were analyzed using a binomial test with a level of statistical significance of p < .05 to statistically compare observed or conditional probabilities (which are computed according to the order of occurrence of the recorded behaviors) with expected or unconditional probabilities (which reflect only the number of occurrences and correspond to the likelihood of chance). Adjusted residuals were calculated to determine the strength of association between behaviors, as we applied the correction established by Allison and Liker (1982) to the binomial test. Lag sequential analysis was performed in both the prospective (positive lags) and retrospective (negative lags) modes to investigate sequences of behaviors that occurred before and after the criterion behavior. Drawing from the experience of many studies conducted in the behavioral and social sciences (Lapresa, Arana, Anguera and Garzón, 2013), it was decided to use just ten lags (lag -5 to lag -1 and lag +1 to lag +5), as patterns appear to become diluted when more are used.

B) Polar coordinate analysis is an elaborate data reduction technique that facilitates the interpretation of data, precisely because of the reduction in the volume of data. It also produces a vectorial image of the complex network of interrelations between categories that make up the different dimensions of an observation instrument. Polar coordinate analysis complements prospective and retrospective sequential analysis (Bakeman, 1978). The first step is to select the main behavior of interest, known as the *focal* behavior, and a series of *conditional* behaviors from among the categories in the observation instrument. The purpose is to investigate the relationships between this central element, the focal behavior, and the other behaviors selected. To do this, it is necessary to have previously calculated the retrospective and prospective adjusted residuals for the focal and conditional behaviors using lag sequential analysis. The retrospective, or "backward" perspective, which incorporates what Anguera (1997) referred to as the concept of "genuine retrospectivity", reveals significant associations between the focal behavior and behaviors that occur *before* this behavior (i.e., negative lags). In our study, this retrospective analysis produced a "mirror-like" image of associations between observation units that occur before the focal behavior; the sequence followed is last, second-last, third-last, etc. In other words, patterns obtained through retrospective sequential analysis reveal patterns formed by categories that lead up to the occurrence of the behavior of interest.

The main objective of polar coordinate analysis is to reduce the volume of conditional probability values calculated previously in the lag sequential analysis, but without losing their significance and interpretative potential. These "results" are processed as data in the polar coordinate analysis and are reduced to a manageable number of significant variables that are presented in an easy-to-interpret vector format that reveals the associations between the different behaviors that make up, in the case at hand, each fencing phrase.

Because the variables produced must adequately reflect the versatility of the situation, the data need to be processed using a powerful data reduction measure. The recommended measure is the Z_{sum} statistic described by Cochran (1954) and subsequently developed by Sackett (1980); it is based on z scores corresponding to relative indices of sequential dependence (Bakeman, 1978). This Z_{sum} is calculated using the following formula $Z_{sum} = \frac{\sum Z}{\sqrt{n}}$, which measures the strength (or associative consistency) between different behaviors. Prospective and retrospective Z_{sum} scores can have a positive or a negative sign, depending on whether the relationship between behaviors is excitatory or inhibitory. These relationships, in turn, can be symmetric or asymmetric, i.e. the focal behavior may be associated with the conditional behavior (or vice versa) in one or both directions. Each conditional behavior is represented by a vector, which, in turn, is located in one of four quadrants (I, II, III, or IV) depending on whether the prospective and retrospective Z_{sum} scores carry a positive or a negative sign. These quadrants indicate whether the focal and conditioned behaviors activate or inhibit each other (see below).

The relationships between behaviors can be represented graphically, with prospective Z_{sum} values on the X axis and retrospective Z_{sum} values on the Y axis. The fact that the conditional behaviors are located in different quadrants according to the relationship with the focal behavior means that it is possible to measure the distance between the origin (0,0) of the Z_{sum} coordinates and the intersection point (or radius), which corresponds to $\sqrt{X^2 + Y^2}$, where X^2 and Y^2 correspond to the Z_{sum} values for the focal and conditioned behavior, respectively. The associated φ angle, $\arcsin e\varphi = \frac{Y}{Radius}$, is calculated based on the number of degrees that are added or subtracted depending on the quadrant in which the conditional behavior is located. The meaning of each vector can then be interpreted objectively as follows:

Quadrant I – The focal and conditional behaviors activate each other.

Quadrant II – The focal behavior inhibits and is activated by the conditional behavior.

Quadrant III – The focal and conditional behaviors inhibit each other.

Quadrant IV – The focal behavior activates and is inhibited by the conditional behavior.

C) Detection of T-Patterns. The assumption underlying the T-pattern detection method is that complex human behaviors have a temporal structure that cannot be fully detected through traditional observational methods or mere quantitative statistical logic. The T-pattern emerges as the fruit of a mathematical process that is automated in the form of an algorithm in the THEME software program, first developed by Magnusson (1988, 1989, 1993, 1996, 2000, 2005, 2006, 2016) around thirty years ago and progressively improved up to the current version (Theme v. 6 Edu). By detecting T-patterns, or "temporal patterns", this method can detect structural analogies across very different levels of organization and enable an important shift from quantitative to structural analysis.

T-patterns detection studies have been conducted in very different scientific domains and also in the field of sport (Aragón, Lapresa, Arana, Anguera and Garzón, 2015; Borrie, Jonsson and Magnusson, 2002; Chaverri, Camerino, Anguera, Blanco-Villaseñor and Losada, 2010; Gutiérrez-Santiago, Prieto,

Camerino and Anguera, 2013; Lapresa, Alsasua, Arana and Garzón, 2013; Lapresa, Álvarez, Arana, Garzón and Caballero, 2013; Tarragó et al, 2015; Zurloni, Cavalera, Diana, Elia and Jonsson, 2014). Since observational records of human behavior have a temporal and sequential structure, an analytical tool that can describe this structure can only enhance the understanding of the target behavior(s). In fencing, for example, T-pattern analysis can reveal the hidden yet stable structures that underlie the interactions that determine what occurs in a competition. The discovery of hidden T-patterns could help fencing coaches to better predict the behaviors of both competitors in a fencing bout thanks to an integrated system that allows for an increased depth of analysis.

For the T-pattern analysis in this study, we used the software program THEME v. 6 Edu, and assigned a constant duration (=1) to each event-type (Lapresa, Alsasua, et al., 2013), as what was important in our analysis was not the duration of each fencing phrase, or the distance between phrases (which is very similar), but rather their internal sequentiality.

The calculations for the methods employed in this study were made using the following software programs, which are all freely available: GSEQ5 (lag sequential analysis) (Bakeman and Quera, 2011), HOISAN v.1.6.3. (lag sequential analysis and polar coordinate analysis) (Hernández-Mendo, López-López, Castellano, Morales-Sánchez and Pastrana, 2012), and THEME v.6 Edu (T-Pattern detection) (Magnusson, 1996, 2000).

The ultimate aim of this empirical study was to analyze the confluence of results from three complementary methods to identify direct relationships between different tactical actions and their efficacy (in terms of score) and to also shed light on successful strategies that could help fencers take better decisions during competitions.

Method

We employed a nomothetic, point, multidimensional observational study design (Anguera, Blanco-Villaseñor, Hernández-Mendo and Losada, 2011) in which we analyzed eight bouts from the Men's Individual Saber Finals of the World Fencing Championships in Budapest, Hungary (2013) and Kazan, Russia (2014). The data were recorded using LINCE v.1.1. (Gabin, Camerino, Anguera and Castañer, 2012), and a total of 407 fencing phrases were recorded using the type II data modality. The images were obtained from broadcasts made available by the *Fédération Internationale d'Escrime* (FIE) and Televisora Venezolana Social (Tves) through the Internet (YouTube).

The study formed part of a larger fencing study approved by clinical research ethics committee of the Catalan Public Sports Authority (2005). According to the ethical requirements specified in the guidelines of the American Psychological Association (American Psychological Association, 2002), the informed consent of the fencers was not required because the study was an observational study of a publicly broadcast event held in a natural setting.

The observation instrument ESGRIMOBS (Tarragó et al., 2015) was adapted for the purpose of the study. The data for the study were systematically recorded and coded using the following instrument (Table 1):

Tabla 1. ESGRIMOBS Observation Instrument (Tarragó et al., 2015)

Dimensions or criteria	Category systems	Code
Pressure	Right pressure	PD
	Left pressure	PI
	No pressure	NP
Preparation	Right preparation	XD
	Left preparation	XI
1st Action	1st action: right offensive	OD
	1st action: left offensive	OI
	1rt action: right defensive	DD
	1st action: left defensive	DI
2nd Action	2nd action: right offensive	DOD
	2nd action: left offensive	DOI
	2nd action: right counteroffensive	DCD
	2nd action: left counteroffensive	DCI
	2nd action: right defensive	DDD
	2nd action: left defensive	DDI
3rd Action	3rd action: right offensive	TOD
	3rd action: left offensive	TOI
	3rd action: right counteroffensive	TCD
	3rd action: left counteroffensive	TCI
	3rd action: right defensive	TDD
	3rd action: left defensive	TDI
Touch	Left touch	TI
	Right touch	TD
	No touch	NT
	Double touch	DT

Each of the six criteria in the observation instrument (shown in the left column of Table 1) corresponds to a dimension of fencing, and each of these dimensions is further broken down into a system of exhaustive and mutually exclusive categories.

The data were recorded by applying the observation instrument to each observation unit. Given the complexity of analyzing fencing tactics, we based our analysis on observation units drawn from concepts described in fencing rules published by the FIE (FIE, 2014). Each unit was formed by a fencing phrase, the aim of which is to gain a touch, and each phrase was analyzed by applying the different criteria that formed the observation instrument.

The reliability of the dataset was duly analyzed prior to subsequent analyses (Blanco-Villaseñor and Anguera, 2000). The construct validity of ESGRIMOBS was guaranteed by the consistency and robustness of the concepts on which it was based, which were drawn from the theoretical framework of fencing and the critical evaluation of the observation instrument by 17 fencing masters. A pilot run of the instrument, with coding by the 17 fencing masters, produced a canonic agreement of 0.81 (Krippendorf, 2004), allowing us to consider the instrument valid.

Results

In view of the primary aim of this study, which was to analyze diachronic relationships underlying the tactics employed by two of the world's top fencers using three complementary techniques,

we have created a table (Table 2) summarizing "favorable" and "unfavorable" tactical sequences shown by each technique for the two fencers. In the T-pattern analysis, a behavior was considered favorable when the set of event-types contained the event "touch", and in the lag sequential and polar coordinate analyses, it was considered favorable when it activated a touch for the fencer or inhibited a touch by his opponent. It should be noted that in the case of antagonistic significances for the different lags in the sequential analysis, lag values of 0 or close to 0 were prioritized.

To our knowledge, this study is the first to undertake a complementary analysis of this type, and it is of particular interest that our interpretation was focused entirely on the analysis of the strategies and tactics employed by the fencers analyzed.

The summary table (Table 2) and the analyses described partially in the results and discussion sections provide detailed insight into the tactics used by the two fencers in each bout, as each of the techniques provides a distinct yet complementary interpretation. Each dataset corresponds to a fencing phrase, which could be considered as a "co-occurrence" of behaviors or actions but that actually presents internal or intra-phrase sequentiality (left-right sequentiality based on the order of transcription). In other words, our analysis shows how behaviors influence other behaviors within each fencing phrase. This relationship links tactical behaviors with specific technical executions, analyzed by sequential analysis at lag 0 and T-pattern analysis.

The strategical analysis is also based on between-phrase relationships, i.e., relationships between each fencing phrase (or datasets arranged from top to bottom). This shows how co-occurrences or events (fencing phrases) influence each other (via T-patterns described in each dendrogram that links two or more phrases with significant associations), and it also shows how different behaviors (actions) influence previous or subsequent behaviors (using sequential analysis of lags -1 to -5 and +1 to +5 and polar coordinate analysis).

The main conclusion of this study is that the complementary use of the three observational methodology techniques—lag sequential analysis, polar coordinate analysis, and T-pattern detection—can provide extremely useful insights that can be used to guide tactical and strategic training in fencing. Our analysis shows that significant conclusions regarding tactical behavior in fencing can be reached using objective analytical techniques. This and similar studies could help to guide training strategies for fencing masters by providing objective data to complement largely subjective judgements based on experience.

Table 2 Complementary evaluation of tactical sequences for fencers 1 and 2 shown by technique

T-Patterns	Lags		Polar coordinate	
Behaviors that FAVOR participant 1 (Fencer on the left)				
(NP,XD,OI,DDD,TI PD,XD,OI)	PD	DDD	PD	OD
(PI,XD,OD,DCI,TD PI,XI,OI,DCD,TI)	XD	DCD	PI	DCD
	OI	TOI	XD	DDI
Behaviors that DO NOT FAVOR participant 1 (Fencer on the left)				
(NP,XI,OD,DCI,TD NP,XI,OD,DDI,TDD)	NP	DDI	NP	
(PD,XD,OI,DDD PD,XD,OD,DCI,TD)	XI	DOI	PI	
(PI,XD,OD,DCI,TD PI,XI,OD,DDI)	DI	TDD	DDD	
(PI,XD,OD,DCI,TD PI,XI,OI,DCD,TI)	OD	TOD	DOD	
	DCI			
Behaviors that FAVOR participant 2 (Fencer on the right)				
((NP,XI,OI,DDD NP,XI,OD,DCI) PI,XI,OD,DCI,TD)	NP	DDI	NP	
(NP,XI,OD,DCI PI,XI,OD,DCI,TD)	PD	DOI	DOI	
(NP,XI,OD,DDI,TD PI,XI,OD,DDI)	OD	TCD	TCD	
(NP,XI,OI,DCD,TDI PI,XI,OD,DDI,TD)	DCI			
(PI,XI,OI,DCD,TI NP,XI,OD,DCI,TD)				
(PI,XI,OI,DDD PI,XI,OD,DCI,TD)				
Behaviors that DO NOT FAVOR participant 2 (Fencer on the right)				
(NP,XI,OI,DDD PI,XI,OI,DCD,TCI,TI)	PI	DDD	NP	
(PD,XD,OD,DCI PI,XI,OI,DOD,TI)	XI	TCI	PI	
	OI	TOD	PD	
	DCD		TOD	

References

Abbott, A. (1990) A primer on sequence methods, *Organization Science*. **1**, 373-392.

Abbott, A. (1995) Sequence analysis, *Annual Review of Sociology*. **21**, 93-113.

Abbott, A. and Hrycak, A. (1990) Measuring resemblance in sequence data: An optimal matching analysis of musicians' careers, *American Journal of Sociology*. **96**(1), 144-185.

Abbott, A. and Tsay, A. (2000) Sequence analysis and optimal matching methods in Sociology, *Sociological Methods & Research*. **29**(1), 3-33.

Allison, P. D. and Liker, J. K. (1982). Analyzing sequential categorical data on dyadic interaction: A comment on Gottman. *Psychological Bulletin,* **93**, 393–403.

American Psychological Association (2002) Ethical principles of psychologists and code of conduct, *American Psychologist*. **57**(12), 1060-1073.

Anguera, M.T. (1997) From prospective patterns in behavior to joint analysis with a retrospective perspective, in *Colloque sur invitation «Méthologie d'analyse des interactions sociales»,* Paris: Université de la Sorbonne.

Anguera, M.T., Blanco-Villaseñor, A., Hernández-Mendo, A. and Losada, J.L. (2011) Diseños observacionales: ajuste y aplicación en psicología del deporte, *Cuadernos de Psicología del Deporte.* **11**(2), 63-76.

Aragón, S., Lapresa, D., Arana, X., Anguera, M.T. and Garzón, B. (2015) Tactical behaviour of winning athletes in major championship 1500-m and 5000-m track finals, *European Journal of Social Sciences.* DOI: 10.1080/17461391.2015.1009494.

Bakeman, R. (1978) Untangling streams of behavior: Sequential analysis of observation data, in G.P. Sackett (ed.), *Observing Behavior*, vol. 2, pages 63-78, Baltimore: University of Park Press.

Bakeman, R. and Quera, V. (2011) *Sequential analysis and observational methods for the behavioral sciences*: Cambridge University Press.

Blanco-Villaseñor, A. and Anguera, M.T. (2000) Evaluación de la calidad en el registro del comportamiento: Aplicación a deportes de equipo, in E. Oñate, F. García-Sicilia and L. Ramallo (eds.), *Métodos numéricos en ciencias sociales*, pages 30-48, Barcelona: Centro Internacional de Métodos Numéricos en Ingeniería (CIMNE).

Borrie, A., Jonsson, G.K. and Magnusson, M.S. (2002) Temporal pattern analysis and its applicability in sport: an explanation and preliminary data, *Journal of Sport Science.* **20**(10), 845 - 852.

Chaverri, J., Camerino, O., Anguera, M.T., Blanco-Villaseñor, Á. and Losada, J.L. (2010) Interaction contexts in soccer: Detection of T-Patterns, *Gymnasium. Revista de Educação Física, Desporto e Saúde.* **2**(1), 69-92.

Cochran, W.G. (1954) Some methods for streghning the common χ^2 test, *Biometrics.* **10**, 417-451.

Comité de ética de investigaciones clínicas de la Administración deportiva de Cataluña (2005) DOGC 4347 de 21/3/2005, orden PRE/99/2005, March 10, pages 6998-6999. Retrieval: http://portaldogc.gencat.cat/utilsEADOP/PDF/4347/417073.pdf

FIE (2014). Reglamentos. Retrieval 2014, June 20: http://www.fie.ch/Fencing/Rules.aspx

Gabin, B., Camerino, O., Anguera, M.T. and Castañer, M. (2012) Lince: multiplatform sport analysis software, *Procedia-Social and Behavioral Sciences.* **46**, 4692-4694.

Gutierrez-Santiago, A., Prieto, I., Camerino, O. and Anguera, M.T. (2013) Sequences of errors in the judo throw Morote Seoi Nage and their relationship to the learning process *Journal of Sports Engineering and Technology.* **227**(1), 57-63.

Iglesias, X., Gasset, A., González, C. and Anguera, M.T. (2010) Interacción competitiva y presión ambiental en deportes de combate: Aplicación de la metodología observacional. *Revista Iberoamericana de Psicología del Ejercicio y el Deporte.* **5**(2), 267-282.

Krippendorff, K. (2004) *Content Analysis. An introduction to its methodology* (2a. ed.). Thousand Oaks, CA: Sage.

Lapresa, D., Alsasua, R., Arana, J. and Garzón, B. (2013). Comparative analysis of *T-patterns* using real time data and simulated data by assignment of conventional durations: the construction of efficacy in children's basketball, *International Journal of Performance Analysis in Sport.* **13**(2), 321-339.

Lapresa, D., Arana, J., Anguera, M.T. and Garzón, B. (2013) Comparative analysis of the sequentiality using SDIS-GSEQ and THEME: a concrete example in soccer, *Journal of Sport Sciences.* **31**(5), 1687-1695.

Lapresa, D., Álvarez, L., Arana, J., Garzón, B. and Caballero, V. (2013) Observational analysis of the offensive sequences that ended in a shot by the winning team of the 2010 UEFA Futsal Championship, *Journal of Sports Science.* **31**(5), 1731-1739.

Magnusson, M.S. (1988). Le Temps et les patterns syntaxiques du comportement humain: Modele, methode et programme THEME. *Revue des Conditions de Travail*, hors série, 284-314.

Magnusson, M.S. (1989). Structure syntaxique et rythmes comportementaux: Sur la detection de rythmes caches. *Science Technical Animal Laboratoire,* **14**(2), 143-147.

Magnusson, M.S. (1993). *THÈME user's manual: With notes on theory, model and pattern detection method*. Reykjavik, University of Iceland.

Magnusson, M.S. (1996) Hidden real-time patterns in intra- and inter-individual behavior: Description and detection, *European Journal of Psychological Assessment.* **12**(2), 112-123.

Magnusson, M.S. (2000) Discovering hidden time patterns in behavior: *T-patterns* and their detection, *Behavior Research Methods, Instruments, & Computers.* **32**(1), 93-110.

Magnusson, M. S. (2005). Understanding social interaction: Discovering hidden structure with model and algorithms. In L. Anolli, S. Duncan, M. Magnusson y G. Riva (Eds.), The hidden structure of interaction: From Genomics to Culture Patterns (pp. 4-24). Amsterdam: IOS Press.

Magnusson, M.S. (2006) Structure and communication in interaction. In G. Riva, M.T. Anguera, B.K. Wiederhold y F. Mantovani (Eds.) *From Communication to Presence: Cognition, Emotions and Culture towards the Ultimate Communication Experience* (pp. 127-146). Amsterdam: IOS Press.

Magnusson, M.S. (2016). Time and self-similar structure in behavior and interactions: From sequences to symmetry and fractals. In M.S. Magnusson, J.K. Burgoon and M. Casarrubea (Coords), *Discovering hiden temporal patterns in behavior and interaction* (pp. 3-35). New York: Springer.

Portell, M., Anguera, M.T., Chacón, S. and Sanduvete, S. (2015). Guidelines for Reporting Evaluations based on Observational Methodology (GREOM). *Psicothema,* **27**(3), 283-289.

Sackett, G.P. (1980) Lag Sequential Analysis as a data reduction technique in social interaction research, in D.B. Sawin, R.C. Hawkins, L.O.Walker and J.H. Penticuff (eds.), *Exceptional infant.*

Phychosocial risksin infant-environment transactions pages 300-340. New York: Brunner/Mazel.

Sánchez-Algarra, P. and Anguera, M.T. (2013). Qualitative/quantitative integration in the inductive observational study of interactive behaviour: Impact of recording and coding predominating perspectives. *Quality & Quantity. International Journal of Methodology,* **47**(2), 1237-1257.

Tarragó, R., Iglesias, X., Michavila, J.J., Chaverri, D., Ruiz-Sanchis, L. and Anguera, M.T. (2015) Análisis de patrones en asaltos de espada de alto nivel, *Cuadernos de Psicología del Deporte.* **15**(1), 149-158.

Zurloni, V., Cavalera, C., Diana, B., Elia, M. and Jonsson, G.K. (2014) Detecting regularities in soccer dynamics: A T-pattern approach, *Revista de Psicología del Deporte.* **23**(1), 157-164.

Session 8A: Parenthood and childhood

Early Parenthood and Inequalities in Family and Work Trajectories.

Experiences of women and men in urban Mexico

Marta Mier-y-Terán * **and Ana Karina Videgain** **

   * Institute of Social Research, National University of Mexico
   ** Metropolitan Autonomous University, Mexico

Extended abstract

During the second half of the twentieth century, the expansion of the educational system in Mexico has achieved an important goal: children and youth increasingly attend and remain in school. Gender inequalities in school attendance and attainment have disappeared but inequalities between socioeconomic sectors remain. Most students only attain low middle education and access to higher levels continues to be restricted to higher socioeconomic sectors.

Growing education levels and a rapid urbanization process are related to an important increase in female labor force participation. However, female participation remains relatively low and labor markets have been growingly unstable, precarious and segmented, and better job opportunities greatly differ by education levels.

At the same time, the fertility transition has taken place in Mexico. Family size has been notably reduced. In spite of this and other profound transformations related to family values, patterns during the first stages of family formation have shown only minor changes in the population as a whole. Most women enter a conjugal union and have the first child at relatively early ages. Only a minority of women delay their marriage and childbearing period. These different behaviors are strongly associated to education inequalities and unequal labor perspectives.

Early ages at birth of first child are associated to economic and social disadvantages. Maternity at early ages is related to low education and participation in

the labor market. When young mothers work, they have more precarious jobs and lower incomes than older mothers. Early timing of fertility can act as a trigger of accumulation of disadvantages in later stages of life. At the same time, having an early birth might be a consequence of poverty because of previous disadvantages young women face before becoming pregnant (Cavenaghy y Rodríguez, 2014). It has been pointed out that the opportunities of Mexican women who become pregnant as teenagers to receive education and progress toward rewarding and fulfilling jobs are limited, regardless of their fertility decisions.

However, most studies on teenage fertility that show the association with economic and social disadvantages are cross-sectional or follow ups for short periods of time. In Mexico, it is unknown if early mothers catch up later on in other spheres of their lives and end having similar life paths as those of women of the same social sector who started their childbearing period at latter ages.

Mexican males' reproductive patterns have been scarcely studied and early fatherhood has not been a matter of concern. However, given the male bread winner traditional role, adopting the responsibility of a child and a family at early ages might also affect their family and work life paths.

In this paper, our objective is to study family and work trajectories of individuals and determine if early transitions to parenthood are associated to adverse consequences in later stages of life, in the sense that previous socioeconomic inequalities would be exacerbated with an early parenthood, and early mothers and fathers would end up in less successful life paths. We study women's as well as men's life paths, given our interest in both experiences and in gender inequalities.

The main contributions of the paper are the use of a life course perspective to observe long term trends; the simultaneous analysis of the interwoven school, work, marital and reproductive trajectories during a period deep social transformations; the inclusion of men's as well as female's experiences.

1.1 Methodology and data source

A Multichannel Sequence Analysis (additive) is applied in order to simultaneously study the trajectories in the family and work domains. A distances matrix is calculated by applying optimal matching analysis with constant costs. In order to group

3

the life paths, an agglomerative hierarchical clustering of the distances matrix data is computed; the Average Silhouette Width is used to define the number of groups.

We analyze school, work, marital and reproductive dimensions. States are in school and out of school for the first dimension. Work states are at home, non-manual high qualification, non-manual low qualification, informal commerce, manual high qualification and manual low qualification. In the marriage dimension, states are previously married, single, in first, second and third or more marital union (marriage or cohabitation). In the reproductive dimension states are the number of children ever born that goes from zero to 5 or more.

To explore the social composition of the life paths groups defined by the cluster analysis, logistic regressions models are estimated. Cluster group is the dependent variable, and birth cohort, social strata and having had an early birth and their interactions are the explanatory variables.

The data source is the Retrospective Demographic Survey (Eder 2011), carried out in urban areas of Mexico. We use annual data on school, work, and family life histories from ages 12 to 41 of 888 women and 851 men of two birth cohorts born in 1951-1953 and 1966-1969.

An indicator of the social strata is created with data on education and economic features of parents of the interviewees at age 15. This variable is particularly relevant given the profound social inequalities that characterize de Mexican society; it reflects the parent's education and the economic conditions of the family household when the individual was about to start his family and work trajectories.

Because of the prevalence of the traditional gender roles, women and men's work and family experiences are analyzed separately. Besides, nuptiality and fertility timing among men take place at later ages than among women and we are especially interested in the effects of the social timing. We consider an early motherhood when occurred before age 18 and an early fatherhood when occurred before age 20. At age 18, young people attain the age of majority, are supposed to finish the upper secondary school which is compulsory, and are legally able to have full time jobs. As mean differences by sex in age at first marriage and first birth are around two years, we suppose that it is an early fatherhood when it occurs before age 20.

MAIN RESULTS

Women. For theoretical as well as for statistical reasons, we chose the five group cluster which suggests the existence of five well differentiated groups in the four simultaneously analyzed domains.

Regarding the four domains, we could say there are more and less successful life paths. Higher education, stable non manual work sequences, stable marriages and small family sizes would be perceived as the most successful in this period of social transformations related to increasing levels of education, growing female labor participation and fertility transition.

Almost all women who had an early birth are in life paths 1 and 5. They are both characterized by low education and big family size. The greatest differences between them is that in the first women remain at home and have stable marital life whereas in life path 5 women work in informal commerce or manual jobs, mainly low qualified, and have more unstable marital life. Life path 1 is the most common, one third of women are in this group.

5
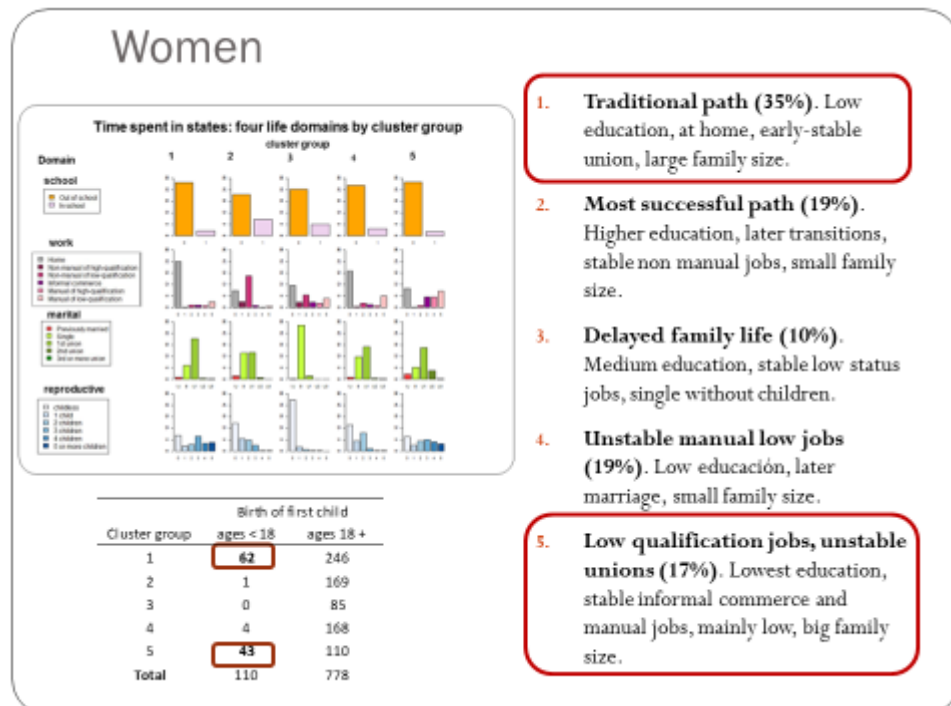
In order to investigate whether within the two cluster groups, early mothers' trajectories in the four domains were similar to those of other women who postpone their first birth, we graphed the trajectories making this distinction. We find that early mothers had in both cases bigger family sizes, and more unstable marital unions in life path 5.
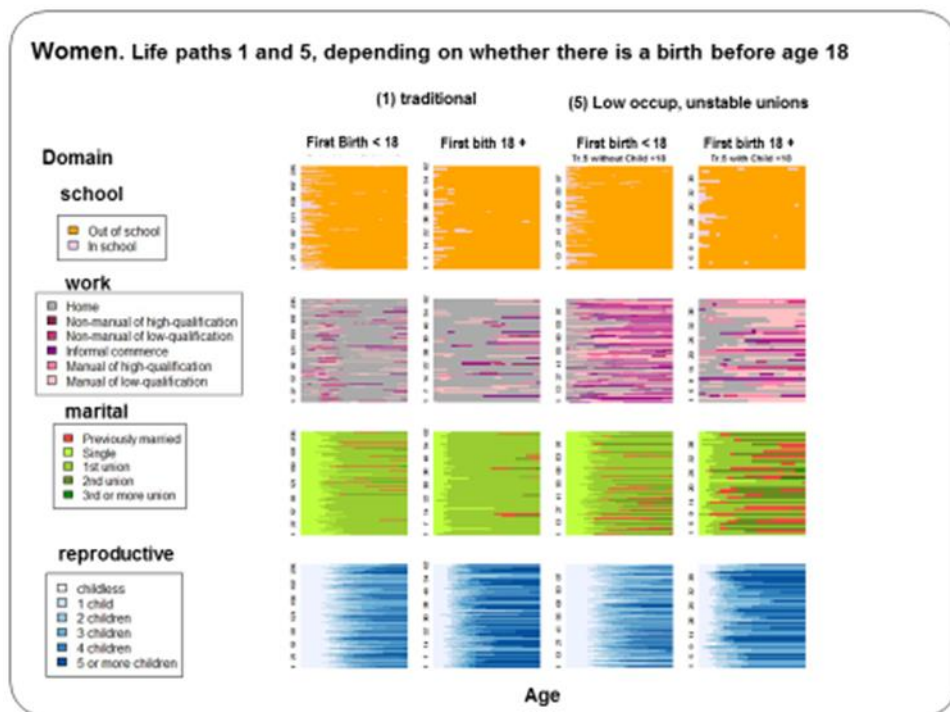
Men. We chose the six group cluster because of theoretical and statistical reasons. In the case of men, the life paths are more unevenly distributed in the groups than in the case of women.

Having an early birth is less common among men than among women. Early fathers are mainly in life paths 3 and 6, both characterized by very low education and manual jobs, and big family size. However, jobs are more precarious, marital life more unstable and family size bigger among the last one.
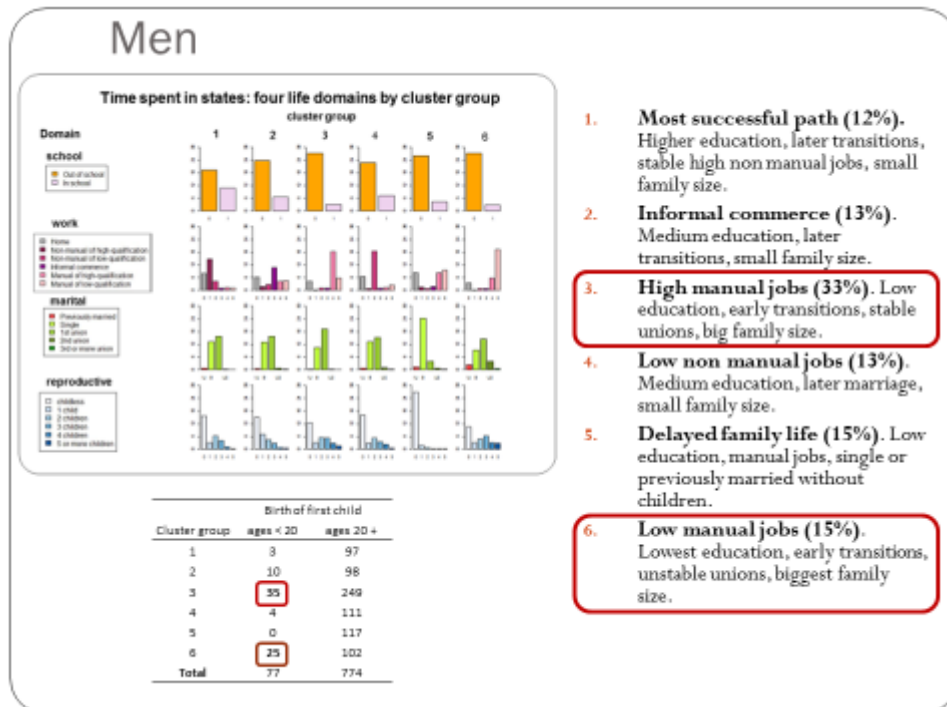
Within life paths 3 and 6, early fathers have higher family size, and more marital instability among the last life path.

7

Men. Life paths 3 and 6, depending on whether there is a birth before age 20
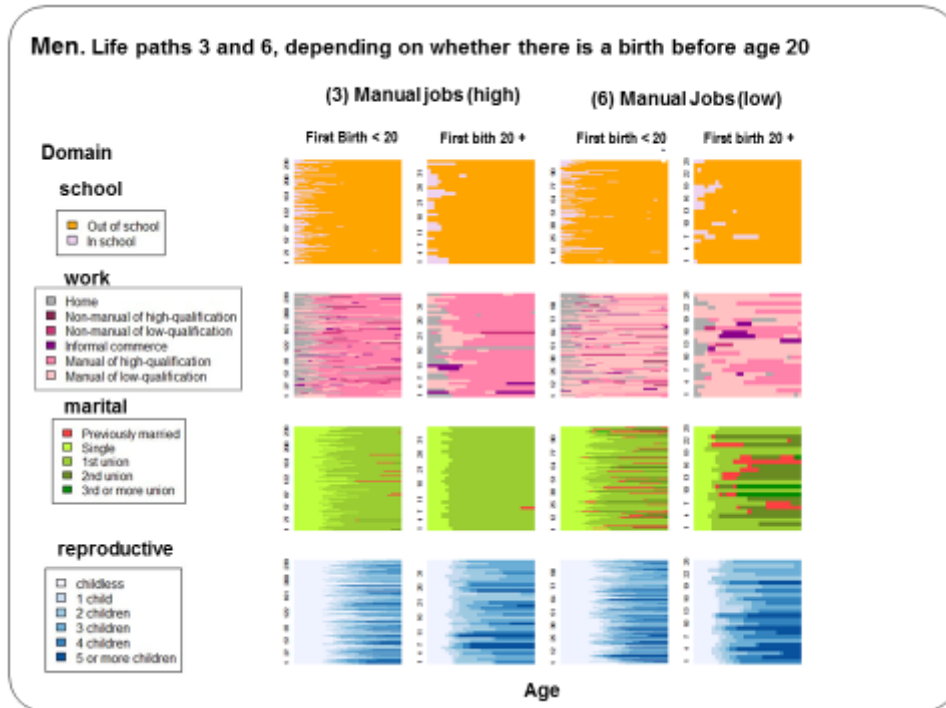
Composition of the cluster groups where early parenthood occurs.

Results of logistic regression models

	Women		Men	
	1. Traditional path. Low education, at home, early-stable union, large family size.	5. Low qualification jobs, unstable unions, big family size.	3. Manual jobs (high). Low education, early transitions, stable unions, big family size.	6. Manual Jobs (low). Lowest education, early transitions, unstable unions, biggest family size.
Intercept		(-) ***	(-) **	(-) ***
Medium Social Strata (MedSS)			(-) *	(-) **
High Social Strata (HighSS)	(-) ***	(-) **	(-) ***	(-) ***
Youngest Birth Cohort (YBC)	(-) *			
Early birth of first child (EBFCh: women < 18; men < 20)	(+) **	(+) **		(+) **
Significant interactions				
MedSS * EBFCh			(+) &	
HighSS* EBFCh				(+) &
YBC * EBFCh		(+) **		

p value *** < .001   ** < .05   * < .1   & < .2

9

References

Abbot, Andrew and A. Tsay (2001) "Sequence Analysis and Optimal Matching Methods in Sociology. Review and Prospect" Sociological Methods and Research 29(1): 3-33.

Ali, Mohamed M y John Clealand (2005) "Sexual and reproductive behaviour among single women aged 15–24 in eight Latin American countries: a comparative analysis" *Social Science & Medicine*, 60 (6): 1175-1185.

Arceo-Gomez, Eva O. and Raymundo M. Campos-Vazquez (2014). Teenage Pregnancy in Mexico: Evolution and Consequences, *Journal of Economics* 51(1), 109-146.

Billari, Francesco C. (2004) "Becoming an adult in Europe: A macro (/micro)-demographic perspective", Demographic Research, Special Collection 3(2):15–44. doi:10. 4054/DemRes.2004.S3.2.

Billari, Francesco C. and Raffaella Picarreta (2005) "Analyzing Demographic Life Courses through Sequence Analysis" Mathematical Population Sudies: An International Journal of Mathematical Demography, 12:81-106. http://dx.doi.org/10.1080/08898480590932287; download at the UNAM April 1rst 2013.

Blossfeld, Hans-Peter and Johannes Huinink (1991) "Human Capital Investments or Norms of Role Transition? How Women's Schooling and Career Affect the Process of Family Formation", American Journal of Sociology, Vol. 97, No. 1 (Jul., 1991), pp. 143-168

Brewster, K.L. and Rindfuss, R.R. (2000) "Fertility and women's employment in industrialized nations", Annual Review of Sociology 26(1): 271–296. doi:10.1146/ annurev.soc.26.1.271.

Budig, Michelle J. (2015). The Fatherhood Bonus & Motherhood Penalty. Parenthood and the Gender Gap in Pay, *Third Way Next Research Report Series*

Cavenaghi Suzana and Jorge Rodriguez (2014), *Adolescent and youth fertility and social inequality in Latin America and the Caribbean: what role has education played?*, GENUS, LXX:1, 1-25

Estrategia Nacional para la Prevención del Embarazo en Adolescentes (ENAPEA) (2015). Gobierno de la República, México

Gauthier, Jacques-Antoine, Eric D. Widmer Philipp Bucher and Cédric Notredame (2009) "How Much Does It Cost? Optimization of Costs in Sequence Analysis of Social Science Data", *Sociological Methods Research* 38(1), 197-231.

Gauthier, Jacques-Antoine, Eric D. Widmer Philipp Bucher and Cédric Notredame (2010). "Multichannel Sequence Analysis Applied To Social Science Data", *Sociological Methodology* (1), 1-38.

Giudici, Francesco and Jacques Antoine Gauthier (2013) "Occupational Trajectories after Childbirth" in René Levy and Eric D. Widmer (eds.), *Gendered Life Course Between Standardization and Individualization. A European Approach Applied to Switzerland*, LIT Verlag Berlin, pp. 93-113.

Hughes, Marion R. (1999). Early Fatherhood: Social Determinants and Life Course Consequences. PhD Dissertation, Department of Sociology, University of North Carolina at Chapel Hill. Copyright 2000 by Bell & Howell Information and Learning Company.

Jalovaara, Marika y Anette Eva Fasang (2015). Are there gender differences in family trajectories by education in Finland? *Demographic Research*: 33(44), 1240-1255.

Kaufman, Leonard y Peter J. Rousseeuw (2005). *Finding Groups in Data: An Introduction to Cluster Analysis*, John Wiley & Sons, Inc., Hoboken, New Jersey

Lesnard, Laurent (2014). Using Optimal Matching Analysis in Sociology:Cost Setting and Sociology of Time *Notes &Documents* n° 2014- 01. Paris, OSC, Sciences Po/CNRS

LeRoux, Brigitte and Henry Rouanet (2010). *Multiple Correspondence Analysis*. Series:Quantitative Applications in the Social Sciences, Sage Publications Inc.

Mier y Terán, Marta (1996) "The Implications of Mexico's Fertility Decline for Women's Participation in the Labour Force" in José Miguel Guzmán et al. (eds.) The Fertility Transition in Latin America, Clarendon Press Oxford, pp. 323-342.

Mier y Terán, Marta, Karina Videgain, Nina Castro and Mario Martínez (in press) "Familia y trabajo: historias entrelazadas en el México urbano" in Marie Laure Coubes, Patricio Solís and María Eugenia Zavala (coord.) *Análisis de datos de la encuesta EDER 2011*. El Colegio de la Frontera Norte and El Colegio de México.

Saraví, Gonzalo (2009). *Transiciones vulnerable. Juventud, desigualdad y exclusión en México.* Publicaciones de la Casa Chata. México. Centro de Investigaciones y Estudios Superiores en Antropología Social

Stange, Kevin (2011). Longitudinal Analysis of the Relationship between Fertility Timing and Schooling, *Demography* 48, 931-956.

Studer, Matthias, Gilbert Ritschard, Alexis Gabadinho and Nicolas S. Müller (2011). "Discrepancy Analysis of State Sequences". Preprint of the article published in *Sociological Methods and Research*. August 2011. 40(3), 471-510.

Urdinola, B.Piedad and Carlos Ospino (2015). "Long term consequences of adolescent fertility: The Colombian case" Demographic Research 12(55): 14871518.

Videgain, Ana Karina (2006). "Análisis de los cambios en la transición a la adultez en mujeres de distintas cohortes en contexto de cambios sociales en el Uruguay contemporáneo." Master in Demography Thesis. México D.F.: El Colegio de México.

Worts, Diana et al. (2013). "Individualization, Opportunity and Jeopardy in American Women's Work and Family Lives: A Multi-state Sequence Analysis", Advances in Life Course Research http://dx.doi.org/10.1016/j.alcr.2013.09.003

Differences in Health between East and West Germans

The "Long Arm of Childhood" under Divergent Political Regimes in Germany

Katharina Loter and Oliver Arránz Becker

Abstract The aim of our study is to investigate the "long arm of childhood" under two divergent political regimes in Germany. Children of the former socialist German Democratic Republic (GDR) grew up in a regime with full-time working mothers and around the clock child care services – in a regime that differed significantly from the German Federal Republic (FRG, also West Germany). GDR, year 1980: Almost 60% children aged 0-3 attend nurseries and more than 90% children aged 3-6 attend all-day kindergartens. In contrast, the respective percentages in the FRG are 1% for nurseries and 65% for predominantly part-time kindergartens. Thus, a great majority of children born and raised in the GDR experienced "equal" educational and nutritional conditions during early childhood regarded as a critical period of development, irrespective of their families' socio-economic situation. Within few years after the German unification health care in East Germany came up to the level of West Germany, nonetheless, for the "former children of the GDR" early childhood influences may continue to affect their adult health in a specific way. Our research question is: Does the childhood experience under a socialist regime play a role in explaining health at subsequent stages of the life course? First, we hypothesize that spending childhood in the GDR, unlike in the FRG, might have an adverse long-term effect on health. Second, we assume that "equal" GDR childhood conditions might attenuate the long-term impact of parental socio-economic status on adult health. To examine these hypotheses we use data from the German Socio-Economic Panel (SOEP) for birth cohorts 1950-1980 and apply latent growth curve analyses. Our preliminary results provide evidence of health disparities according to the kind of socialization. Further, we observe different patterns in the social health gradient for East and West Germans.

Katharina Loter · Oliver Arránz Becker
Institute of Sociology, Martin Luther University Halle-Wittenberg, Germany
e-mail: katharina.loter@gmail.com

The Impact of Marital Separation on Childbearing: A Test for the Selection Hypothesis

Sergi Vidal and Yara Jarallah

Abstract The extent to which childbearing occurs within marital unions has decreased dramatically over recent decades. While a wealth of studies examined the recent patterns of childbearing out-of-wedlock and premarital childbearing, less systematic has been research on deciphering childbearing patterns after marital separation. Evidence from related research on step-families and multi-partner fertility suggests that people may also have children after first marital unions to fulfil parenthood desires and to cement new relationships. In this paper, we investigate a further mechanism on the selectivity of childbearing after marital separation. We argue that childbearing can be a function of individual-specific unobserved factors that lead people to dissolve unions, re-partner, and build or grow their families. To assess this mechanism, we use hazard regression models for first and second order childbearing episodes in the framework of multilevel multiprocess event history analysis. The sample is restricted to women aged 16 to 40 from the panel study Household, Income and Labour Dynamics in Australia (HILDA) Survey. Results of this study will contribute to the understanding of contemporary fertility patterns, by shedding light on the conditions of fertility variations across partnerships over the life course.

Sergi Vidal · Yara Jarallah
The University of Queensland
e-mail: s.vidal@uq.edu.au

Session 8B: Relational sequence network and combined sequence-survival analysis

Relational sequence networks as a tool for studying gendered mobility patterns

Klaus Hamberger (Ecole de Hautes Etudes en Sciences Sociales, Paris, klaus.hamberger@ehess.fr)

Abstract

This paper uses relational sequence networks to study the gendered differences of migration biographies. Starting from an integrated model of kinship and migration relations as parts of a single bi-modal network of individuals and events, sequence networks are constructed by classifying mobility events according to the social (kinship or other) relation between the individuals they link together as migrants and hosts. Itineraries thus are conceived of as walks in a space of relational positions.

Using data from 508 migration biographies collected in rural South-east Togo between 2010-2015, we show that male and female trajectories do not so much differ in their degree of mobility as in the topology of the social spaces they traverse and in the structure of the social sequences they trace. While both rest on a basic kinship axis (linking an "internal" parent pole and an "external" extended-kinship pole), male networks tend to evolve through a succession of multiple but structurally isolated non-kinship links, whereas female networks develop into complex and integrated multifocal networks sewn together by marital and affinal ties. Since marital ties are precisely the ties that link male and female networks together, many of the differences between these networks can thus be traced back to their mutual relation (in particular to the fact that women move to and from men but not vice versa). Rather than just confirming the macro-tendencies for male and female mobility patterns stated in the demographic literature, sequence network analysis yields insight into the relational logics that bring these tendencies about.

Data have been analyzed with the open source software Puck 2.2., which implements the model presented in the paper.

Keywords: Social networks, Sequence networks, Gender, Migration, Kinship, West Africa

Extended abstract

Method

The recent applications of network theory to social sequence analysis (Bison 2014; Fitzhugh, Butts, and Pixley 2015; Cornwell et Watkins 2015; Cornwell 2015) have focused on the connections created between individuals through the sharing of similar events, similar sequence motifs or similar sequence network structures. However, events also directly link individuals in a face-to-face relation. The familiar two-mode network linking individuals and events is not just representing an affiliation network, but actually the incidence network where event nodes actually correspond to hyper-edges linking several individuals in complex, polyadic relations. Examples for this type of events are mobility events, where migrants are linked to hosts, co-migrants or financers, but also kinship events, where children are linked to mothers and fathers. In fact, we can conceive of kinship and mobility as two closely related and structurally similar subnetworks within an integrated social space-time made up of individuals and events (Hamberger and Sohler 2014):
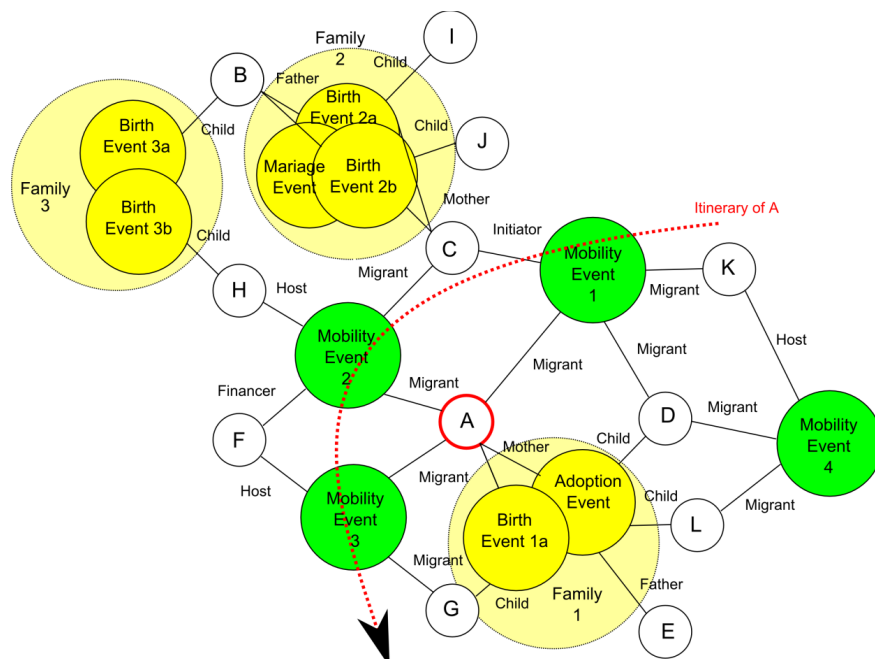

Figure 1: An integrated kinship-mobility network

From this integrated two-mode network are derived the one-mode networks usually studied in migration and life course studies: networks of events linked to each other by the involved individuals, and networks of individuals linked to each other by shared events. For a given individual (ego), we can thus derive both the event sequence that constitutes his or her itinerary, and the personal network that has shaped (or has been shaped by) this itinerary.

Individual itineraries are linear paths unless we merge events into higher-order event classes. This could be done by reference to some attribute (such as the reason for migration, the destination of the move, and so on). Adopting a thoroughly relational approach, we shall, however, derive all attributes characterizing individuals or events from the network itself: thus, individuals that make up ego's migration network will be characterized by the relational (kinship or other) chains that link them to

ego, and the events that constitute his or her itinerary will accordingly be characterized by the chains that link ego to the individuals that play a given role (such as host) in this event. More precisely, we shall characterize both individuals and events according to the type of *relational circuit* (Hamberger 2011) that emerge as the mobility link combines with chains of other links (kinship, friendship, employment, and so) – for example, individuals C and H in figure 1 are linked to each other both as migrant and host (via mobility event 2), but also as brother's wife and husband's sibling (via birth events 3a and 3b and the marriage event 2). As a consequence, mobility event 2 can be characterized as being of the "brother-in-law as host" type.

We will thus partition the ego-network according to the relational circuit types linking ego and alter, and then use the clusters of this partition to classify the events of ego's itinerary. Merging the event nodes of the same cluster then yields a directed network of social relation types linked by temporal adjacency: a relational sequence network.

The morphological similarity of these relational sequence networks (measured by the number of shared arcs rather than that of shared nodes) can then be used to construct a continuous typology of itineraries (represented e.g. by phylogenetic trees), but also to evaluate the extent to which the *similarity* of two given itineraries corresponds to their actual *intersection* in the total event network. Indeed, since one and the same event represents different relational types for the various individuals involved, people who have largely *shared* itineraries not necessarily have *similar* itineraries in relational terms, unless their positons in social space are structurally similar.

By considering the way degree to which ego's itinerary resembles and/or coincides with the itineraries of each of the alters of his or her personal network, we can finally study the way both the similarity and the contiguity between itineraries is related to certain relational positions. Thus, parents and children may have itineraries that are largely coincident but different, friend's itineraries may be similar but show little coincidence, and sibling's itineraries tend to be both similar and coincident, at least at the beginning. Relational sequence network analysis thus constitutes a tool of studying both the similarity and the contiguity relations that events establish between individuals, inasmuch as these relations can be conceived of as relations between sequence networks. In this capacity to analyze the perspectival structure of social space as a network of networks lies the (still largely unexplored) potential of relational sequence network analysis.

<u>Software tools</u>

All analyses used in this paper have been effectuated with the open source software Puck (Hamberger et al. 2014), which can be downloaded at www.kintip.net (source code at http://sourceforge.net/projects/tip-puck/). Initially developed for the study of kinship networks, Puck contains, from its 2.2 version onwards, a package for the study of spatiotemporal networks, including census data and migration biographies.

Visualization has been done with Pajek (Nooy, Mrvar, et Batagelj 2011) from files produced by Puck, and a variant of the software Geneaquilt (Bezerianos et al. 2010) implemented in Puck.

<u>Data</u>

In this paper we analyze the gendered patterns of West-African migration biographies, based on data collected in a field survey from 2010 to 2015 in rural South-east Togo. The zone of inquiry is located in the region of West Africa with the highest intra-regional migrations. The dataset includes extended genealogies and detailed migration biographies of 508 individuals (adults and children) that form part

of the personal networks of 60 inhabitants initially drawn at random from the 2005 population of the rural town Afagnan-Gbléta (about 4800 inhabitants in 2015).

Migration biographies were collected through retrospective semi-directive interviews. We recorded all migratory events leading to a change of residence for more than three months from birth to the present (time of the interview, last update of information in 2015). For each migratory event, the interviewees were asked to describe the context, motive and course of the event. In particular, we collected information (name, relation and contact) on the persons who received ego (*hosts*), accompanied ego (*co-migrants*), initiated the displacement (*initiators*) and financed the journey/means of transport (*financers*). All other persons mentioned in the context of the migration event were equally noted (*others*). This set of names was complemented by a complete list of the person's *parents*, *spouses* and *children*. This dataset is completed by a large genealogical and residential dataset (about 50.000 individuals including the deceased) collected during three subsequent censuses of the village of Afagnan-Gbléta (2005, 2010 and 2015). Most of the remote kinship relations (such as "father's mother's brother's daughter") and mediated non-kinship relations (such as "friend's employer") were thus not (only) directly reported by interviewees but computed from data stemming from numerous different oral sources. The anonymized dataset will be available on the platform www.kinsources.net, where two previous versions (2008 and 2011) can be already accessed.

Since the snowball method leads to the systematic overrepresentation of mobile and connected persons in the total network, most of our analyses are restricted to the initial random sample of 60 itineraries. We thus do use the remaining 448 itineraries to construct the ego-centered social spaces and sequence networks of these 60 independent cases. For the purpose of illustration, four of these cases are considered in more ethnographic detail.

Results

In order to construct the relational circuits that result by linking the participants of a mobility event by a chain of other (non-mobility) links, we have considered seven main types of relations: *genealogical kinship* (chains that can be retraced in the genealogical network), *non-genealogical kinship* (chains were some or all of the kinship links are not retrievable in the genealogy), *friendship*, *apprenticeship*, *initiation*, *employment* (employer-employee) and *rent* (landlord-lodger). Depending on the analysis, kinship relations have be further differentiated into various types of direct (parent, child, sibling, spouse) and remote (agnatic, uterine, cognatic, affinal) relations.

After a general analysis of the frequencies of each of these relation types in male and female itineraries, we consider their structural importance in the personal networks and the relational sequence networks.

Their structural role in ego's personal network is assessed indirectly by considering the relation type of the most central alters, as well as the relation types that dominate in each of the network's connected components. Figure 2 shows four example networks, where relation types are indicated by different colors, central alters have been marked by bold borders, connected components have been encircled.
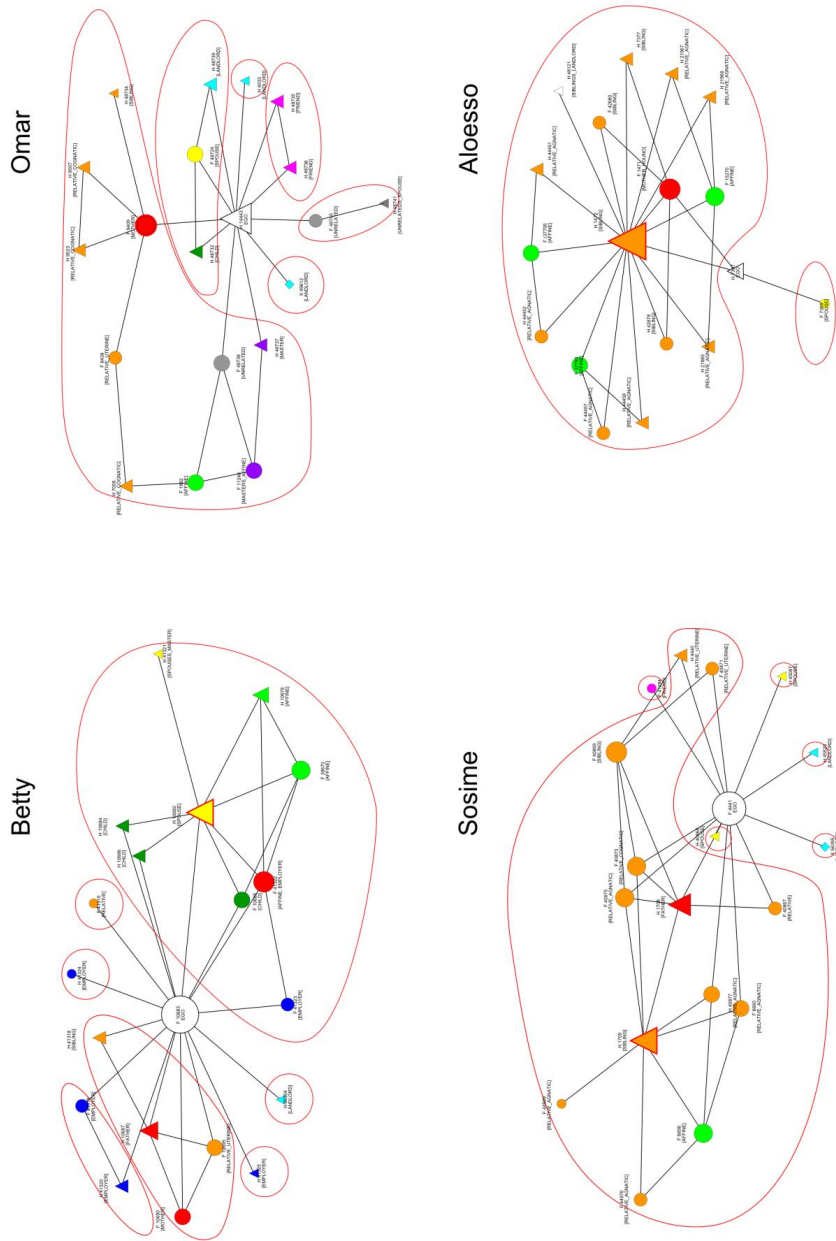
Figure 2: Four case examples – social networks with direct social relations (central alters are highlighted by bold borders, components are encircled). Color code (also valid for figures 5, 6, 7 and 9): red = parents, orange = relative, yellow = spouse, light green = affine, green = child, dark green = own property, grey = unrelated, light blue = rent, blue = public accommodation, marine blue = state, dark blue = employer, violet = master, rose = friend, light red = houno, black = death, white = unknown

A comparison of the 60 personal networks already reveals that relations of different types operate in a different way in organizing ego's social space. While parental ties function as centers and nuclei of large connected components, relations mediated by money tend to form numerous small peripheral components. Between these two extremes, whose co-presence is largely characteristic for the social space of men, we find the intermediary case of marriage relations, which, due to the high matrimonial mobility in this region, give rise to multiple cohesive components of moderate size and centrality which are typical for the social space of women. As a consequence, we find that on average, female networks are more cohesive than male networks, while at the same time, ego's own betweenness centrality is generally lower for women then for men.

After having examined in a synchronic manner the social environments shaped by (and shaping) people's itineraries, we turn to the relational sequence networks that are constructed by successively linking the relational types that characterize ego's relation to his or her host in the adjacent mobility events of his or her itinerary. Figure 3 shows these networks for the four example individuals. Each relational type is now represented by a node (colored according to the same code as the individual nodes in figure 2). We have not numbered the arcs, but each of the four itineraries starts at the position "south pole" of the network, since the "host" of the first "mobility event" (birth) is always the mother.

To visualize the individual and temporal structure of the sequence underlying the network, we have added to the right of each sequence graph the bimodal network of the corresponding individuals and events (ordered by age and time order) in the quilt format developed by Bezerianos et al. (2010). Individuals are listed on the vertical axis, events on the horizontal axis, and dots indicate the involvement of individuals in events. Ego's itinerary can easily be distinguished as a continuous horizontal line of dots.

Figure 4 represents the "aggregate" network that results from projecting all sequence networks of men and women respectively in a single network, where line values represent the numbers of individual itineraries in which a given event sequence occurs.
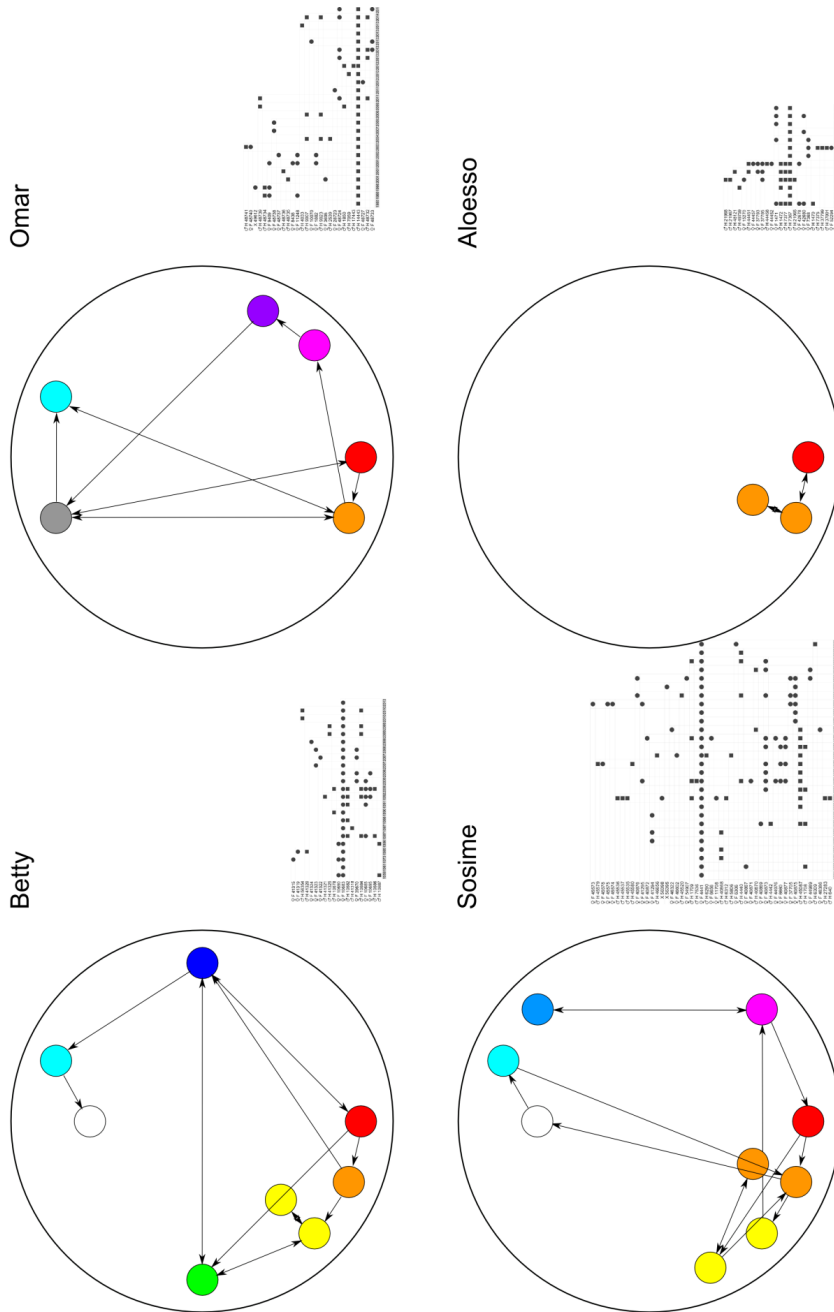
Figure 3: Four case examples - social sequence networks and individual-event networks (in quilt format)
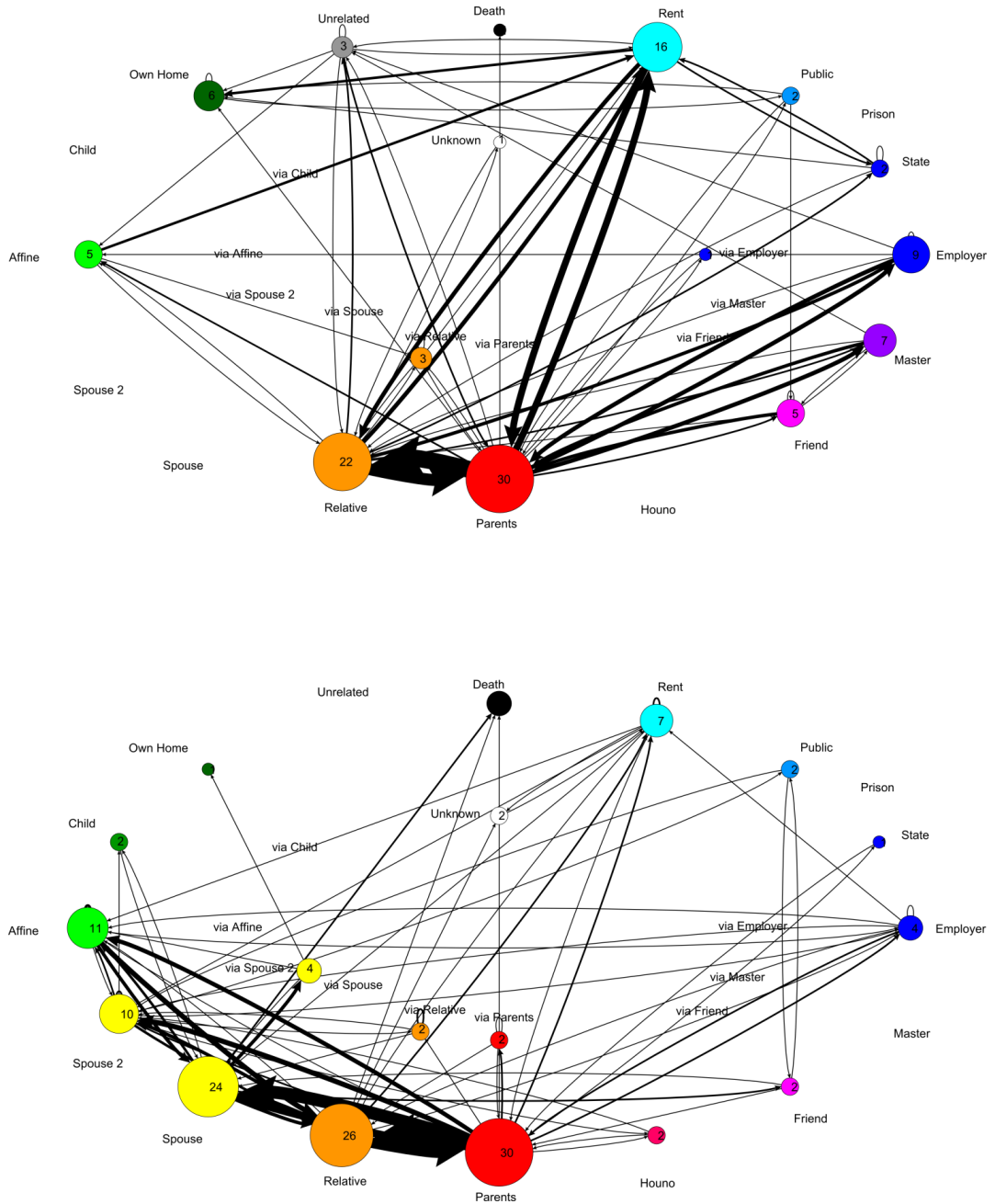
Figure 4: "Aggregate" sequence networks of men and women (initial random sample, n = 60)

As can be seen from fig. 4 as well as from a comparison of node degrees and arc frequencies in male and female aggregate networks, there are marked differences (but also some important similarities) between gendered mobility patterns in terms of social relations. Both male and female networks are centered around a fundamental axis linking parents and non-parental relatives. For both genders, relatives outside the immediate family circle are at least as important as parents in the role of hosts – with a slight preference of male and female for agnatic and uterine relatives, respectively. For women, they are even more important.

However, while men enter two times as frequently into relations with landlords than women, they are five times as frequently to pass from a stay with relatives to a rented apartment. The most significant difference is, of course, the importance of marital and affinal homes for female itineraries, which have no equivalent in their male counterparts. The virilocal orientation of residence is unambiguously brought out (almost all adult women have at least once moved to their husbands or affines, while the reverse is a small minority). While the male network is largely organized around a triangle formed by the parental home, relatives and the residential market, the female network contains in addition a spouse node, which may become the nucleus of an entire marital subnetwork.

The problem with this kind of comparative macro-analysis is, of course, that it already presupposes that gender is a pertinent classification criterion for merging individual sequence networks, instead of deriving the classification criteria from a comparative analysis on the micro-level of individual sequence networks.

This latter perspective characterizes the approach of optimal matching analysis (Abbott 1995) and its more recent network-analytic developments, where sequence matching is replaced by graph similarity (Butts et Pixley 2004; Butts 2008; Fitzhugh, Butts, et Pixley 2015). Since in our relational sequence networks each node type appears, by construction, exactly once in each network (the only quasi-exception being first and subsequent spouses, which we have treated as distinct types), edit distance between networks boils down to the number of differently connected pairs of nodes (that is, the number of different cells in the adjacency matrices of the two graphs, all of which have equivalent node sets). Based on this graph distance matrix, we use the neighbor-joining algorithm (Saitou et Nei 1987) to plot the similarities and differences of the itineraries in form of a phylogenetic tree. Figure 5 shows the tree of the 60 individuals of the initial random sample, where the corresponding individual sequence networks have been plotted into the nodes, and border colors indicate gender (blue for male and red for female).
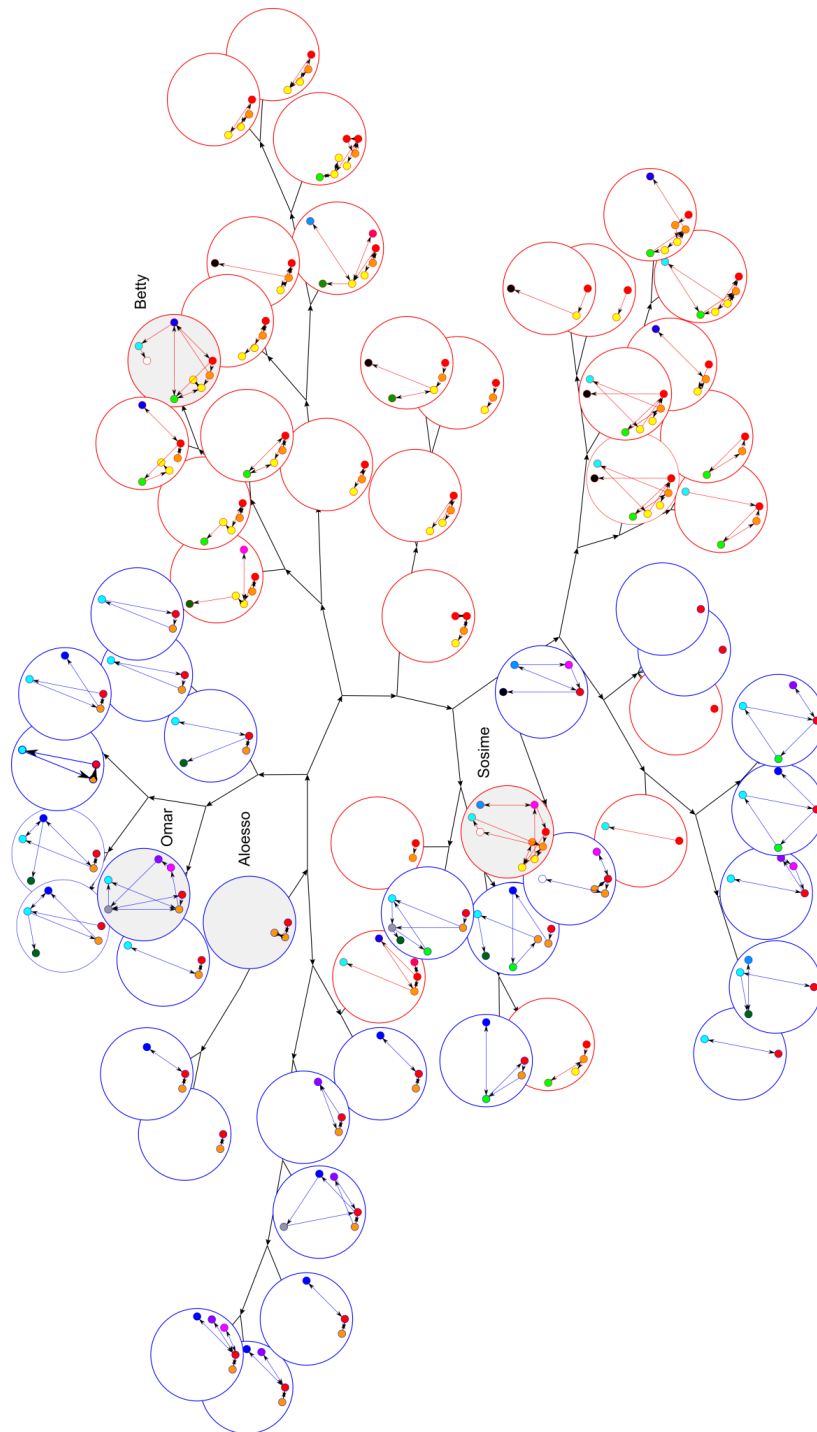
Figure 5: Phylogenetic tree of 60 sequence networks (random seed sample)

While the gender difference is neatly brought out by the graph (roughly speaking, female itineraries constitute the right, male itineraries the left side, with a "mixed" zone of complex itineraries in the middle), a closer look shows that, beyond the great divide introduced by virilocality (the presence of absence of "yellow" marital nodes in the sequence network), both male and female itineraries are differentiated into several branches of the tree.

Apart from the three single-node cases in the middle bottom, the most simple sequence networks, consisting mainly of the parent-relative axis, are located in the upper left, "male" part of the tree. This branch may develop into two more complex forms, as the basic axis combines either with rent (as for the urban migrant workers) or with hosting by the master or employer (as for the rural itinerary masons or woodcutters). A still more complex variant combines rent with lodging by the state, which is the characteristic feature of teachers' itineraries. A quite different version of typically male itineraries is given at the bottom, where the parental node is directly linked to the rent node without passing through a relative's home. This minority pattern is reserved to men who have migrated for professional reasons later in life, and to the few boys whose parents could afford paying them rented apartments for school.

On the "female" side, we find a large group of highly similar sequences completing the parent-relative-axis by one or more marital or affinal nodes, thus giving rise to several distinct groups according to the complexity of the marital subnetwork. This pattern may further evolve by integrating stays with landlords or with employers (mainly for domestic work). By contrast, we almost never find a female pattern without the "relative" node it. Women who leave their parental home directly for a rented apartment are even more rare then men.

The complex patterns in the "mixed" middle of the figure combine "male" and "female" features, such as young men staying with affines (e.g. sisters' husbands) or women leaving for transnational migration with friends.

While the tree of fig. 5 has been constructed from 60 independently drawn examples, the (dis)similarities between individuals' mobility patterns are generally not independent of their mutual interconnections. Indeed, both the similarities and the interconnections between itineraries may be related to their respective roles in the mobility events linking them, as well as to the (kinship or other) relation existing between them, in other words: to the type of relational circuit(s) which ties them together. The logic of these interdependencies is still largely unexplored. We can, however, get an idea of them by linking the individuals that make up a given ego's personal network according to the similarity and intersection of their respective relational sequence networks.

Figures 6 and 7 do so for the four example networks. In figure 6, arc thickness indicates to the degree of intersection of itineraries (measured as the percentage of events which one individual shares with the other). In figure 7, edge thickness indicates the degree of similarity between itineraries (measured as the number equivalently linked or non-linked node-pairs in the respective event sequence networks, normalized by the maximal number of pairs that can occur in either network).
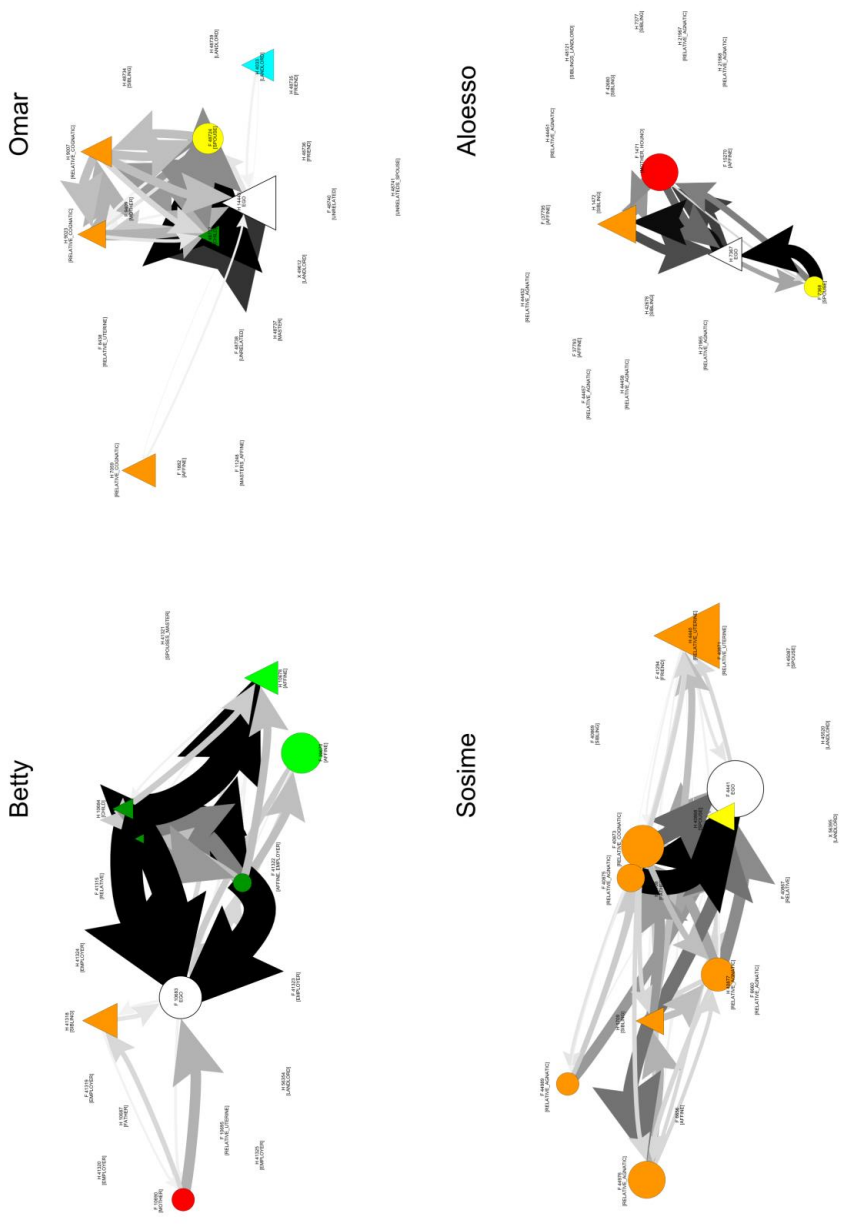
Figure 6: Four case examples – social networks with arc thickness indicating itinerary intersection rates
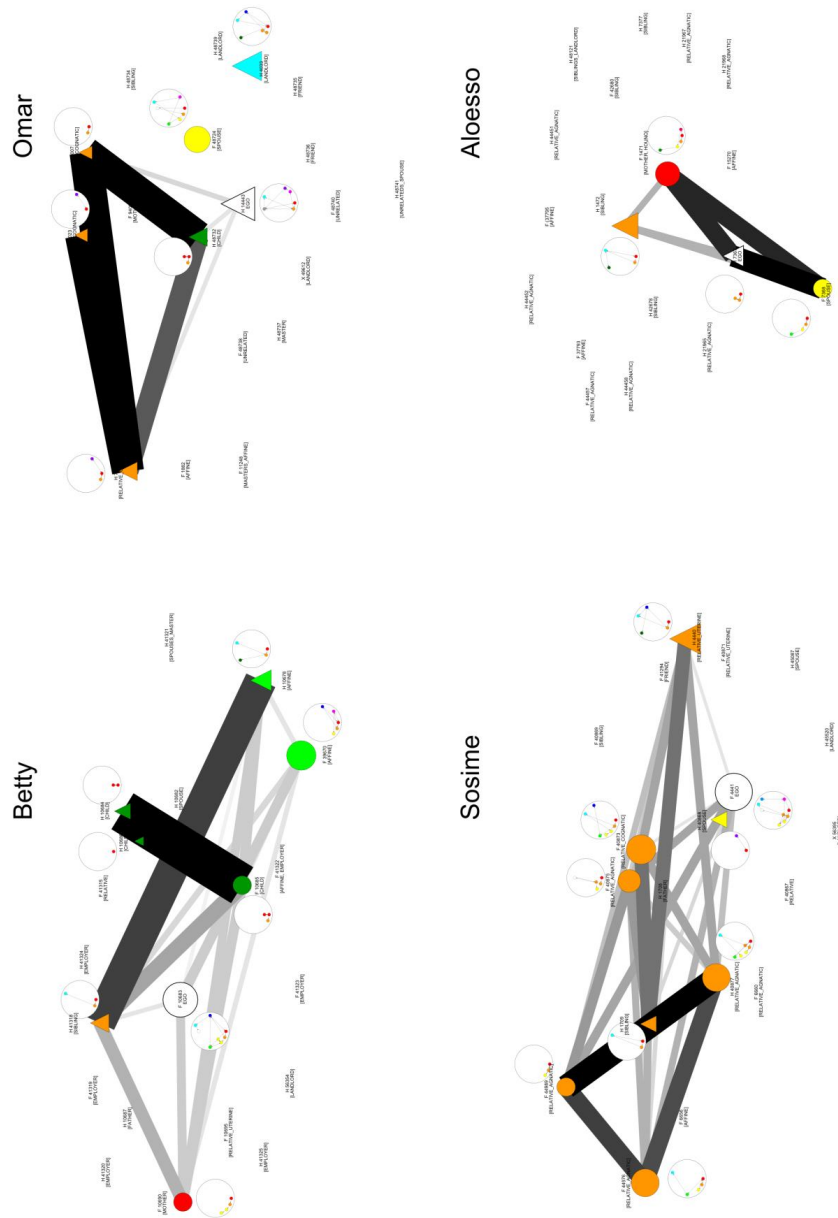
Figure 7: Four case examples – social networks with line thickness indicating itinerary similarity degrees

As one can see, similarity and contiguity structures are as a rule quite different from each other. Thus, the two important male persons in Betty's network – her brother and her father-in-law – have similar itineraries, though they never intersect. To the contrary, the itineraries of Omar and his wife, though overlapping, show no similarity with each other. For itineraries to be both overlapping and similar, it is required that both egos either hold a structurally equivalent position in social space – such as siblings – or have very reduced itineraries – such as children.

The fact that both the intersections and the similarities between two individuals' relational sequence networks are rooted in their mutual relation invites us to reconsider the results of our analysis with a view to a relational conception of the gender difference. Our results show that the topology of both male and female sequence networks rests on a basic kinship axis (linking an "internal" parent pole and an "external" extended-kinship pole). However, while male networks tend to evolve through a succession of multiple but structurally isolated non-kinship links, female networks develop into complex and integrated multifocal networks sewn together by marital and affinal ties. Now, the marital tie is precisely the tie that puts together male and female itineraries. In other words, the central source of difference between the sequence networks of men and women is precisely the relation between them.

Rather than just confirming the macro-tendencies for male and female mobility patterns (as stated in the demographic literature) at the micro-level of individual trajectories, sequence network analysis yields insight into the relational logics that bring these tendencies about. It serves not only to study the differences between gendered social networks, but also to understand the gender relation itself as a relation between networks, that is, not just as an attribute of individuals, but as a structural trait of social space-time. In a more general perspective, the beginning integration of network and sequence analysis may be the first step towards a full-fledged social topology.

References

Abbott, Andrew. 1995. « Sequence Analysis: New Methods for Old Ideas ». *Annual Review of*

Sociology 21 (1): 93‑113.

Bezerianos, Anastasia, Pierre Dragicevic, Jean-Daniel Fekete, Juhee Bae, et Ben Watson. 2010. « GeneaQuilts: A System for Exploring Large Genealogies ». *IEEE Transactions on Visualization and Computer Graphics, Institute of Electrical and Electronics Engineers (IEEE),* 16 (6): 1073‑81.

Bison, Ivano. 2014. « Sequence as network: an attempt to apply network analysis to sequence analysis ». In *Advances in sequence analysis : theory, method, applications*, édité par Philippe Blanchard, Felix Bühlmann, et Jacques-Antoine Gauthier, 231‑48. Cham: Springer.

Butts, Carter T. 2008. « A Relational Event Framework for Social Action ». *Sociological Methodology* 38 (1): 155‑200.

Butts, Carter T., and Joy E. Pixley. 2004. « A Structural Approach to the Representation of Life History Data ». *The Journal of Mathematical Sociology* 28 (2): 81‑124.

Cornwell, Benjamin. 2015. *Social Sequence Analysis: Methods and Applications*. Cambridge: Cambridge University Press.

Cornwell, Benjamin, and Kate Watkins. 2015. « Sequence-Network Analysis: A New Framework for Studying Action in Groups ». In *Advances in Group Processes*, 32:31‑63. Advances in Group Processes 32. Emerald Group Publishing Limited.

Fitzhugh, Sean M., Carter T. Butts, and Joy E. Pixley. 2015. « A life history graph approach to the analysis and comparison of life histories ». *Advances in Life Course Research* 25 (septembre): 16‑34.

Hamberger, Klaus. 2011. « Matrimonial circuits in kinship networks: Calculation, enumeration and census ». *Social Networks* 33 (2): 113‑28.

Hamberger, Klaus, Cyril Grange, Michael Houseman, et Christian Momon. 2014. « Scanning for patterns of relationship: analyzing kinship and marriage networks with Puck 2.0 ». *The History of the Family* 19 (4): 564‑96.

Hamberger, Klaus, and Karin Sohler. 2014. « Integrated kinship-mobility networks. A West African case study ». Paper presented at the 1st European Conference on Social Networks, Barcelona (UAB).

Nooy, Wouter de, Andrej Mrvar, and Vladimir Batagelj. 2011. *Exploratory social network analysis with Pajek*. Royaume-Uni, Royaume-Uni de Grande-Bretagne et d'Irlande du Nord: Cambridge University Press.

Saitou, N., and M. Nei. 1987. « The Neighbor-Joining Method: A New Method for Reconstructing Phylogenetic Trees ». *Molecular Biology and Evolution* 4 (4): 406‑25.

Childhood co-residence structures and home-leaving

A combination of survival and sequence analyses

Florence Rossignon (1), Matthias Studer (2), Jacques-Antoine Gauthier (1) and Jean-Marie Le Goff (1)

(1)   NCCR LIVES, University of Lausanne

(2)   NCCR LIVES, University of Geneva

Abstract   The aim of this study is to examine whether the co-residence structures in which young adults grew up is likely to affect their propensity of leaving the parental home. The empirical research was based on the LIVES Cohort study, a panel survey that started in autumn 2013 in Switzerland. Two longitudinal statistical methods were used as complementary approaches. First, sequence and cluster analyses were conducted to identify typical trajectories of childhood co-residence structure. Second, event history analysis was used to estimate whether these aforementioned structures influence home-leaving. Analyses show that it is not only the occurrence of an event that increases the risk of experiencing another event, but also the order in which various states occurred. What is more, it seems that two features have a significant influence on the departure from the parental home, which are the co-residence structures and the arrival or departure of siblings from the parental home.

1 Introduction

According to Marini (1984), the transition from youth to adulthood may be marked by several interlinked events which induce a movement from economic dependence to economic independence. Likewise, it has also been stated that it might be marked by the departure from the family of origin to found one's own family. Five major transitions are generally enumerated as the main markers of adulthood: leaving school, entering the labour force, leaving the parental home, marrying and becoming a parent (Modell, Furstenberg, & Hershberg, 1976). Thereby, the transition to adulthood can be seen as "*a series of ordered stages through which an individual passes in his or her life and which are associated from one stage to the next with age*" (Hogan & Astone, 1986, p. 110). However, the aim of this article is not to study the transition to adulthood as a set of experiences and statuses changes, but to focus on one of them: leaving the parental home.

The recent and rapid increase in divorces and remarriages has led to a growing complexity of the household's composition (Goldscheider & Goldscheider, 1998). Indeed, new co-residence structures have emerged in the past few decades, such as stepfamilies and single-parent households. As a result, there are a growing number of children who do no longer grow up in a home with their two biological parents. According to the aforementioned researchers, these increasingly common co-residence patterns are likely to affect the ways in which the parents invest in their children and, thus, the parents-children relations. Consequently, the aim of this paper is to examine whether the previously mentioned changing co-residence structures, that are likely to affect the roles and statuses of numerous family members, may influence the decisions children make when they become adults, such as leaving the family of origin. Leaving the parental home can be considered as a prerequisite to achieve other life transitions, such as getting married and becoming a parent (Mulder, 2009). In this way, this analysis aims to give an illustration of the importance of the co-residence structure prevailing during childhood as a sig-

3

nificant determinant of the transition toward stable and successful work and family trajectories (Goldscheider & Goldscheider, 1998). Until now a significant number of studies have examined the impact of having ever experienced a parental disruption during childhood or of having ever lived with a stepparent on the probability of leaving home (Holdsworth, 2000; Bernhardt, Gähler, & Goldscheider, 2005). Other studies also focused on the co-residence structure into which a young adult lived in a specific moment of his/her life, often at the time of the youth's final home-leaving (Mitchell et al., 1989; Chiuri & Del Boca, 2010). In this way, even though many sociological theories assert that age at leaving home might be linked with the whole individual trajectory preceding that moment, only few studies have developed methodological framework able to examine this perspective. At least two reasons can be cited. First, only few studies collected detailed life history records of the co-residence structures during childhood, making it difficult to tackle this issue (Aquilino, 1991; Goldscheider & Goldscheider, 1998; Blaauboer & Mulder, 2010). Second, we lack a proper method to estimate the influence of the previous co-residence trajectories on age at leaving the parental home. Hence, the aim of this paper is twofold. On a sociological level, we aim to provide a better understanding on how co-residence trajectories influence the probability of leaving home. On a methodological level, we propose a new framework to estimate this influence.

What is more, it has been reported that the transition to adulthood has become late (Billari & Liefbroer, 2010). According to Galland (1996), the norm regarding the age at the entry into adulthood has been altered. Indeed, a norm of precocity has given way to a norm of delay. Nothing anymore induces young people to hasten their departure from the parental home. As a consequence, we could wonder who are the young adults who decide or who are forced to become independent and to assume the responsibilities that go along with adulthood. In other words, another aim of this paper is to identify the factors that lead to an early departure from the parental home and to examine whether lone-parenthood or stepparenthood contribute to an early emancipation.

The structure of this paper is as follows. In section 1, we start by a detailed review of the literature on the links between the childhood co-residence structures on the one hand and the departure from the parental home on the other hand. Section 2 describes our dataset and the measures we use. We then turn in section 3 to the presentation of our methodological framework. In section 4, we apply our methodological framework to the study of the effect of the past co-residence trajectories on the risk of leaving home. Finally, section 5 summarizes and discusses our findings.

2 Conceptual framework

2.1 Why study the departure from the parental home?

In the past few decades, a growing number of researchers have devoted particular attention to the departure from the parental home. Some of the reasons for this growing devotion are that the departure from the parental home is likely to have significant consequences for important areas of policy, such as the demand for housing (Ermisch & Di Salvo, 1997) and the risk of poverty among young people (Iacovou & Aassve, 2007). Furthermore, leaving home is one of the main and, very often, one of the first components of the transition to adulthood (Schizzerotto & Lucchini, 2004). As a consequence, it has been stated that "*both the destination and the timing of young people's home-leaving are likely to be crucial in determining later life opportunities*" (Buck & Scott, 1993, p. 863). Indeed, there is a common belief according to which age norms define the appropriate timing at which major life events should occur (Billari & Liefbroer, 2007). They also provide guidance and regulations throughout the life course of individuals (Heckhausen, 1999). Nevertheless, Aassve, Arpino and Billari (2013) have demonstrated that differences in age norms exist both between and within countries. Divergences in terms of earnings, employment rates, education system,

5

state welfare system and social norms can be cited to explain this heterogeneity. As a result, in each country, there is a distinct definition of when it is too early or too late to leave home, even though some variations can also be observed within each society. In this way, there is some evidence that the home-leaving patterns which do not respect the age norms are likely to have negative impact on the professional and co-residence trajectories of young adults that will, in turn, threaten their subsequent success and stability (Goldscheider & Goldscheider, 1998). As an illustration, it has been shown that leaving home too early is likely to reduce education aspirations and attainments (Goldscheider & Goldscheider, 1993). This could stem from the fact that young adults who leave home before the end of high school tend to forgo education for work (Mitchell, Wister, & Burch, 1989). Conversely, leaving the parental home at a later age might delay marriage and childbearing (Chiuri & Del Boca, 2010). Regarding women, a higher age at first birth may have a negative influence on the total number of children, but it might also affect birth weights and birth defects (*Ibid*.). Concerning men, a protracted transition to adulthood might also have negative consequences on the household's division of labour. This could be explained by the fact that the little experience of sharing household chores with a partner they have accumulated over the years might negatively impact their wives' labour supply, career and fertility; in particular in countries where the child care services are less widespread and/or more expensive (Brodmann, Esping-Andersen, & Güell, 2007). For instance, a recent study has demonstrated that husbands from Southern Europe participate less equally to housework tasks and that this excessive burden on women is strongly associated with lower fertility rates (Rosina, 2005).

2.2 Leaving the parental home in Switzerland

Even though leaving home is considered in many countries as one of the main life events that define the concept of adulthood (Billari & Liefbroer, 2007), it has been shown that its process may vary from one country to another. For this reason, it is

necessary to take into account the specific national context into which the present study has been conducted.

In Switzerland, leaving home tends to occur early and it often happens simultaneously with the first integration into the labour market (Thomsin, Le Goff, & Sauvain-Dugerdil, 2004). For example, Schumacher, Spoorenberg and Forney (2006) have indicated that, in Switzerland, the median age at home-leaving for the cohort born in 1976-1987 is equal to 23. This situation has to be seen in the light of the Swiss education system, which, as in Germany, is a largely apprenticeship-based system of education (Thomsin et al., 2004). In Switzerland, almost 70 per cent of every cohort of students who achieves a compulsory education enters a vocational education and training program (Meyer, 2003). This later is also known as the *"dual education system"*. While the apprentice spends most of his/her time working for an approved company, he/she attends a vocational school for 1-2 days per week. As the apprentice is simultaneously studying and working, he/she receives a salary (though it is modest). Also, the professional stabilisation of the young apprentices is quite quickly attained. Indeed, young adults enter such an education at age 15. As the vast majority of such vocational trainings lasts 3 or 4 years, many of them can fully enter the labour market from the age of 18, or even 15 if one considers the apprenticeship as integration into the labour market. Furthermore, it has been shown that unmarried cohabitation has progressively emerged as the most frequent form of living arrangements (Thomsin et al., 2004). Consequently, the Swiss model of leaving appears as a combination of two other European models as defined by Cavalli and Galland (1993). On the one hand, it shares similitudes with the Northern model, which is characterized by an extension of the extra-marital life. On the other hand, the Swiss model is close to the British system, which is marked by a precocious entry into the labour force and by the extension of the unmarried cohabitation without children.

Nonetheless, we have to keep in mind that leaving home does not necessarily lead to a neglect of family ties and to a lack of parental care (Zorlu & Mulder, 2011).

7

Indeed, geographic distances are rather small in Switzerland, even more so for the migrant population who tend to be concentrated in large urban centres. Consequently, living away from home, but at a small distance, enables young adults to escape from the daily parental surveillance, but, at the same time, it also allows them to benefit regularly from parental support.

2.3 Childhood coresidence structures and leaving home

There are some reasons to believe that the co-residence structures tend to expose young individuals to different options regarding family formation, because they provide different social and economic resources that can have an influence on the transition to adulthood (Sandefur, Eggerling-Boeck, & Park, 2008). As an illustration, a significant number of studies have demonstrated that the co-residence structures in which young people grew up has a significant influence on the propensity of young adults to leave home (Mitchell et al., 1989; Aquilino, 1991). Indeed, the decision to leave home cannot be understood as an individual choice, rather as the result from the characteristics of the co-residence structures in which the person grew up (Blaauboer & Mulder, 2010). Based on an analysis of the literature, several co-residence structures can be distinguished.

First and foremost, even though the number of divorces has experienced a strong increase in Switzerland over the past 40 years (Swiss Federal Statistical Office, 2015)[1], growing up with two biological parents is still the most common form of living arrangements. As reported by many social researchers, closer family bonds and the physical presence of both biological parents often induce a delayed departure from the parental home (Mitchell et al., 1989; Aquilino, 1991; Mitchell, 1994; Goldscheider & Goldscheider, 1998). As such, young adults who spent most of their childhood in such co-residence structures are expected to be among the last to leave home.

Secondly, the single-parent household can be considered as an alternative form of co-residence structures ensuing mainly from the increase in divorces. In this situation, the custodial parent (in many cases the mother) often has to increase his/her activity rate in order to compensate for the economic loss that generally results from divorce (Acock & Demo, 1994). As a consequence, the time he/she spends with his/her child/ren is reduced and this/these latter is/are likely to suffer from a lack of support and attention. This deteriorated co-residence environment can reduce the attractiveness of prolonging one's stay in the parental home. The parental disruption may also lead the individual to think of him/herself as an independent unit from the family. Therefore, it might hasten his/her transition to adulthood. What is more, it is commonly agreed that one of the major difficulties encountered by those families are financial. It is thus not surprising that young people who grow up in this environment are by far the most economically disadvantaged. As a result, a significant number of studies conducted in many countries consistently show that children of divorced parents leave home at a younger age than those from intact families (Goldscheider & Goldscheider, 1998; Cherlin, Kiernan, & Chase-Lansdale, 1995; Juang, Silbereisen, & Wiesner, 1999; Holdsworth, 2000; Bernhardt, Gähler, & Goldscheider, 2005). Nonetheless, as stated by Mitchell et al. (1989), this ascertainment is more linked to the family socio-economic status than to the absence of one of the parental figures. Indeed, the presence at home of young adults can be considered as a financial burden for the lone parent. Thus, their departure from the parental home might reduce this strain (Mitchell, 1994). Regarding young adults who have grown up in a single-parent household from birth, Aquilino (1991) has demonstrated that their likelihood of leaving home does not differ from that of those who have grown up in an intact household. Consequently, in addition to the type of co-residence structures, we could assume that the stability of the co-residence structures during childhood also has an impact on the timing of home-leaving.

---

[1] The divorce rate has, though, been slightly decreasing since 2005.

9

Thirdly, children who have been raised in a stepparent household are more likely to leave home sooner than their counterparts who have grown up with a lone-parent or with both of their parents (Mitchell et al., 1989; Aquilino, 1991; Kiernan, 1992; Goldscheider & Goldscheider, 1998). Having to welcome a new parental figure and often step-siblings and/or half-siblings into one's home may make young adults feel that leaving home would lead to an enhancement of their situation in comparison to remaining at home (Goldscheider & Goldscheider, 1998). As an illustration, they might not tolerate having to share the attention, love and material support that once were theirs with complete strangers. Accordingly, severe conflicts and disagreements within stepfamilies have been enumerated as playing a significant role in early nest-leaving (Gähler & Bernhardt, 2000; Gossens, 2001). Likewise, having stepchildren has been enumerated as one of the major sources of marital instability for remarried couples (White & Booth, 1985). As such, stepparents might be strongly motivated to push their children toward early independence.

Fourthly, there might be some circumstances in which both intact and non-intact families may no longer be able to maintain their household. In such situations, both children and parents might seek shelter in someone else's household, in most cases into the house of the grandparents (Aquilino, 1991). This type of co-residence structure is often referred to as "*extended family*". Therefore, as having to move back with relatives is most of the time the result from financial difficulties, it might push children to establish earlier an independent household.

To summarise, we could say that the differences in dynamics related to staying or leaving home between young adults from intact and dissolved families can be the result of divergences regarding economic factors and quality of relations. However, it might also be linked to parental investment. Indeed, concerning single-parent households, sociologists and developmental psychologists have shown that divorce is likely to reduce parental skills and time investments. This

can stem from the fact that, because single-parents tend to experience more stress, their capacity to support and nurture their children may diminish (Furstenberg & Kiernan, 2001). Regarding stepfamilies, although they may seem more similar to intact families in terms of monetary resources and availability of two parental figures, there is some evidence that parents in stepfamilies devote less time to their children and to their children's activities than parent from intact households (Morrison, Moore, Blumenthal, Coiro, & Middleton, 1994). Moreover, it has been shown that stepchildren tend to receive less parental support for the pursuit of their higher education (Zvoch, 1999). This lower level of parental investment may be the result of ambiguity regarding parental role and kinship obligations (Morrison et al., 1994). The absence of biological bonds between stepchildren and stepparents could explain the lower level of emotional support from stepparents. Conversely, *"parents who wed and remain together have greater material resources from the start, have more human capital, are better able to collaborate, are more likely to be embedded in a system of social support, and probably have greater cognitive and social skills as well"* (Furstenberg & Kiernan, 2001, p. 448). As a consequence, intact families are better able to keep their children longer at home. This can be considered as a mechanism to afford higher education, to pursue low paying or no-paying internships that boost their children's capital in the labour market, or to save for a stronger launch when young adults leave home. It can also be seen as a way for middle-class families to support their children while they explore options.

2.4 Other explaining factors

Even though the impact of the co-residence structures on leaving home has been repeatedly demonstrated, it is also known that nest-leaving is associated with other variables, such as sex, labour force participation, geographical location, ethnic origin, socio-economic background, educational level and presence of siblings in the household. As a consequence, these factors need to be integrated into a model

which studies the relationships between the co-residence structures and the departure from the parental home.

Firstly, there is some evidence that sex has a significant discriminating influence on the departure from the parental home (Thomsin et al., 2004). Indeed, it has been shown that women leave home at an earlier age than men. For example, Billari, Philipov and Baizán (2001) have shown that while the median age at first home-leaving for women is equal to 19.2 in Switzerland, that of men is slightly higher (21.5). This observation can result from the fact that, in agreement with Mitchell (1994, p. 666), *"the socialisation process may perpetuate and reproduce traditional behaviours for each sex, so that some women place a greater value on family life than young men and marry at an earlier age"*. In this way, the difference in age at first home-leaving by sex can be without any doubt explained by the difference in age at first marriage (Chiuri & Del Boca, 2010). Another reason for which women leave home at an earlier age might be that leaving home is a good means to escape the closer surveillance and control that weigh on them when they are still living at home.

Furthermore, as far as the co-residence structures is concerned, it has been shown that the stepfamily effect has a divergent influence on home-leaving according to sex. For instance, having an involved stepfather can be considered as a benefit for young boys, whereas stepdaughters encounter much more difficulties when their stepfathers attempts to get involved in child-rearing (Aquilino, 1991; Buck & Scott, 1993; Cooney & Mortimer, 1999). In other words, while daughters seem to adjust better to a family environment where divorced mother do not remarry, sons tend to benefit from the acquisition of a stepfather. Lastly, there is some evidence that living in an extended family[2] has only an effect on women (Aquilino, 1991). Thus, while young girls from extended family structure are expected to leave home at an earlier age, we make the assumption that boys will not be affected by

this household's environment. To summarise, young women who have either grown up with two biological parents, in a stepparent household or in an extended family are expected to leave home at an earlier age than boys. Nonetheless, the opposite effect is presumed in a single-parent household.

Second, if one assumes that leaving home requires at least a minimum amount of financial resources, economic independence may be seen as a significant prerequisite for moving out of the parental home (Nilsson & Strandh, 1999; Aassve, Billari, & Ongaro, 2001; Jacob & Kleinert, 2008; Couppié & Gasquet, 2009). Nonetheless, obtaining employment might also cause the departure from the parental home (Couppié & Gasquet, 2009). Indeed, if a young adult finds employment in a different city than the one in which he/she is currently living, he/she will have to move out in order to live closer to his/her work place. Alternatively, people residing in isolated areas may also be forced to move out from the parental home in order to access better work opportunities.

Thirdly, residential location is also a determining factor for home-leaving because of its influence on the availability of educational and work opportunities, and housing markets (Mitchell, 1994; Mulder & Hooimeijer, 1999). Consequently, as mentioned beforehand, people living in isolated areas may be forced to move out from their hometown in order to benefit from better job and education opportunities. They are, thus, more likely to leave the parental home at an earlier age.

Fourthly, the ethnic origin of young adults is also expected to have a significant effect on their propensity to leave home. Indeed, as asserted by Giuliano (2007), the second-generation immigrants are more likely to follow the patterns of leaving home that are dominant in the home country of their parents than those who are typical of their host country, independently of their economic and educational

---

[2] As a reminder, an extended family is a household that goes beyond the nuclear family. It is often composed of grandparents, aunts, uncles or cousins, all living

13

backgrounds. As a result, it has been shown that children of Italian and Spanish migrants tend to leave home later than children of Swiss natives (Bolzman, 2007). According to a certain number of studies conducted in Switzerland, two factors can explain the behaviour of these specific national communities. Firstly, a delayed departure from the parental home may be due to a lack of economic resources. Indeed, as a significant number of families from a migratory background belong to lower classes, they often cannot afford to pay several rents at the same time. Secondly, it has been demonstrated that the values conveyed by the parents tend to vary according to the country in which they were raised. For instance, there is some evidence that, in migrant families, the departure from home is only considered when children acquire economic independence and are, thus, able to found their own household (Bolzman, Fibbi, & Vial, 2003). This requirement probably stems from the first argument which is that these families cannot afford to pay simultaneously several rents. These aforementioned observations corroborate the model developed by Reher (1998) on family ties. He makes a distinction between a Nordic family system with weak ties, where the individual and individual values have priority over everything else, and a Southern family system with strong ties, in which the family group dominates the individual. According to Granovetter (1973, p. 1361), *"the strength of a tie is a (probably linear) combination of the amount of time, the emotional intensity, the intimacy (mutual confiding) and the reciprocal services which characterize the tie"*. Consequently, the definition of the Nordic family system as a system with weak ties does not mean that there are no relationships among family members, but that they are less strong than in the Southern family system. Indeed, at a general level, family ties are one of the strongest social ties, though some cultural variations can exist. What is more, as demonstrated by the study of Luetzelberger (2014) on high-educated students in Italy and Germany, Reher's family is just as topical as ever. Concerning the population hailing from the Balkan Peninsula, Mandic (2008) shows that people from Eastern Europe present home-leaving patterns that are quite similar to

---

under the same roof.

their Southern peers, even though they leave the parental home at a slightly higher age than the latter. As a consequence, we make the assumption that the propensity to leave home will diverge according to the ethnic origin of young adults, even if they grew up in the same country. Nonetheless, I suppose that the main distinctions will be found between the Swiss natives and the second-generation immigrants from Southern or Eastern Europe, with the latter slightly less likely to leave the parental home than their Southern peers. The rest of the population from a Northern European and Northern American background will probably not significantly differ from their native counterparts as their cultural systems are not very dissimilar.

Fifthly, there is some evidence that having a tertiary education is positively associated with the probability of still living at home, both for daughters and sons (Chiuri & Del Boca, 2010). Nonetheless, it could also be assumed that young adults who pursue a higher education might be more likely to leave home, because institutions of higher education are mainly concentrated in urban centres (Mulder & Hooimeijer, 2002; Bernhardt et al., 2005).

Lastly, it has been shown that the number of siblings living in the same household is likely to affect the probability of young adults leaving the parental home (Mitchell et al., 1989; Aquilino, 1991; Gierveld, Liefbroer, & Beekink, 1991; Avery et al., 1992 ; Buck & Scott, 1993). This may be explained by the fact that individuals who grow up with a large number of siblings have a higher risk of feeling "overcrowded" in their parental home and of suffering from a lack of physical space for privacy. For this reason, they have a higher likelihood to leave home than individuals who grow up alone or with a limited number of siblings. First-born children have a higher likelihood of leaving home at an earlier age than any other children, except if they are only children (Bianchi, 1987). Indeed, Holdsworth (2000) has also shown that only children tend to stay longer at home in order to take care of their parents.

15

3 Methodology

3.1 Data

The analyses used data from the LIVES Cohort Study[3], a panel survey whose first wave was conducted from mid-October 2013 to the end of June 2014 (Elcheroth & Antal, 2013). The sample was composed of 1691 respondents, among which 415 were Swiss and 1276 were from a foreign background. Various criteria had to be fulfilled in order to be eligible, such as being a Swiss resident and being aged 15-24 on January 1st 2013. Also, respondents had to have begun attending a Swiss school before the age of 10. Regarding people of foreign origin, only those whose parents were born in a foreign country and arrived in Switzerland after the age of 18 were taken into consideration. What is more, whether naturalized or not, the second-generation immigrants were over-represented and a particular attention was paid to offspring of low- or middle-skilled migrants who mainly hailed from Southern Europe or from the Balkan Peninsula. The aim of this study is to follow those people over at least ten years in order to study their transition to adulthood.

The sampling process of this survey was very similar to to the respondent-driven sampling, meaning that an initial randomly chosen sample serves as a primary contact to assess a particular type of population (Heckathorn, 1997). In a respondent-driven sampling, the initial subjects are asked to provide the names of a specific number of individuals who fulfil the research criteria. Then, these individuals are approached and asked whether they want to participate in the study. Each person who agrees is asked to give a fixed number of supplementary names. This procedure continues for as many stages as desired. This method is often used to contact hidden populations who are hard to reach, such as drug addicts. Accordingly, the sampling process of the Cohort study can be divided in two stages. First-

---

[3] PRN LIVES. (2013). *Enquête de cohorte* [Data file]. Lausanne : MIS Trend.

ly, the first sampling stage was very similar to a random sampling with unequal probabilities. The use of unequal probability in sampling was first suggested by (Hansen & Hurwitz, 1943). The aim of this method is to randomly select individuals, though the probabilities of selection of each stratum are unequal. Concerning the LIVES Cohort survey, the Federal Statistical Office selected around 4000 people from the Swiss Federal Resident Registration[4]. Then, from these latter, a random sampling with unequal probability of respondent selection was generated which means that the second-generation immigrants were more likely to be selected. In order to be more likely to reach this type of population, the selection process depended on various criteria. Indeed, individuals who met those criteria had a higher probability of being part of the survey. First, people who held the nationality of one of the following countries - Bosnia-Herzegovina, Croatia, Spain, Italy, Kosovo, Macedonia, Montenegro, Portugal, Serbia and Turkey – or who were born in one of these aforementioned countries were eight times more likely to be selected. The resident permit was also a selection criterion. Indeed, it was assumed that a holder of a B or C resident permit was more likely to be a second-generation immigrant than someone who had the Swiss nationality and had, thus, eight times more chances of being selected in the sample. Lastly, the people residing in one of the thirty Spatial Mobility (SM) regions[5] with the highest percentage of foreign-born residents - such as Lausanne, Geneva, Lugano etc. – were two times more likely to be selected in the sample. The second stage was very similar to a random snowball sampling {Citation}. During this stage, the selected respondents had to indicate the name of the people with whom they remembered having had a conversation at least once a week for the last three months. At first, four potentially eligible individuals were randomly selected from the network of each respondent, though second-generation immigrants were four times more likely to be selected than other eligible members. Secondly, the same procedure was applied with the

---

[4] Einwohnerkontrolle

17

exception that only two potentially eligible members were randomly selected from the network of each respondent. In the final step, respondents transmitted the contact information of the selected individuals.

3.2 Methods

Over the past few years, the life course research has known a great development leading to many methodological improvements in longitudinal data analysis (Aisenbrey & Fasang, 2010). These developments can be classified in two broad classes of methods.

First, event history analysis is a probabilistic approach that focuses on the study of events and discrete transitions. The main aims are to analyse the distribution of the timing of the occurrence of an event and to examine the influence of different factors, time-varying or not and related to the respondents or to the context in which they live, on this distribution (Aalen, Borgan, & Gjessing, 2008; Mills, 2011; Allison, 2014). In our case, the event under study is the departure from the parental home.

A second set of methods based on sequence analysis allows studying life trajectories in a holistic perspective. Among several advantages, this approach takes into account much more complex dynamics than a single change of status would (i.e., transition or event). This is a very useful innovation because these dynamics are often nonlinear, disordered, reversible, long-lasting and complex (Martin, Schoon, & Ross, 2008) and they should, thus, be studied in continuity. Accordingly, sequence analysis analyses the timing, duration, order and reversibility of states changes (*Ibid*.). Sequence analysis was initially developed by molecular biologists

---

[5] Switzerland is divided into 106 so-called spatial mobility (SM) regions elaborated on the basis of commuter flows. For more information about these analytical regions, please refer to (Martin, Dessemontet, & Joye, 2005)

whose aim was to compare DNA and protein sequences and to determine distance between two DNA strands (Kruskal, 1983). Andrew Abbott (1983) re-applied it in social sciences for his work on the careers of musicians. This method functions by comparing sequences of states, such as states of living arrangements, and by identifying typical patterns among them (Abbott, 1995). Sequence analysis is divided in three steps. First, sequences of states are created. Second, a pairwise distance matrix describing how different each sequence is from the others is formed. Finally, the closest sequences are gathered into clusters and the resulting clustering can be used as a dependent or an independent variable (*Ibid*.). Sequence and clusters analyses can be conducted with the TraMineR package (Gabadinho, Ritschard, Mueller, & Studer, 2011), a library for sequence analysis in R.

To sum up, both approaches have undeniably divergent objectives. Indeed, while event history analysis predicts life course transitions, sequence analysis aims to compare individuals and to emphasize their resemblances. Likewise, the first approach studies the probabilistic risk of the occurrence of an event, whereas the second one concentrates on the distance between individual trajectories. However, we wish to develop a combination of both approaches in order to study how complex past trajectories influence the probability of leaving home. This combination functions as follows. We used a discrete-time representation of our yearly data[6]. For each individual *i* at each time point *t (from age 0 to the end of the observation period)*, we reconstructed the past co-residence trajectories from age 0 until year *t-1*. These trajectories can therefore be interpreted as trajectories until all possible present times. We thereby have *t* trajectories for each individual that are of varying length. In order to include these past trajectories in our subsequent analysis, they are clustered into ideal types of past trajectories. These clusters are time-specific, meaning that an individual classified into a given cluster at age 10 may belong to another one at age 18. The aim is here to construct a time-varying co-

---

[6] This is coherent with our data collection method, but also with sequences analysis, which uses a discrete-time representation.

variate representing typical trajectories of past co-residence structures and to test their effect on the chance/risk of leaving the parental home, using a discrete-time event analysis framework. This strategy prevents us from many common problems resulting from the combination of these two sets of methods. Indeed, using sequences of states to explain the occurrence of an event can appear problematic if the time frame of the sequence analysis overlaps that of the event history analysis. We cannot estimate the probability of an individual leaving the parental home by taking into account changes in the co-residence structures that occur later in life.

3.3 Operationalization

Dependent variable: Departure from the parental home

As postulated by Holdsworth (2000, p. 201), "*the process of leaving home is viewed as an integral part of establishing economic and emotional independence from the parental home*". As a consequence, as long as the respondents were living with a member of their family of origin (siblings excepted), this situation was defined as "dependent". On the other hand, the co-residency with siblings, children, partner or friends was considered as "independent". In this way, we assumed that the financial dependence from one's parents ends when one leaves the parental home. In this way, a spouse who was cohabiting with his or her partner, but who was not working, was, nonetheless, considered as economically independent. Also, there could be some situations in which students were living by their own for education purposes, even though they were still economically supported by their parents. Indeed, we did not have any information on the financial support provided by parents to their children. As a consequence, some of the respondents could have been considered as economically independent even though they were not completely. Nevertheless, even if the parents often continue to support their children when they leave home to get a higher education, a significant number of students work besides their studies (Mileti, Plomb, & Henchoz, 2015). It has also been stated that living away from the parents can constitute a symbolic independ-

ence, even though frequent return trips may be needed to refuel financially and emotionally before moving out again (Corijn, 2001). As such, the departure from the parental home to pursue higher education could be considered as a transitional period toward economic independence. We created a variable "status", equal to zero when the event had not yet occurred and to 1 when it had. All the episodes following the occurrence of the event have been removed from the database because we were only interested in the first departure from the parental home. As a consequence, the aim of this contribution is to show how independent factors affect the likelihood of leaving the parental home for the first time rather than staying at home. The assumption was made that the individuals enter the risk period of experiencing the event at the age of 15. As a result, two departures from the parental home were not taken into account in the analyses. Consequently, during the observation period, 147 people experienced the event studied. This means that only 9% of the sample had already established an independent household at the time of survey. This low value could come from the fact that the respondents are very young. Indeed, the median age of the sample is equal to 19. It is maybe due to the fact that young adults still living at their parental home were easier to contact and had therefore a higher propensity to participate to the study. We also have to remember that the second-generation immigrants are overrepresented in the sample and that they are more likely to leave home at a later age than the Swiss natives.

Independent variable: Childhood co-residence structures

According to Martinson and Wu (1992), a significant number of studies of childhood co-residence structures are based on "snapshots" which only focus on a particular age, most often age 14. In our case, the LIVES Cohort study collected very detailed life history records of the composition of the respondents' co-residence structure at each age. Accordingly, the previously introduced methodological framework was applied to the data. Eight groups of typical trajectories of co-residence structures were constructed (Cf. Figure 1). The following clustering procedure was used. In order to emphasize the importance of the ordering of states

21

within the trajectories, we used the optimal matching on the distinct states sequences (states sequences without timing information). Then, two groups representing specific cases were manually constructed: a whole trajectory spent with both parents either with or without siblings. Each category represents respectively 40.5 per cent and 4.1 per cent of the sample. Even though the last group is not very frequent in the data, the decision was made to keep it as being an only child is expected to have a significant influence on the risk of leaving the parental home. The co-residence structure with both parents and siblings was also emphasized because it is the most frequent co-residence structure in the sample. For the clustering procedure, a partitioning method, which divides the database into a pre-defined number of groups, was used. To do so, the PAM ("Partitioning Around Medoids") algorithm was selected. Its aim is to obtain the best partitioning for a data set into a predefined number k of groups (Studer, 2013). In other words, the objective of this algorithm is to identify the k best representatives of groups, called "medoids". The medoids can be defined as the observation of a group that has the smallest weighed sum of distances from the other observations of this group. As a consequence, this algorithm seeks to minimize the weighted sum of distances from the medoid. The measures of the quality of a partition help in choosing the best partition among a set of possibilities. As a consequence, according to the "ASW"[7] index, a solution into six groups seemed the best (Cf. Annexes, Table 1). Therefore, the final clustering is divided into eight clusters.

1. Both parents & siblings (40.5%) – As its name indicates, this cluster designates people who grew up with their both parents and siblings. This category was used as the category of reference in the logit regression.

2. Both parents (4.1%)– This category concerns individuals who spent all the observation time with their both parents without any siblings.

---

[7] Average Silhouette Width is based on the coherence of assignments of an observation to a given group. High coherence indicates high between-group distances and strong within-group homogeneity (Kaufman & Rousseeuw, 2005)

3. Late departure of siblings (3.1%)- This category is characterised by young adults who lived the departure of their (probably older) siblings. Thus, they are certainly younger children.

4. Early arrival of siblings (28.7%) – This group is composed of oldest children who experienced the arrival of younger siblings during their teens.

5. Both parents to one parent (with siblings) (10.4%) – This group is characterised by individuals who went from a bi-parental to a lone-parent household, in both cases in the presence of siblings.

6. Early arrival of siblings & parental separation (6.2%) – This cluster is charac- terised by older siblings who experienced the arrival of younger siblings during their teens. A parental disruption occurred subsequently.

7. One parent to both parent (with siblings) (2.5%) – Individuals belonging to this group started their life by living with one parent only and siblings. The second parent joined the household later.

8. Both parent to one parent (without siblings) (4.5%) – These young adults are only children who experienced the parental disruption of their parents.

One of the main problems of the life history calendar used in this survey is that it did not enable us to distinguish the extended family from the stepparent house- hold. What was known is that the respondents, at a certain point in time, were liv- ing with one of their parents and other relatives, but the nature of the family ties between those relatives and the respondents were unknown. This could be a grandparent, an aunt or any other family members, but it could also be a steppar- ent. Also, as few people lived with only one parent and other relatives or with one parent, siblings and other relatives (respectively 0.009% and 2.4% of the observa- tions), those aforementioned situations were defined as living with one parent in the first case and as living with one parent and siblings in the second case.
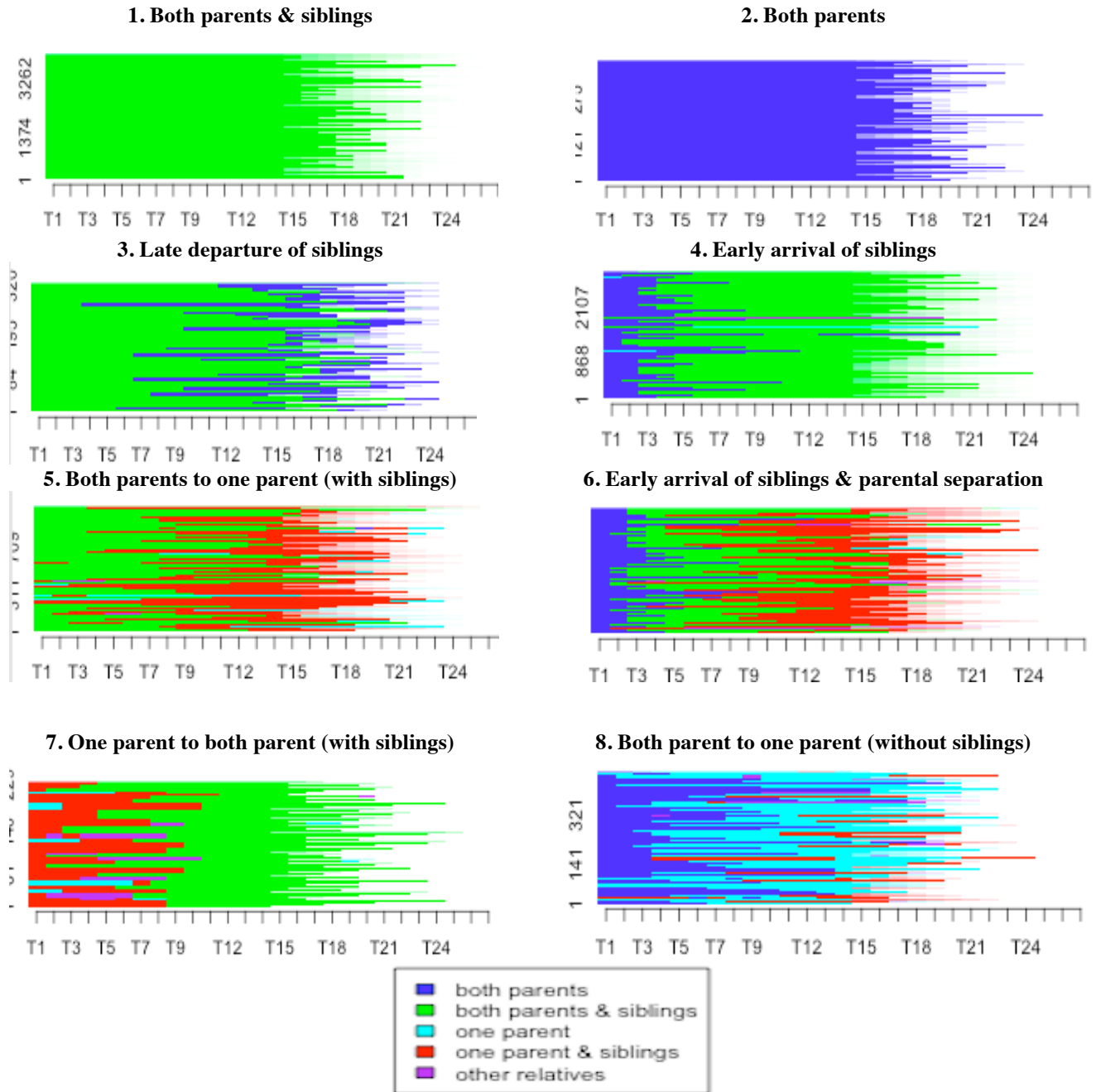
Fig. 1: Clusters of trajectories of past co-residence structures.

Control variables

Some control variables were introduced in the model as their influence on home-leaving had been demonstrated in previous researches.

First, *age* was included in the model, since people are expected to be more likely to leave home as they grow older. A variable indicating the elapsed number of years since the beginning of the risk period was first created. Nonetheless, when event history data are in the form of a discrete-time process and the dependent variable is binary, it is necessary to account for duration dependence (Box-Steffensmeier & Jones, 2004). In other words, it is important to take into consideration the fact that the conditional probability of experiencing the event is likely to vary over time or age. For this reason, a variable representing the *natural logarithm of age* has been included in the model.

Second, it has been asserted that the timing of home-leaving is likely to diverge according to sex. Thus, se*x* was added to the analyses and men were defined as the category of reference.

Third, it has been shown that the *ethnic origin* of young adults is likely to have an impact on the choices they make regarding their transition to adulthood. Thereby, a categorical variable regarding their ethnic origin was created. In order to distinguish the Swiss natives from the second-generation immigrants, we referred to the place of birth of their parents. In some research conducted by the National Institute of Demographic Studies (INED) and the Centre for Studies and Research on Qualifications (CEREQ), the respondents were considered as second-generation immigrants if at least one of their parents was not born in the host country (Santelli, 2004). We also decided to use this definition and the origin country of the foreign parents was used as the benchmark to define the ethnic origin of the respondents. Concerning mixed unions, namely marriages between people with different national origins (Swiss not included), we always emphasized the native

25

country of the mother. Indeed, it has been previously shown that, among the population from a foreign background, the departure from the parental home is more an issue of socialisation than a lack of opportunities. As the role of socialisation is principally endorsed by the mother, we only considered her native country in case of mixed unions. In some circumstances, the information about the country of birth of the parents was missing. In this case, the ethnic origin was deduced from the respondent's first nationality. As it was a self-assessed nationality, if "Swiss" was mentioned as the first nationality, we verified that the respondent did not mention a second foreign nationality. If he or she had, the respondent was considered as a second-generation immigrant and his or her foreign nationality was used to assess his or her ethnic origin. Five categories were created: Switzerland, Eastern Europe, South-western Europe, North-western Europe and Northern America, and other continents.

Besides, it has been shown that *the labour market integration* may act as an incentive to leave home. For this reason, we added to the model a dummy variable coded 0 when the respondents were out of the labour market and 1 after their first integration into the labour market. The apprenticeship was considered as an entry into the labour market.

Furthermore, the Cohort study provided information concerning the different trainings achieved over the years. Thus, *education trajectories* have been reconstructed and added to the analyses. Accordingly, respondents could have attended one or some of the following education programs: compulsory education (category of reference), a 10th year or an au pair/residential language courses, an apprenticeship, a professional formation, a higher secondary education, a higher professional education and training, a university or a university of applied sciences. The 10th year, or bridge-year courses, refers to all transitory offers that are provided for young people with educational deficits at the end of compulsory education (Swiss Media Institute for Education and Culture, 2011d). The aim is to support young adults in their decisions regarding their career prospects, to ease their integration

into the labour market or to prepare them for vocational education and training (VET) or to schools offering general upper-secondary education. The apprenticeship concerns two types of training: a two-years vocational and training VET programme with Federal VET Certificate and a three- or four-year VET programme with Federal VET Diploma (*Ibid*.). The professional formation designates a Federal Vocational Baccalaureate programme leading to a Federal Vocational Baccalaureate Certificate. This latter may be seen as an extended general education to supplement the three- or four-year VET programme for adolescents with higher learning performance. It can be completed either during the three- or four-year VET or by attending a corresponding educational institution. As a consequence, bridge-year courses, apprenticeships and professional formations have been gathered into one group entitled "vocational education and training". Furthermore, the higher secondary education refers to baccalaureate schools that prepare students for further education at tertiary level, namely at a university (Swiss Media Institute for Education and Culture, 2011a). They can be matura schools or gymnasiums that prepare young adults for the university or general training schools that give access to universities of applied sciences or universities of teacher education. Accordingly, higher secondary education and general training schools have been gathered into the same category titled "higher secondary education". Lastly, higher professional education and training (HPET), and universities and universities of applied sciences have been gathered into one group: tertiary education. HPET is a Swiss speciality. It is a type of tertiary education, but it provides programmes for demanding occupational fields and leadership positions (Swiss Media Institute for Education and Culture, 2011b). On the other hand, universities and universities of applied sciences are the traditional academic institutions for higher education. While studies at the university have a scientific approach, university of applied sciences supplement the university education with professionally-oriented programmes (Swiss Media Institute for Education and Culture, 2011c).

In addition, as it has been previously postulated, people living in or close to a big urban centre benefit from greater educational and work opportunities and are,

thus, less likely to leave home than people living in rural municipalities. Therefore, we created a variable indicating the *place of residence* of the respondents at age 14. This factor is composed of 6 modalities: big centres (category of reference), middle and big centres, periurban and metropolitan centres, periurban and pendular municipalities, tourist municipalities, and outlying municipalities. This classification results from the typology of municipalities in 22 categories developed by Martin, Dessemontet and Joye (2005) The recoding of this typology is presented in Table 2 in the appendix. The classification developed by the aforementioned researchers is based on a model centre-periphery, meaning that the municipalities are classified in different categories according to their belonging to a metropolitan agglomeration, to a non-metropolitan agglomeration or to a rural municipality. The other criteria used to construct this typology are variables related to employment, structure of buildings, wealth, tourism, structure of the population and centrality. Concerning the sample, there were also a small number of people who were living abroad when they were 14 years old (n=5). Because this number was very small, these cases have been recoded as missing.

Likewise, in the section regarding the description of the sample, it has been reported that second-generation immigrants have been over-represented in this survey and that, for this reason, the selection process was based on various criteria such as the place of birth, the nationality, the residence permit, the place of residence and the size of social network. As a result, in order to avoid biases in the analyses, the inclusion of these factors was a necessary step. Nonetheless, almost all these criteria designated the situation of respondents at the time of the survey, namely in 2013. However, most of the people who left the parental home did it before 2013 and, methodologically speaking, one cannot explain the probability of an event occurring by factors that refers to a subsequent time period. As a consequence, only the variables that referred to the time period preceding the beginning of the risk period have been kept, namely the place of birth. Indeed, the nationality, the residence permit, the place of residence and the network size are all time-varying variables that can change over time. Moreover, nationality is already par-

tially taken into account in the analyses through the ethnic origin variable. Accordingly, a variable indicating the place of birth of respondents has been created and divided in two modalities: over-represented places of birth and under-represented places of birth. The first modality designated individuals who were born in Bosnia-Herzegovina, Croatia, Spain, Italy, Kosovo, Macedonia, Montenegro, Portugal, Serbia or Turkey because they had a higher probability of being selected in the sample. All other given answers have been gathered into the second modality.

Moreover, it could be asserted that the previously presented childhood co-residence structures mainly measure a distinction between the intact families and the disrupted households. Accordingly, it might be reasonable to think that a variable recording the *occurrence of parental disruption* could give the same result, though by using a much easier variable to construct and to interpret. For this reason, we constructed a time-varying variable that indicated at each age whether the respondents had experiences a parental disruption. More precisely, it was coded 0 if the parents were still together and 1 in case of marital disruption.

Lastly, we mentioned in the theoretical section that the presence of siblings in the household might have an influence on young adults' risk of leaving home. This information is taken into account in the independent variable. Nonetheless, the information recorded in the previously mentioned variable is more complete since we do not only record the presence or absence of siblings in the household, but we also examine whether the arrival of younger siblings or the departure of siblings from the parental home affect those who are still living at home. As such, in order to demonstrate the higher value of constructing childhood co-residence structures, we created a variable simply recording the presence or the absence of siblings and compared its results with those obtained with the independent variable.

29

4 Results

In order to test the previously presented hypotheses, logit regressions were run in order to estimate the impact of past co-residence structures on the probability of leaving the parental home (Cf. Table 1). Four models were built. The aim was to progressively add factors from the theoretically most central variable to the control variables in order to better understand their separate effects. Thus, only the variable regarding the childhood co-residence structures and the age were selected in the first model. Then, all the control variables were introduced in the second model. The third model was completed by the variables measuring the occurrence of parental disruption and the presence of siblings. In the last model, the independent variable was removed. The aim was to examine the effects of the "divorce" and "siblings" variables without including the "childhood co-residence structures" variable.

What is more, we only selected the individuals for whom we had information in every variable. As such, the number of individuals was equal to 1637. It means that 52 respondents had missing data in at least one of the variables that were included in the models. Likewise, 147 individuals out of 1637 left the parental home over the observation period. As it can be noticed, the education variable was not included in the table presented below. The reason is that its inclusion led to the exclusion of 607 additional individuals from the database. What is more, preliminary analyses have shown that education does not have a strong impact on the different pathways out of the parental home (Cf. Annexes, Table 3). For example, only pursuing a tertiary education increases the odds of leaving home and this effect is only significant at the 0.1 level.

The analyses show that young adults who only grew up with their both parents only are as likely to leave home as those who spent their childhood with their both parents and siblings. Conversely, it appears that staying in the parental home after

the departure of their siblings increases the probability of leaving home. Regarding older children who experienced the birth of younger siblings during their teens, it seems that their likelihood of leaving home is higher than that of young adults who grew up with their both parents and siblings from birth. Besides, there is some evidence that having experienced a parental separation – in the absence of siblings – leads to a higher risk of leaving home. Similarly, when parental disruption occurs in the presence of siblings, it also increases the odds of leaving the parental home, though the increase is smaller than in the previous case. What is more, it seems that older children are less affected by parental separation. As an illustration, even though being the oldest sibling and having experienced divorce positively influence the departure from the parental home, this effect is only significant at the 0.1 level in the second model. Lastly, the results show that young individuals who started by living in a lone-parenthood with siblings before moving to a biparental household are as likely to leave home as those who spent their entire childhood in a biparental household.

What is more, we can see that, even after the inclusion of the variables measuring the occurrence of divorce and the presence of siblings in the third model, the effects of the childhood co-residence structures are still statistically significant, though being the oldest child and having experienced a parental disruption do no longer increase the risk of leaving the parental home. Concerning the variable recording the occurrence of divorce, its effect is not statistically significant (whether included with the childhood co-residence structures or not). As far as siblings are concerned, it seems that their presence in the household foster the departure from the parental home. However, this ascertainment is only true for the third model, where all the variables were included.

Regarding control variable, consistent with what was assumed, young adults seem to be more likely to found their own household as they grow older. Conversely, men's and women's risk of leaving home surprisingly do not differ. In addition, the outcomes show that the departure from the parental home is significantly in-

fluenced by the ethnic origin. For instance, it has been demonstrated that second-generation immigrants from the Balkan Peninsula or from Southern European have lower odds of leaving home than their Swiss counterparts. As for second-generation immigrants from North-western Europe, they have a higher tendency to leave the parental home than their Swiss counterparts. Furthermore, as expected, the integration into the labour market increases the probability of leaving the parental home. Lastly, regarding the place of residence, although residing in an outlying municipality or in a periurban and pendular municipality (in the third model) during childhood seems to increase the odds of leaving home, this effect is only significant at the 0.1 level.

Lastly, the Akaike information criterion (AIC)[8] and the Bayesian information criterion (BIC)[9] can be used to compare the quality of each model. They are a measure of the relative quality of statistical models for a given set of data. Given a collection of models for the data, these criteria estimate the quality of each model, relative to each of the other models. They thus provide a means for model selection. They offer an estimation of the trade-off between the goodness of fit of the model and its complexity. Concerning the BIC, Raftery (1995) asserts that, when calculating the BIC for event history data, N can refer to tree different notions: the number of observations (person-period), the number of individuals or the number of events. They suggest using the last option, which is the last conservative. This last option is also coherent with the calculation of the BIC in the case of survival continuous time models (i.e. Cox models) in which N represents the number of observed events. Accordingly, we can see that, according to the both criteria, the best one is the third model, namely the model which is the most complete. However, the differences between the second and the third model are very small.

---

[8] AIC= 2k – 2ln(L), where k represents the number of parameters and -2ln(L) is equal to the deviance

[9] BIC = -2ln(L) + ln(N)*k, where L is the likelihood, -2ln(L) is equal to the deviance, ln is the logarithm and k represents the number of parameters (i.e. coefficients).

Table 1. Logit models predicting probability of first home-leaving

Covariates		Model 1			Model 2			Model 3			Model 4		
		Coef.	Std. Err.	Sig.	Coef.	Std. Err.	Sig.	Coef.	Std. Err.	Sig.	Coef.	Std. Err.	Sig.
Intercept		-10.59	0.75	***	-11.61	1.02	***	-12.87	1.19	***	-10.77	1.00	***
Co-residence structures	Both parents & siblings (ref.)	-	-		-	-		-	-		-	-	
	Both parents	0.27	0.76		0.20	0.79		1.18	0.92				
	Late departure of siblings	1.52	0.42	***	1.28	0.44	**	2.27	0.64	***			
	Early arrival of siblings	0.67	0.34	*	0.78	0.35	*	0.81	0.35	*			
	Both parents to one parent (without siblings)	1.96	0.45	***	1.86	0.50	***	2.14	0.61	***			
	Early arrival of siblings & parental separation	0.62	0.47		0.91	0.49	+	0.78	0.68				
	One parent to both parents (with siblings)	0.33	1.06		1.36	1.12		1.44	1.13				
	Both parents to one parent (with siblings)	1.12	0.36	**	1.36	0.38	***	1.61	0.39	***			
Age (ln)		3.45	0.36	***	3.55	0.40	***	3.63	0.40	***	3.65	0.40	***
Sex	Men (ref.)				-	-		-	-		-	-	
	Women				0.31	0.24		0.34	0.25		0.22	0.24	
Ethnic origin	Switzerland (ref.)				-	-		-	-		-	-	
	Eastern Europe				-1.06	0.41	**	-1.11	0.41	**	-1.15	0.40	**
	South-western Europe				-0.89	0.40	*	-0.85	0.40	*	-0.95	0.40	*
	North-western Europe & northern America				1.25	0.41	**	1.26	0.41	**	1.22	0.41	**
	Other continents				0.24	0.38		0.18	0.38		0.49	0.37	
Labour market integration					0.69	0.29	*	0.72	0.29	*	0.55	0.30	+
Place of residence	Big centres (ref.)				-	-		-	-		-	-	
	Periurban & metropolitan centres				0.35	0.48		0.33	0.48		0.47	0.47	
	Touristic municipalities				0.45	0.65		0.51	0.64		0.54	0.63	
	Middle & little centres				0.32	0.31		0.32	0.31		0.40	0.31	
	Periurban & pendular municipalities				0.72	0.46		0.79	0.47	+	0.52	0.46	
	Outlying municipalities				0.65	0.37	+	0.64	0.37	+	0.63	0.37	+
Place of birth	Overrepresented places of birth (ref.)				-	-		-	-		-	-	
	Underrepresented places of birth				0.12	0.44		0.17	0.47		0.54	0.63	
Divorce								0.20	0.53		0.53	0.33	
Siblings	No siblings (ref.)							-	-		-	-	
	Siblings							1.03	0.47	*	-0.06	0.29	
Nb obs.	8796												
Nb ind.	1637												
Nb events	147												
Deviance			652.2			606.8			601.16			627.8	
AIC			670.2			648.8			647.2			659.8	
BIC			669.6			652.3			651.0			662.5	

+ p> 0.1; * p < 0.05; ** p < 0.01; *** p < 0.001

33

4 Discussion

The main results show that punctual events of the past life history (such as the occurrence of divorce and the arrival of siblings etc.) do not play a strong role on the departure from the parental home. Conversely, there are some reasons to believe that childhood co-residence patterns influence the ways in which young adults leave the parental home. More precisely, it is rather the ways in which various states occur, as well as their sequencing, which has an effect on home-leaving. For instance, it seems that two features have a significant and strong impact on the departure from the parental home, which are the childhood co-residence structures and the siblings.

Regarding the co-residence structures, it has been demonstrated that having spent some years in a lone-parent household has a positive impact on the risk of leaving home. It does not seem to matter much if it occurred in the presence of siblings or not, as both situations lead to an increase in the likelihood of leaving home. Though, we can see that, in the case of parental separation, the presence of other siblings tends to buffer the propensity to leave the parental home. This might also come from the fact that, in general, divorce occurs earlier in couples with only children. As a consequence, we could say that the influence of non-normative changes in the co-residence configurations tends to vary according to the timing of its onset. What is more, there is some evidence that, in households of more than one child, oldest children are less affected by divorce than younger children. Lastly, the outcomes show that there are no significant differences in the risk of leaving the parental home between young adults who grew up in a biparental household and those who experienced the re-partnering of their parents at a later age, in both cases in the presence of siblings. This could stem from the fact that both groups reach the same destination state; namely a household with two parents. Thereby, we could assume that this shift cancels out the negative effect of the non-standard co-residency structure.

Concerning siblings, it appeared that what matters is not only whether or not young adults grew up with other siblings, but also their birth order. Indeed, it has

been shown that only children have the same probability of leaving home than children who grew up with siblings, though our hypothesis was that only children tend to stay longer at home in order to take care of their parents (Bianchi, 1987; Holdsworth, 2000). As a result, the opposite result could indicate that taking of their parents is no longer a necessary requirement. What is more, it might be possible that parents invest more in their children's future when they are only children and this may foster their departure from the parental home. Second, contrarily to what would have been expected, it seems that the departure from the parental home of siblings (most probably older siblings) incites the other siblings – who stayed at home - to leave home. Indeed, it could have been assumed that young individuals who attended the departure of their siblings would be less likely to leave home than those who are still living with their both parents and siblings, because they do no longer have to share space, parental attention and support with other siblings. On the contrary, it seems that young individuals look upon their siblings and are very likely to reproduce their behaviours. This could also be an age effect. Indeed, young adults with siblings who have already left home are more likely to have reached the ages in which the departure from the parental home is the most frequent. Finally, in consonance with the previous assumptions, oldest children who spent their first years as only children before the birth of younger siblings are more likely to leave home than youngest children. The reason is that they might have encountered difficulties in sharing the attention, love and support that was once only theirs. They may also be more likely to suffer from a lack of physical space for privacy, as they used to live alone with their parents before the arrival of their siblings.

As a consequence, as leaving home very early might have significant consequences on later life opportunities, the findings draw attention to the fact that the past household structure is a significant determinant of the transition toward a stable and successful work and family trajectory.

In addition, in concordance with our expectations, most of the control variables have an influence on the event studied. First of all, in accordance with the hypoth-

35

eses, the probability of leaving home increases with age. Nevertheless, contrarily to what was expected, women have the same risk of leaving home than their male counterparts. It thus indicates that socialisation processes do no longer seem to contribute to the reproduction of traditional behaviours for each sex. In other words, this might mean that young women do not necessarily place a greater value on family life than young men. Moreover, the ethnic origin of young adults was expected to have a significant effect on their propensity to leave home. More precisely, while second-generation immigrants from Eastern or Southern Europe were expected to leave home at an older age, no different in the age at leaving home was supposed between second-generation from North-western Europe and Swiss natives. The present outcomes partially confirm these assumptions. As an illustration, while it is true that second-generation immigrants from Eastern or Southern Europe are less likely to leave home than Swiss natives, it nonetheless has been demonstrated that second-generation immigrants from North-western Europe or Northern America have higher odds of leaving home than Swiss natives. Besides, in accordance with the hypotheses, economic independence leads to a greater likelihood of leaving the parental home. Likewise, the results also showed that residing in an outlying municipality increases the likelihood of leaving home, even though this effect is really small. This result may be explained by the fact that, although most of the institutions of higher education are concentrated in some Swiss agglomerations (e.g. Zurich, Geneva, Basel, Bern and Lausanne) and, as mentioned beforehand, commuting between the place of residence and these metropolitan areas has become easier and quicker for young individuals thanks to the development of railroad and road networks (Viry et al., 2009). Also, as Switzerland is a small country, the distances between cities are not too big. As a result, tough some young adults who want to pursue higher education might be forced to move out from their parental home, this proportion is likely to be more limited in Switzerland than in other European countries.

Limitations of the study and further contributions

One relative weakness of this study is that it did not enable us to integrate variables regarding the socioeconomic status of the respondent's family, because this information will only be asked in the second wave of the Cohort study. However, the higher probability of children from lone-parent families of leaving home could stem from the fact that those families often encounter financial difficulties and that this situation is positively related to the tendency of children to leave home (Bianchi, 1987). As a consequence, a further contribution of this study will be to integrate those variables into our analysis as soon as they are available.

Another weakness of our study is that the link between the household structure and the departure from the parental home may be explained by another factor, which is the quality of relations within the household. Indeed, the higher probability of children from stepfamilies leaving home could come from the fact that conflicts and disagreements are more frequent in this household environment (Gähler & Bernhardt, 2000). Unfortunately, there is no variable in the survey that might enable us to verify this assumption. As a result, longitudinal data supplemented with more detailed qualitative accounts of the quality of family relations might provide useful information that could fill the gap.

Lastly, as previously mentioned, the sample used for the present study was quite young. As a result, a small number of respondents had already left the parental home at the end of the observation period. Accordingly, it is possible that the results previously presented might concerned principally the early departure from the parental home. A further contribution would be to apply the same methodological framework to an older sample in order to verify whether the present outcomes can be generalised to other populations.

To conclude, the results obtained with the combination of survival analysis and sequence analysis provided results that would not have been obtained if each method had been used separately. Nonetheless, the proposed framework may have a much broader field of application. As a consequence, further investigations have

37

to be conducted in order to examine whether other life course events or transitions are influence by past trajectories. For instance, this methodological framework could allow us to study how previous professional trajectories are linked with the risk of dying.

References

Aassve, A., Arpino, B., & Billari, F. C. (2013) Age norms on leaving home: Multilevel evidence from the European Social Survey. *Environment and Planning A*. **45**(2), 383–401.

Aassve, A., Billari, F. C., & Ongaro, F. (2001) The impact of income and employment status on leaving home: evidence from the Italian ECHP sample. *Labour*. **15**(3), 501–529.

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

Abbott, A. (1995) Sequence analysis: new methods for old ideas. *Annual Review of Sociology*, **21**, 93–113.

Acock, A. C., & Demo, D. H. (1994) *Family diversity and well-being*. California: Sage Publications, Inc.

Aquilino, W. S. (1991) Family structure and home-leaving: A further specification of the relationship. *Journal of Marriage and the Family*. **53**(4), 999–1010.

Avery, R., Goldscheider, F., & Speare, A. (1992) Feathered nest/gilded cage: Parental income and leaving home in the transition to adulthood. *Demography*. **29**(3), 375–388.

Bernhardt, E., Gähler, M., & Goldscheider, F. (2005) Childhood family structure and routes out of the parental home in Sweden. *Acta Sociologica*. **48**(2), 99–115.

Bianchi, S. M. (1987) Living at home: Young adults' living arrangements in the 1980s. Presented at the Annual meeting of the American Sociological Association, Chicago.

Billari, F. C., & Liefbroer, A. C. (2007) Should I stay or should I go? The impact of age norms on leaving home. *Demography*. **44**(1), 181–198.

Billari, F. C., & Liefbroer, A. C. (2010) Towards a new pattern of transition to adulthood? *Advances in Life Course Research*. **15**(2), 59–75.

Billari, F. C., Philipov, D., & Baizán, P. (2001) Leaving home in Europe: The experience of cohorts born around 1960. *International Journal of Population Geography*, **7**(5), 339–356.

Blaauboer, M., & Mulder, C. H. (2010) Gender differences in the impact of family background on leaving the parental home. *Journal of Housing and the Built Environment*. **25**(1), 53–71.

Bolzman, C. (2007) Familles espagnoles et italiennes en Suisse et transition vers la vie adulte. *Enfances, Familles, Générations*. (6).

Bolzman, C., Fibbi, R., & Vial, M. (2003) Que sont-ils devenus? Le processus d'insertion des adultes issus de la migration. In H. R. Wicker H.R., R. Fibbi & W. Haug (eds), *Les migrations et la Suisse: résultats du Programme national de recherche "Migrations et relations interculturelles"*, pages 434–459, Zürich: Seismo.

Box-Steffensmeier, J. M., & Jones, B. S. (2004) *Event history modeling: A guide for social scientists*. Cambridge University Press.

Brodmann, S., Esping-Andersen, G., & Güell, M. (2007) When fertility is bargained: Second births in Denmark and Spain. *European Sociological Review*. **23**(5), 599–613.

Buck, N., & Scott, J. (1993) She's leaving home: But why? An analysis of young people leaving the parental home. *Journal of Marriage and the Family*. **55**(4), 863–874.

39

Cavalli, A., & Galland, O. (1993) *L'allongement de la jeunesse en Europe*. Arles: Actes Sud.

Cherlin, A. J., Kiernan, K. E., & Chase-Lansdale, P. L. (1995) Parental divorce in childhood and demographic outcomes in young adulthood. *Demography*. **32**(3), 299–318.

Chiuri, M. C., & Del Boca, D. (2010) Home-leaving decisions of daughters and sons. *Review of Economics of the Household*. **8**(3), 393–408.

Cooney, T. M., & Mortimer, J. T. (1999) Family structure differences in the timing of leaving home: Exploring mediating factors. *Journal of Research on Adolescence*. **9**(4), 367–393.

Corijn, M. (2001) Transitions to adulthood in Europe: Sociodemographic factors. In M. Corijn & E. Klijzing (eds), *Transition to adulthood in Europe*, pages 1-25, Netherlands: Springer.

Couppié, T., & Gasquet, C. (2009) Insertion professionnelle et départ du domicile parental: une relation complexe. In C. Perez, S. Divay, J-L. Outin, T. Couppié, J-F. Giret, Y. Grelet & P. Werquin (eds), *Les cheminements longs: données, métodes et apports pour les analyses du marché du travail*, pages 227–258, Marseille: Céreq.

Elcheroth, G., & Antal, E. (2013) *Echantillon de cohorte LIVES. Première vague* (pp. 1–7). Lausanne: University of Lausanne.

Ermisch, J., & Di Salvo, P. (1997) The economic determinants of young people's household formation. *Economica*. **64**(256), 627–644.

Furstenberg, F. F., & Kiernan, K. E. (2001) Delayed parental divorce: How much do children benefit? *Journal of Marriage and Family*. **63**(2), 446–457.

Gähler, M., & Bernhardt, E. (2000) The impact of parental divorce, family reconstitution, and family conflict on nest-leaving in Sweden (pp. 6–8). Presented at the Workshop on Leaving Home–a European Focus, Max Planck Institut für Bevölkerungsforschung, Rostock.

Galland, O. (1996) L'entrée dans la vie adulte en France. Bilan et perspectives sociologiques. *Sociologie et Sociétés*. **28**(1), 37–46.

Gierveld, J. D. J., Liefbroer, A. C., & Beekink, E. (1991) The effect of parental resources on patterns of leaving home among young adults in the Netherlands. *European Sociological Review*. **7**(1), 55–71.

Giuliano, P. (2007) Living arrangements in western europe: Does cultural origin matter? *Journal of the European Economic Association*. **5**(5), 927–952.

Goldscheider, F., & Goldscheider, C. (1993) Whose nest? A two-generational view of leaving home during the 1980s. *Journal of Marriage and the Family*. **55**(4), 851–862.

Goldscheider, F. K., & Goldscheider, C. (1998) The effects of childhood family structure on leaving and returning home. *Journal of Marriage and the Family*. **60**(3), 745–756.

Gossens, L. (2001) Transition to adulthood: Developmental factors. In M. Corijn & E. Klijzing (eds), *Transition to adulthood*, pages 27–42, Netherlands: Springer.

Granovetter, M. S. (1973) The strength of weak ties. *American Journal of Sociology*. **78**(6), 1360–1380.

Hansen, M. H., & Hurwitz, W. N. (1943) On the theory of sampling from finite populations. *The Annals of Mathematical Statistics*. **14**(4), 333–362.

Heckathorn, D. D. (1997) Respondent-driven sampling: a new approach to the study of hidden populations. *Social Problems*. **44**(2), 174–199.

Heckhausen, J. (1999) *Developmental regulation in adulthood. Age-normative and sociostructural constraints as adaptive challenges*. New York: Cambridge University Press.

Hogan, D. P., & Astone, N. M. (1986) The transition to adulthood. *Annual Review of Sociology*.

12, 109–130.

Holdsworth, C. (2000) Leaving home in Britain and Spain. *European Sociological Review*. **16**(2), 201–222.

Iacovou, M., & Aassve, A. (2007) *Youth poverty in Europe*. York, UK: Joseph Rowntree Foundation.

Jacob, M., & Kleinert, C. (2008) Does unemployment help or hinder becoming independent? The role of employment status for leaving the parental home. *European Sociological Review*. **24**(2), 141–153.

Juang, L. P., Silbereisen, R. K., & Wiesner, M. (1999) Predictors of leaving home in young adults raised in Germany: A replication of a 1991 study. *Journal of Marriage and the Family*. **61**(2), 505–515.

Kaufman, L., & Rousseeuw, P. J. (2005) *Finding groups in data: an introduction to cluster analysis*. New Jersey: John Wiley & Sons, Inc.

Kiernan, K. E. (1992) The impact of family disruption in childhood on transitions made in young adult life. *Population Studies*. **46**(2), 213–234.

Kruskal, J. B. (1983) An overview of sequence comparison. In D. Sankoff & J. B. Kruskal (eds), *Time wraps, string edits, and macromolecules-The theory and practice of sequence comparison*, pages 1–44, Ontario: Addison-Wesley Publishing Company.

Luetzelberger, T. (2014) Independence or interdependence: Norms of leaving home in Italy and Germany. *European Societies*. **16**(1), 28–47.

Mandic, S. (2008) Home-leaving and its structural determinants in Western and Eastern Europe: An exploratory study. *Housing Studies*. **23**(4), 615–637.

Marini, M. M. (1984) The order of events in the transition to adulthood. *Sociology of Education*. **57**(2), 63–84.

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

Martin, S., Dessemontet, P., & Joye, D. (2005) *Les niveaux géographiques de la Suisse*. Neuchâtel: Swiss Federal Statistical Office.

Martinson, B. C., & Wu, L. L. (1992) Parent Histories Patterns of Change in Early Life. *Journal of Family Issues*. **13**(3), 351–377.

Meyer, T. (2003) When being smart is not enough: institutional and social access barriers to upper secondary education and their consequences on successful labour market entry. The case of Switzerland. Presented at the "Competencies and Careers" of the European Research Network on Transitions in Youth (TIY), Funchal, Madeira. Retrieved from http://www.coreched.ch/publikationen/meyer.pdf

Mileti, F. P., Plomb, F., & Henchoz, C. (2015) De la socialisation financière à l'autonomie économique: processus d'acquisition des compétences et des représentations liées à l'argent auprès d'étudiants vivant en Suisse. *Pensée Plurielle*. **37**(3), 53–65.

Mitchell, B. A. (1994) Family structure and leaving the nest: A social resource perspective. *Sociological Perspectives*. **37**(4), 651–671.

Mitchell, B. A., Wister, A. V., & Burch, T. K. (1989) The family environment and leaving the parental home. *Journal of Marriage and the Family*. **51**(3), 605–613.

Modell, J., Furstenberg, F., & Hershberg, T. (1976) Social change and transitions to adulthood in

41

historical perspective. *Journal of Family History*. **1**(1), 7–32.

Morrison, D. R., Moore, K. A., Blumenthal, C., Coiro, M. J., & Middleton, S. (1994) *Parent-Child Relationships and Investments of Parental Time in Children: Are Children in Stepfamilies at a Disadvantage?*. Washington DC: Child Trends, Inc.

Mulder, C. H. (2009) Leaving the parental home in young adulthood. In A. Furlong (eds), *Handbook of youth and young adulthood: new perspectives and agendas*, pages 203–210, New York: Routledge.

Mulder, C. H., & Hooimeijer, P. (1999) Residential relocations in the life course. In L. van Wissen & P. Dykstra, *Population issues: An interdisciplinary focus*, pages 159–186, New York: Kluwer Academic Publishers.

Mulder, C. H., & Hooimeijer, P. (2002) Leaving home in the Netherlands: Timing and first housing. *Journal of Housing and the Built Environment*. **17**(3), 237–268.

Nilsson, K., & Strandh, M. (1999) Nest leaving in Sweden: The importance of early educational and labor market careers. *Journal of Marriage and the Family*. **61**(4), 1068–1079.

Raftery, A. E. (1995) Bayesian model selection in social research. *Sociological Methodology*. **25**, 111–164.

Reher, D. S. (1998) Family ties in Western Europe: persistent contrasts. *Population and Development Review*. **24**(2), 203–234.

Rosina, A. (2005) Family formation and fertility in Italy. In G. Dalla Zuanna & G.A. Micheli (eds), *Strong Family and Low Fertility: A Paradox? New perspectives in interpreting contemporary family and reproductive behaviour*, pages 23–43, Netherlands: Springer.

Sandefur, G. D., Eggerling-Boeck, J., & Park, H. (2008) Off to a good start? Postsecondary education and early adult life. In Jr. R. A. Settersten, Jr. F. F. Furstenberg & R. G. Rumbaut (eds), *On the frontier of adulthood: Theory, research and public policy*, pages 292–319, Chicago: University of Chicago Press.

Santelli, E. (2004) De la «deuxième génération» aux descendants d'immigrés maghrébins. Apports, heurts et malheurs d'une approche en termes de génération. *Temporalités. Revue de Sciences Sociales et Humaines*. (2), 29–43.

Schizzerotto, A., & Lucchini, M. (2004) Transitions to Adulthood. In R. Berthoud & M. Iacovou (eds), *Social Europe: Living standards and welfare states*, pages 46–68, Cheltenham, UK: Edward Elgar.

Schumacher, R., Spoorenberg, T., & Forney, Y. (2006) Déstandardisation, différenciation régionale et changements générationnels. Départ du foyer parental et modes de vie en Suisse au XXe siècle. *European Journal of Population/Revue Européenne de Démographie*. **22**(2), 153–177.

Studer, M. (2013) WeightedCluster Library Manual: A practical guide to creating typologies of trajectories in the social sciences with R. *LIVES Working Papers*, (24). http://doi.org/10.12682/lives.2296-1658.2013.24

Swiss Federal Statistical Office. (2015) Components of population change – Data, indicators - Divorces and divortiality. Retrieved April 20, 2015, from http://www.bfs.admin.ch/bfs/portal/fr/index/themen/01/06/blank/key/06.html

Swiss Media Institute for Education and Culture (2011a) Baccalaureate schools. Retrieved November 13, 2015, from http://swisseducation.educa.ch/en/baccalaureate-schools

Swiss Media Institute for Education and Culture (2011b) Higher professional education and

training. Retrieved November 13, 2015, from http://swisseducation.educa.ch/en/higher-professional-education-and-training

Swiss Media Institute for Education and Culture (2011c) Institutions for higher education. Retrieved November 13, 2015, from http://swisseducation.educa.ch/en/institutions-higher-education

Swiss Media Institute for Education and Culture (2011d) Vocational education and training in Switzerland. Retrieved November 13, 2015, from http://swisseducation.educa.ch/en/vocational-education-and-training-0

Thomsin, L., Le Goff, J.-M., & Sauvain-Dugerdil, C. (2004) Genre et étapes du passage à la vie adulte en Suisse. *Espace Populations Sociétés. Space Populations Societies*. **11**(1), 81–96.

White, L. K., & Booth, A. (1985) The quality and stability of remarriages: The role of stepchildren. *American Sociological Review*. **50**(5), 689–698.

Zorlu, A., & Mulder, C. H. (2011) Ethnic differences in leaving home: Timing and pathways. *Demography*. **48**(1), 49–72.

Zvoch, K. (1999) Family type and investment in education: A comparison of genetic and stepparent families. *Evolution and Human Behavior*. **20**(6), 453–464.

Annexes

Table 1. Measures of quality of different partitions

	PBS	HG	HGSD	ASW	ASWw	CH	R0.242	CHsq	R2sq	HC
Cluster 2	0.41	0.64	0.58	0.50	0.50	1722.03	0.24	2376.86	0.30	0.23
Cluster 3	0.56	0.80	0.73	0.56	0.56	1718.57	0.38	2627.01	0.49	0.11
Cluster 4	0.61	0.83	0.78	0.59	0.59	1769.22	0.49	2494.38	0.57	0.09
Cluster 5	0.60	0.83	0.79	0.61	0.61	2014.69	0.59	2608.70	0.65	0.12
Cluster 6	0.66	0.92	0.89	0.69	0.69	2301.24	0.67	3116.81	0.74	0.06
Cluster 7	0.68	0.94	0.91	0.71	0.71	2290.66	0.71	3341.62	0.78	0.04
Cluster 8	0.69	0.96	0.93	0.74	0.74	2262.54	0.74	3303.17	0.81	0.03
Cluster 9	0.69	0.96	0.94	0.76	0.76	2229.23	0.76	3332.85	0.83	0.03
Cluster 10	0.69	0.97	0.94	0.77	0.77	2179.48	0.78	3272.93	0.84	0.02

Table 2. Recoding of the variable "place of residence" according to the typology of municipalities of Martin, Dessemontet and Joye (2005)

	Typology of municipalities	Recoding
1	Big centres	Big centres
9	Employment municipalities from metropolitan regions	
10	Suburban municipalities from metropolitan regions	
11	Periurban municipalities from metropolitan regions	Periurban & metropolitan centres
5	High-income municipalities	
6	Toursitic municipalities	Touristic municipalities
7	Semi-touristic municipalities	
8	Municipalities with collective institutions	
2	Middle centres	Middle & little centres
3	Small centres	
12	Employment municipalities from non-metropolitan regions	
13	Suburban municipalities from non-metropolitan regions	
14	Periurban municipalities from non-metropolitan regions	Periurban & pendular municipalities
15	Pendular municipalities of allochtons	
16	Pendural municipalities of autochtons	
4	Centre of peripheral regions	Outyling municipalities
17	Industrial and tertiary municipalities	
18	Industrial municipalities	
19	Agro-industrial municipalities	
20	Agro-tertiary municipalities	
21	Agricultural municipalities	
22	Municipalities in strong demographic decline	

45

Table 3. Logit models predicting probability of first home-leaving, controlled for education

Covariates		Model 1			Model 2			Model 3			Model 4		
		Coef.	Std. Err.	Sig.	Coef.	Std. Err.	Sig.	Coef.	Std. Err.	Sig.	Coef.	Std. Err.	Sig.
Intercept		-10.59	0.75	***	-11.47	1.17	***	-12.67	1.32	***	-10.41	1.15	***
Household structure	Both parents & siblings (ref.)	-	-		-	-		-	-		-	-	
	Both parents	0.27	0.76		0.29	0.80		1.23	0.92				
	Late departure of siblings	1.52	0.42	***	1.32	0.44	**	2.29	0.64	***			
	Early arrival of siblings	0.67	0.34	*	0.78	0.35	*	0.81	0.36	*			
	Both parents to one parent (without siblings)	1.96	0.45	***	1.90	0.50	***	2.12	0.61	***			
	Early arrival of siblings & parental separation	0.62	0.47		0.97	0.49	*	0.79	0.68				
	One parent to both parents (with siblings)	0.33	1.06		1.46	1.12		1.54	1.13				
	Both parents to one parent (with siblings)	1.12	0.36	**	1.39	0.38	***	1.62	0.92	***			
Age (ln)		3.45	0.36	***	2.99	0.45	***	3.09	0.45	***	3.11	0.45	***
Sex	Men (ref.)				-	-		-	-		-	-	
	Women				0.34	0.25		0.38	0.25		0.22	0.24	
Ethnic origin	Switzerland (ref.)				-	-		-	-		-	-	
	Eastern Europe				-1.03	0.41	*	-1.06	0.42	*	-1.15	0.40	**
	South-western Europe				-0.89	0.42	*	-0.83	0.42	*	-0.95	0.40	*
	North-western Europe & northern America				1.23	0.42	**	1.26	0.42	**	1.22	0.41	**
	Other continents				0.34	0.39		0.31	0.39		0.49	0.37	
Labour market integration					0.62	0.30	*	0.65	0.30	*	0.55	0.30	+
Education	Compulsory education (ref.)				-	-		-	-		-	-	
	VET				1.01	0.77		0.92	0.77		0.93	0.77	
	Higher secondary education				0.21	0.84		0.14	0.84		0.03	0.83	
	Tertiary education				1.34	0.81	+	1.27	0.81		1.17	0.81	
Residency	Big centres (ref.)				-	-		-	-		-	-	
	Periurban & metropolitan centres				0.48	0.48		0.48	0.49		0.47	0.47	
	Touristic municipalities				0.38	0.66		0.47	0.65		0.54	0.63	
	Middle & little centres				0.33	0.31		0.33	0.32		0.40	0.31	
	Periurban & pendular municipalities				0.69	0.48		0.75	0.47		0.52	0.46	
	Outlying municipalities				0.66	0.38	+	0.66	0.38	+	0.63	0.37	+
Place of birth	Overrepresented places of birth (ref.)				-	-		-	-		-	-	
	Underrepresented places of birth				0.08	0.44		0.12	0.45		0.54	0.63	
Divorce								0.24	0.53		0.60	0.34	+
Siblings	No siblings (ref.)							-	-		-	-	
	Siblings							1.01	0.47	*	-0.11	0.29	
Nb obs.	5362												
Nb ind.	1025												
Nb events	85												
Deviance			652.2			598.79			593.42			619.98	
AIC			670.2			646.79			645.42			657.98	
BIC			669.56			645.09			643.58			660.50	

A New Tool for Old Questions:
The Sequence-Analysis Multistate Model
to Study Relationships Between
Time-Varying Covariates and
Trajectories.

Matthias Studer, Emanuela Struffolino and Anette E. Fasang

Abstract The relationship between processes and time-varying covariates is of central theoretical interest in many social sciences research questions. On the one hand, event history analysis has been the method of choice to study these relationships. However, it is limited to outcomes that can meaningfully be specified as simple instantaneous events or transitions. On the other hand, sequence analysis (SA) has made increasing inroads into the social sciences to analyze trajectories as holistic "process outcomes". However, it does not allow for studying their relationship with time-varying covariates.

We propose the sequence-analysis multistate model (SAMM) that combine the advantages of both approaches. SAMM models the relationship between time-varying covariates and trajectories specified as processes outcomes that unfold over time. it proceeds in two steps. First, we use an adapted sequence analysis to identify typical sequencing and spacing between main transitions in trajectories. Second, we adapt multistate models to estimate the chances to follow each kind of the identified typical sequence. The usefulness of SAMM is illustrated with an example from life course sociology on how 1) time-varying family status is associated with women's employment trajectories in East and West Germany, and 2) how the German reunification affected these trajectories in the two sub-societies.

Matthias Studer
Faculty of Social Sciences, University of Geneva, Switzerland, e-mail: `matthias.studer@unige.ch`

Emanuela Struffolino
Demography and Inequality Research Group, WZB Berlin Social Science Center, Germany

Anette E. Fasang
Demography and Inequality Research Group, WZB Berlin Social Science Center, Germany
Department of Social Sciences, Humboldt-University Berlin, Germany

1 Introduction

Many theoretical questions in the social sciences address the relationship between time-varying covariates and processes. Life course and career researchers are interested how changes in one life domain affect trajectories in another life domain, e.g. family and employment (Aisenbrey et al., 2009). How changing economic condition or family policies shape the transition to adulthood is also a central question in life-course researches (Shanahan, 2000).Social policy analysis is concerned with processes of policy development that can be altered by specific events, such as wars or a change of government (Abbott, 1995; Frank et al., 2000). Related research questions are at the center of historical comparative sociology and institutional analysis. Similarly, social movement scholars study how social movements unfold over time in response to trigger events (Minkoff, 1995; Olzak, 1989). Organizational ecology researchers examine how industries and organization develop over time. The relationship between these developmental processes and time-varying factors, for instance the introduction of new technologies, is of core theoretical interest in this field as well (Carroll et al., 1993). This list of selected examples could easily be extended to other sub-fields of sociology and related disciplines. They have in common that they are not only interested in processes of metric outcomes, such as income or IQ, but more often than not, in processes that consist of categorical states, including family trajectories, specific policy programs, or stages of group behavior until the outbreak of violent protest.

Two broad families of methodological strategies have been used to study the relationship between time-varying covariates and outcomes on the one hand, and trajectories of categorical states, on the other hand. The first strategy focuses on the occurrence of events or transitions (Allison, 1984; Yamaguchi, 1991; Therneau and Grambsch, 2000) and relies on event history analysis (EHA) for estimating the effect of time-varying covariates on the risk to observe an event. However, event history analysis is limited to modeling instantaneous changes and looses sight of the trajectory as a whole (Billari, 2005). The second strategy emphasizes the holistic nature of trajectories or processes of categorical states by relying on sequence analysis (SA) (Abbott, 1995; Studer and Ritschard, 2015). SA considers change and multiple transitions as lasting over longer time spans rather than being instantaneous. However, within the traditional sequence analysis framework it is not possible to study the relationship between time-varying covariates and trajectories.

We propose an original combination of these two approaches called Sequence Analysis Multistate Model (SAMM). Combining sequence analysis and multistate models, SAMM offers several advantages for studying processes. First, it allows for modeling the relationship between time-varying covariates and patterns of change within processes that unfold over longer periods of time. This methodological approach closely corresponds to the theoretical concepts of trajectories as "process outcomes" (Abbott, 2005). Second, studying trajectories holistically allows us to uncover potential interdependencies

between states and transitions within trajectories. The social meaning of a given situation often depends on previous but also later events, which may be known in advance by the actors involved. For instance, a woman may start a new job even (or because) she knows that it will be only temporary. Finally, SAMM can handle censored observations. This allows for the inclusion of trajectories that are only partially observed, which is not possible within the traditional sequence analysis framework.

The contribution of SAMM is demonstrated with an original example application in life course sociology. Two central theoretical principles in the life course paradigm include that individual life courses are multidimensional, e.g. family and employment, and that macro-structural, historical change strongly shapes individual life courses trajectories (Elder et al., 2003). We employ the historically unique social experiment of the German reunification to exemplify how SAMM can contribute to a better understanding of these two core life course principles. Specifically, we assess how 1) time-varying status in the family domain are associated with women's employment trajectories in East and West Germany (multidimensionality of lives), and 2) how the German reunification affected women's employment trajectories in the two sub-societies (impact of macro-structural change). In 1990, the former communist East Germany with a centrally planned economy (German Democratic Republic, GDR) was abruptly absorbed into the democracy and social market economy of the West (Federal Republic of Germany, FRG). The German reunification thereby provides an ideal case for studying how abrupt and profound macro-structural change was associated with change in individual life course trajectories that unfold over time. Beyond previous research on the impact of the German reunification (Bonin and Euwals, 2002; Hauschild, 2002; Trischler and Kistler, 2010; Klammer and Tillmann, 2001), our example application uses data for more recent cohorts that allows us to track difference and similarity in women's employment trajectories in East and West Germany not only in the immediate transition period, but until 20 years after the reunification.

References

Abbott, Andrew. 1995. "Sequence Analysis: New Methods for Old Ideas." *Annual Review of Sociology* 21:93–113.

Abbott, Andrew. 2005. "The Historicality of Individuals." *Social Science History* 29:1–13.

Aisenbrey, Silke, Marie Evertsson, and Daniela Grunow. 2009. "Is there a Career Penalty for Mothers' Time Out? A Comparison of Germany, Sweden and the United States." *Social Forces* 88:573–606.

Allison, Paul D. 1984. *Event history analysis, Regression for longitudinal event data*. QASS 46. Beverly Hills and London: Sage.

Billari, Francesco C. 2005. "Life Course Analysis: Two (Complementary) Cultures? Some Reflections With Examples From the Analysis of the Transition to Adulthood."

Advances in Life Course Research 10:261–281.

Bonin, Holger and Rob Euwals. 2002. "Participation Behavior of East German Women after German Unification." SSRN Scholarly Paper ID 297087, Social Science Research Network, Rochester, NY.

Carroll, Glenn R., Peter Preisendoerfer, Anand Swaminathan, and Gabriele Wiedenmayer. 1993. "Brewery and Brauerei: The Organizational Ecology of Brewing." *Organization Studies* 14:155–88.

Elder, Glen H., Monica Kirkpatrick Johnson, and Robert Crosnoe. 2003. "The Emergence and Development of Life Course Theory." In *Handbook of the Life Course*, edited by JeylanT. Mortimer and MichaelJ. Shanahan, Handbooks of Sociology and Social Research, pp. 3–19. Springer US.

Frank, David John, Ann Hironaka, and Evan Schofer. 2000. "The Nation-State and the Natural Environment Over the Twentieth Century." *American Sociological Review* 65:96–116.

Hauschild, Christine. 2002. "Die empirische Typisierung von Versichertenbiografien." *Deutsche Rentenversicherung* 57:539–589.

Klammer, Ute and Katja Tillmann. 2001. "Erwerbsbiografien als Mosaik-Längsschnittergebnisse zur Zusammensetzung von Erwerbsbiografien und ihren Veränderungen, zum Einkommenserwerb im Lebensverlauf und den Folgen für die Absicherung im Alter." *Flexicurity: Soziale Sicherung und Flexibilisierung der Arbeits-und Lebensverhältnisse, Düsseldorf: WSI und Hans Böckler Stiftung* pp. 141–223.

Minkoff, Debra C. 1995. *Organizing for Equality: The Evolution of Women's and Racial-Ethnic Organizations in America, 1955-1985.* Rutgers, NJ: Rutgers University Press.

Olzak, Susan. 1989. "Analysis of Events in the Study of Collective Action." *Annual Review of Sociology* 15:119–41.

Shanahan, Michael J. 2000. "Pathways to Adulthood in Changing Societies: Variability and Mechanisms in Life Course Perspective." *Annual Review of Sociology* 26:pp. 667–692.

Studer, Matthias and Gilbert Ritschard. 2015. "What matters in differences between life trajectories: A comparative review of sequence dissimilarity measures." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* Published ahead of print.

Therneau, Terry M. and Patricia M. Grambsch. 2000. *Modeling Survival Data: Extending the Cox Model*. Springer.

Trischler, F. and E. Kistler. 2010. "Gute Erwerbsbiographien–Arbeitspapier 1: Erwerbsverläufe im Wandel." *Bericht an die Hans-Böckler-Stiftung, Stadtbergen (im Internet)* .

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

Session 9A: Care

Physical occupational exposures and healthy life expectancy in a French occupational cohort

Loretta G. Platts[*,1]**, Jenny Head**[2]**, Sari Stenholm**[3,4]**, Holendro Singh Chungkham**[5]**, Marie Zins**[6,7] **& Marcel Goldberg**[6,7]

* Corresponding author.

[1] Stress Research Institute, Stockholm University, Stockholm, Sweden

[2] Department of Epidemiology and Public Health, University College London, London, UK

[3] Department of Public Health, University of Turku and Turku University Hospital, Finland

[4] School of Health Sciences, University of Tampere, Tampere, Finland

[5] Indian Statistical Institute, North-East Centre, Tezpur, India

[6] INSERM U1018, Centre for Research in Epidemiology and Population Health, Epidemiology of Occupational and Social Determinants of Health, F-94807 Villejuif, France

[7] Université de Versailles Saint-Quentin, UMRS 1018, F-94807 Villejuif, France

Abstract   To examine the relationships of strenuous and hazardous working conditions and night work with healthy life expectancy (HLE). The sample contained male gas and electricity workers from the French GAZEL cohort (n=13654). Six measures of physical working conditions were examined: Self-reports from 1989 and 1990 of ergonomic strain, physical danger, night work and perceived physical strain; company records of workplace accidents and a job-exposure matrix of chemical exposures. Partial HLEs (age 50–75) relating to 1) self-rated health and 2) chronic health conditions, obtained from annual questionnaires (1989–2014) and company records, were estimated using the multi-state life table method and microsimulation. The analyses were adjusted for social class and occupational grade. Participants who reported more physically demanding and dangerous work did not have shorter partial life expectancy but had shorter healthy life expectancies in terms of both chronic illness and self-rated health. No differences were observed in relation to night shift work. Strenuous and hazardous work may contribute to ill health in later life, which has implications for individuals' quality of life as well as healthcare use and labour market participation.

Taking Turns or Halving It All: Care Strategies of Dual-Caring Couples

Helen Eriksson

Abstract The birth of a child induces parents to substitute family work for paid work. As responses to child care time demands remain remarkably gendered, division of child care in relation to parents' work is key to understanding reproduction of gender inequality in both the family and the labor market. This study conceptualizes a 'care strategy' as a trajectory of time allocations by mothers and fathers to primary child care, i.e., care that requires absence from work, over the child's early years. It makes use of detailed data on claims for parental leave and sequence analysis to identify care strategies. Results show that half of all Swedish couples realize dual-caring strategies in which each parent serves as primary caregiver for substantial periods of the child's early life. Despite a uniquely flexible parental leave system that allows egalitarian couples to share care on a daily basis, the dominant dual-caring strategy consists of 'taking turns' in serving as primary caregiver where the mother takes leave to care for an initial period, followed by a period of solo-father care. One fourth of dual-caring couples, about 10 percent of all couples, use a 'halving it all' strategy in which primary care is shared in every point in time.

Helen Eriksson
Stockholm University
e-mail: helen.eriksson@sociology.su.se

Family structures and the organization of care for children in Italy

Sequences of time use by caregivers and activity

Tiziana Nazio

Abstract Using time diary data from Italy (pooled time use surveys 2002-03 and 2008-09), on a sample of around 10.000 households with children younger than 14 years of age, this paper assesses the amount of differences (in any) in the time mothers and fathers devote to unpaid childcare in differently shaped family structures (marital or cohabiting union, single parent and blended family), and how in turn family structure reflects in children's time use (addressing several activities from children's time diaries, as well as the overall amount of time children spend in presence of their parents or other caregivers). Time-use data obtained from daily diaries allow examining both the overall amounts spent on activities by day and, through sequence analysis, the patterns though which parents and children in different family structures engage in specific activities. Results show that family structure matters, in Italy, on how children time is organized around activities. Whereas children in cohabiting families seem to have, if at all, an advantage over children from married parents in Italy, those from single parents and blended families, with a larger recourse to more structuring of their time and activities out-side the family (more formal provision of education and courses) and stronger demands on their parents' time, organizational and economic resources, would largely benefit from the provision of high quality services. Parental education (i.e. a tertiary degree) plays a substantial role also in the provision of childcare, especially for the 'engaged' type and in the case of mothers' tertiary education.

Tiziano Nazio
Collegio Carlo Alberto & Dept. of Culture, Politics and Society, University of Turin
e-mail: tiziana.nazio@unito.it

Session 9B: Methods II

Missingness and truncation in sequence data: A non-self-identical missing state

Brendan Halpin

Abstract

Missingness in sequence data is a problem that has not received a great deal of attention. Since longitudinal data is more vulnerable to missingness than cross-sectional data, this is a lacuna. Two approaches are suggested in the literature: treating missingness as a separate state in its own right Gabadinho, Ritschard, Studer and Müller, 2009, and multiple imputation Halpin, 2015. The former is simple to implement but has shortcomings, not least in treating a missing–missing combination as a match; the latter is somewhat onerous. In this paper I propose a new concept: a non-self-identical missing state, where missing–missing combinations are treated as mismatches. I apply this to both normal missingness (gaps) and truncation of sequences (where entry or exit may happen at different times, yielding sequences of different length).

General missingness (gaps) and truncation are related but practically and conceptually different phenomena. If general missingness is treated as a special state, there is a risk that sequences will cluster according to how much missingness they contain.

Truncation is usually dealt with (if using the optimal matching distance measure) by deletion, but this may lead to sequences being disproportionately clustered according to length. In circumstances where sequences either begin or end on a specified event (e.g., begin with the first labour market participation spell, or end with retirement) the length of the sequence may be informative (e.g., late entry may be associated with higher education) and we may wish to differentiate between similarity represented in the length and similarity represented as matches between the observed portion. For instance, in both Halpin and Chan, 1998 and Bukodi, Goldthorpe, Halpin and Waller, 2016, class-career sequences are padded on the left by a "pre-entry" state to represent left-censoring, but (in the latter in particular) the con-

Brendan Halpin
University of Limerick, e-mail: brendan.halpin@ul.ie

cern is with how education affects later career: if we treat the pre-entry state as a non-self-identical missing, are the resulting distances more independent of sequence length than with OM-style deletion?

Using simulation and analysis of real data I show that a non-self-identical missing state can yield better results than the conventional approach for general missingness, but that locating a self-identical or non-self identical missing state in a maximally neutral location (as similar to all other states as is consistent with metricity) is also important. While multiple imputation functions better, it is more demanding in terms of setup and computation.

As regards truncation, OM's built-in capacity to compare sequences of different length is shown to work surprisingly well.

References

Bukodi, E., Goldthorpe, J. H., Halpin, B. & Waller, L. (2016). *Is education now class destiny? Class histories across three British birth cohorts*. Nuffield College, Oxford.

Gabadinho, A., Ritschard, G., Studer, M. & Müller, N. S. (2009). *Mining sequence data in R with the TraMineR package: a user's guide for version 1.2*. University of Geneva.

Halpin, B. (2015). *MICT: multiple imputation for categorical time-series* (Working Paper No. WP2015-02). Dept of Sociology, University of Limerick. Ireland. Retrieved from http://www.ul.ie/sociology/pubs/wp2015-02.pdf

Halpin, B. & Chan, T. W. (1998). Class careers as sequences: an optimal matching analysis of work-life histories. *European Sociological Review*, *14*(2).

Normalization of Distance and Similarity in Sequence Analysis

C. (Cees) H. Elzinga, M. (Matthias) Studer

Abstract We explore the relations between the notion of distance and a feature set based concept of similarity and show that this concept of similarity has a spatial interpretation that is complementary to distance: it is interpreted as "direction". Furthermore, we show how proper normalization leads to distances that can be directly interpreted as dissimilarity: closeness in normalized space implies and is implied by similarity of the same objects while remoteness implies and is implied by dissimilarity. Finally, we show how, in research into de-standardization of the life course, properly normalizing may drastically and unequivocally change our interpretation of inter-cohortal distances.[1]

Key Words: distance, similarity, normalization, sequence analysis.

C. (Cees) H. Elzinga
Netherlands Interdisciplinary Demographic Institute, The Hague,
VU University Amsterdam, the Netherlands e-mail: `c.h.elzinga@vu.nl`

M. (Matthias) Studer
Faculty of Social Sciences, University of Geneva
NCCR Program "LIVES", Switzerland e-mail: `matthias.studer@unige.ch`

2   C. (Cees) H. Elzinga, M. (Matthias) Studer

1 Introduction

In science, the concept of distance is used in two, quite different ways. First and oldest, in physics, distance or length is one of the fundamental dimensions that are used to express physical quantities (see e.g. Pfanzagl, 1968; Krantz et al., 1971). These quantities in turn are used to formulate models of natural phenomena. For example, the gravitational force that two masses exert on each other is inversely proportional to the squared distance between the masses.

In the social and behavioral sciences, spatial concepts are abundantly applied in different methodologies like multidimensional scaling (Borg and Groenen, 2005), multivariate linear statistics (Rencher and Cristensen, 2012), classification (Shawe-Taylor and Cristianini, 2004), clustering (Hennig et al., 2015) and sequence analysis (Blanchard et al., 2014). Spatial concepts are used in fields as diverse as clinical psychology, social demography and political science. However, in these sciences, distance is not one of the fundamental dimensions of theory. Instead, social scientists and psychologist calculate distances between personalities, political programs or labour market careers with the objective to judge or gauge the similarity between their objects of interest, eventually sorting or grouping these objects into more or less homogeneous subsets or projecting the objects onto a limited number of dimensions. In doing so, these colleagues assume that similarity is the opposite of distance: the more distant, the less similar and vice versa. This way of using distances is not unique for the social sciences; for example, chemists calculate distances between graph-representations of molecules (see e.g. Borgwardt, 2011) with the objective of finding similar but cheaper or more effective variants.

However, the relation between distance and similarity is not obvious as distance relations derive from spatial considerations and similarity relations derive from considering common and non-common feature sets (see e.g. Tversky, 1977). For a qualitative review of the concept of similarity, the reader is referred to Decock and Douven (2011).

So, grouping objects on the basis of distance may not result in groups of objects that are similar. Peeking around the corner at the issues dealt with in the next sections, we plotted distances versus similarities for the same pairs of objects in Figure 1. Indeed, distance and similarity are not simply anti-monotone and equal distance does not imply nor is implied by equal similarity. On the other hand, given a distance d, we can always calculate a similarity $s = h(d)$ and the reverse is also possible: given a similarity s, we can construct[2] a distance d from it through $s = h'(d)$. However, these calculations are not very intuitive. Fortunately, when we properly convert a distance d into a normalized distance D, we will obtain a normalized similarity S through the intuitive $S = 1 - D$. This latter transformation is the most simple and straightforward expression that formalizes the widely held beliefs that close objects are similar, that dissimilar objects are remote, etc. But apparently, we first have to normalize either the distance or the similarity before we can apply such an intuitive transform. In Figure 2 we illustrate the conceptual relations between (normalized) distance and similarity.

In this paper, we build on what was discussed in e.g. Chen et al. (2009); Elzinga et al. (2011) and in Elzinga (2014a). The discussion is quite general in the sense that it pertains to all measures of distance, whichever the context or application. However, we will focus on sequence analysis and to illustrate, we will use a publicly available data set and a measure of distance that has become popular in that context. For an introduction to sequence analysis, the reader is referred to Martin and Wiggins (2009) or to Cornwell (2015); an overview of distance measures used in this context is provided in

---

[2] It should be noted that $h' \neq h^{-1}$. It will appear that neither h nor h' is a 1-1 function: different distances may be mapped onto the same similarity and different similarities may be mapped onto the same distance.

Normalization of Distance and Similarity in Sequence Analysis   3

Fig. 1 Upper panel: scatter plot of distances (horizontal axis, metric: OMspell) versus similarities (vertical axis). Pearson's $r = -.29$. Lower panel: scatter plot of distances (horizontal axis) versus normalized distance. Pearson's $r = .75$. Darker areas contain more points. The data used in both panels were published in McVicar and Anyadike-Danes (2002).
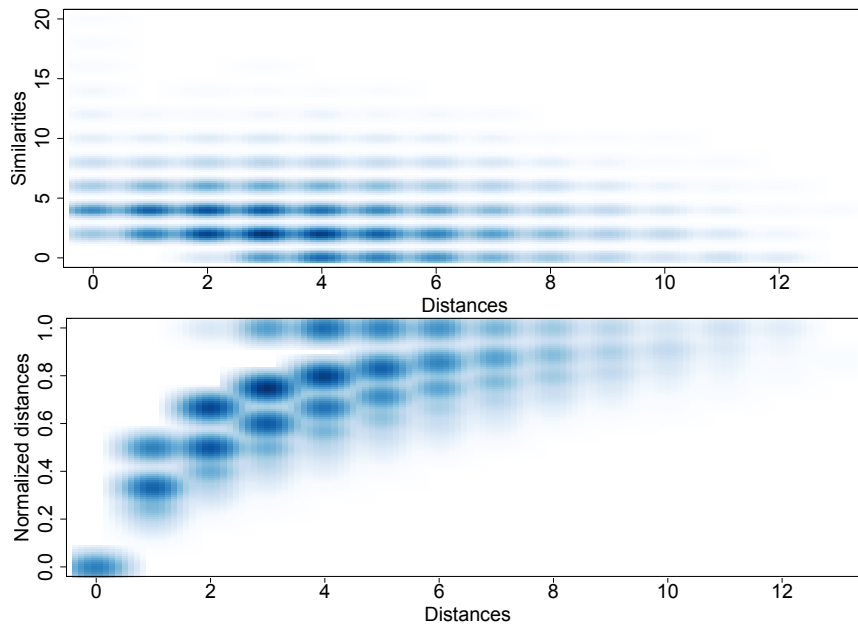
Studer and Ritschard (2015). For now, it suffices to define a sequence as an ordered set of labeled states or events that, depending on the specific application, may have an associated""time-stamp". In the case of states like "living single" or "being unemployed", the time-stamp is interpreted as "duration" and in case of events like "becoming a parent" or "starting parental leave", the time-stamp is interpreted as a date or amount of lapsed time.

 The purpose of this paper is threefold. First, we will explore the relations between distance as a spatial concept and similarity as a feature set based concept. We will conclude that the spatial interpretation of similarity is "direction" or angle, similar objects being located in (almost) the same direction, relative to some arbitrary but fixed reference object. This interpretation allows us to consider distance and similarity as complementary in the spatial structuring of a set of objects.

Second, we will discuss normalization as a means to transform the object space in such a way that distance and similarity become opposites in the sense that similar objects are close and remote objects are dissimilar, etc..

Third, we will show that normalization may affect the way that we interpret differences between life courses of different age-cohorts.

The paper is structured as follows: In the next Section 2, we discuss the concepts of distance and

Fig. 2 Formal relations between distance d, similarity s and their normalized versions D and S. Neither of the mappings f, g, h or h' is 1-1. Therefore, we do not write $h^{-1} = h'$.

$$S \;=\; 1 \;-\; D$$

$g : s \mapsto S$ $\qquad\qquad$ $f : d \mapsto D$

s $\qquad\qquad$ d

$h : d \mapsto s,$

$h' : s \mapsto d$
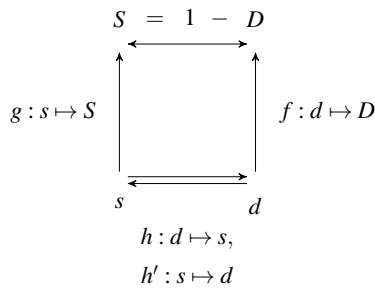
similarity and illustrate them by discussing classes of distance measures that have become popular in the context of sequence analysis. In its final subsection 2.4, we deal with the spatial interpretation of similarity through discussing transforms from distance to similarity and vice versa. In Section 3, we discuss normalization and in Section 4, we will apply normalization and discuss some of its effects. In Section 5, we make some concluding remarks.

2 Distance and Similarity

2.1 Distance

We briefly discuss the formal concept of distance. A function $d : X \times X \mapsto \mathbb{R}$ that maps pairs of objects from an object set X onto nonnegative (real) numbers is called a "distance" or, equivalently, a "(distance) metric" if it satisfies the following conditions or axioms for all triples of elements from the object set:

D1: $d(x,x) = 0$ $\qquad$ "one location per object"
D2: $d(x,y) > 0$ $\qquad$ "one object per location"
D3: $d(x,y) = d(y,x)$ $\qquad$ symmetry
D4: $d(x,y) \leq d(x,z) + d(z,y)$ triangular inequality (TI)

The first three axioms are quite intuitive to most readers so we only elaborate on D4, the triangular inequality. Axiom D4 says that if two objects x and y are close to a third object, say z, the objects cannot be far apart. At the same time, this is a way of saying that the space, i.e. the set of objects structured by d, cannot be irregular in the sense that all distances are confined to boundaries imposed by other distances. From the TI, it follows that, for all triples of objects (x,y,z), we have

Normalization of Distance and Similarity in Sequence Analysis 5

$$|d(x,z) - d(z,y)| \leq d(x,y) \leq d(x,z) + d(z,y). \tag{1}$$

Hence, the distance between a particular pair (x,y) is bounded by the distances $d(x,z)$ and $d(z,y)$ for *all* possible objects z in $X - \{x,y\}$ (see e.g. Elzinga and Studer, 2015). Without such boundaries, actually gauging distances between objects would not be very informative since, when presented with a new object, we would have no idea about its possible distance to the objects already known, i.e. d would not impose much structure on the object set.

The reader notes that the axioms D1-D4 only restrict the method of actually measuring distances, they do not specify or favour any method in particular; they just formalize our intuitions about a particular spatial relation.

2.2 Distances for Sequences

Here, we discuss two broad classes of distance measures that are often used in the context of sequence analysis. For a detailed overview of such measures, the reader is referred to Deza and Deza (2014); Studer (2012); Studer and Ritschard (2015).

The first class, and by far the most popular one in applications of sequence analysis, is the class of so-called edit-distances: an edit distance is a function that counts the minimum number of (weighted) edit-operations that is necessary to turn one sequence into a perfect copy of the other sequence (see e.g. Navarro, 2001). The smaller this number, the smaller the distance between the sequences. In the social sciences, edit-distances were introduced by Abbott and Forrest (1986) and are known as "OM-distances"; "OM" being the acronym for "Optimal Matching". Quite a variety of edit distances has been proposed, each variant defining a distinct way of weighing the edit-operations and the pairs of characters involved. Here, we will denote an edit-distance by writing d_E.

An alternative class of distance functions for sequences has been proposed by Elzinga and Studer (2015) and derives from representing sequences as non-negative, infinite-dimensional vectors and defines the distance between the sequences as the Euclidean distance between the vector-representations. Formally, let $X\{x,y,\ldots\}$ denote a set of sequences and let $\mathbf{X} = \{\mathbf{x}, \mathbf{y}, \ldots\}$ denote the set of vectors, representing the sequences in X. Furthermore, let $\mathbf{x}'\mathbf{y}$ denote an inner product of the vectors in $\mathbf{X}$. Then the distance d_K is defined (see e.g. Golub and Van Loan, 2013) as the Euclidean vector-distance

$$d_K(x,y) = \sqrt{\mathbf{x}'\mathbf{x} + \mathbf{y}'\mathbf{y} - 2\mathbf{x}'\mathbf{y}}. \tag{2}$$

Within this class, distance measures differ because of the different methods used to define the vectors and/or the inner vector-product. The algorithms required to evaluate the inner products are called "kernels" (see e.g. Shawe-Taylor and Cristianini, 2004) and hence, the distances obtained from such algorithms are subscripted as d_K.

2.3 Similarity

Similarity too is a function that maps pairs of objects to the real numbers but its properties do not derive from spatial relations. Instead, they derive from the general notion that similarity between objects is determined by the number of features that the objects share. A function $s : X \times X \mapsto \mathbb{R}$ that maps pairs[3] of objects to non-negative real numbers is a similarity if it satisfies, for *all* triples of objects $x, y, z \in X$:

S1	$s(x,y) \geq 0$	nonnegativity
S2	$s(x,y) \leq \min\{s(x,x), s(y,y)\}$	"bounded by self-similarity"
S3	$s(x,y) = s(y,x)$	symmetry
S4	$s(x,y) + s(z,z) \geq s(x,z) + s(z,y)$	covering inequality (CI)

Here, we elaborate on S2 and S4. If similarity somehow depends on sets of common and non-common features, S2 states that the set of features that x and y have in common cannot be bigger than the smallest of the feature sets of each of x and y. S4 is an inequality that, like TI, regularizes or smoothes the similarity space since similarities are bounded by other sums of similarities. Together, axioms S2 and S4 imply, for all triples $x, y, z \in X$, that $s(x,y)$ satisfies

$$\min\{s(x,x), s(y,y)\} \geq s(x,y) \geq s(x,z) + s(z,y) - s(z,z). \tag{3}$$

Two issues deserve some comment. The first pertains to "self-similarity" $s(x,x)$. Unlike D1: $d(x,x) = 0$, the system S1-S4 only limits $s(x,x)$ in S2 and allows for $0 < s(x,x) \neq s(y,y)$. At first sight, this may seem counter-intuitive: "x is more (or less) similar to itself than y". In spatial models of similarity (see e.g. Torgerson, 1965), this representational issue does not arise since $d(x,x) = 0$ for *all* objects. The system S1-S4 is based on the counting of feature sets and the result is the possibility of non-identical self-similarities: similarity depends on the set of common features. If however we interpret such counts as "description lengths" or "complexities" (Elzinga, 2010; Gabadinho et al., 2010), unequal self-similarities become quite natural and if x has more features than y, we have that $s(x,x) > s(y,y)$. We will come back on this in our discussion of the spatial interpretation of similarity (subsection 2.5) and the issue will be resolved when we discuss normalized similarity in subsection 3.2.

The second issue pertains to S3, the symmetry of similarity. We know that this axiom often fails when s is interpreted as *perceived* similarity, perhaps due to "context-switching" (see e.g Gärdenfors, 2000). Here we consider models of cognitive similarity as irrelevant although we are aware of the fact that the concept of similarity should not be fully decoupled from its role in our everyday experience.

For a more detailed account of the axiomatisation of the concept of similarity, the reader is referred to Chen et al. (2009) or Elzinga (2014a). We confine to discussing two simple similarities, one edit-based and one kernel-based.

---

[3] Similarity between more than two objects is dealt with in e.g. Elzinga et al. (2011)

Normalization of Distance and Similarity in Sequence Analysis 7

2.3.1 An edit-based similarity

The concept of similarity has been widely ignored among those applying sequence analysis, probably because of two reasons. First, most sequence analysis software generates distances, not similarities. Second, there is a generally held belief that remoteness implies dissimilarity and that closeness implies similarity. But despite this belief, it is not clear if and how similarity and distance are related. However, when we calculate the length ℓ of the longest common subsequence (lcs), then (Wagner and Fisher, 1974, Section 5)

$$\ell(x,y) = \frac{1}{2}(|x| + |y| - d_E(x,y)), \tag{4}$$

and $\ell(x,y)$ satisfies the similarity axioms S1-S4. Here, $|x|$ denotes the length of sequence x and $d_E(x,y)$ denotes the edit-distance based on unit insertion- and deletion-cost. It is almost immediate that ℓ satisfies axioms S1-S3 but it is not trivial to see that ℓ satisfies S4 too. A proof of this claim is presented in the Appendix 1. Hence, although there are many faster algorithms to calculate ℓ (see e.g. Bergroth et al., 2000, 2003), it can be considered as an edit-based similarity. Later we will generalize to the case of a metric edit-cost structure.

2.3.2 Kernel-based similarity

Kernel algorithms evaluate inner products of vectors, i.e. they evaluate $\mathbf{x}'\mathbf{y} = \sum_i x_i y_i$ for vectors $\mathbf{x} = (x_1, x_2, \ldots)$ and $\mathbf{y} = (y_1, y_2, \ldots)$ without directly using the vectors themselves. Thus, if we want to use kernels to define and/or evaluate similarities, we first have to represent objects by vectors, so called "feature-vectors". Let $\mathcal{F} = \{u, v, \ldots\}$ denote a set of features and when an object x possesses a feature u, we denote this fact by writing $u \sqsubseteq x$. To construct feature vectors, we first index the features by assigning to each and every feature u a unique positive integer $r(u)$ that is as small as possible. So, with n distinct features, the features will be indexed by the integers $1, \ldots, n$. Now we may construct a vector representation $x \mapsto \mathbf{x} = (x_1, x_2, \ldots)$ for an object x by assigning to each coordinate a non-negative value through

$$x_{r(u)} = \begin{cases} c(u) \text{ if } u \sqsubseteq x \\ 0 \text{ otherwise} \end{cases} \tag{5}$$

This in effect means that we assign a value $c(u)$ to the ith coordinate if the object x possesses the ith feature u and if x does not possess that feature, the coordinate is set to zero. If, for example $c(u) = 1$ for all features, the resulting binary vector shows which features are or are not possessed by the object represented. The inner product of such binary vectors is then a count of the features that the two objects have in common. If $c(u)$ varies with the features, the inner product evaluates a weighted count of common features. In all cases where the coordinate values *only* depend on the features, the resulting inner product $\mathbf{x}'\mathbf{y}$ is a similarity, i.e. $s(x,y) = \mathbf{x}'\mathbf{y}$ satisfies the similarity axioms S1-S4. In the context of sequence analysis, the set of features can be taken to be the set of all possible subsequences (see e.g. Elzinga and Studer, 2015), in which case an inner product counts the number of distinct common subsequences.

To actually calculate the inner product of feature vectors that use the possible subsequences as coordinates, kernel algorithms have been designed that allow for calculation times that are roughly proportional to the third power of the length of the sequences involve. These kernels have been described and analysed in e.g. Elzinga and Wang (2013); Elzinga and Studer (2015) and in Shawe-Taylor and Cristianini (2004). If however the coordinate values not only depend on the features as such, but also depend on the way these features are "embedded" in the pertaining sequences, i.e. when

$$\mathbf{x}_{r(u)} = \begin{cases} c(u,x) \text{ if } u \sqsubseteq x \\ \\ 0 \text{ otherwise} \end{cases}, \tag{6}$$

the inner products do not necessarily satisfy the axioms of similarity. For example, if $c(u,x)$ is a value that depends on the embedding frequency or the duration of u in x, $\mathbf{x}'\mathbf{y}$ will *not* satisfy the axioms of similarity[4]. Therefor, in the next subsection, we discus how to construct a similarity given a distance and vice versa.

2.4 From Distance to Similarity and Back

We already discussed a particular construction of a similarity from a distance: in Equation 4 , we transformed an edit-based distance d_E into an edit-based similarity $\ell(x,y)$. Unfortunately, there is no general and explicit solution to this problem, i.e. there is no (known) general specification of the map $h : d \mapsto s$, nor is there a general solution for the reverse problem - constructing a similarity from a distance. Some quite broad classes of solutions were derived by Chen et al. (2009) and some results of Yianilos (2002) are useful too. Here we discuss simple instances of these classes: one instance of $h(d) = s$ and one instance of $h'(s) = d$. For most applications, these examples will suffice.
First we discuss a solution to $s = h(d)$: let $X = \{x,y,z,r,\ldots\}$ denote an object set and d a metric on $X \times X$. Then, for an arbitrary but fixed reference object r,

$$s(x,y) = d(x,r) + d(y,r) - d(x,y) \tag{8}$$

is a similarity[5]. To see that this is true indeed, we check if the axioms S1-S4 hold.
Because of $d(x,y) \leq d(x,r) + d(y,r)$ (TI), $s(x,y) \geq 0$ so S1 holds. Using Inequality 1 and the above Equation 8, we derive

---

[4] As an example, we define $c(u,x)$ as the embedding frequency of the subsequence u of sequence x. Consider the sequences $x = aabab$ and $y = aabb$. Then, for example, we have that $c(ab,x) = 5$ and $c(abb,y) = 4 = c(abb,x)$. Now the inner product $\mathbf{x}'\mathbf{y}$ can be computed through

$$\mathbf{x}'\mathbf{y} = \sum_u c(u,x)c(u,y). \tag{7}$$

By constructing a table that lists all common subsequences of x and y and their embedding frequencies, the reader will discover that $\mathbf{x}'\mathbf{y} = 47$, that $\mathbf{x}'\mathbf{x} = 83$ and that $\mathbf{y}'\mathbf{y} = 35$. Taking the inner product for a similarity would violate axiom S2 since $\mathbf{x}'\mathbf{y} > \mathbf{y}'\mathbf{y}$

[5] We should take the reference object r into account in our notation by writing $s_r(\cdot,\cdot)$, but we do not since in this paper, the simpler notation does not lead to ambiguities.

Normalization of Distance and Similarity in Sequence Analysis   9

$$s(x,y) \leq d(x,r) + d(y,r) - |d(x,r) - d(y,r)| \tag{9}$$

$$= 2\min\{d(x,r), d(y,r)\} \tag{10}$$

$$= \min\{s(x,x), s(y,y)\}, \tag{11}$$

so $s(x,y)$ is bounded by self-similarity (S2). Symmetry is trivial (S3) and the CI (S4) follows from the TI. Apparently, s as constructed through Equation 8 is a similarity.

A way to construct a distance d from a similarity s is through (Chen et al., 2009)

$$d(x,y) = s(x,x) + s(y,y) - 2s(x,y). \tag{12}$$

d is a distance that depends on the size of the symmetric set difference $|\triangle(\mathcal{X}, \mathcal{Y})|$. That it is a proper distance is checked by verifying the distance axioms D1-D4. Clearly, $d(x,x) = 0$ and if $d(x,y) = 0$, we must have that $x = y$ because of S2. S2 also ensures that $d(x,y) > 0$ (D2) and symmetry (D3) is trivial. That d satisfies the TI immediately follows from the CI.

However, this formal reasoning does not clarify the conceptual relations between the spatial concept of distance and the set-based concept of similarity. Therefore, in the next subsection, we will provide for a spatial interpretation of similarity and show how the spatial and set-based concepts touch through binary feature vectors.

2.5 Similarity as angle or direction

To spatially interpret the concept of similarity, Equation 8 is a good starting point and to explore its behavior, we constructed Figure 3. Figure 3 shows a plane[6] in which we fixed two objects r and y at

Fig. 3 The similarity $s(x,y)$ is determined by the (fixed) distances $d(x,r) + d(y,r)$ and the *obtuse* angle φ between the lines $L(r,y)$ and $L(r,x)$. See text for a detailed explanation.

---
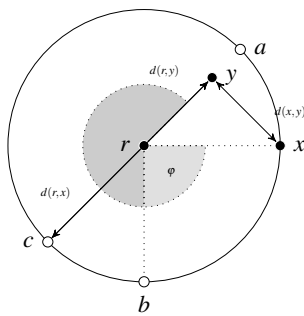
[6] not necessarily a coordinated plane

distance $d(r,y)$. Furthermore, we presumed a third object x at distance $d(r,x)$. This latter assumption determines a circle around r that includes x. Having drawn that circle, we fixed an x on it, therewith fixing $d(y,x)$ too. The picture then contains enough information to determine $s(x,y)$ according to Equation 8. Furthermore, we determined the *obtuse* angle φ between the lines $L(r,x)$ and $L(r,y)$. When we now move the object x along the circle in the direction of location b, φ will decrease and $s(x,y)$ decreases since $d(x,r)$ and $d(y,r)$ are fixed and $d(x,y)$ increases. Once x has reached location c, φ will attain its minimum-value: passing c will cause the obtuse and the acute angle to flip. Also, $d(x,y)$ will reach its maximum value and $s(x,y) = 0$ since, at c, $d(x,y) = d(x,r) + d(r,y)$. Moving x clockwise from c will increase the *obtuse* φ and, since $d(x,y)$ will get smaller, $s(x,y)$ will increase too. At a, φ will reach its maximum value and $s(x,y)$ will be maximal too.

So it seems that φ is monotone with $s(x,y)$: when $s(x,y)$ is maximal, φ is maximal and when $s(x,y)$ is minimal, then x and y are opposite relative to r and φ is minimal. However, in Figure 3, φ is anti-monotone with $d(x,y)$ but this is not necessary. In Figure 4, we fixed the objects x and y and move the reference relative to the line $L(x,y)$. We see that when the reference moves away from the line $L(x,y)$, the obtuse angle φ_i between the lines $L(r_i,x)$ and $L(r_i,y)$ gets bigger and, simultaneously, $s(x,y)$ gets bigger since $d(r_i,x)$ and $d(r_i,y)$ get bigger while $d(x,y)$ remains fixed.

From Figures 3 and 4, we conclude that $s(x,y)$ is monotone with the angle φ: the bigger it gets, the more similar the objects. Equivalently, we could say that x and y are similar to the degree that the *directions* of the lines $L(r,x)$ and $L(r,y)$ coincide. The difference $d(x,r_i) + d(y,r_i) - d(x,y) = s_{r_i}$

Fig. 4 The similarity $s_{r_i}(x,y)$ is measured as the distance from the reference r_i to the line L between x and y, i.e. as monotone with the obtuse angle between the lines $L(r_i,x)$ and $L(r_i,y)$. See text for a detailed explanation.
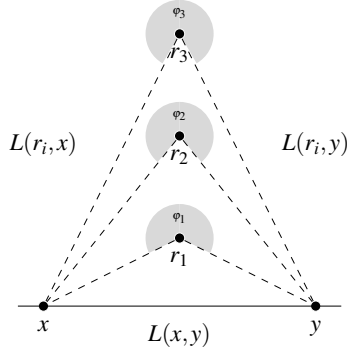

measures the distance of r_i to the line L between x and y. If this distance equals zero, x and y are opposite relative to r_i and hence $s_{r_i} = 0$. The greater the distance of r_i to L, the greater the similarity s_{r_i} the more the directions of x and y, relative to r_i coincide.

In the above interpretation, we did not assume anything about the properties of the object-space (X,d). However, when we assume that the space is a Euclidean coordinate space and we take $\mathbf{0} = (0,0,\dots)$ as our reference, Equation 8 implies that

$$s(x,y) = \|\mathbf{x}\| + \|\mathbf{y}\| - \|\mathbf{x} - \mathbf{y}\|, \tag{13}$$

where $\|x\|$ denotes the "length" of the vector $\mathbf{x}$, or, equivalently, the distance of $\mathbf{x}$ to the reference $\mathbf{0}$, i.e. for $\mathbf{x} = (x_1, x_2, \ldots)$, $\|\mathbf{x}\| = \sqrt{\sum_i x_i^2}$. Hence in a Euclidean vector space, $s(x, y)$ equals the sum of the lengths of $\mathbf{x}$ and $\mathbf{y}$ in as far as this sum is *not* due to the distance between $\mathbf{x}$ and $\mathbf{y}$ (see Figure 5). When

Fig. 5 Illustration of $s(x,y) = \|\mathbf{x}\| + \|\mathbf{y}\| - \|\mathbf{x} - \mathbf{y}\|$, which holds when sequences x and y are represented in Euclidean vector space and $\mathbf{0}$ is used as the reference object. In such cases, the similarity is interpreted as the sum of the lengths of the representations $\mathbf{x}$ and $\mathbf{y}$ in as far as this sum is not due to the distance between the two.
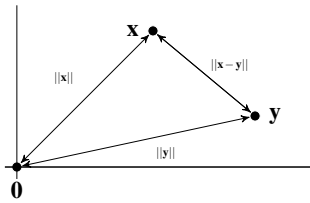
the vectors are binary feature vectors, the spatial and set-based interpretation of similarity coincide. For let $\{\mathcal{X}, \mathcal{Y}, \ldots\}$ denote the feature sets of the objects $\{x, y, \ldots\}$ and $\{\mathbf{x}, \mathbf{y}, \ldots\}$ their representing vectors, then we have that $\|\mathbf{x}\| = \sqrt{|\mathcal{X}|}$ and $\|\mathbf{x} - \mathbf{y}\| = \sqrt{|\triangle(\mathcal{X}, \mathcal{Y})|}$ with $\triangle(X, Y) = (X - Y) \cup (Y - X)$ and hence that

$$s(x, y) = \sqrt{|\mathcal{X}|} + \sqrt{|\mathcal{Y}|} - \sqrt{|\triangle(\mathcal{X}, \mathcal{Y})|}. \tag{14}$$

So, for binary feature vectors, the similarity defined by Equation 8 depends in a nonlinear way on the size of the set of features common to both x and y.

3 Normalizing Distance and Similarity

As will appear later, once we have constructed a normalized distance, denoted as D, the evaluation of a normalized similarity S simply amounts to evaluating the quantity $S = 1 - D$. So, once we have distances available and know how to normalize them, the calculation of similarity is easy.

The simple relation between normalized distances and similarities is the main, but not the only reason to discuss the subject of normalizing distances. The other reason is that normalizing greatly facilitates the interpretation of the numbers generated by the algorithm or procedure that generates the distance measurements.

3.1 The Normalization of Distance

Normalizing a distance(-scale) not only requires that we divide out units and map to a closed interval but also that the normalized scale still adheres to the metric axioms D1-D4. Simultaneously imposing these three requirements is not trivial.

A simple normalization[7] of distance is provided by the transform

$$D_r(x,y) = \frac{d(x,y)}{(d(x,y) + d(x,r) + d(y,r))/2}.$$

(15)

The choice of r determines the scale D_r. The effect of normalization can be understood through looking at the quantities $D_r(x,r)$: according to Equation 15, we have that $D_r(x,x) = 1$ since $d(r,r) = 0$. Because this holds for all $x \neq r$, normalization according to Equation 15 has the effect of projecting all objects other than r on a unit sphere with r at the center and gauging distances between the objects through evaluating distances between the projections. This geometrical interpretation is illustrated in Figure 6.

The effect of remoteness from the reference object can also be understood as follows: assume that both $d(x,r)$ and $d(y,r)$ are equal to, say, $\frac{1}{2}a$. Then $D_r(x,y)$ reduces to

$$D_r(x,y) = \frac{d(x,y)}{(d(x,y) + a)/2}.$$

(16)

In Figure 7, we plot this function for an arbitrary but fixed distance $d(x,y)$ and $a \geq d(x,y)$. The plot shows that the normalized distance D_r gets smaller, the more remote x and y are from r, i.e. the bigger a gets.

Fig. 6 Normalization amounts to projecting all objects on an r-centered unit-sphere. The reader notes that D_r may produce a different ordering of pairs of objects than d. Normalization is a way to control for variation among the distances to the reference r.

----
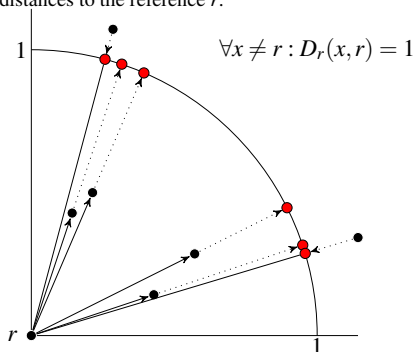
[7] In the sequel, we will write D instead of d to denote a normalized distance

Normalization of Distance and Similarity in Sequence Analysis

13

A proof that D_r satisfies the axioms D1-D4 is presented in the Appendix. Summarizing, we have that D_r satisfies the axioms of distance D1-D4 and $0 < D_r(x,y) \leq 1$ if and only if $x \neq y$.

Finally, we note that D_r is compressive with respect to d in the sense that differences between big d-distances tend to be smaller on the D_r-scale (see Figure 8).

Fig. 7 The effect of normalizing a fixed distance $d(x,y)$ with respect to a reference at a distance $\frac{1}{2}a$ from x and y: the bigger a gets, the smaller the normalized distance $D(x,y) = \frac{d(x,y)}{(d(x,y)+a)/2}$ (vertical axis) gets.
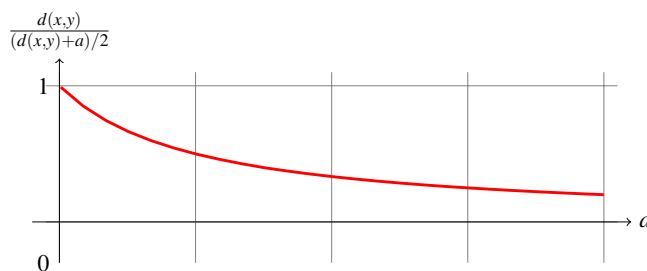
We now have discussed normalisation of distances, thereby confining ourselves to just one technique: the one embodied in Equation 15. The reader might be interested in different ways of normalizing distance and indeed, alternatives to Equation 15 can be found in Chen et al. (2009) or in Elzinga (2014a). The behavior of the resulting scales is roughly the same as the properties that we discussed here, so we leave it to the interested reader to explore these alternatives. We now turn our attention to the normalization of similarity.

3.2 The Normalization of Similarity

There are two ways to obtain a normalized similarity, the direct and the indirect way. The indirect route is via a normalized distance according to the simple formula

$$S_r(x,y) = 1 - D_r(x,y), \tag{17}$$

and the direct way uses "raw" or non-normalized similarity through one of the many transforms that yield a normalization. Here we start discussing the indirect way, i.e. via Equation 17, since it is simple and because it directly connects similarity to *any* normalized distance. Thereafter, we will just discuss just one of the many possibilities to normalize a similarity, a generalization of the Tanimoto-coefficient. Again, we confine ourselves to just one technique because we believe that discussing more different techniques will not be very productive for social science sequence analysts and again, the interested reader is referred to Elzinga (2014a) for different techniques and more

Fig. 8 Plot of D_r vs d, showing that normalizing is compressive. Because of the triangular inequality, the domain (horizontal axis) is limited to $d(x,r) + d(y,r)$.

$$D_r(x,y) = \frac{d(x,y)}{(d(x,y)+d(x,r)+d(y,r))/2}$$
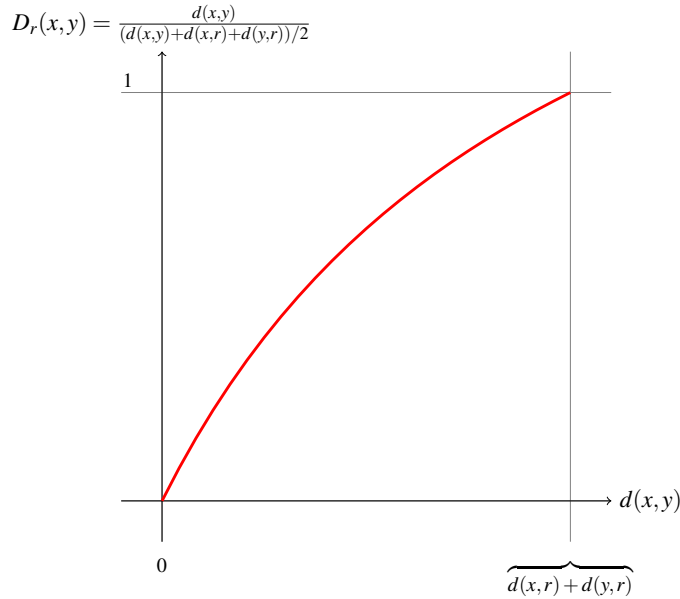
literature.

Using Equation 17, the axioms D'1-D'4 can be written in terms of S_r:

S'1 $S_r(x,x) = 1$ for all x,
S'2 $1 > S_r(x,y) \geq 0$ if and only if $x \neq y$,
S'3 $S_r(x,y) = S_r(y,x)$,
S'4 $S_r(x,y) + 1 \geq S_r(x,z) + S_r(z,y)$.

Comparing the system S'1-S'4 with the system S1-S4 reveals that S'1-S'4 is a special case of S1-S4: S2 states that a similarity can never exceed the self-similarity of the pertaining sequences while the same is implied in S'1-S'2. Only S'1-S'2 is more explicit through specifying a uniform upper boundary for all self-similarities and this is reflected in the difference between S'4 and S4. So we conclude that $S_r(x,y)$ as defined in Equation 17 is a normalized similarity: it adheres to the axioms of similarity and it is dimensionless and bounded in $[0,1]$ because D_r is dimensionless and bounded in the same interval.

The reader notes that, since $D_r(x,r) = 1$ for all $x \neq r$, $S_r(x,r) = 0$ for all $x \neq r$, again expressing that normalizing with respect to a reference implies "controlling for differences in distance/similarity to the reference object r".

Next we turn our attention to normalizing similarity without explicitly using a reference object. This

Normalization of Distance and Similarity in Sequence Analysis 15

is possible, given a non-normalized similarity *s*, through evaluating

$$S(x,y) = \frac{s(x,y)}{s(x,x) + s(y,y) - s(x,y)}. \tag{18}$$

This formulation generalizes Tanimoto's coefficient (see e.g. Duda et al., 2001) for similarity in biological taxonomy: if $s(x,y)$ is interpreted as a count of common features, $S(x,y)$ expresses this count as a fraction of the total number of features of *x* and *y*.
In Appendix A3, we show that *S* as defined in Equation 18 satisfies the Covering Inequality S'4. Normalizing similarity does not involve a reference object, at least not according to Equation 18. But Equation 18 does not say how the similarity *s* was constructed; if *s* does not directly derive from a (weighted) set count, *s* will be derived from a scale that derives from some spatial concept. But then the concept of "direction" will require a reference object: direction is always relative to some fixed point in space.
We discuss one example: we set $s(x,y) = llcs(x,y)$ and hence obtain

$$S(x,y) = \frac{llcs(x,y)}{|x| + |y| - llcs(x,y)} \tag{19}$$

since $llcs(x,x) = |x|$, the length of the sequence *x*. This formulation shows that normalized similarity can be computed in an edit-oriented sequence analysis. Using Equation 4, we elaborate the above expression to

$$S(x,y) = \frac{|x| + |y| - d_{OM}(x,y)}{|x| + |y| + d_{OM}(x,y)}, \tag{20}$$

expressing the normalized coefficient directly in terms of OM-distances. Distances, plural, since $|x| = d_{OM}(x,\lambda)$, the sum of unit deletion costs when transforming *x* into λ, the empty sequence. However, when a cost-metric other than the standard unit-cost is used, we have that $d_{OM}(x,\lambda) = \sum_i c(x_i,-)$ where $c(x_i,-)$ denotes the deletion cost of the *i*-th character of *x*. This leads to a generalization of *llcs* to the general OM-context with any metric cost-structure. Thereto, we define $\delta(x) = d_{OM}(x,\lambda)$ and introduce the concept of the "cost of a most expensive common subsequence", abbreviated as *cmcs*. Formally,

$$cmcs(x,y) = \max\{\delta(u) : u \sqsubseteq (x,y)\}. \tag{21}$$

Clearly, $cmcs(x,\lambda) = \delta(x)$ and $\delta(x) = |x|$ in case of unit-indel cost. Just like *llcs*, *cmcs* is a similarity and it can be computed through

$$cmcs(x,y) = \frac{1}{2}(\delta(x) + \delta(y) - d_{OM}(x,y)) \tag{22}$$

or through a variant of an algorithm to evaluate $llcs(x,y)$. To normalize *cmcs*, we apply

$$S(x,y) = \frac{cmcs(x,y)}{\delta(x) + \delta(y) - cmcs(x,y)}. \tag{23}$$

3.3 The use of normalization in sequence analysis

In the sequence analysis literature, little attention has been paid to normalization although it has been used by several authors. We mention a few examples.

In a paper on gendered trajectories on the labor market and household status, Levy et al. (2006) applied a form of normalization to compensate for the very different lengths of their sequences and thus to avoid clustering on the basis of the lengths of the trajectories: for each pair of trajectories, they they divided the distance by the length of the longest sequence:

$$d_E^*(x,y) = \frac{d_E(x,y)}{\max\{|x|,|y|\}}.$$ (24)

This is not a proper normalization since it does not map to $[0,1]$ and it does not satisfy TI. Furthermore, normalization should not be used to mask or compensate for unequal censoring of the trajectories: if sequences differ in length due to differences in censoring, there is a problem caused by missing data and that should be handled by appropriate imputation methods. Unfortunately, the frequently occurring problem of missing data has hardly received attention (but see Halpin, 2012).

Elzinga and Liefbroer (2007) and Bras et al. (2010) proposed to use

$$s_K^*(x,y) = \frac{\mathbf{x}'\mathbf{y}}{\sqrt{\mathbf{x}'\mathbf{x} \cdot \mathbf{y}'\mathbf{y}}}$$ (25)

to compare average similarities of family formation trajectories from different countries or different epochs. Indeed, this index is bounded by $[0,1]$ but it is not a proper similarity since it fails to satisfy the Covering Inequality (see Elzinga, 2014a).

Gabadinho et al. (2011) proposed to use

$$d^*(x,y) = \frac{d(x,y)}{d(x,r) + d(y,r)}$$ (26)

with $r = \lambda$. This maps to $[0,1]$ but it fails to satisfy TI unless d is a Euclidean distance (see Yianilos, 2002).

Finally, in a comparative study of distance metrics for sequence analysis, Ritschard et al. (2014) proposed to use Equation 15.

How can we use proper normalization in sequence analysis, other than for creating a scale that can be interpreted both as a similarity and as a distance? We see three possibilities.

First, normalization can be used when analyzing sequences while ignoring durations, i.e. analyzing the sequence of distinct subsequent subsequences (DSS). As Studer and Ritschard (2015) demonstrate, such analysis can be very useful when one wants to focus on order of states or events. However, such sequences will be of very different lengths. For example, a sequence on the labor market that just consists of uninterrupted unemployment then has only one element while a sequence that consists of an alteration of employment and unemployment may consist of many states and thus constitutes a (much) longer sequence. Normalization may be used to weigh the differences by the length of such DSS's by choosing the empty sequence as the reference object.

Second, normalization can be used to compare scales. For example, when we want to compare average distances or similarities between object from different regions or different epochs as was done, for example, in the study of Bras et al. (2010).

Finally, normalization can be used to focus on the deviations from a particular template sequence, e.g. uninterrupted employment or a career that uses the full length of legally facilitated parental leave (see e.g. Zhelyazkova, 2015). Using such a template will enlarge the various kinds of ways in which the objects differ from the reference. An alternative is to pick a medoid sequence[8] as the reference object, therewith simultaneously enlarging the various ways in which the sequences differ from this medoid. In the next subsection, we will demonstrate some of the effects mentioned. However, the reader should be aware that such effects, with different data, may not show up when the topological structure of the distance matrix is equal to or very close to the topological structure of the similarity matrix.

4 Balancing Distance and Similarity through Normalizing

Table 1 Average OM-distances, similarities (Equation 8) and normalized distances between the household trajectories of three different cohorts. The columns labeled with "se" show the corresponding bootstrapped standard errors of these quantities.

Cohortal Birthyear	Distance	se	Similarity	se	Normalized Distance	se
<1955	2.24	.052	4.90	.053	.43	.008
1955-1964	2.93	.045	4.95	.056	.51	.006
1965-1974	3.44	.047	5.21	.062	.54	.005

In this section we demonstrate how normalizing can be used to compare data from different cohorts. Thereto, we use data on family formation in The Netherlands as collected in the 2008-wave of the Family Formation Research Program (CBS, 2008) by Statistics Netherlands. The data pertain to the retrospective household histories of 5287 Dutch, born between 1945 and 1989. We encoded these histories using 8 different states: single and childless living in the parental home, living in the parental home in all other circumstances, living single, single with child, cohabiting with child(ren), cohabiting without child(ren), married with child(ren), married without child(ren). Ignoring durations, we used the OM-metric with unit-cost to calculate the distance between trajectories.

We distinguished three birth-cohorts as shown in the leftmost column of Table 1. If indeed de-standardization of the life course (see e.g. Brückner and Mayer, 2005) took place, one would expect that, on the average, distances between the household histories of younger Dutch are greater than the distances between the trajectories of the older ones. Table 1 shows the averages and bootstrapped

---

[8] A medoid is an observed object of which the sum of distances to all other sequences is minimal; thus a medoid may not be unique - the centroid may not have been observed or even may not be an observable object. On the other hand, given a full distance matrix, the distances to the centroid are always computable (see Elzinga et al., 2011, section 5.3).

standard errors[9] of these estimates. We also calculated similarities (according to Equation 8) between the household trajectories, the averages of which are also shown in the middle column of Table 1. According to the same hypothesis of de-standardization, one would expect that average similarity would increase with cohortal age.

Surprisingly, Table 1 shows that both the average distance and the average similarity decrease with cohortal age. One would expect an *in*crease of similarity but a *de*crease of distance. How can we explain these results? How can we explain their apparent contradiction? Stated in terms of common and non-common features, the results show that both, the number of common and the number of non-common features increase with younger cohorts, implying that the *total* number of feature also increased. In other words, we observe an increase in the number of common features (revealed by the increase of similarity) in a general context of an increasing number of features. This result is interesting. The de-standardization could have been the result of the emergence of new phenomena like unmarried cohabitation with or without children and/or because divorce has become much more common than it was in the past (Shanahan, 2000). These phenomena apparently have resulted in longer, more complex sequences, with, on the average, more features. Indeed, the average complexity index (see e.g. Gabadinho et al., 2010) of the sequences significantly increases (figures not shown) for younger cohorts.

By normalizing the distances according to Equation 15 and using the empty sequence as the reference, we take the total number of features and therefore the complexity of the sequences into account. Average normalized distance measures the evolution of the number of non-common features, while taking into account that the maximum (potential) number of features, which is not necessarily the same in every context. The results are shown in the rightmost column of Table 1. Now it appears that the average normalized distances, and thus the normalized similarities, again are significantly different in the direction predicted by the assumption of de-standardization. This implies that the increase of the number of common features was less important than the increase of the number of non-common features.

5 Conclusions

In the previous sections, we have explored the relation between the concepts of distance and similarity and shown that adopting the axiomatic definition of similarity as presented here, leads to a spatial interpretation of similarity as "direction", complementary to distance.

We also discussed normalisation as a means to create distances that may directly be interpreted as similarities in the sense that similarity is linearly anti-monotone with distance: the more remote, the less similar and vice versa.

---

[9] The standard errors have been computed as follows: We took 500 samples (with replacement) of 1000 individuals each and for each of these samples, we computed the average distance. The standard deviation of these estimates is the standard error reported here. By sampling individuals, the estimated se's are bigger than when sampling distances from the original distance matrix since therein, the distances are correlated; sampling individuals instead of distances avoids this problem. The time required for this procedure is limited to the sampling time as the required distances, given the sampled individuals, can be taken from the original distance matrix. The computation time for this procedure is negligible. For an introduction to bootstrapping techniques, the reader is referred to Davison and Hinkley (1997).

Now there are a number of questions that can be put up. First of all the question of why we need this axiomatic concept of similarity at all. And if we really need this concept of similarity, does it really matter, wether or not we apply it, will it lead to new interpretations and new insights about our data? Below, we will discuss these matters one by one.

In research into choice behavior and decision making, similarity of (sets of) stimuli has long been dealt with through spatial representation of the objects and analyzing distances as if these distances were actually gauging similarity. This long tradition originates from the seminal work of Torgerson (1965), Guttman (1968) and Lingoes (1968) on multidimensional scaling. It rests on the idea that, given the multidimensional representation is precise enough, the number of significant dimensions is limited and the dimensions can be assigned a rigorous substantive interpretation in terms of meaningful properties of the stimulus- or item-set, similarity and (lack of) distance coincide. Some even considered the Minkowski-metrics, in particular for the exponent values 1, 2 and ∞, as formalisations of psychological composition rules for the generation of preferences, similarity judgments and item-correlations, to mention only a few of the variety of object sets and applications studied in this tradition.

However, in the context of sequence analysis, the dimensions of the spatial representations created through applying an edit-based or kernel-based algorithm are hard to isolate. Edit-distances are not defined through an equation that relates to a notion of dimension. Kernel-based algorithms evaluate inner-products in high-dimensional vector-spaces where each and every feature, e.g. a particular subsequence, defines its own dimension. In general, it does not make much sense to try to substantively and separately interpret these dimensions. Therefore, directly interpreting distance as similarity is hard to justify in the context of sequence analysis applications.

The alternative is a feature-based approach and it would have been appealing to adopt an axiomatisation like Tversky's. However, Tversky's axioms allow for a asymmetric similarity, therewith cutting the possibility of a structural bond between distance and similarity. Therefore we adopted an approach embodied in the axioms S1-S4: it retains Tversky's matching condition, stating that similarity is a function of both common and non-common feature sets and it has a direct connection to spatial representations. However, we do not claim that the axioms S1-S4 constitute a unique solution to the problem. Perhaps a relaxation of the system, for example by replacing S2 with the weaker $s(x,y) \leq \max\{s(x,x), s(y,y)\}$, would be interesting in some applications. Here, we will not pursue such an alternative.

Several researchers have published interesting results despite the fact that they used distance measures or normalizations that (potentially) violate the triangular inequality. Understandably, the question has been put up why we need such formal restrictions at all when, without them, we seem to have interesting results too. Some (e.g. Lesnard, 2006; Gauthier, 2015) even doubt that a formal, axiomatic approach is necessary at all. We believe that an axiomatic approach helps us to better understand the methods that we apply, either to explore data structures for patterns or to test hypothesis about these patterns. Axioms like the triangular inequality (D4) or the covering inequality (S4) allow us to map relations onto real or rational numbers and assure that the spatial structure is "smooth" in the sense that new data points are restricted by the ones that we already observed (Elzinga and Studer, 2015). Applying methods that violate such axioms not necessarily are invalid but we cannot trust them to be generalizable beyond the data collected and the representation constructed from them.

Appendix 1

Here, we are concerned with the length of the longest common subsequence as a similarity in the sense of the axioms S1-S4. A set of sequences may have many distinct lcs's, see e.g. Elzinga (2014b), but their length is a unique integer. Let us write $\mathcal{L}(x,y,\ldots)$ to denote any lcs of the sequences $\{x,y,\ldots\}$ and let $\ell(x,y,\ldots)$ denote the length of such an lcs. Here we prove that ℓ satisfies the covering inequality S4.

To prove this, we write S4, using ℓ as

$$\ell(x,y)+\ell(y,z) \leq \ell(x,z)+\ell(y,y) \tag{27}$$

and derive its correctness. First, we consider the right hand side of Inequality 27. Clearly, we have that $\ell(y,y) = |y|$. Next, we write $\ell_y(x,z)$ for the length of the longest common subsequence $\mathcal{L}_y(x,z)$ that only consists of elements that are common to y. Similarly, we write $\mathcal{L}_{\bar{y}}(x,z)$ for a longest common subsequence that does not contain any character that is common to y and we let $\ell_{\bar{y}}(x,z)$ denote its length. Since and $\mathcal{L}_y(x,z)$ and an $\mathcal{L}_{\bar{y}}(x,z)$ have no common subsequences, we must have that $\ell(x,z) = \ell_y(x,z) + \ell_{\bar{y}}(x,z) = \ell(x,y,z) + \ell_{\bar{y}}(x,z)$. Hence the right side of Inequality 27 is equivalent to

$$\ell(x,y,z)+\ell_{\bar{y}}(x,z) + |y|. \tag{28}$$

Next, we consider the left had side of Inequality 27: we have that $\ell(x,y) = \ell_z(x,y) + \ell_{\bar{z}}(x,y)$ and $\ell(y,z) = \ell_x(y,z) + \ell_{\bar{x}}(y,z)$ and we know that $\ell_z(x,y) + \ell_x(z,y) \leq \ell(x,y,z)$ and that $\ell_{\bar{z}}(x,y) + \ell_{\bar{x}}(z,y) \leq \ell(y,y)$. Therefore, the right hand side of Inequality 27 cannot exceed its right hand side and this proves the claim.

Appendix 2

To see that D_r as defined in Equation 15 is a distance, we have to investigate whether or not it satisfies the axioms D1-D4. Here we confine ourselves to showing that D_r satisfies the triangular inequality D4. We first present the proof and then comment:

$$\begin{aligned} D_r(x,y) &= \frac{d(x,y)}{(d(x,y)+d(r,x)+d(r,y))/2} \\ &\leq \frac{d(x,z)+d(z,y)}{(d(x,z)+d(z,y)+d(r,x)+d(r,y))/2} \tag{29} \\ &\leq \frac{d(x,z)}{(d(x,z)+d(r,x)+d(r,y))/2} + \frac{d(z,y)}{(d(z,y)+d(r,x)+d(r,y))/2} \tag{30} \\ &= D_r(x,z)+D_r(z,y) \end{aligned}$$

In inequality 29, we replaced $d(x,y) = a$ by one of its upper bounds: $d(x,z)+d(z,y) = b$. Writing $c = d(r,x)+d(r,y)$, inequality 29 states $\frac{a}{a+c} \leq \frac{b}{b+c}$ and we know this to be true for all positive a, b

Normalization of Distance and Similarity in Sequence Analysis

21

and c with $a \leq b$.

In the next step, we split the sum of ratio's and then remove a positive term from each of the denominators: this must yield bigger ratio's and hence inequality 30 is correct.

The reader notes that we used the premiss that d is a distance, i.e. that d satisfies the triangular inequality in generating inequality 29.

Appendix 3

To prove that S as defined in Equation 18 satisfies S'4, we use an equivalent of S4: $s(x,y) \geq s(x,z) + s(z,y) - s(z,z)$.

Using Equation 18 in the left hand side of S'4, we write

$$\frac{s(x,y)}{s(x,x) + s(y,y) - s(x,y)} + 1$$

$$\geq \frac{s(x,z) + s(z,y) - s(z,z)}{s(x,x) + s(y,y) - (s(x,z) + s(z,y) - s(z,z))} + 1 \tag{31}$$

$$= \frac{s(x,x) + s(y,y)}{s(x,x) + s(y,y) - (s(x,z) + s(z,y) - s(z,z))} \tag{32}$$

$$= \frac{s(x,z) + s(y,y) - s(x,z) + s(z,y) + s(x,x) - s(z,y)}{s(x,x) + s(y,y) - s(x,z) - s(z,y) + s(z,z)} \tag{33}$$

$$= \frac{s(x,z) + s(y,y) - s(z,y)}{s(x,x) + s(y,y) - s(x,z) - s(z,y) + s(z,z)}$$

$$+ \frac{s(z,y) + s(x,x) - s(x,z)}{s(x,x) + s(y,y) - s(x,z) - s(z,y) + s(z,z)} \tag{34}$$

$$\geq \frac{s(x,z)}{s(x,x) + s(z,z) - s(x,z)} + \frac{s(z,y)}{s(z,z) + s(y,y) - s(z,y)} \tag{35}$$

$$= S(x,z) + S(z,y),$$

as required.

We concisely comment on the above steps as follows: In going from Equation 31 to 31, we used S4 to replace $s(x,y)$ by the smaller quantity $s(x,z) + s(z,y) - s(z,z)$. In Equation 33 we added $s(x,z) - s(x,z)$ and $s(z,y) - s(z,y)$ to the numerator of Equation 32 and in Equation 34, we split the ratio of Equation 33 into two appropriately chosen ratio's. Then, in passing from Equation 34 to 35, we used the general principle that, for nonnegative numbers p, q and a with $q > p$, we have that $\frac{p+a}{q+a} \geq \frac{p}{q}$.

References

Abbott, A. and Forrest, J. (1986). Optimal matching methods for historical sequences. *Journal of Interdisciplinary History*, XVI(3):471–494.

Bergroth, L., Hakonen, H., and Raita, T. (2000). A survey of longest common subsequence algorithms. In *Proceedings of the Seventh International Symposium on String Processing and Information Retrieval (SPIRE'00)*, pages 39–48. IEEE Computer Society.

Bergroth, L., Hakonen, H., and Väisänen, J. (2003). New refinement techniques for longest common subsequence algorithms. In Nascimento, M. A., Silva de Moura, E., and Oliveira, A. L., editors, *String Processing and Information Retrieval, 10th International Symposium,SPIRE 2003*, pages 287–303. Springer, New York.

Blanchard, P., Bühlmann, F., and Gauthier, J.-A. (2014). *Advances in Sequence Analysis: Theory, Method, Applications*. Life Course Research and Social Policies. Springer, New York.

Borg, I. and Groenen, P. J. F. (2005). *Modern Multidimensional Scaling: Theory and Applications*. Springer Series in Statistics. Springer, 2 edition.

Borgwardt (2011). Kernel methods in bioinformatics. In Horng-Shing Lu, H., Schölkopf, B., and Zhao, H., editors, *Handbook of Statistical Bioinformatics*, Springer Handbooks of Computational Statistics, chapter 15, pages 317–334. Springer, New York.

Bras, H., Liefbroer, A. C., and Elzinga, C. H. (2010). Standardization of pathways to adulthood? An analysis of Dutch cohorts born between 1850 and 1900. *Demography*, 47(4):1013–1034.

Brückner, H. and Mayer, K.-U. (2005). De-standardization of the life course: What might it mean? and if it means anything, whether it actually took place. In *The Structure of the Life Course: Standardized? Individualized? Differentiated?(Advances in Life Course Research, Volume 9)*, pages 27–53. Elsevier, Amsterdam.

CBS (2008). *Onderzoek Gezinsvorming 2008*. Data Archiving and Network Services (DANS, https://easy.dans.knaw.nl/ui/datasets/id/easy-dataset:39585), The Hague.

Chen, S., Ma, B., and Zhang, K. (2009). On the similarity metric and the distance metric. *Theoretical Computer Science*, 410(24-25):2365–2376.

Cornwell, B. (2015). *Social Sequence Analysis: Methods and Applications*. Structural Analysis in the Social Sciences. Cambridge University Press, New York.

Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap Methods and their Applications*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge.

Decock, L. and Douven, I. (2011). Similarity after goodman. *Review of Philosophy and Psychology*, 2(1):61–75.

Deza, M. M. and Deza, E. (2014). *Encyclopedia of Distances*. Springer, Berlin, 3rd edition.

Duda, R. O., Hart, P. E., and Stork, D. G. (2001). *Pattern Classification*. Wiley, Chichester (UK).

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

Elzinga, C. H. (2014a). Distance, similarity and sequence comparison. In Blanchard, P., Bühlmann, F., and Gauthier, J.-A., editors, *Advances in Sequence Analysis: Theory, Method, Applications*, number 2 in Life Course Research and Social Policies, chapter 4, pages 51–73. Springer, New York.

Elzinga, C. H. (2014b). Sequence A152072. The On-Line Encyclopedia of Integer Sequences (2014), published electronically at http://oeis.org.

Elzinga, C. H. and Liefbroer, A. C. (2007). De-standardization and differentiation of family life trajectories of young adults: A cross-national comparison using sequence analysis. *European Journal of Population*, 23(3-4):225–250.

Elzinga, C. H. and Studer, M. (2015). Spell sequences, state proximities and distance metrics. *Sociological Methods & Research*, 44(1):3–47.

Elzinga, C. H. and Wang, H. (2013). Versatile string kernels. *Theoretical Computer Science*, 495:50–65.

Elzinga, C. H., Wang, H., Lin, Z., and Kumar, Y. (2011). Concordance and consensus. *Information Sciences*, 181:2529–2549.

Gabadinho, A., Ritschard, G., Matthias, S., and Müller, N. M. (2010). Indice de complexité pour le tri et la comparaison de séquences catégorielles. *Revue des Nouvelles Technologies de l'Information*, E-19:61–66.

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

Gärdenfors (2000). *Conceptual Spaces: The geometry of thought.* MIT Press.

Gauthier, J.-A. (2015). Comment: How to make a long story short. *Sociological Methodology*, 45(1):1–3.

Golub, G. H. and Van Loan, C. H. (2013). *Matrix Computations*. Johns Hopkins Studies in the Mathematical Sciences. The Johns Hopkins University Press, Baltimore, 4th edition.

Guttman, L. (1968). A general nonmetric technique for finding the smallest coordinate space for a configuration of points. *Psychometrika*, 33:341–353.

Halpin, B. (2012). Multiple imputation for life-course sequence data. Technical report, University of Limerick, Dept. of Sociology, http://hdl.handle.net/10344/3639.

Hennig, C. M., Meila, M., Murtagh, F., and Rocci (Eds.), R. (2015). Handbook of cluster analysis. Chapman & Hall, London.

Krantz, D. H., Luce, R. D., Suppes, P., and Tversky, A. (1971). *Foundations of Measurement Volume I: Additive and Polynomial Representations*. Academic Press, San Diego.

Lesnard, L. (2006). Optimal matching and social sciences. Technical Report 01, Institut National de la Statistique et des Etudes Economiques.

Levy, R., Gauthier, J.-A., and Widmer, E. (2006). Entre contraintes institutionelle et domestique: les parcours de vie masculins et féminins en Suisse. *Canadian Journal of Sociology*, 31(4):71–92.

Lingoes, J. C. (1968). The multivariate analysis of qualitative data. *Multivariate Behavioral Research*, 3(1):61–94.

Martin, P. and Wiggins, R. D. (2009). Optimal matching analysis. In Wiliams, M. and Vogt, W. P., editors, *Sage Handbook of Methodological Innovations*. Sage.

McVicar, D. and Anyadike-Danes, M. (2002). Predicting successful and unsuccessful transitions from school to work by using sequence methods. *Journal of the Royal Statistical Society. Series A*, 165(2):317–334.

Navarro, G. (2001). A guided tour to approximate string matching. *ACM Computing Surveys*, 33(1):31–88.

Pfanzagl, J. (1968). *Theory of measurement*. Wiley, Oxford, UK.

Rencher, A. C. and Cristensen, W. F. (2012). *Methods of Multivariate Analysis*. Wiley Series in Probability and Statistics. Wiley, New York.

Ritschard, G., Bürgin, R., and Studer, M. (2014). Exploratory mining of life event histories. In McArdle, J. and Ritschard, G., editors, *Contemporary Issues in Exploratory Data Mining in the*

Social Sciences, Quantitative Methodology Series, chapter 9, pages 221–254. Routledge, New York.

Shanahan, M. J. (2000). Pathways to adulthood: Variability and mechanisms in life course perspective. *Annual Review of Sociology*, 26:667–692.

Shawe-Taylor, J. and Cristianini, N. (2004). *Kernel Methods for Pattern Recognition*. Cambridge University Press, Cambridge.

Studer, M. (2012). *Analyse de données séquentielles et application à l'étude des inégalités sociales en début de carrière académique*. PhD thesis, Faculté des sciences économiques et sociales de l'Université de Genève http://archive-ouverte.unige.ch/vital/access/manager/Repository/unige:22054.

Studer, M. and Ritschard, G. (2015). What matters in differences between life trajectories: A comparative review of sequence dissimilarity measures. *Journal of the Royal Statistical Society Series A*. published ahead of print.

Torgerson, W. S. (1965). Multidimensional scaling of similarity. *Psychometrika*, 30(4):379–393.

Tversky, A. (1977). Features of similarity. *Psychological Review*, 84(4):327–352.

Wagner, R. A. and Fisher, M. J. (1974). The string-to-string correction problem. *Journal of the ACM*, 21(1):168–173.

Yianilos, P. N. (2002). Normalized forms for two common metrics. Technical Report 91-082-3-9027-1 Rev 7/7/2002, NEC Research Institute, Princeton, NJ.

Zhelyazkova, N. (2015). *Work-Family Reconciliation and Use of Parental Leave in Luxembourg. Empirical Analysis of Administrative Records*. PhD thesis, Maastricht University.

Acknowledgements

Do State Policies Generate Different Life Courses?

An Empirical Study of the Case of the Two Germanys via a Statistical Assessment

Tim F. Liao, University of Illinois

Anette E. Fasang, Humboldt University Berlin & WZB

Abstract

1. Introduction

The life course is informative for social scientists to study because we can see the effects of individual characteristics such as education and occupation.[1] For sociologists, it is even more fascinating to analyze life courses because the state also impacts and structures them (Mayer 1989). The state may impact the life course in various ways: It may affect individual life courses by social structure such as educational and occupational systems (Shanahan 2000); it may directly influence life courses through institutional regulations and social policies (Mayer 2009; Neyer 2013).

As a quasi-experiment, the German reunification in 1990 offers an exceptional event for studying the imprints of the state on individual life courses (e.g. Goldstein and Kreyenfeld; Schnettler and Klüsener 2014). In this paper, we aim to formally assess how East and West German life courses differ by analyzing life course sequence data from a recent national survey (NEPS) with a retrospective history component, along the lines of life course research focused on contexts specified according to space and time (Mayer 2004). Specifically, we are interested in how Germans who were born and grew up in the former East and West Germanys lived

1

different life courses in both the family and the employment dimensions even though they now
live in the reunified Germany.

Life course and employment history observations often provide sequence data, a type of
data more complex than the observation of a single aspect of the life course (e.g., Blanchard,
Bühlmann, and Gauthier 2014; Gabadinho et al 2011). The timing and the duration of a life
course event, such as the timing of first birth or the duration between completing education and
first labor market entry, are examples of such single aspects. Statistical assessments of timing
and duration variables are straightforward because mean timing and durations can be compared
across samples, and their expected values can be used in further analysis.

With sequence data, however, there is no obvious single aspect to construct a variable for
such statistical assessments. A life course sequence contains information on the timing and the
duration of various life course events as well as on their sequence order (Brückner and Mayer
2005). Therefore, in this paper, we have twin objectives. In addition to the substantive goal
above, we propose a medoid-based method for formally assessing the difference between sets of
life course sequences, where the medoid is defined as the sequence in a given set of sequences
that has the minimum sum of distances to all other sequences in the set. Using this method, we
compute the Bayesian Information Criterion (BIC) and the Likelihood Ratio Test (LRT) statistic
in the context of sequence comparisons. While the LRT is a common statistical test in many
disciplines, the BIC has become a popular method for comparing and selecting models in
sociology since Raftery's (1986) introduction into our discipline.

In the following, we first review the literature on the life course in the East and West
Germanys, especially how their social policies and institutional regulations may differ. We then
present the medoid-based method for analyzing sequence differences. The method is set up for

2

the construction of the BIC and the LRT statistics for comparing sequence data. We then apply

the method to the data from the 2009/2010 National Education Panel (NEPS), cohorts 1955-1965,

to both illustrate the efficacy of the method and to assess the difference between East and West

German life courses. We conclude by drawing some general conclusions about how the East and

West German life courses differ initially and how they converge over cohorts and about the

appropriateness of the proposed method.

2. Social Policies and Life Courses in East and West Germanys

Between 1949 and 1990 Germany was divided into two sub-societies with notable

contextual differences in the welfare state, economic system and family policies between the

communist German Democratic Republic (GDR) in the East and the democratic social market

economy in the Federal Republic of Germany (FRG) in the West.

The GDR had a state socialist system, and a centrally planned economy. The regulative

regime aimed at reconciling the political priorities of economic and population growth. To this

end women were expected to work equal to men in a dual earner model and the family ideology

was strongly *pro-natalist* (Engelhardt, Trappe & Dronkers, 2002). Family policies in the GDR

conditioned access to state-controlled resources, such as housing, on marriage and parenthood to

incentivize fertility. Normative pressure to have children in the early 20s was coupled with

generous financial incentives for parenthood and practically universal public child care that

widely enabled mothers' employment. As a result female labour market participation was around

90% (Huinink et al., 1995). Employment was practically universal but wages were rather low. In

practice, marriage and parenthood often remained the quicker and more viable route to obtain

housing and generous loans from the state than trying to accrue such privileges through labour market activity.

In contrast, in the period following World War II, the FRG had a democratic multiparty parliament, a market economy, and a corporatist conservative welfare state (Rosenfeld, Trappe, & Gornick, 2004). The arguably purest implementation of the male breadwinner model (Prince Cooke & Baxter, 2010) resulted in female labour market participation around only 50%, of which much was part time. Family policies where foremost *pro-traditional* and social policies comprehensively set strong incentives for a traditional male breadwinner–female carer household division of labour and the financial dependence of women on men (Brückner, 2004; Prince Cooke, 2011). These include joint taxation of married couples that discouraged employment of a second earner and the absence of public child care particularly for children under the age of three (Aisenbrey, 2009; Prince Cooke & Baxter, 2010).

In 1990, Germany was reunified by adapting the FRG model to the former GDR, which ceased to exist. Especially during the 1990s the reunification process was accompanied by severe economic recession in the East but also in the West that peaked in the mid-1990s. However, the West German institutional model was applied to a population with markedly different compositional features in the former East. For instance joint taxation for married spouses makes a large difference particularly when earnings between husbands and wives are very unequal. Joint taxation sets few incentives for marriage for couples with similar earnings, which was much more often the case in the former East also after reunification (Bastin, Kreyenfeld & Schnor, 2012). Further, the public child care infrastructure remained largely intact in the East maintaining a much more conducive environment for women's employment which remains less affected by motherhood and marriage (Matysiak and Steinmetz, 2008). East Germany is one of

4

the only regions in Western societies that shows no motherhood penalty in wages, whereas this penalty remains among the highest in the West at 32 percent after in the early 2000s (Budig, Misra & Boeckmann, 2012). For an overview of institutional differences relevant to family formation around the reunification see Goldstein and Kreyenfeld (2011) and Fasang (2014).

Initial expectations of a quick convergence of the two Germanies thus did not materialize (Schneider, Naderi & Ruppenthal, 2012). Even in 2008, 18 years after the reunification, East Germany still had significantly lower rates of property ownership, lower average earnings, higher rates of female employment and a higher proportion of children in public care (Goldstein and Kreyenfeld, 2011). East Germany further continues to be one of the most secularized regions of Europe with 74 percent reporting no religious affiliation in 2008 compared to only 16 percent in the West (ibid: 457). While total fertility rates have converged (Goldstein and Kreyenfeld 2011), within Europe the two sub-societies remain on opposite ends of the continuum for a number of demographic indicators, including a proportion of nonmarital birth above 70 percent in the East compared with only around 12 percent in the West in 2007 (Klüsener, Perelli-Harris, Sanchez Gassen, 2013), as well as much higher rates of cohabitation in the East. Recently, Klüsener and Goldstein (2014) suggested that the long-standing divide in non-marital fertility in East and West Germany is attributable to regional variation that preceded the war in 1945 and was merely intensified during the division. Therefore the two subsocieties might plausibly never converge or at least not within a short time frame on some deeply rooted demographic and economic differences. Overall, there is more research on similarities and differences in family formation than on employment trajectories in East and West Germany. Studies on family formation are mostly located in demography and likely stimulated by the extreme patterns of difference in the two German sub-societies.

Precisely because we see notable convergence in some indicators of family formation and employment trajectories, but persistent and sizeable differences in others, it is particularly useful to examine holistic life course trajectories. In addition to a detailed view on single outcomes this enables us to broaden our analytical scope and assess to what extent overall differences emerge form diverging patterns in single indicators. In this paper we therefore analyze whether and how men's and women's employment and family life courses were different and remain different in divided and reunified Germany conceptualizing them as longitudinal "process outcomes" (Abbott 2005).

3. A Medoid-Based BIC Method for Sequence Assessment

Comparing sequences across samples is essentially a problem of assessing differences across groups. Such assessments can be conducted by computing distance measures both within and between sets of sequences belonging in specific groups. On the one hand, distances within specific groups can be computed summarizing the degree of homogeneity or standardization of the life courses in this set of sequences. On the other hand, distance between sequences belonging to different groups can be computed (Fasang 2014). The discrepancy analysis proposed by Studer et al (2011) is a good method for analyzing differences between sets of sequences. However, as reported later, a discrepancy analysis of the East and West German life course data can be very insensitive to any variation in the degree of differences between sets of sequences, as also shown through our simulation study discussed later, thus failing to distinguish between cohorts for either men or women family formation and employment history data.

In this section, we consider the BIC and the LRT as an alternative method that is also based on distance measures. There are many viable candidates as distance measures for making

statistical comparisons, including optimal matching, dynamic hamming matching or the so called

subsequence metrics (e.g. Aisenbrey and Fasang 2010 for an overview of several distance

measures). One can compute the classical Levenshtein or another statistical distance between all

possible pairs of sequences of the groups under comparison. Alternatively, one can compute a

distance of each sequence to the medoid sequence of the sequence set, a popular method used for

complexity reduction in recent research on sequence visualization and comparison (Aassve et al

2007; Fasang and Liao 2013; Gabadinho et al 2011; Piccarreta 2012). One important advantage

of the medoid based approach is that there is only one distance value for each sequence to the

medoid instead of $N(N\text{-}1)/2$ pariwise comparisons thus yielding higher efficacy. For reason that

we detail below, we use the medoid for computing distances within and between groups.

Let s_i denote the sum of squared distance of each sequence group i to the sequence group

medoid of group i:

$$s_i = \sum_{j=1}^{n_i} q_{ij}{}^2 \tag{1}$$

where q_{ij} is the distance between each sequence j in the ith sequence group and the ith sequence

group medoid, which can be understood as the best representative observed sequence in the ith

group (i.e., giving the shortest distances to all group members overall). Note that a set of

sequences may have several – i.e. tied – medoid sequences, where more than one sequence share

a minimum distance to all other sequences in the group. The statistic in (1) can be computed for

all G number of sequence groups.

Because a group medoid is the observed sequence that has the shortest distance to all the

sequences in the same group, it is theoretically possible that the medoid for a sequence from

another group may have a shorter distance than its own medoid, a rare situation we have

empirically verified. To avoid this empirical problem because of a single unexpected better

7

sequence, we use a selection of best representative sequences from a group, or collective

medoids, in (1) for computing BIC and LRT statistics. For example, we may select a small

percentage or number of best representative sequences, such as *m* number of or percent

sequences from a sequence group. While any small percentage or number of representative

sequences may work, we have found in our simulation study that using 5% can sufficiently or

smaller avoid misrepresentation (for each of the groups with the *m* for the combined group as the

sum of the *m*s). The ultimate choice of *m* depends on the heterogeneity and unevenness of the

sequence groups to be compared.

3.1 BIC

Assuming the distances in (1) are normally distributed – an assumption we test below – we

follow Burnham and Anderson (1998, 2004) in expressing least squares as likelihoods for

constructing BICs as well as the closely related Akaike Information Criterion).[2] Thus, –2 log-

likehood for (1) can be expressed as

$$-2 \text{ log-likelihood}_i = n_i \log\left(\frac{s_i}{n_i}\right) \qquad (2)$$

where n_i denotes the number of sequences in sequence group *i*. The BIC for the *i*th sequence

group can then be expressed as

$$\text{BIC}_i = n_i \log\left(\frac{s_i}{n_i}\right) + K \log(n_i) \qquad (3)$$

For the *i*th sequence group, *K*=1 because the only degree of freedom involved is that associated

with the *i*th medoid. For comparing the *h*th and the *i*th sequence groups, or more generally, *G*

number of groups, we propose a nested model approach, which is consistent with the typical

application of BIC when models are nested. We define BIC_A as the BIC based on the sum of

8

squared distances using the overall medoid and BIC_G as the BIC based on the sum of squared distances using group-specific medoids:

$$\text{BIC}_A = \frac{n_A}{w}\log\left(\frac{s_A}{n_A}\right) + k\log\left(\frac{n_A}{w}\right) \qquad (4)$$

$$\text{BIC}_G = \frac{n_A}{w}\log\left(\sum_{i=1}^{G}\frac{s_i}{n_i}\right) + k\log\left(\frac{n_A}{w}\right) \qquad (5)$$

where n_A is the total number of distance computations between the sequences in all groups combined and the reprentative sequences, and w is a weight that equals to $(e \times \log(s) \times n_A)^{0.5}$ and is applied once in each of the additive items where n_A appear. This weight is applied to the entire BIC because the distance computation of (1) is unnecessarily increased by the multiplicative factor of the number of episodes, e, and the number of states, s, and the number of sequences n_A (the number of distance computations between sequences and their representatives) though the contribution by s can be reasonably conceived as a modest nonlinear function.[3] The parameter K gives the number of degrees of freedom, which equals to 1 in (4) because only the overall medoid (or an overall set of representative sequences) is used and equals to G in (5) because G medoids (or G sets of representative sequences) are used in the computation.[4] The BIC difference between (4) and (5), or (4)–(5), forms the criterion for assessing differences between sets of (life course) sequences.

To assess the statistical properties of the BIC difference based on (4) and (5), we conducted a simulation study, as explained and reported in Appendix A. The simulation results show that the statistic of BIC differences using (4) and (5) can capture actual differences between two sets of life course sequences very well though it still is, to a degree, sensitive to sample sizes. The original BIC is sensitive to sample size increases; the weight in (4) and (5) reduces some of that sensitivity. We also applied discrepancy analysis methods to the simulation

9

data. Regardless of the degree of difference between the two simulated sets of sequences, the *F*-test in the discrepancy analysis is always significant at least at the 0.001 level. Thus, we do not report the results of the simulation study regarding the behavior of the discrepancy analysis.

3.2 LRT

While the BIC should provide a proper statistical assessment of differences between groups or samples of sequences, it does not give a significance test. Additionally, we can construct a Likelihood Ratio Test (LRT) using (2). Let ll_i stand for the -2 log-likelihood of the *i*th group. Following Liao's (2002, 2004) discussion of the LRT for generalized linear or logit models, the LRT for testing the null hypothesis that all groups are equal in their sequences is obtained as the difference between the *ll* from the restricted model and the *ll* from the unrestricted model. In the current application of comparing *G* number of sequence groups, the restricted model is the situation when we assume all groups are equal, with a single medoid (i.e., the global medoid), and the unrestricted model is the case where all groups are considered unique, with their individual medoids. Following Liao (2002: (6.13) & (6.14)), we further define ll_R as the -2 log-likelihood from the restricted model and ll_U as the -2 log-likelihood from the unrestricted model where

$$ll_U = \sum_{i=1}^{G} ll_i \qquad\qquad (6)$$

Therefore, the LRT for testing sequence group differences is given as

$$\text{LRT} = ll_R - ll_U = ll_R - \sum_{i=1}^{G} ll_i \sim \chi^2 \qquad\qquad (7)$$

with *G*-1 degrees of freedom that follows the chi-squared distribution. Applying (7) to sequence data by using (4) and (5), we obtain

$$\text{LRT} = \frac{n_A}{w} \log\left(\frac{s_A}{n_A}\right) - \frac{n_A}{w} \log\left(\sum_{i=1}^{G} \frac{s_i}{n_i}\right) \sim \chi^2 \qquad (8)$$

Note that the LRT of (8) is also based on the nested model principle, just like the BIC difference given by (4) and (5). Thus, it can be considered a significance test equivalent to the nested BIC assessment. Because of the similarity between the LRT in (8) and the BIC difference based on (4) and (5), we did not conduct another simulation study to verify its statistical behavior. The results in Appendix A should apply here (as the actual data analysis confirms later).

4. Data

We use the newly released German National Education Panel data (NEPS), starting from cohort 6 (NEPS) (Leopold, Skopek, & Raab, 2011). This part of NEPS contains retrospective life course information for 11,649 individuals born between 1944 and 1986 who were surveyed in 2009/2010. The survey instruments contain detailed questions about education, work and work interruptions, as well as family formation, including the formation and dissolution of marital and cohabiting unions. We use data for family and employment trajectories from ages 15 to 40 of East and West German men and women born 1944-1970 measured in monthly intervals. We thereby start with the oldest available cohort born in 1944 and end with the cohort born in 1970, which is the last birth cohort that we can observe until age 40. Examining sequences until age 40 is important to assume that individuals have reached occupational maturity (Aisenbrey and Brückner 2008) and have largely completed the active family formation phase. For these cohorts we have 6,578 sequences of complete family and employment information from ages 15 to 40, 5,432 for West Germany and 1,146 for East Germany. For the cohort comparison, we group birth cohorts in 4 year-intervals: 1944-1949, 1950-1953, 1954-1957, 1958-1961, 1962-1965, 1966-1970. The oldest and youngest birth cohort groups are slightly larger and comprise 5 birth years.

This grouping provides sufficient case numbers per cohort group for East and West Germany and is sufficiently fine grained to detect change over time in the birth cohorts' life courses.

Figure 1 shows a Lexis diagram that places our study cohorts in historical context. The cohorts born 1944-1949 experienced their entire life courses until age 40 in divided Germanys. The cohorts born 1950-1953 and 1954-1957 had already largely completed family formation and reached occupational maturity when they experienced the reunification in the 30ies. For these cohorts childhood and young adulthood happened during very different state systems that had been clearly established in the 1960s and 1970s, whereas the 1940s and 1950s were arguably foremost characterized by rather universal post-war situation of scarcity and then an emerging new economy. Finally the cohorts born in the 1960s already experienced their active family formation phase and labor market entry after the reunification in 1990.

Figure 1: Lexis diagram on study cohorts

The family sequences are specified with nine states of "single, no child" (SNC), "single, with child" (SC), "cohabiting, no child" (CNC), "cohabiting, with child" (CC), "married, no child" (MNC), "married with child" (MC). Being single is defined as not being in a cohabiting relationship and thus includes persons who were never married as well as divorcees. Only biological children are included. The percentage of reported adopted and foster children is very low (below 1 percent each) and legislation and practice of adoption and foster parenting differed substantially in East and West Germany.

The employment sequences are operationalized based on different stages out of the labor force and Erickson-Goldthorpe-Portocarero (EGP) classes of employment. EGP classes are hierarchical but categorical in nature and thus ideally suited to examine in a sequence analysis framework while maintaining the ability to distinguish upward and downward mobility. We

specify 13 employment states including "out of the labor force/gap"(OLF), "unemployment" (UE), "military" (M) "education" (EDU)", and "parental leave" (PL). The EGP classes are included as "Higher grade professionals [I]", "Lower grade professionals [II]", "Routine non manual employees, higher grade [IIIa]", "Routine non manual employees, lower grade [IIIb], "Small Proprietors, farmers [IVa,b,c], "Lower grade technicians; supervisors of manual workers [V]", "Skilled manual workers [VI]", "Manual worker in primary production [VIIa,b]". Further, to increase rigor in our East-West comparison, we only included respondents in the East West samples, who were born in the respective region and where still living there at the time of the interview in 2009/2010. We excluded foreign-born persons, persons who migrated between East and West and people living in Berlin. This was necessary because the survey did not distinguish between the former East and West regions of Berlin at the time of the interview in 2009/2010 and thus they could not unambiguously be assigned to either the East or West regions.The data are weighted using a calibrated design weight provided in the NEPS. This weight includes a sampling design weight and a calibration factor (multiplier) to adjust the sample to the means of the German Microcensus 2009 (Aßmann & Zinn, 2011).

Note that in addition to this substantive application it could be useful to test the proposed methods on simulated sequence data. However, to date it is not clear how to simulate appropriate life course sequences. Simply randomly assigning orders of states in random duration with transitions at random timing are not appropriate benchmarks, because patterning in life course sequences is generally high and will always be more structured that such completely randomly simulated sequences. If random assignment of the core properties – order, duration timing – is not an option, this requires specifying other properties. This is difficult to do in a sensible way

for the combination of order duration and timing, especially considering that sequence analysis is embedded in an algorithmic data modeling culture that abstains from parametric assumptions.

4.2 Visual description of the example data: relative frequency sequence plots

The sample sizes of the life course sequences measured over 25 years are too large to plot in conventional sequence index plots, due to overplotting that distorts the visual impression of the graphs. Therefore we use relative frequency sequence plots to graphically display the sequences (Fasang & Liao 2014). For the relative frequency sequence plots, first the sequences of each subgroup are sorted according to a score of the first factor derived with multidimensional scaling (see also Piccareta and Lior 2010). The multidimensional scaling is based on a sequence distance matrix that was generated with Optimal Matching with constant substitution costs of 2 and indel costs of 1 (see MacIndoe and Abbott 2004). Then each sample of the four comparison groups is divided into 100 equally sized frequency groups. For each of these frequency groups, the medoid sequence, also calculated using optimal matching with constant substitution costs of 2 and indel costs of 1, is chosen as the representative of these medoids (see Fasang an Liao 2014 for details).

Figure 2 shows the relative frequency plots for men and women in East and West Germany for the family life courses. Figure 3 displays the employment trajectories for men and women in East and West Germany. The distance to medoid plots on the right indicate the average distance from the chosen medoid that is displayed in the RF sequence plot.

Figure 2: RF Sequence Plots of family trajectories for East and West German men and women, k=100 sorted by score of first factor derived with multidimensional scaling

14

The two upper panels in Figure 2 for East Germany show a polarization between individuals who are married with children at an early age, visualized by the blue colors. Notably, this early marriage with child pattern for East Germany is highly standardized, as indicated by the very low distances in the distance to medoid-plots for this region of the plot. Marriage and parenthood are highly coupled with only short periods of childless marriage for both men and women. This pattern accounts for the majority of East and West Germany men in our sample (about 2/3). At the bottom of the two upper panels, we see a very different pattern of unmarried parenthood in cohabitation for men and women (dark purple), and single motherhood for women in East Germany (green). The family life courses represented in the lower region of the RF Sequence plots are far more heterogeneous with larger distances from the medoid.

The two lower panels of figure 2 show family life courses for West German men and women. We see an orderly pattern of short cohabitation, relatively long childless marriage followed by parenthood, which suggests clear "normative clocks" for structuring family formation. Men and women who marry at a later age on average show longer periods of cohabitation before. Overall, the coupling of marriage and parenthood is weaker in terms of timing compared to East Germany. In contrast to East Germany, the prevalence of parenthood out of wedlock is much lower in the West reflecting the strong male breadwinner norms for our study cohort. In West Germany family life courses characterized by long periods of cohabitation are the most heterogeneous, indicated by the distance to medoid plot.

Figure 3: RF Sequence Plots of employment trajectories for East and West German men and women, k=100 sorted by score of first factor derived with multidimensional scaling

Figure 3 shows the RF plots for the corresponding employment sequences. The EGP classes are depicted using heat colors, such that light yellow represents the lowest class and red

15

represents the highest class. For all four comparison groups employment trajectories were most orderly and homogenous in the highest prestige occupations visualized in red. It is a known feature of the EGP class scheme that the lowest classes primarily comprise male occupations, which is also visible in our graphs. In West Germany we see more uninterrupted episodes of employment or non-employment (women) compared with more frequent movements between classes and in and out of the labor force in East Germany. This likely reflects the strong insider outsider segmentation and high employment protection in the conservative corporatist West Germany welfare model. We now turn to assessing quantifying and assessing the difference between these for groups and assess how differences between East and West German men and women developed across birth cohorts. In particular, we are interested in observing a converging, diverging or stable difference between family and employment life courses in East and West Germany based on the proposed BIC and LRT statistics.

4.3 Assessing the normality assumption and making adjustment

Equation (2) given earlier relies on the normality assumption for the distances between individual sequences and the medoid. That is, to reasonably consider sequence distances as errors in a linear model, normal distribution of such distances is assumed. To assess the assumption visually, a common plot, known as the quantile-quantile (QQ) plots of the distance data can be employed. A QQ plot presents the standardized observed data against the standard normal distribution. There are also statistical tests for testing distributions, such as the Anderson-Darling test, the Kolmogorov-Smirnov test, and the Shapiro-Wilk test.[5]

Through a series of preliminary analysis, we found a mild degree of deviation from normality in the German data. To deal with the general problem of deviation from normality

regardless of its severity, we propose an adjustment to the BICs from (3), (4), and (5) by using

the Shapiro-Wilk test statistic. Of four common statistical tests of normality, the Shapiro-Wilk

test is the most powerful (Razali and Wah 2011). The BIC can be adjusted by applying a

function of the Shapiro-Wilk statistic to obtain aBIC, or adjusted BIC:

$$aBIC = \frac{BIC}{1 - \log_{10}(SW_y)} \qquad (8)$$

where SW_y is the Shapiro-Wilk test statistic of the distribution of variable y. We conducted a

simulation study over a range of sample sizes and degree of deviation from normality, and found

(8) an effective way of compensating BIC in the presence of deviation from normality.[6]

5. Findings from the Comparion of East and West German Life Course Sequences

We report in Tables 1 and 2 the results from an application of discrepancy analysis on the East

and West German life course sequence data. As shown clearly in all the statistics presented in the

two tables, discrepancy analysis is very sensitive to any differences between the two data sources,

producng rather small pseudo-R^2s and highly significant pseudo-F tests for either the

employment history or the family history life course comparisons between the East and the West.

Table 1: Pseudo-R^2, pseudo-F, and Levene statistics analyzing the German employment history data

Table 2: Pseudo-R^2, pseudo-F, and Levene statistics analyzing the German family history data

Because of this lack of differentiation in comparing the two sequence groups, we conducted BIC

and LIT analysis of the East and West German life course data, and present the findings in

Tables 4 and 5. As summarized in Table 3, a BIC differences between 0 and 2 indicates

negligible differences between two sets of sequences, 2 to 6 suggests moderate differences

whereas BIC difference above 6 can be considered strong and above 10 very strong.

Table 3: BIC comparison guide

Table 4: Bayesian Information Criterion differences and Likelihood Ratio statistics of the German family history data, the East vs the West (nested model approach, BIC_a, BF_a, and LR_a are the adjusted BIC, BF, and LR using Shapiro-Wilke statistic)

Table 5: Bayesian Information Criterion differences and Likelihood Ratio statistics of the German employment history data, the East vs the West (nested model approach, BIC_a, BF_a, and LR_a are the adjusted BIC, BF, and LR using Shapiro-Wilke statistic)

Overall we find significant differences between east and West German family (Table 4) and employment life courses (Table 5). The difference indicated by BICa is larger for family life courses (27.3) than for employment life courses (17.7). Moreover on both life course dimension, East and West German women are more different from one another than East and West German men as is visible in the larger BICa for women.

In addition to these overall differences, we find notable cohort variation that support divergence after the life courses of the 1944-1949 cohort followed by convergence for the post-reunification cohorts for family and employment life courses. BIC differences are largest for the cohorts born 1950-1953 and remain sizeable for the cohorts born 1954-1958. Beginning with the cohorts born 1960 differences between east and West German life course diminish and are no longer significant for the youngest birth cohorts. We visualize the adjusted BIC differences by cohort in figures4 (based on values displayed in tables 4 and 5). Also, men's life courses appear to differ less between the East and the West than women's life courses. This pattern is true for both the family formation and the employment history comparisons.

Figures 4: Visualization of BICa

18

6. Discussion

One may believe that the 1944-1949 birth cohort should exhibit the greatest difference in BIC because these women and men lived almost their entire lives between age 15 and 40 in the divided countries. This understanding, however, is incorrect. According to prior life course research, a period of planful orientation during mid-adolescence (around ages 14 and 15) is important for one's later life course because this planful orientation helps with realistic decisions about adult roles and relationships (Shanahan 2000).

The building of the Berlin Wall started in August, 1961 though improved wire fences were added during the three years of 1962 to 1965 and improved concrete walls were added after that. Thus, even the youngest of the 1944-1949 cohort did not spend their mid-adolescence in an entirely closed-off and separated Germanys. Perhaps that is why the differences for the oldest cohort are rather modest in size, even showing statistically significant (at the 0.05 level) LRTs for men's family history and employment history comparisons.

The middle cohorts of either men or women show the greatest differences between the East and the West for both their family formation and employment history life course comparisons. This is not surprising because these women and men spent most of their post-mid-adolescent years in the separate Germanys, with their different social welfare systems and availability of childcare facilities. The extent to which men's and women's life courses differ is rather small for these cohorts in either the family formation or the employment history dimension.

The youngest male or female cohort, however, had a different life course experience from their older compatriots. These women and men spent their entire early, mid- and late adolescent years in the separate Germanys. However, when they became adults, the two Germanys had been unified; similar employment and welfare opportunities began to become available to them.

19

Indeed, these women and men lived much of their adult years between 18 and 40 in the unified Germany, with their life courses, in terms of either family formation or employment history, shaped by the new unified social structure.

To return to our research question of whether state policies generate differences in life courses, we offer a rather affirmative answer because the middle cohorts demonstrate consistent differences in family formation and in employment history for both men and women of the East and the West origins. The smaller differences in the oldest as well the youngest cohorts further confirm this answer. Whereas similar macro-situations during mid-adolescent years helped reduce men's life course differences among those of the oldest cohort, the identical state policies promoted a divergence of the earlier differing life courses between those from the former East Germany and those from the former West Germany, especially among men.

7. Conclusion

We have shown in this paper that statistically comparing samples of life course sequences is a difficult task: Comparing sets of single value measurement such as income would require simply testing mean or log-mean differences of the two samples. Life course sequences, on the other hand, contain a set of complex measurements including the timing of life course events, the duration of such events as well as their ordering. In this proper, we have proposed a method based on distances of sequences from a set of representative medoid sequences (instead of a single medoid) for generating BICs and LRTs for properly assessing differences between sets of sequences statistically.

Because distances can be a function of sequence length, the number of qualitative states, and the number of comparisons being made, we devised a weight for compensating the inflation effect of this function. Our simulation study assessed the relative adequacy of the weighted BIC

20

and LRT computations. Such weighted BICs and LRTs were used in the assessment of the differences in family and employment history sequences between East and West German men and women.

The analysis of the German life course sequence data provided some clear support for macro-level, institutional effects on micro-level, individual life course differences. A pattern of diverging trends between the two Germanys in individual life courses took shape with the erecting of the Berlin Wall and converging trends in these individual life course sequences occurred with the tearing down of the Berlin Wall and the reunification, based on our analysis of the six birth cohorts of East and West German data. Such diverging and converging trends are more observable among men than among women. This gender difference suggests that men's life courses are more moldable than women's by societal and institutional changes exemplified by the postwar German history of two Germanys and their reunification.

References

Aassve, Arnstein, Francesco C. Billari and Raffaella Piccarreta. 2007. Strings of Adulthood: A Sequence Analysis of Young British Women's Work-Family Trajectories. *European Journal of Population* 23: 369-388.

Aisenbrey, Silke, and Anette E. Fasang. 2010. "New Life for Old Ideas: The 'Second Wave' of Sequence Analysis. Bringing the 'Course' Back into the Life Course." *SociologicalMethods & Research* 38:420-62.

Aßmann, C., & Zinn, S. 2011. NEPS Data Manual. Starting Cohort 6: Adult Education and Lifelong Learning [Supplement C] Weighting. *National Eucation Panel Study Data Center (NEPS) Documentation, University of Bamberg*.

Blanchard, Philippe, Felix Bühlmann and Jacques-Antoine Gauthier (eds). 2014. *Advances in Sequence Analysis: Methods, Theories and Applications*. Spinger.

Brückner, Hannah and Karl Ulrich Mayer. 2005. De-standardization of the Life Course: What it might mean? And if it means anything, whether it actually took place? *Advances in Life Course Research* 9: 27-53.

Burnham, Kenneth P. and David R. Anderson. 1998. *Model Selection and Inference: A Practical Information-Theoretical Approach*. New York: Springer-Verlag.

_____. 2004. "Multimodel Inference: Understanding AIC and BIC in Model Selection." *Sociological Methods & Research* 33: 261-304.

Fasang, Anette E. 2014. New Perspectives on Family Formation: What Can we Learn from Sequence Analysis? In: Blanchard, Philippe, Felix Bühlmann and Jacques-Antoine Gauthier (eds). 2014. *Advances in Sequence Analysis: Methods, Theories and Applications*. Spinger.

Fasang, Anette E. and Tim F. Liao. 2014. "Visualizing Sequences in the Social Sciences: Relative Frequency Sequence Plots." *Sociological Methods & Research* 43(4): 643-676.

Gabadinho, Alexis, Gilber Ritschardm Nicolas Müller and Matthias Studer. 2011. Analyzing and Visualizing State Sequences in R with TraMineR. *Journal of Statistical Software* 40(4): 1-37.

Goldstein, Joshua and Michaela Kreyenfeld. 2011. Has East Germany Overtaken West Germany? Recent Trends in Order-specific Fertility. *Population and Development Review* 37(3): 453-472.

Kass, Robert E., and Adrian E. Raftery. 1995. "Bayes Factor." *Journal of the American Statistical Association* 90: 773-795.

Leopold, T., Skopek, J., & Raab, M. (2011). NEPS Data Manual. Starting Cohort 6: Adult Education and Lifelong Learning. *National Eucation Panel Study Data Center (NEPS) Documentation, University of Bamberg*.

Liao, Tim F. 2002. *Statistical Group Comparison*. New York: Wiley.

Liao, Tim F. 2004. "Comparing Social Groups: Wald Statistics for Testing Equality among Multiple Logit Models." *International Journal of Comparative Sociology* 45: 3-16.

Oh, Man-Suk, and Adrian E. Raftery. 2001. "Basyesian Multidimensional Scaling and Choice of Dimension." *Journal of the American Statistical Association* 96: 1031-1044.

Piccarreta, Raffaella. 2012. "Graphical and Smoothing Techniques for Sequence Analysis." *Sociological Methods & Research* 41: 362-380.

Marshall, Victor W., and Margaret M. Mueller. 2003. "Theoretical Roots of the Life-Course Perspective." Pp. 3-32 in *Social Dynamics of the Life Course: Transitions, Institutions,*

and Interrelations, edited by Water R. Heinz and Victor W. Marshall. New York: Aldine de Gruyter.

Mayer, Karl Ulrich, and Urs Schoepflin. 1989. "The State and the Life Course." *Annual Review of Sociology* 15: 187-209.

Mayer, Karl Ulrich. 2004. "Why Lives? How History, Societies, and Institutions Define and Shape Life Courses." *Research in Human Development* 1(3): 161-187.

_____. 2009. "New Directions in Life Course Research." *Annual Review of Sociology* 35: 413-433.

Neyer, Gerga. 2013. "Welfare State, Family Policies, and Fertility in Europe." Pp. 29-53 in *The Demography of Europe*, edited by Gerda Neyer, Gunnar Andersson, Hill Kulu, Laura Bernardi, and Christopher Büller. London: Springer.

Raftery, Adrian E. 1986. "Choosing Models for Cross-Classifications." *American Sociological Review* 51(1): 145-146.

Razali, Nornadiah Mohd, and Yap Bee Wah. 2011. "Power Comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling Tests." *Journal of Statistical Modeling and Analytics* 2(1): 21-33.

Schnettler, S., & Klüsener, S. (2014). Economic stress or random variation? Revisiting German reunification as a natural experiment to investigate the effect of economic contraction on sex ratios at birth. *Environmental Health*, 13(1), 117.

Shanahan, Michael J. 2000. "Pathways to Adulthood in Changing Societies: Variability and Mechanisms in Life Course Perspective." *Annual Review of Sociology* 26: 667-692.

Studer, Matthias, Gilbert Ritschard, Alexis Gabadinho, and Nicolas S. Müller. 2011. "Descrepancy analysis of State Sequences." *Sociological Methods & Research* 40(3): 471-510.

Yazici, Berna, and Senay Yolacan. 2007. "A Comparison of Various Tests of Normality." *Journal of Statistical Computation and Simulation* 77(2): 175-183.

Appendix A

We conducted a simulation study comparing two random samples of family history sequences drawn with replacement from the early births and late first births life history sequence samples from the 1966 women birth cohorts sorted by the timing of first births at varying sample sizes of different mixing percentages. These two birth-timing groups were chosen to ensure their life courses are different enough. The purpose of the simulation study is to investigate the effect of two variations—that of different mixing proportions from the two samples of origin and that of sample size on BIC differences (because BIC is a function of sample size). Without doubt, we are primarily interested in finding out whether BIC differences can reflect true sample differences in their life course sequences.

Each pair of samples sum up to a combined sample size of 100, 200, 500, 1,000, and 2,000, with the mixing percentages from the west and the east samples of 5% and 95%, 10% and 90%, 20% and 80%, 30% and 70%, 40% and 60%, and 50% and 50%, respectively. Each of the data situations of the mixing percentages at the various sample sizes is simulated 1,000 times.

Figure A.1 presents the simulation results. Clearly, BIC differences, to a degree, are still a function of sample size, with larger sample sizes generating greater BIC differences (going down the rows of the violin plots), despite of the weight w applied in (4) and (5). It is reassuring to know that these BIC differences truly reflect the actual differences in the percentages of the samples drawn from each of the sources. That is, a 50%-50% should give rise to the greatest BIC difference while a 5%-95% should generate an insignificant amount of difference. The rather small spread shown by the violin plots indicate statistical efficiency of BIC difference caculations.

---Figure A.1 about here---

Table A.1 shows the median values of the BIC difference simulations when comparing these two groups of samples. These median values indicate when a BIC difference is above, 2, 4, 6, or 10, the various guideline threshold values, given the different mixtures of the two birth-timing cohorts and the sample size.

---Table A.1 about here---

The simulation study shows that the computation of BIC difference discussed in the methods section provides an efficient method for assessing sequence differences. Relatively speaking, small sample sizes are less efficient than larger sample size BIC calculations.

26

Figures and Tables

Figure 1: Lexis Diagram of Life Course from Ages 15 to 40 of the Study Cohorts (1944-1970)
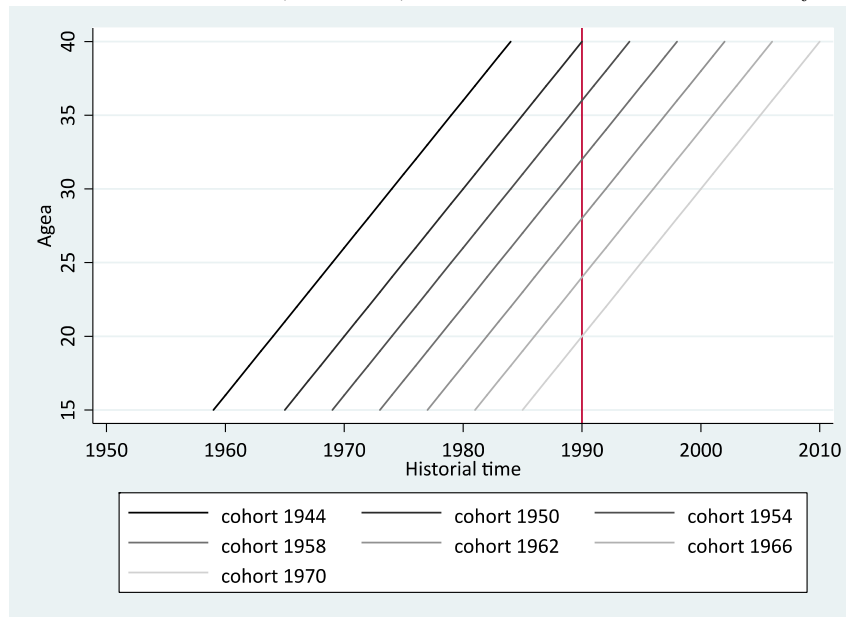*Placed in Historical Time (1959-2009), red line vertical marks German reunification*

Figure 2: RF Sequence Plots of family trajectories for East and West German men and women, k=100 sorted by age of first childbirth
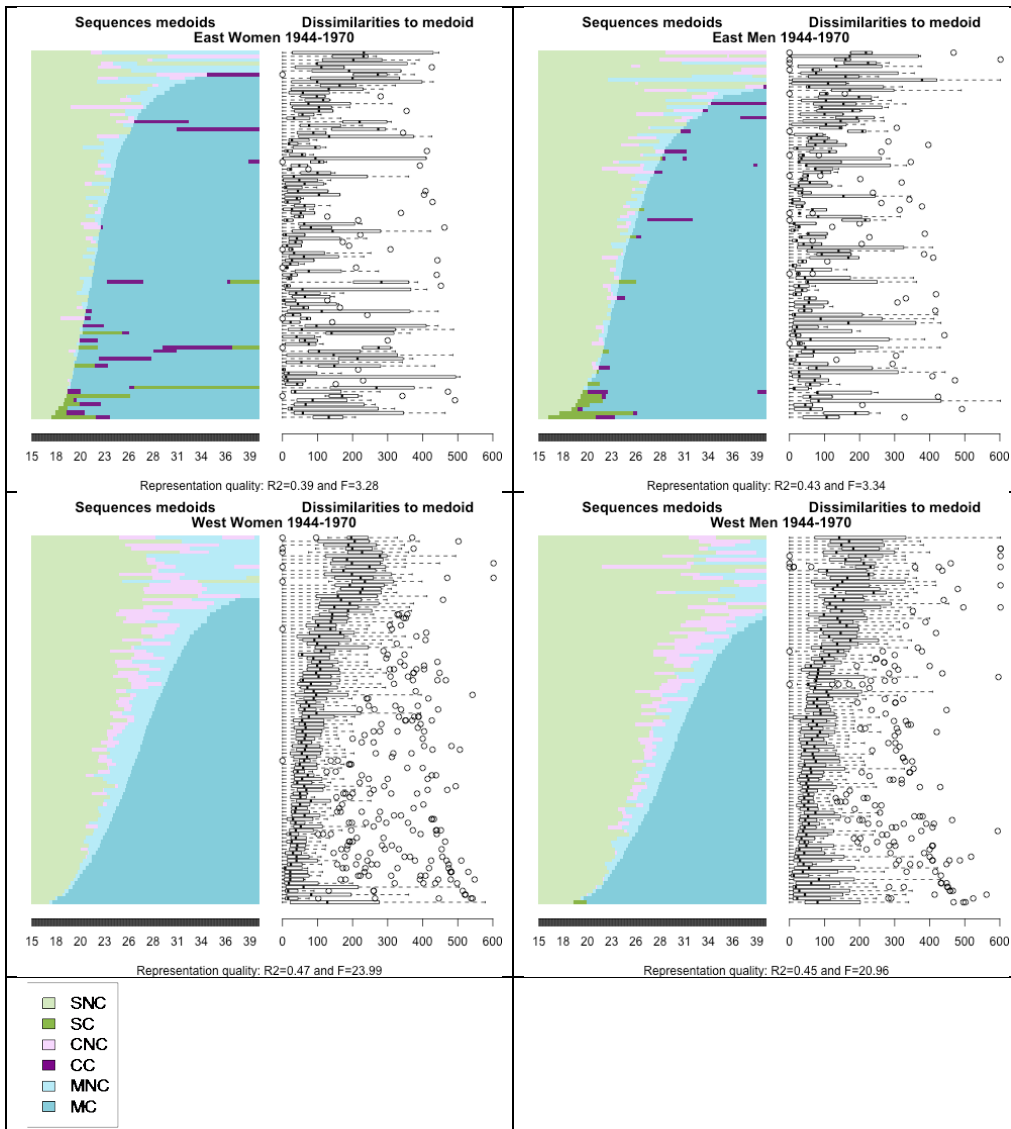
Figure 3: RF Sequence Plots for employment trajectories for East and West German men and women, k=100 sorted by age of first job
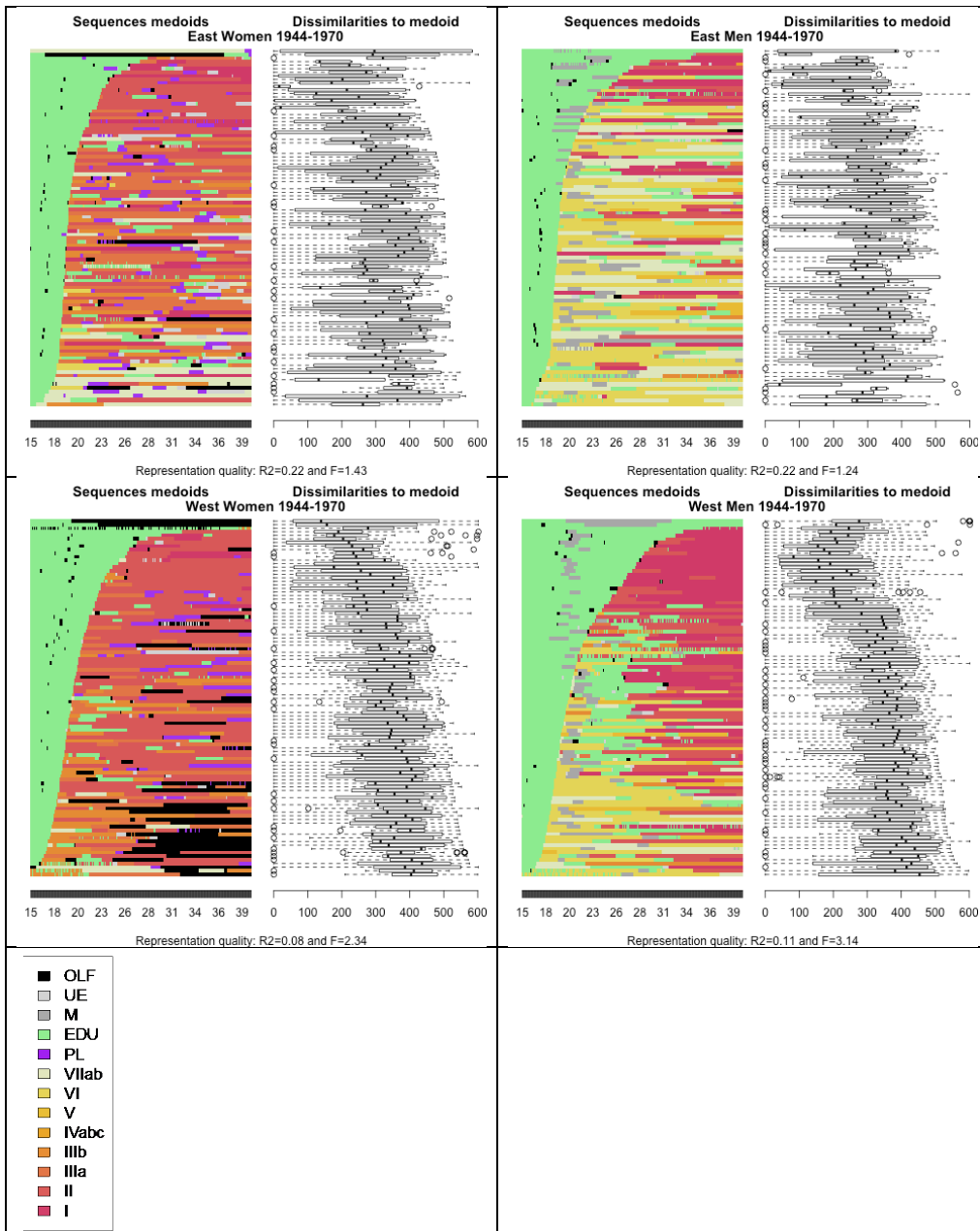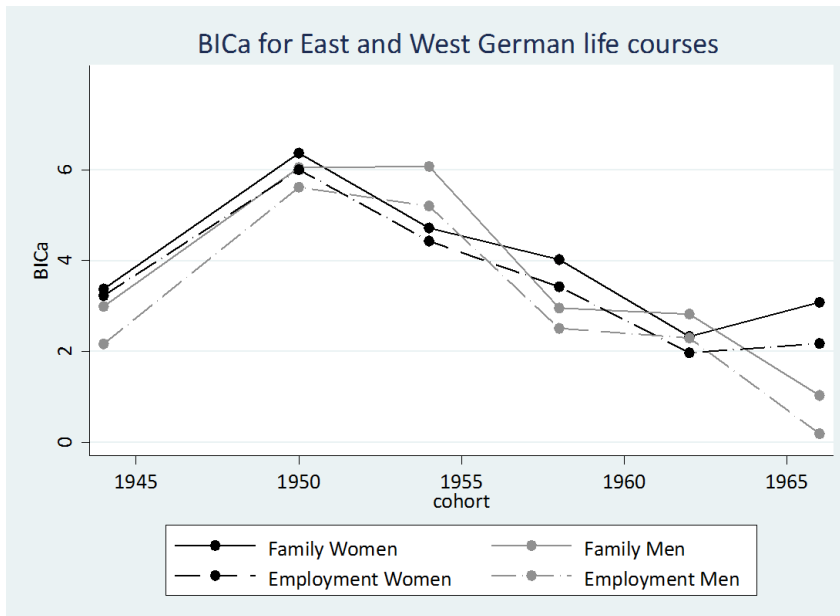
Figure 4: Line Graph of BIC difference

Table 1: Pseudo-R^2, pseudo-F, and Levene statistics analyzing the German employment history data

East–West	Pseudo R^2	Pseudo F	Sig. level	Levene	Sig. level
overall	0.005	30.187	0.001	0.812	0.380
men	0.008	23.918	0.001	7.407	0.008
women	0.007	23.906	0.001	0.048	0.806
cohort 44, m	0.004	2.005	0.018	3.559	0.054
cohort 50, m	0.004	2.005	0.018	3.559	0.054
cohort 54, m	0.013	6.377	0.001	14.285	0.001
cohort 58, m	0.012	7.642	0.001	5.983	0.014
cohort 62, m	0.007	4.529	0.001	0.451	0.487
cohort 66, m	0.008	4.287	0.001	0.375	0.535
cohort 44, w	0.008	4.287	0.001	0.375	0.535
cohort 50, w	0.008	4.287	0.001	0.375	0.535
cohort 54, w	0.008	4.287	0.001	0.375	0.535
cohort 58, w	0.008	4.287	0.001	0.375	0.535
cohort 62, w	0.008	4.287	0.001	0.375	0.535
cohort 66, w	0.008	4.287	0.001	0.375	0.535

Table 2: Pseudo-R^2, pseudo-F, and Levene statistics analyzing the German family history data

East–West	Pseudo R^2	Pseudo F	Sig. level	Levene	Sig. level
overall	0.019	128.267	0.001	19.277	0.001
men	0.018	58.57	0.001	1.730	0.170
women	0.021	73.253	0.001	21.423	0.001
cohort 44, m	0.021	73.253	0.001	21.423	0.001
cohort 50, m	0.021	9.234	0.001	5.506	0.015
cohort 54, m	0.029	14.74	0.001	4.859	0.028
cohort 58, m	0.022	14.109	0.001	0.071	0.795
cohort 62, m	0.028	17.311	0.001	0.342	0.576
cohort 66, m	0.010	5.094	0.001	6.275	0.009
cohort 44, w	0.020	9.035	0.001	10.52	0.001
cohort 50, w	0.016	6.525	0.001	6.311	0.021
cohort 54, w	0.018	10.000	0.001	7.397	0.007
cohort 58, w	0.028	20.777	0.001	1.218	0.272
cohort 62, w	0.026	18.922	0.001	0.491	0.472
cohort 66, w	0.026	15.647	0.001	0.030	0.866

Table 3: BIC comparison guide

Evidence	BIC difference	Bayes factor
Not worth a bare mention	0 to 2	1 to 3
Positive	2 to 6	3 to 20
Strong	6 to 10	20 to 150
Very strong	>10	>150

Table 4: Bayesian Information Criterion differences and Likelihood Ratio statistics of the German family history data, the East vs the West (nested model approach, BIC_a, BF_a, and LR_a are the adjusted BIC, BF, and LR using Shapiro-Wilke statistic)

East–West	BIC	BIC_a	BF	BF_a	LR	Sig. level	LR_a	Sig. level
overall	28.014	27.292	1211309	843871.3	31.486	0.000	30.678	0.000
men	15.061	14.920	1863.655	1737.08	17.808	0.000	17.641	0.000
women	24.284	23.376	187613.8	119107.5	27.094	0.000	26.08	0.000
cohort 44, m	2.212	2.206	3.022	3.014	3.173	0.075	3.165	0.075
cohort 50, m	6.994	6.957	33.018	32.415	7.594	0.006	7.554	0.006
cohort 54, m	5.971	5.925	19.793	19.344	6.708	0.010	6.656	0.010
cohort 58, m	2.272	2.248	3.115	3.077	3.425	0.064	3.389	0.066
cohort 62, m	3.293	3.235	5.189	5.042	4.42	0.036	4.343	0.037
cohort 66, m	0.197	0.191	1.103	1.100	1.26	0.262	1.226	0.268
cohort 44, w	4.592	4.521	9.933	9.59	5.351	0.021	5.269	0.022
cohort 50, w	7.326	6.796	38.969	29.898	7.842	0.005	7.275	0.007
cohort 54, w	5.592	5.470	16.380	15.411	6.563	0.010	6.420	0.011
cohort 58, w	4.955	4.696	11.912	10.463	6.228	0.013	5.902	0.015
cohort 62, w	2.120	2.072	2.887	2.817	3.44	0.064	3.360	0.067
cohort 66, w	3.563	3.448	5.939	5.606	4.673	0.031	4.521	0.033

Note: The computation used a multiple representative medoid approach, in this case, 5% representative medoids.

Table 5: Bayesian Information Criterion differences and Likelihood Ratio statistics of the German employment history data, the East vs the West (nested model approach, BIC, BICₐ, BFₐ, BFₐ, and LRₐ are the adjusted BIC, BF, and LR using Shapiro-Wilke statistic)

East–West	BIC	BIC$_a$	BF	BF$_a$	LR	Sig. level	LR$_a$	Sig. level
overall	15.032	14.746	1837.476	1592.316	18.325	0.000	17.670	0.000
men	11.5	11.391	314.24	297.455	14.068	0.000	13.934	0.000
women	12.972	12.84	655.926	613.998	15.602	0.000	15.443	0.000
cohort 44, m	2.696	2.617	3.850	3.700	3.494	0.062	3.391	0.066
cohort 50, m	6.604	6.468	27.170	25.38	7.024	0.008	6.879	0.009
cohort 54, m	6.378	6.24	24.269	22.647	6.936	0.008	6.786	0.009
cohort 58, m	3.586	3.499	6.006	5.753	4.559	0.033	4.449	0.035
cohort 62, m	2.883	2.803	4.227	4.062	3.831	0.05	3.725	0.054
cohort 66, m	0.577	0.567	1.335	1.328	1.461	0.227	1.435	0.231
cohort 44, w	3.453	3.384	5.621	5.431	4.033	0.045	3.952	0.047
cohort 50, w	6.443	6.24	25.063	22.645	6.780	0.009	6.567	0.010
cohort 54, w	4.895	4.854	11.557	11.325	5.686	0.017	5.639	0.018
cohort 58, w	4.268	4.215	8.447	8.227	5.361	0.021	5.295	0.021
cohort 62, w	2.527	2.509	3.538	3.506	3.667	0.056	3.641	0.056
cohort 66, w	3.135	3.055	4.795	4.607	4.065	0.044	3.962	0.047

Note: The computation used a multiple representative medoid approach, in this case, 5% representative medoids.

35

Figure A.1: Simulation Adjusted BIC Results of Comparing Varying Percentages of Samples 1 and 2 of a Total Sample Size of 100 to 2000 Using Random Samples of the 1966 Early and Late Birth Women's Cohort Sequences with Replacement
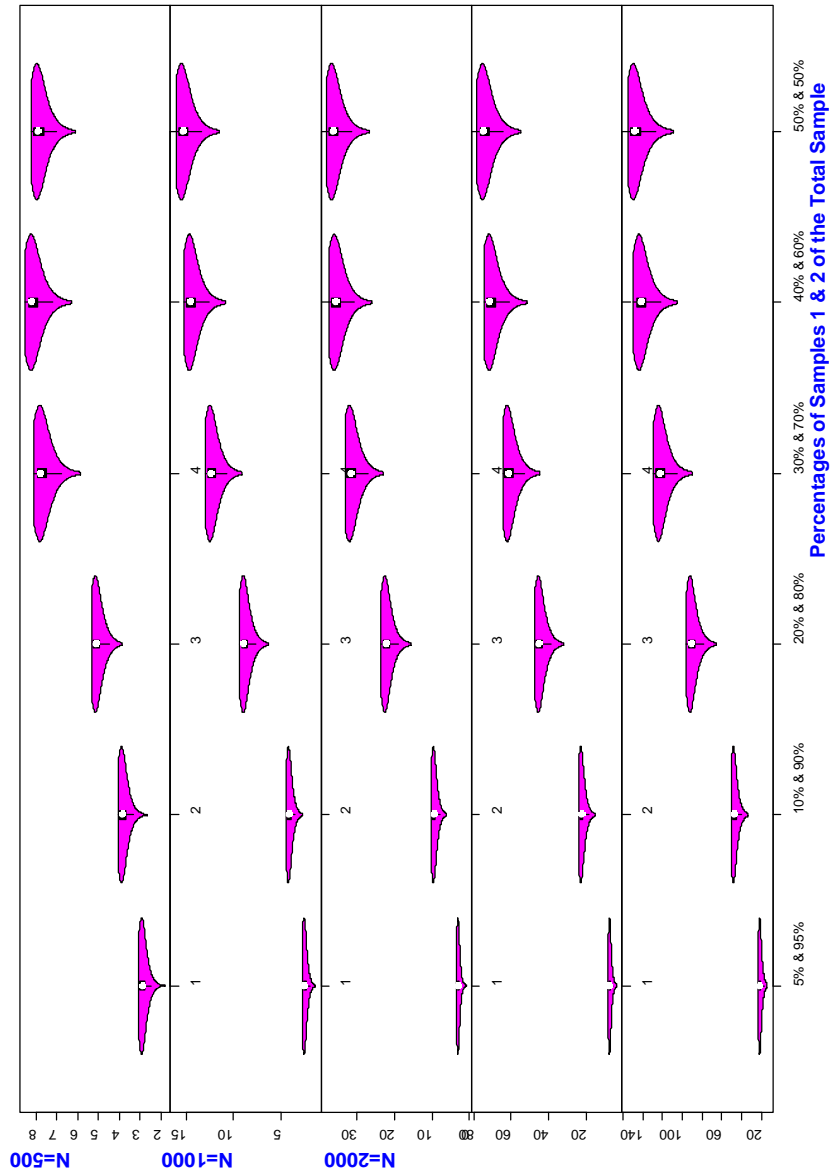
37

Table A.1: Medians of the Simulated BICs Reported in Figure A.1

N	5% & 95%	10% & 90%	20% & 80%	30% & 70&	40% & 60%	50% & 50%
100	2.908040	3.874113	5.125869	7.782945	8.210573	7.929981
200	2.521137	4.169957	8.932371	12.454760	14.622118	15.463125
500	3.079883	9.587429	22.545704	31.932319	35.956990	36.619731
1,000	7.538608	22.734413	45.291397	61.664360	71.330921	74.745039
2,000	21.636605	48.013100	91.704994	123.490443	142.945227	148.732730

Endnotes

[1] Indeed, the status attainment model that focuses on education and occupation is one of the origins of life course research (Marshall and Mueller 2003).

[2] In a similar tradition, Oh and Raftery (2001) adapted BIC for assessing dimension choice in multidimensional scaling.

[3] While increasing the number of states theoretically increases then chance of variability exponentially, which can be further multiplied by the number of sequences in the sample, empirically many observations tend to change states similarly, thus decreasing the total variability. This is reflected by the natural logarithm function and the square root function of the overall weight.

[4] The treatment of a set of representative sequences as a single medoid actually makes statistical sense. Say we use 5% representative sequences for both the combined and the separate samples. As a result the actual number of representative sequences used in the overall, combined sample and the number of representative sequences used in the separate (say the East and the West) samples are equal but in actuality the latter situation provides a better fit to the data. Thus, here the collective representation set replaces the single medoid.

[5] For a general comparison of various normality tests, see Yazici and Yolacan (2007).

[6] In the interest of space, we do not present the simulation results, which can be obtained upon request.

Session 10A: Applied sequence anaylsis

Intergenerational Patterns of Family Formation in East and West Germany

Zachary Van Winkle[1], Humboldt-University Berlin & WZB Berlin Social Science Center

Anette Fasang, Humboldt-University Berlin & WZB Berlin Social Science Center

Marcel Raab, University of Mannheim & WZB Berlin Social Science Center

Paper Prepared for LaCOSA II –International Conference on Sequence Analysis and Related Methods

Abstract:

Why is intergenerational transmission of family formation weaker in some country contexts than in others? This paper employs the historically unique situation of the German division to study country context effects on intergenerational regularities in family formation. We use the German Socio-Economic Panel (SOEP) to analyze the longitudinal family formation trajectories from age 15-35 of children born 1953-1978 and their mothers. Findings show that East German mother-child family formation trajectories are more dissimilar than West German mother-child family formation trajectories. Further, East German mother-child dyads are more likely to be categorized as patterns of intergenerational contrast, whereas West German mother-child dyads are more likely to display strong transmission. To account for these differences in intergenerational transmission of family formation between East and West Germany, we propose to combine multichannel sequence analysis, multinomial logistic modeling and decomposition methods for nonlinear probability models. This new methodological approach enables us to show that differences in parental education and children's educational mobility in East and West Germany mediate the strength of intergenerational transmission and contribute to explaining differences in intergenerational patterns of family formation in the two contexts. We conclude that the proposed approach is promising to disentangle cross-national differences in intergenerational regularities in family formation.

[1] Corresponding Author:
Humboldt-University Berlin
Faculty of Humanities and Social Sciences
Institute for Social Sciences
Unter den Linden 6
D-10099 Berlin
Telephone: +49 (0)30 2093 4336
Fax: +49 (0)30 2093 4430
E-mail: vanwinkz@hu-berlin.de

I. Introduction

The transmission of family demographic behavior, e.g. marriage, fertility and divorce, from one generation to another has been extensively studied. Research has demonstrated that mothers' age at first birth is significantly associated with their children's age of birth in the United States (Barber 2001). Similarly, intergenerational transmission of fertility patterns have been established in Britain (Booth and Kee 2009) and other European countries (Murphy and Wang 2001; Fasang 2015; Murphy 2013). Vast research has also shown the association between parental divorce and children's likelihood to marry and divorce in the United States (Amato 1996; Amato and DeBoer 2001; Wolfinger 2003) as well as in other countries, such as Finland (Erola, Härkönen, and Dronkers 2012). However, the degree of intergenerational transmission of fertility (Murphy 2013) and divorce (Härkönen and Dronkers 2006) varies considerably across countries.

A number of studies have investigated the extent of intergenerational transmission of family behavior in East and West Germany. Fasang (2015) found rather strong evidence for intergenerational fertility transmission in West Germany, but none in the former German Democratic Republic (GDR). Similarly, Engelhardt and coauthors (2002) demonstrated that the extent of intergenerational divorce transmission is lower in East Germany, and that this difference is mediated by marriage age, age at first birth and religiosity in the two sub-societies. These studies conceptualize the separation and later the unification of East and West Germany as a natural experiment, which can be exploited to investigate the micro mechanisms that account for intergenerational transmission within different macro-structural contexts in terms of social policies and economic development while holding important cultural and normative factors constant.

Several studies pertaining to the intergenerational transmission of family formation has shifted from "point in time" towards "process" outcomes (Abbott 2005). The intergenerational transmission of family formation trajectories or patterns adds important information to the study of single indicators, e.g. completed fertility. Parents and their children might strongly resemble one another in one indicator, e.g. age at first marriage but vastly differ on others, such as completed fertility or subsequent union dynamics of separation and partnering. Studies on the intergenerational transmission of single indicators of family formation are unable to show, whether multiple behaviors as well as their timing and sequencing – holistic family formation patterns - are transmitted from one generation to the next. However, from a theoretical point of view, it is

1

particularly information about different potential mechanisms of intergeneration transmission, whether children emulate the entire process of family formation they observe in their parents or simply happen to be similar on one or two indicators. Liefbroer and Elzinga (2012) analyzed similarities of parents and children's longitudinal family formation sequences and concluded that there is intergenerational continuity in the family demographic trajectories in the United States. Fasang and Raab (2014) went beyond investigating the similarity between family formation trajectories of parents and their children, and established three salient patterns of intergenerational transmission of longer-term family formation processes between ages 20 and 40 for middle class American families: strong transmission, moderated transmission and intergenerational contrast. Using the conceptual and analytical framework of transmission patterns, they were able to identify different predictors associated with specific patterns of transmission.

The aim of this paper is to determine the extent of intergenerational regularities in family formation in East and West Germany, and further to investigate the mechanisms that account for the transmission differential between East and West.

We make three contributions to the existing literature. First, we apply the conceptual and analytical framework of intergenerational patterns of family formation as a longitudinal process outcome to nationally representative data in East and West Germany and thereby systematically compare them in different macro-structural contexts. Specifically, we use the German Socio-Economic Panel (SOEP) to analyze family formation trajectories from age 15-35 of children born 1953-1978 and their mothers. In East Germany, most of the mother child dyads experienced their active family formation phase under vastly different macro-structural conditions, where the mothers established families during the early years of the communist GDR, whereas the daughters experienced most of their family formation in reunified Germany after 1990. Using multichannel sequence and cluster analysis, we establish three patterns of intergenerational transmission that correspond with those identified by Fasang and Raab (2014) for their selected sample of middle class American families: strong transmission, moderate transmission and intergenerational contrast. The family formation patterns of mothers and their children that display strong transmission experience the same process at the same speed, e.g. entering marriage and parenthood at the same age. Moderate transmission is exemplified by children that have similar family formation processes as their mothers, but delay family formation. Mothers and children with different family formation processes follow patterns of intergenerational contrast.

2

Second, we demonstrate that intergenerational transmission of family formation trajectories is not only quantitatively lower in East Germany compared to West Germany, but that the qualitative patterns of transmission also differ. East German children are significantly more likely to display intergenerational contrast or moderate transmission rather than strong transmission compared to West German children. We show that children's gender, parental educational level, children's educational mobility, and size of family of origin are significantly associated with the likelihood to display intergenerational contrast or moderate transmission rather than strong transmission.

Third, we propose to combine the sequence analysis approach with the Karlson, Holm and Breen (2010; Breen, Karlson, and Holm 2013) decomposition method (KHB method) for nested nonlinear probability models to uncover factors that mediate the differential patterns of intergenerational transmission in East and West Germany. We demonstrate that the increased likelihood for intergenerational contrast or moderate transmission in East Germany is accounted for by compositional differences in parents' educational level and children's educational mobility as well as the size of the family of origin, which are related to the different education and mobility systems, as well as family policies in East and West Germany for our study cohorts. The higher likelihood of contrast or moderate transmission patterns in the East may result from educational policies in the former GDR that increased parental education levels more than in the Federal Republic of Germany (FRG), while both states fostered upwards educational mobility for children.

2. The German Comparison

Between 1949 and 1990 Germany was divided into two sub-societies with marked contextual differences in the welfare state, economic system and family policies between the communist German Democratic Republic (GDR) in the East and the democratic social market economy in the Federal Republic of Germany (FRG) in the West. These differences set the stage for family formation of the mother generation in our analyses that experienced family formation in the early and middle years of the German division. In 1990 Germany was reunified by adapting the institutional model of the FRG to the former GDR. The child generation in our study experienced most of their active family formation phase in reunified Germany. Note that even 20 years after the reunification, a number of major macro-structural differences persist in the former East and West that remain relevant for family formation of the child generation (Goldstein and Kreyenfeld 2011).

3

However, the mother-child dyads in our East German sample all experienced a sudden break in political and economic regime. As a result the early adult life courses of mothers and children developed in vastly different opportunity structures. In contrast, there is far more continuity and gradual change in the macro-structural contexts in which the family life courses of mothers and their children in West Germany were situated for our study generations.

The state socialist system and centrally planned economy in the GDR promoted a classless society in which access to education and employment was strongly regulated by the state. In addition, mothers of our study generation were expected to combine work and family in a dual earner model (Engelhardt, Trappe, and Dronkers 2002). Family policies were strongly pro-natalist and conditioned access to state-controlled resources, including housing, on marriage and parenthood (ibid.). Normative pressure to have children was coupled with generous financial incentives for parenthood and practically universal public child care. Female labour market participation was around 90% (Huinink et al. 1995), but wages were rather low. In practice, marriage and parenthood often remained the quicker and more viable route to obtain housing and generous loans from the state than trying to accrue such privileges through labour market activity.

In contrast, the mother generation in the FRG experienced their active family formation phase in a society that was governed by a democratic multiparty parliament, functioned as a social market economy, and provided a corporatist conservative welfare state (Rosenfeld, Trappe, and Gornick 2004). The FRG represents a fairly unequal stratification system that was strongly transmitted through education from one generation to the next (Huinink and Mayer 1995; Pollak 2011). Social policies comprehensively set strong incentives for a traditional male breadwinner–female carer household division of labour and the financial dependence of women on men (Brückner 2004; Prince Cooke 2011). Family policies were foremost *pro-traditional*, including joint taxation of married couples that discouraged employment of a second earner and the absence of public child care particularly for children under the age of three (Prince Cooke and Baxter 2010; Aisenbrey, Evertsson, and Grunow 2009). The arguably purest implementation of the male breadwinner model (Prince Cooke and Baxter 2010) resulted in female labour market participation around only 50%, of which much was part time for the mother generation in our study.

In 1990 the West German institutional model was abruptly and unexpectedly applied to a population with markedly different compositional features in the East compared to the West. For

instance joint taxation for married spouses remained a much stronger incentive for marriage in the West, where earnings between husbands and wives were more unequal than in the East with its legacy of a dual-earner model. Initial expectations of a quick convergence of the two sub-societies did not materialize (Schneider, Naderi, and Ruppenthal 2012). Even in 2008, 18 years after the reunification, East Germany still had significantly lower rates of property ownership, lower average earnings, higher rates of female employment and a higher proportion of children in public care (Goldstein and Kreyenfeld 2011). While the intergenerational transmission of social status has increased in East Germany after the reunification, it remains notably lower than in the former West (Pollak 2011). Further, the former East continues to be one of the most secularized regions of Europe with 74 percent reporting no religious affiliation in 2008 compared to only 16 percent in the West (ibid: 457). While total fertility rates have converged (Goldstein and Kreyenfeld 2011), within Europe the two sub-societies remain on opposite ends of the continuum for a number of family demographic indicators, including a proportion of non-marital birth above 70 percent in the East compared with only around 12 percent in the West in 2007, and higher rates of cohabitation in the East. As a result opportunity structures of family formation remained markedly different for the child generation in our study in the former East and West. However, the sudden and massive change in institutional context, often referred to as the shock of reunification, was uniquely experienced by the mother child dyads in the former East, but not in the West (see also Fasang 2015).

3. Theoretical Background

In this section, we introduce three micro mechanisms that are theorized to facilitate the intergenerational transmission of family demographic behavior. Further, we theorize on how these micro mechanisms may generate different levels of intergenerational transmission within different macro-structural contexts, specifically the former GDR and FRG.

First, intergenerational transmission of family formation patterns may be a byproduct of intergenerational status transmission (McLanahan and Bumpass 1988). It has been shown, at least in western societies, that children "inherit" the socio-economic status of their parents to a certain degree (Breen 2004). Children are embedded in similar contexts and opportunity structures as their parents during their early life course, including similar partner and marriage markets as well as

similar socioeconomic conditions for childbearing. *Social status inheritance or immobility is expected to be associated with a pattern of strong intergenerational transmission.*

The 20th century was characterized by unprecedented structural shifts in national economies and educational systems, which facilitated upward mobility rather than mere socioeconomic inheritance or immobility for many children. Upward mobility is associated with delayed family formation, because individuals exit the educational system at later ages (Fussell and Furstenberg 2005). Further, long durations in tertiary education may also foster alternative family demographic behavior associated with post-materialism and the second demographic transition (Lesthaeghe 2010). *Upward status mobility is expected to be associated with a pattern of moderate intergenerational transmission or intergenerational contrast.*

Levels of intergenerational status transmission were lower in the "classless society" of the GDR (Huinink et al. 1995) and continue to remain lower than in West Germany even after reunification (Pollak 2011). *If the transmission of family formation is a byproduct of status transmission, then intergenerational transmission of family formation is expected to be stronger in West Germany.*

Second, parents may socialize their children and transmit values, norms and expectations specific to family formation. Parental influence through socialization may occur unconsciously when children internalize parental roles, or more explicitly through parental control of their children's actions by setting incentives (Barber 2000; 2001; Bernardi 2003). A central mode of value transmission with regard to "ideal" family formation patterns lies in the realm of religious practice and belief (Heaton and Goodman 1985; Berghammer 2012). *Religious parents are expected to be associated with strong patterns of intergenerational transmission.*

During state socialism, East Germany was the most secularized country in the world (Froese and Pfaff 2005) and remains so after the reunification (Goldstein and Kreyenfeld 2011) with the highest percentage of professed atheists. *If the central mode of family formation transmission occurs through the transmission of religious values, then intergenerational transmission of family formation is expected to be stronger in West Germany.*

Third, family structure during childhood may influence whether children decide to replicate the family structure during adulthood or choose alternative paths (Merz 2012). While some argue that socially accepted family structures are more likely to be transmitted from one generation to the

6

next, the literature on divorce transmission indicates that "deviant" family behavior is more likely to be transmitted to the next generation (Wolfinger 1999). Fertility transmission has been found to be strongest for higher-parity families, i.e. families with three or more children (Booth and Kee 2009; Fasang 2015), which can be conceived as an nonstandard family structure in societies characterized by low fertility (Boehnke, Hadjar, and Baier 2007). *Strong patterns of intergenerational family formation transmission are expected to be associated with higher-parity families of origin.*

Although the total fertility rate was higher in East compared to West Germany during state socialism due to pro-natalist policies (Fodor et al. 2002; Kreyenfeld 2004), higher-parity families were more common in West Germany (Goldstein et al. 2010). *If strong family formation transmission results from high-parity family structure during childhood, then intergenerational transmission of family formation is expected to be stronger in West Germany.*

4. Data & Methods

4.1 Definition of Family Formation Sequences

We use data from the German Socio-Economic Panel (SOEP) to operationalize the family formation trajectories of children and their mothers as sequences. The SOEP is a nationally representative longitudinal survey of households in the Federal Republic of Germany, which began collecting a variety of prospective and retrospective social, economic and demographic information in 1984 in West Germany and in 1990 in East Germany (Wagner, Frick, and Schupp 2007). We can utilize SOEP data to match the family formation processes of mothers and their children, because children born in SOEP households are followed after they leave the parental home and asked to found new, second generation SOEP households.

We identify 6,140 children born before 1978 to 3,995 SOEP mothers that can be potentially observed until age 35. Unfortunately, only 2,167 of these children participate in SOEP in their adult lives, and our analysis sample is further reduced to 1,524 (29%) mother-child dyads due to missing values on key variables (see Table 1). We use SOEP marital and birth biographical data to construct family formation sequences for each child and their mother. These sequences consist of 20 consecutive, annual states from age 15 to 35. Each state is defined as either single (S), married (M)

7

or divorced (D) without children or single with children (SC), married with children (MC) or divorced with children (DC). We restrict the sample to children born after 1952 to keep the cohort range relatively narrow. This allows for a more rigorous comparison between East and West Germany, because we ensure that parents experienced active family formation in divided Germany.

4.2 Dependent Variables – Sequence Distance & Cluster Analysis

The first aim of our paper is to determine the extent of family formation transmission between children and their mothers in East and West Germany. To this end, we first calculate pairwise *distances* between children's family formation sequences and the sequences of their mothers. We use dynamic Hamming's distance (DHD) to measure the dissimilarity between sequences. DHD assigns substitution operations time-dependent costs, which are inversely proportional to state transition frequencies (Lesnard 2010). Thus DHD pairwise distances will emphasize the timing of sequence states and their temporal order, which is important to capture intergenerational differences between family trajectories that stem from delayed family formation of children. Small distances indicate strong intergenerational transmission and large distances indicate weak intergenerational transmission.

Fasang and Raab (2014) demonstrated that different mechanisms are associated with different patterns of intergenerational transmission. A continuous measure of transmission that is based on comparing means for subgroups, such as distance, cannot account for the distribution of distance and may hide important qualitative differences in the transmission of family formation between children and their mothers. This would in turn make it difficult to identify factors that are associated with specific patterns of intergenerational transmission, if similar mean values in fact conceal very different distributions of distance or distinct qualitative patterns of transmission, e.g. early marriage and high parities or delayed and protracted family formation as two possible patterns of family formation. We use multichannel sequence analysis to compare every mother-child dyad with all other mother-child dyads (Pollock 2007; Gauthier et al. 2010) and generate a pairwise DHD distance matrix. We then enter this matrix in hierarchical cluster analysis, specifically the Ward method. According to cluster solution quality criteria, the mother-child dyads are reasonably well structured into three clusters (Kaufman and Rousseeuw 2008; ASW = 0.29). These cluster match those established by Fasang and Raab (2014) as *strong transmission*, *moderate transmission* and *intergenerational contrast*.

Table 1: Summary Statistics of Sample

	West	East	Germany
Mother-Child	38.98	44.73	40.40
DHD Distance	(19.85)	(18.02)	(19.57)
Cluster			
Membership			
Contrast	0.406	0.422	0.410
	(0.491)	(0.495)	(0.492)
Moderate	0.314	0.393	0.333
Transmission	(0.464)	(0.489)	(0.472)
Strong	0.280	0.185	0.256
Transmission	(0.449)	(0.389)	(0.437)
Parental	10.90	12.73	11.35
Education	(2.281)	(2.373)	(2.433)
Educational	1.697	1.020	1.530
Mobility	(2.684)	(2.510)	(2.657)
Religious	0.441	0.150	0.370
Attendance	(0.497)	(0.358)	(0.483)
Family Parity	3.116	2.364	2.930
	(1.744)	(1.219)	(1.662)
N*	1203	321	1,524
(%)	(75.34)	(24.66)	

Note: Averages and standard deviations displayed; *Frequencies not weighted;
Data weighted

4.3 Independent Variables & Analytical Strategy

The second aim of our paper is to investigate factors that account for the different degree of intergenerational transmission of family trajectories in East and West Germany. Mother-child dyads are considered *East* German if children were located in the former GDR in 1989 and West German if they were located in FRG in 1989.

We measure status transmission as highest *parental education* in years and children's *educational mobility* as the difference between children's education and their parent's highest education in years. Socialization and value transmission of traditional family demographic behavior is measured through an indicator of mother's *religious attendance*, which is one if mothers attend religious

9

services at least weekly and zero if mothers attend religious services at least monthly or less.[2] Family structure during childhood is measured as *family parity*, i.e. number of children in the family of origin. We also include the child's gender and birth year as additional control variables in the statistical analyses.

Our analytical strategy consists of three steps. First, we descriptively analyze the distribution of distance and cluster membership in East and West Germany to establish the extent of intergenerational transmission of family formation in both regions of Germany. Second, we estimate OLS regressions on mother-child sequence distance and multinomial logistic regressions on cluster membership to estimate the effects of parental education, educational mobility, mother's religious attendance and family parity on intergenerational transmission.

Finally, we use the KHB decomposition method for nonlinear probability models (Karlson, Holm, and Breen 2010; Breen, Karlson, and Holm 2013) to identify the factors that mediate the differential intergenerational transmission pattern membership in East and West Germany. Specifically, we use the KHB method to test whether compositional differences between East and West Germany with regard to parental education, educational mobility, religious attendance and family parity account for the differential likelihood of pattern membership between East and West Germany. While mediation effects can be estimated easily using OLS regressions by adjusting regression models stepwise, unobserved heterogeneity and constant error terms in logit and probit models prohibit the estimation of mediation effects in a similar fashion (Mood 2010). The KHB method enables the decomposition of unadjusted effects in multinomial logistic regression models into two effects: a direct effect, i.e. the adjusted effect, and an indirect effect, i.e. the difference between the unadjusted and adjusted effect. Studies commonly combine sequence analysis and multinomial regression analysis to explore associations between life course patterns and individual characteristics, which theoretically arise on account of specific institutional arrangements. We take advantage of the fact that institutional and cultural differences between societies generate the compositional differences between societies that are associated with specific life course patterns. The KHB decomposition method provides hitherto unknown insights on the institutions foremost associated with societal differences in life course patterns.

---

[2] We use the mode value if mother's religious attendance was observed more than once. If there were two modes, then we prioritize more frequent religious attendance.

5. Results[3]

5.1 Descriptive Results – Intergenerational Transmission in East & West Germany

There are substantial differences in the distribution of dyadic sequence distances between East and West Germany, as displayed in Figure 1. On average, distances between children's family sequences and the sequences of their mothers are larger in East Germany, which indicates weaker intergenerational transmission. Further, the distribution of mother-child sequence distance in West Germany resembles a bimodal distribution, where large portions of the population either exhibit very strong or very weak intergenerational transmission. In East Germany, the distribution is strongly negatively skewed, indicating that only a small proportion of the population exhibit strong transmission patterns.
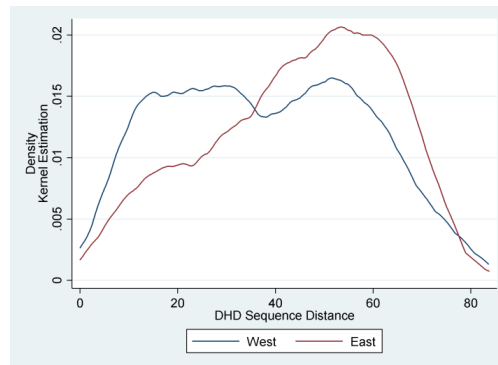
Fig. 1: Distribution of Mother-Child Sequence Distance in East and West Germany

The different distance distributions in East and West Germany are reflected in the distribution of cluster membership. The family formation sequences of mothers and children found in the three

---

[3] We use the `TraMineR` (Gabadinho et al. 2011) package to calculate sequence distances and the `WeightedCluster` (Studer 2013) package to perform cluster analyses on the sequence based distance matrixes in R, version 3.2.0. The decompositions are calculated using the `khb` (Kohler, Karlson, and Holm 2011) command in STATA, version 14.

clusters are displayed in Figure 2 as relative frequency sequence index plots that select representative sequences from each cluster to avoid visual distortions through over-plotting the graphs with too many sequences (Fasang and Liao 2013).[4] The family trajectories of the mothers in all clusters are very similar pointing to a strong standardization of family formation for the mother generation in line with previous studies (Brückner and Mayer 2005). They are all characterized by early marriage coupled with parenthood within marriage. Single parenthood and divorce are not common. The family trajectories of the children in the strong transmission cluster resemble the family formation patterns of their mothers, but single parenthood and divorce are somewhat more common. More than half of the children in the moderate transmission cluster marry, and most have children within marriage, but later than their parents. The cluster also comprises a notable share of childless marriage and single parenthood. The family trajectories of children in the intergenerational contrast cluster exhibit a dominant pattern of living single without children.

Strong Transmission

---
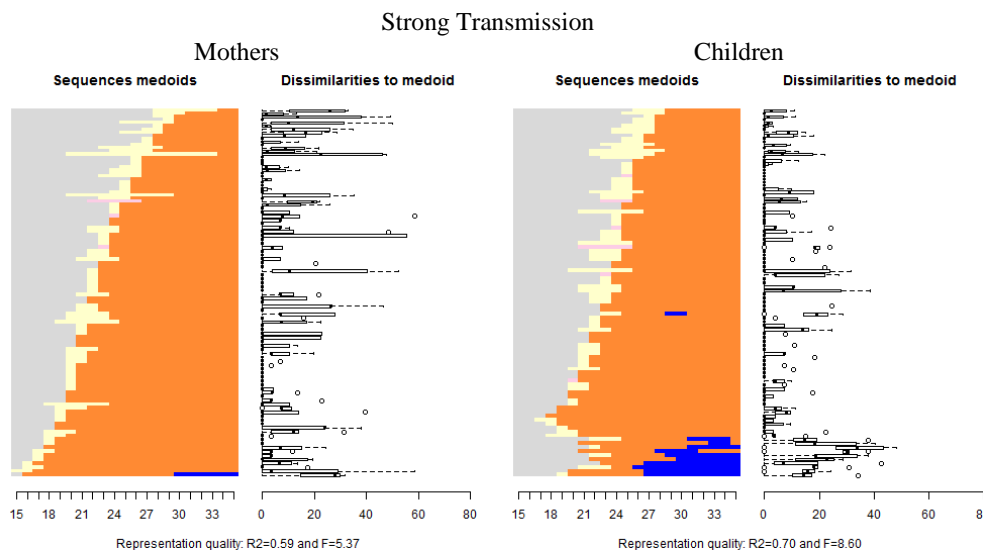
[4] Relative frequency sequence index plots solve the problem of "over plotting" individual sequences, i.e. different sequences occupy the same plot space, by displaying a representative subset of sequences. Relative frequency sequence index plots are generated by 1) sorting the sequences, 2) dividing the sorted sample into subgroups, 3) choosing mediod sequences from the subgroups, 4) plotting the mediod sequences and dissimilarities from the mediod sequences as box-plots. R^2 and F statistics that evaluate the goodness of fit of the chosen mediod sequences are additionally displayed. We sort the sequences using the first factor derived from multidimensional scaling and divide the sample into 100 subsamples. Mediod sequences are chosen using DHD distance. The plots were created with the `seqplot.rf` function developed by Matthias Studer, Anette Fasang and Tim Liao using R, version 3.2.0.
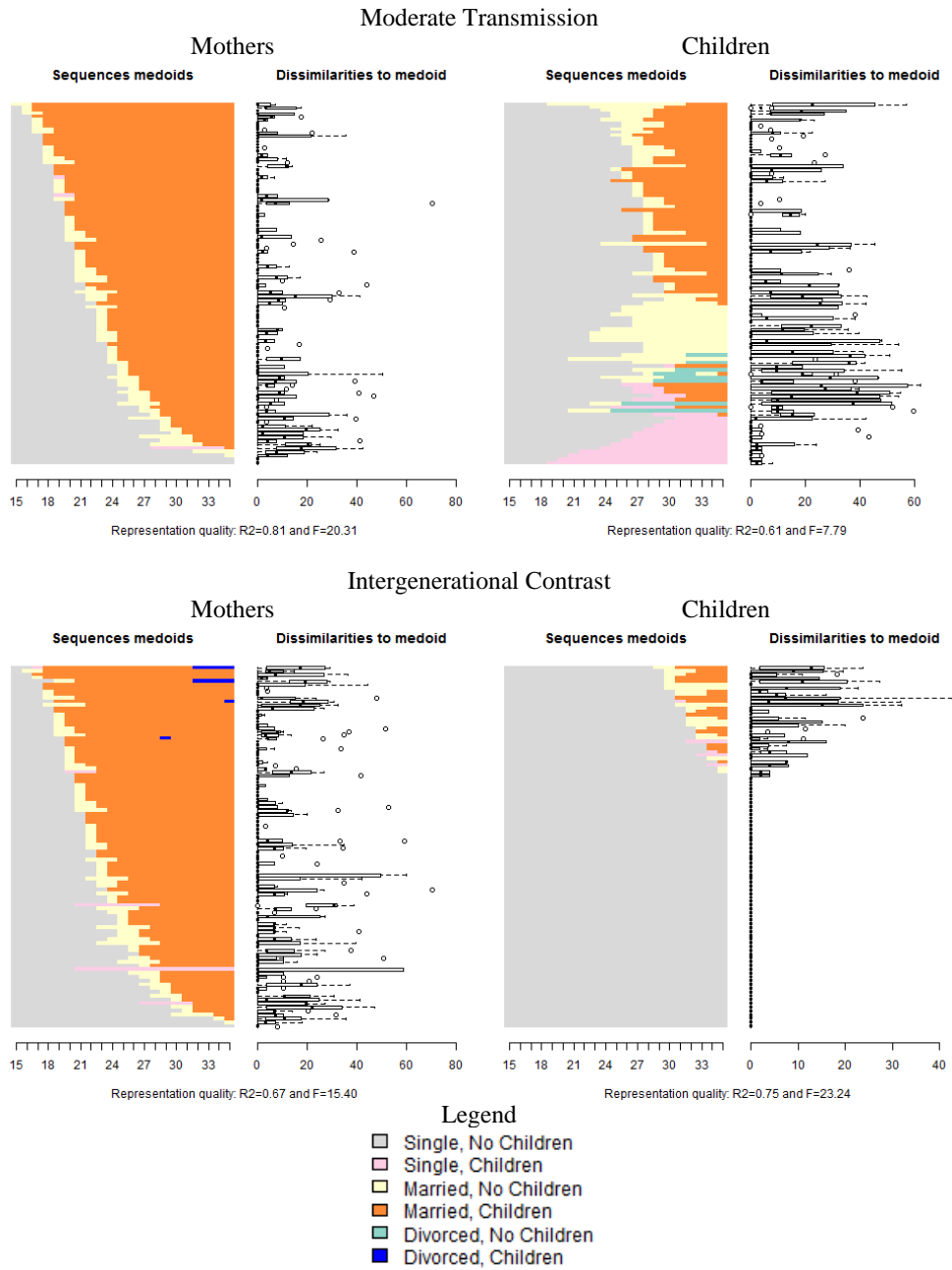
Fig. 2: Relative Frequency Sequence Index Plots of Strong, Moderate and Contrast Patterns of
Intergenerational Family Formation Transmission

The strong transmission pattern has the lowest average mother-child sequence distance (see Figure 3) and is much more common in West Germany than East Germany. While an estimated 28% of mother-child dyads display a pattern of strong intergenerational transmission in West Germany, only 18.5% of East German dyads belong to this cluster (see Table 1). Patterns of moderate transmission and intergenerational contrast are more common in East Germany than in West Germany, however the difference is quite small with regard to the intergenerational contrast pattern.
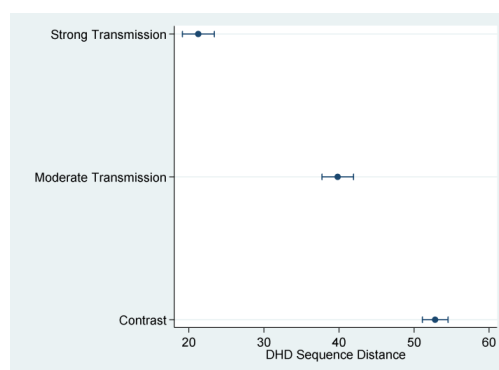
Fig. 3: Average Mother-Child Sequence Distance of Strong, Moderate and Contrast Patterns of Intergenerational Family Formation Transmission

5.2 Regression Results – Predicting Intergenerational Transmission

The OLS regression results on mother-child sequence distance are displayed in Table 2. The results corroborate the descriptive findings and indicate that, when adjusting only for children's gender and birth year, East German mother-child family formation sequences are significantly more dissimilar than West German dyads. Mother-son dyads are significantly more distant than mother-daughter dyads, and mother-child dyads have become significantly more dissimilar over time.

14

Table 2: OLS Regression on Mother-Child DHD Sequence Distance

	(1)	(2)	(3)	(4)	(5)
Region	5.318**	5.471**	3.757*	5.180**	3.938*
(Ref.: West)	(1.774)	(1.741)	(1.821)	(1.810)	(1.786)
Gender	5.442***	5.409***	5.730***	5.464***	5.715***
(Ref.: Female)	(1.530)	(1.528)	(1.495)	(1.522)	(1.486)
Birth Year	0.309*	0.321*	0.291*	0.305*	0.309*
	(0.149)	(0.149)	(0.144)	(0.151)	(0.145)
Parental		-0.029			-0.078
Education		(0.378)			(0.373)
Educational		-0.282			-0.209
Mobility		(0.406)			(0.384)
Educational		0.116*			0.114*
Mobility^2		(0.058)			(0.058)
Religious			-5.563**		-5.654**
Attendance			(1.771)		(1.769)
Family Parity				-0.198	-0.085
				(0.523)	(0.504)
Constant	-572.120+	-595.959*	-534.816+	-563.359+	-569.781*
	(294.073)	(292.376)	(283.305)	(297.985)	(284.388)
N	1524	1524	1524	1524	1524
R^2	0.039	0.044	0.057	0.040	0.063

Note: Sig: + p < .10, * p < .05, ** p < .01, *** p < .001; Unstandardized coefficients and robust standard errors in parentheses displayed; Data weighted

Children's educational mobility and religious attendance are significantly associated with mother-child sequence dissimilarity. In line with our expectations, educational mobility is nonlinearly associated with dyadic distance when adjusted for parental education: dyads are least distant for educationally immobile children, somewhat more distant for downwards mobile children and most distant for upwards mobile children. Further, mother-child dyads are significantly less distant for children whose mothers attended religious services at least weekly. The significant difference between East and West German dyads remains, although reduced, in a fully specified model.

15

Table 3: Multinomial Logistic Regression on Intergenerational Transmission Pattern

Membership

Contrast					
Region	0.495*	-0.043	0.392	0.315	-0.207
(Ref.: West)	(0.240)	(0.243)	(0.245)	(0.241)	(0.254)
Gender	1.284***	1.303***	1.311***	1.349***	1.367***
(Ref.: Female)	(0.207)	(0.210)	(0.205)	(0.208)	(0.212)
Birth Year	0.019	-0.008	0.018	0.014	-0.009
	(0.018)	(0.019)	(0.018)	(0.019)	(0.019)
Parental Education		0.444***			0.413***
		(0.076)			(0.079)
Educational Mobility		0.267***			0.257***
		(0.052)			(0.055)
Religious Attendance			-0.371		-0.397
			(0.227)		(0.244)
Family Parity				-0.265***	-0.139*
				(0.057)	(0.060)
Constant	-38.353	9.514	-36.170	-27.810	13.623
	(35.940)	(37.473)	(35.783)	(36.491)	(37.831)
Moderate Transmission					
Region	0.669**	0.208	0.687**	0.579*	0.210
(Ref.: West)	(0.258)	(0.257)	(0.258)	(0.259)	(0.263)
Gender	0.535*	0.551*	0.534*	0.566*	0.558*
(Ref.: Female)	(0.226)	(0.224)	(0.227)	(0.225)	(0.225)
Birth Year	0.004	-0.019	0.004	0.002	-0.019
	(0.019)	(0.020)	(0.019)	(0.019)	(0.020)
Parental Education		0.359***			0.343***
		(0.072)			(0.075)
Educational Mobility		0.194***			0.186***
		(0.053)			(0.054)
Religious Attendance			0.050		0.026
			(0.226)		(0.230)
Family Parity				-0.114*	-0.026
				(0.057)	(0.062)
Constant	-7.622	33.480	-7.380	-3.077	34.115
	(37.765)	(38.966)	(37.764)	(37.847)	(39.021)
N	1524	1524	1524	1524	1524
Pseudo R^2	0.034	0.083	0.038	0.047	0.091

Note: Sig: + p < .10, * p < .05, ** p < .01, *** p < .001; Unstandardized coefficients and robust standard errors in parentheses displayed; Strong transmission as base category; Data weighted

The results of multinomial logistic regressions on cluster membership are displayed in Table 3, which can be interpreted as a multiplicative increase in odds when the coefficients are

16

exponentiated. When adjusted for children's gender and birth year, the odds that East German mother-child dyads display intergenerational contrast rather than strong transmission increase by 64% compared to West German dyads. The odds that East German mother-child dyads display moderate transmission rather than strong transmission are 95% higher. The odds to display intergenerational contrast rather than strong transmission are 261% higher for mother-son dyads compared to mother-daughter dyads. For mother-son dyads, the odds increase by 70% to display moderate transmission. Patterns of intergenerational contrast and moderate transmission have not become significantly more or less likely than a pattern of strong transmission over time.

Parental education, children's educational mobility and family parity are significantly associated with transmission pattern membership, and the respective predicted probabilities are displayed in Figure 4.[5] The probability to exhibit a pattern of strong intergenerational transmission is highest for educationally downwards mobile children, those with low parental education and those who grew up with a large number of siblings. Intergenerational contrast patterns are most probable for educationally upwards mobile children, those with highly educated parents and those who grew up as single children. Patterns of moderate transmission are also more probable for educationally upwards mobile children, those with highly educated parents, but become slightly more probable for those who grew up with a large number of siblings.
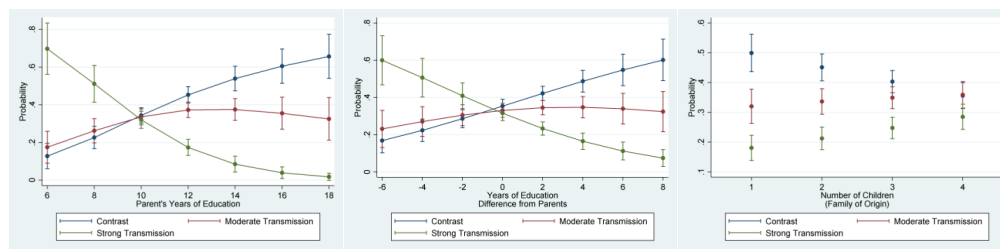

Fig. 4: Predicted Probability of Pattern Membership by Parental Education, Children's
Educational Mobility and Family Parity

---

[5] The predicted probabilities are all estimated using global sample means, expect for children's educational mobility. Educational mobility is modelled independent of parental education, although children of parents with university education cannot be upwards mobile and children of parents without any education cannot be downwards mobile. To account for this, we estimated the predicted probabilities for children's educational mobility using the corresponding mean of the observed parental education values.

5.3 KHB Decomposition Results – Accounting for Transmission in East & West Germany

In the models adjusted for parental education and children's educational mobility, East German mother-child dyads are not significantly more likely to display intergenerational contrast or moderate transmission rather than strong transmission. Further, East German dyads are not significantly more likely to exhibit intergenerational contrast rather than strong transmission if the models are adjusted for family parity. However, it would be a premature conclusion to state that these factors mediate the East/West effect on cluster membership. It is statistically possible that these results stem from unobserved heterogeneity or constant error terms rather than compositional differences. Therefore, we use the KHB decomposition method to test whether these insignificant East/West effects are attributable to compositional differences.

Table 4: KHB Decomposition Results of Multinomial Logistic Regression on Intergenerational Transmission Pattern Membership for East/West Effect

Effect	Education & Educational Mobility		Mother's Religious Attendance		Parity of Family of Origin	
	Moderate Transmission	Contrast	Moderate Transmission	Contrast	Moderate Transmission	Contrast
Total	1.684*	1.966*	1.642*	1.959**	1.648*	1.932**
	(0.438)	(0.523)	(0.392)	(0.505)	(0.393)	(0.492)
Direct	0.958	1.231	1.480	1.987**	1.371	1.783*
	(0.232)	(0.316)	(0.363)	(0.513)	(0.331)	(0.462)
Indirect	1.757***	1.597***	1.110	0.986	1.202**	1.083$^+$
	(0.283)	(0.226)	(0.074)	(0.062)	(0.080)	(0.050)
N	1524	1524	1524	1524	1524	1524

Note: Sig: + p < .10, * p < .05, ** p < .01, *** p < .001; Exponentiated coefficients and robust standard errors in parentheses displayed; Strong transmission as base category; Data weighted

The results of the KHB decomposition for the East/West effect are displayed in Table 4. The total, i.e. unadjusted, East/West effect is significant in all models. This means that East German dyads are significantly more likely to display intergenerational contrast or moderate transmission rather than strong transmission in all unadjusted models. The direct effects, i.e. East/West effect adjusted

18

for additional factors, are insignificant in all models, with the exception of intergeneration contrast adjusted for mother's religious attendance as well as parity in the family of origin. The significant indirect effects signify that at least a portion of the East/West effect is mediated through the adjusted factors. These effects are strongest for the models that adjusted for parental education and children's educational mobility.

6. Discussion & Conclusion

Our aim was to determine the extent of intergenerational family formation regularities in East and West Germany, and to investigate the factors that account for the transmission differential between East and West. We applied the conceptual and analytical framework of intergenerational patterns of family formation (Fasang and Raab 2014) to nationally representative data. Specifically, we used the German Socio-Economic Panel (SOEP) to analyze the family formation trajectories from age 15-35 of children born 1953-1978 and their mothers. We demonstrated that East German mother-child family formation trajectories are more dissimilar than West German mother-child family formation trajectories. Further, East German mother-child dyads are more likely to be categorized as intergenerational contrast and moderate transmission, whereas West German dyads are more likely to display strong transmission.

Based on a theoretical discussion of micro mechanisms that account for the intergenerational transmission of family formation, we developed the following expectations: First, social status immobility is associated with a pattern of strong intergenerational transmission and upward status mobility is expected to be associated with a pattern of moderate intergenerational transmission or intergenerational contrast. Second, religious parents are associated with strong patterns of intergenerational transmission. Third, strong patterns of intergenerational transmission are associated with higher-parity families of origin. According to these expectations, the results indeed show that upward educational mobility is associated with more distant mother-child family formation trajectories and a higher likelihood of intergenerational contrast and moderate transmission patterns. The family formation trajectories of mothers who attend religious services at least weekly are less distant, but the attendance of religious services is not associated with any pattern of intergenerational transmission. Large families of origin are not associated with the

distance between children's family trajectories and their mother's, but do increase the likelihood of strong intergenerational transmission.

We also discussed theoretically how differences between East and West Germany may account for differences in the extent of transmission. First, if the transmission of family formation is a byproduct of status transmission, then intergenerational transmission of family formation is higher in West Germany due to more mobility. Second, if family formation transmission occurs through the transmission of religious values, then intergenerational transmission of family formation is higher in West Germany due to higher levels of religiosity. Finally, if strong family formation transmission results from high-parity family structure during childhood, then intergenerational transmission of family formation is higher in West Germany due to a higher prevalence of high-parity families. Our results demonstrate that the East/West effect on mother-child family trajectory distance can be partially mediated by compositional differences in educational mobility and parity of family of origin. Further, we demonstrate that parental education and children's educational mobility completely mediate the East/West effect on intergenerational transmission pattern membership. Surprisingly, differences in parental religiosity between the two German regions did not contribute to the mediation of the differential patterns of transmission between East and West Germany.

We make three contributions to the existing literature. First, we applied the conceptual and analytical framework of intergenerational patterns of family formation transmission (Fasang and Raab 2014) to nationally representative data. Second, we demonstrate that intergenerational transmission of family formation trajectories is not only quantitatively lower in East Germany compared to West Germany, but that the qualitative patterns of transmission also differ. Third, we utilize the Karlson, Holm and Breen (2010; Breen, Karlson, and Holm 2013) decomposition method (KHB method) for nested nonlinear probability models to uncover factors that mediate the differential patterns of intergenerational transmission in East and West Germany. We believe that the proposed approach is promising to disentangle cross-national differences in life course domains more broadly.

In conclusion, the higher likelihood of contrast or moderate transmission patterns in the East may result from educational policies in the former GDR that increased parental education levels more than in the Federal Republic of Germany (FRG), while both states fostered upwards educational

mobility for children. Comparative life course sociologists have long pointed to differences in educational systems, but also to welfare state regimes and labor market institutions, to account for cross-national differences in the prevalence of life course patterns. Our results indicate that educational systems may account for cross-national differences in intergenerational regularities in family formation.

6. Works Cited

Abbott, Andrew. 2005. "The Idea of Outcome in U.S. Sociology." In *The Politics of Method in the Human Sciences*, edited by George Steinmetz, 393–426. Duke University Press.

Aisenbrey, Silke, Marie Evertsson, and Daniela Grunow. 2009. "Is There a Career Penalty for Mothers' Time Out? A Comparison of Germany, Sweden and the United States" 88 (December): 573–606.

Amato, Paul R. 1996. "Explaining the Intergenerational Transmission of Divorce." *Journal of Marriage and Family* 58 (3): 628–40.

Amato, Paul R., and Danelle D. DeBoer. 2001. "The Transmission of Marital Instability across Generations: Relationship Skills or Commitment to Marriage?" *Journal of Marriage and Family* 63 (4). National Council on Family Relations: 1038–51. http://www.jstor.org/pss/3599812.

Barber, Jennifer S. 2000. "Intergenerational Influences on the Entry into Parenthood: Mothers' Preferences for Family and Nonfamily Behavior." *Social Forces* 79 (1): 319–48. doi:10.1093/sf/79.1.319.

———. 2001. "The Intergenerational Transmission of Age at First Birth among Married and Unmarried Men and Women." *Social Science Research* 30 (2): 219–47.

Berghammer, Caroline. 2012. "Family Life Trajectories and Religiosity in Austria." *European Sociological Review* 28 (1): 127–44.

Bernardi, Laura. 2003. "Channels of Social Influence on Reproduction." *Population Research and Policy Review* 22 (5/6): 527–55. doi:10.1023/B:POPU.0000020892.15221.44.

Boehnke, Klaus, Andreas Hadjar, and Dirk Baier. 2007. "Parent-Child Value Similarity: The Role of Zeitgeist." *Journal of Marriage and Family* 69 (3): 778–92. doi:10.1111/j.1741-3737.2007.00405.x.

Booth, Alison L., and Hiau Joo Kee. 2009. "Intergenerational Transmission of Fertility Patterns." *Oxford Bulletin of Economics and Statistics* 71: 183–208. doi:10.1111/j.1468-0084.2008.00524.x.

Breen, Richard. 2004. *Social Mobility in Europe*. Oxford: Oxford University Press.

Breen, Richard, Kristian Bernt Karlson, and Anders Holm. 2013. "Total, Direct, and Indirect Effects in Logit and Probit Models." *Sociological Methods & Research* 42 (2): 164–91. doi:10.1177/0049124113494572.

Brückner, Hannah. 2004. *Gender Inequality in the Life Course: Social Change and Stability in West Germany 1975-1995*. Transaction Publishers.

Brückner, Hannah, and Karl Ulrich Mayer. 2005. "The De-Standardization of the Life Course: What It Might Mean and If It Means Anything Whether It Actually Took Place." In *The Structure of the Life Course. Standardized? Individualized? Differentiated?*, edited by Ross Macmillan, 27–54. Amsterdam: Elsevier.

Engelhardt, Henriette, Heike Trappe, and Jaap Dronkers. 2002. "Differences in Family Policy and the Intergenerational Transmission of Divorce : A Comparison between the Former East and West Germany." *Demographic Research* 6: 295–324.

Erola, Jani, Juho Härkönen, and Jaap Dronkers. 2012. "More Careful or Less Marriageable?

22

Parental Divorce, Spouse Selection and Entry into Marriage." *Social Forces* 90 (4): 1323–45. doi:10.1093/sf/sos073.

Fasang, Anette Eva. 2015. "Intergenerationale Fertilitätstransmission in Ost- Und Westdeutschland." *KZfSS Kölner Zeitschrift Für Soziologie Und Sozialpsychologie* 67 (S1): 11–40. doi:10.1007/s11577-015-0314-9.

Fasang, Anette Eva, and Tim F. Liao. 2013. "Visualizing Sequences in the Social Sciences: Relative Frequency Sequence Plots." *Sociological Methods & Research* 43 (4): 643–76. doi:10.1177/0049124113506563.

Fasang, Anette Eva, and Marcel Raab. 2014. "Beyond Transmission: Intergenerational Patterns of Family Formation Among Middle-Class American Families." *Demography*, August. doi:10.1007/s13524-014-0322-9.

Fodor, Eva, Christy Glass, Janette Kawachi, and Livia Popescu. 2002. "Family Policies and Gender in Hungary, Poland, and Romania." *Communist and Post-Communist Studies* 35 (4): 475–90. doi:10.1016/S0967-067X(02)00030-2.

Froese, Paul, and Steven Pfaff. 2005. "Explaining a Religious Anomaly: A Historical Analysis of Secularization in Eastern Germany." *Journal for the Scientific Study of Religion* 44 (4): 397–422. doi:10.1111/j.1468-5906.2005.00294.x.

Fussell, Elizabeth, and Frank Furstenberg. 2005. *The Transition to Adulthood during the 20th Century: Race, Nativity, and Gender. On the Frontier of Adulthood: Theory, Research, and Public Policy*. Chicago: University of Chicago Press.

Gabadinho, Alexis, Gilbert Ritschard, Nicolas S. Müller, and Matthias Studer. 2011. "Analyzing and Visualizing State Sequences in R with TraMineR." *Journal of Statistical Software* 40 (4): 1–37.

Gauthier, Jacques-Antoine, Eric D. Widmer, Philipp Bucher, and Cédric Notredame. 2010. "Multichannel Sequence Analysis Applied To Social Science Data." *Sociological Methodology* 40 (1): 1–38. http://doi.wiley.com/10.1111/j.1467-9531.2010.01227.x.

Goldstein, Joshua R, and Michaela Kreyenfeld. 2011. "Has East Germany Overtaken West Germany? Recent Trends in Order-Specific Fertility." *Population and Development Review* 37 (3): 453–72. http://www.ncbi.nlm.nih.gov/pubmed/22167811.

Goldstein, Joshua R, Michaela Kreyenfeld, Johannes Huinink, Dirk Konietzka, and Heike Trappe. 2010. "Familie Und Partnerschaft in Ost- Und Westdeutschland. Ergebnisse Im Rahmen Des Projektes 'Demographic Difefrences in Life Course Dynamics in Eastern and Western Germany.'" Rostock.

Härkönen, Juho, and Jaap Dronkers. 2006. "Stability and Change in the Educational Gradient of Divorce . A Comparison of Seventeen Countries." *European Sociological Review* 22 (5). doi:10.1093/esr/jcl011.

Heaton, Tim B, and Kristen L Goodman. 1985. "Religion and Family Formation." *Review of Religious Research* 4 (4): pp. 343–59. http://www.jstor.org/stable/10.2307/3511049.

Huinink, Johannes, and Karl Ulrich Mayer. 1995. "Gender, Social Inequality and Family Formation in West Germany." In *Gender and Family Change in Industrialized Countries*. Oxford: Oxford University Press.

Huinink, Johannes, Karl Ulrich Mayer, Heike Solga, Aage Sørensen, and Heike Trappe. 1995.

Kollektiv Und Eigensinn. Lebensverläufe in Der DDR Und Danach. Berlin: Akademie Verlag.

Karlson, Kristian B., Anders Holm, and Richard Breen. 2010. "Comparing Regression Coefficients Between Models Using Logit and Probit: A New Method." 0003. *CSER Working Paper Series*.

Kaufman, Leonard, and Peter J. Rousseeuw. 2008. *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley.

Kohler, Ulrich, Kristian Bernt Karlson, and Anders Holm. 2011. "Comparing Coefficients of Ested Nonlinear Probability Models." *The Stata Journal* 11 (3): 420–38. doi:The Stata Journal.

Kreyenfeld, Michaela. 2004. "Fertility Desicions in the FRG and GDR: An Analysis with Data from the German Fertility and Family Survey." *Demographic Research* 10 (SUPPL. 3): 275–318. doi:10.4054/DemRes.2004.S3.11.

Lesnard, L. 2010. "Setting Cost in Optimal Matching to Uncover Contemporaneous Socio-Temporal Patterns." *Sociological Methods & Research* 38 (3): 389–419. doi:10.1177/0049124110362526.

Lesthaeghe, Ron. 2010. "The Unfolding Story of the Second Demographic Transition." *Population and Development Review* 36 (2): 211–51.

Liefbroer, Aart C., and Cees H. Elzinga. 2012. "Intergenerational Transmission of Behavioural Patterns: How Similar Are Parents' and Children's Demographic Trajectories?" *Advances in Life Course Research* 17 (1). Elsevier Ltd: 1–10. doi:10.1016/j.alcr.2012.01.002.

McLanahan, Sara, and Larry Bumpass. 1988. "Intergenerational Consequences of Family Disruption." *The American Journal of Sociology* 94 (1): 130–52.

Merz, Eva-Maria. 2012. "Fertility Intentions Depend on Intergenerational Relations: A Life Course Perspective." *Family Science* 3 (3-4). Routledge: 237–45. doi:10.1080/19424620.2013.789976.

Mood, Carina. 2010. "Logistic Regression: Why We Cannot Do What We Think We Can Do, and What We Can Do about It." *European Sociological Review* 26 (1): 67–82. doi:10.1093/esr/jcp006.

Murphy, Michael. 2013. "Cross-National Patterns of Intergenerational Continuities in Childbearing in Developed Countries." *Biodemography and Social Biology* 59 (2): 101–26. doi:10.1080/19485565.2013.833779.

Murphy, Michael, and Duolao Wang. 2001. "Family-Level Continuities in Childbearing in Low-Fertility Societies." *European Journal of Population* 17 (1): 75–96. doi:10.1023/A:1010744314362.

Pollak, R. 2011. "Soziale Mobilität." In *Datenreport 2011. Ein Sozialbericht Für Die Bundesrepublik Deutschland*, edited by Statistisches Bundesamt and WZB, 180–87. Bonn: Bundeszentrale für politische Bildung.

Pollock, Gary. 2007. "Holistic Trajectories: A Study of Combined Employment, Housing and Family Careers by Using Multiple-Sequence Analysis." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 170 (1): 167–83. http://doi.wiley.com/10.1111/j.1467-985X.2006.00450.x.

24

Prince Cooke, Lynn. 2011. *Gender-Class Equality in Political Economies*. New York, NY: Routledge.

Prince Cooke, Lynn, and Janeen Baxter. 2010. "'Families' in International Context: Comparing Institutional Effects Across Western Societies." *Journal of Marriage and Family* 72 (3): 516–36. doi:10.1111/j.1741-3737.2010.00716.x.

Rosenfeld, Rachel A, Heike Trappe, and Janet C Gornick. 2004. "Gender and Work in Germany: Before and After Reunification." *Annual Review of Sociology* 30: 103–24. doi:10.1146/annurev.soc.30.012703.110531.

Schneider, Norbert F, Robert Naderi, and Silvia Ruppenthal. 2012. "Familie in Deutschland Nach Dem Gesellschaftlichen Umbruch." *Journal of Family Research* 9: 29–53.

Studer, Matthias. 2013. "WeightedCluster Library Manual: A Practical Guide to Creating Typologies of Trajectories in the Social Sciences with R." *LIVES Working Papers, 24*.

Wagner, Gert G, Joachim R Frick, and Jürgen Schupp. 2007. "The German Socio-Economic Panel Study (SOEP) – Scope , Evolution and Enhancements." *Schmollers Jahrbuch* 127.

Wolfinger, Nicholas H. 1999. "Trends in the Intergenerational Transmission of Divorce." *Demography* 36 (3): 415–20. doi:10.2307/2648064.

———. 2003. "Parental Divorce and Offspring Marriage: Early or Late?" *Social Forces* 82 (1). Oxford University Press: 337–53. doi:10.1353/sof.2003.0108.

Enduring contexts

Persistent segregation by affluence through the life course

Maren Toft

Department of sociology and human geography, University of Oslo

Abstract   This paper adheres to calls for incorporating time and place as important aspects in shaping contemporary inequality. Drawing on Sharkey's (2013) concept of 'contextual mobility'– understood as the social composition of one's neighbourhood over time – this paper seeks to investigate enduring structures of context in the Oslo region by following three complete birth cohorts left their parental home in 1989 and measure their social surroundings onwards until adulthood in 2012. The context of each individual is recorded annually with particular emphasis on the extremes in the social environments, i.e. the affluent versus the deprived areas. Utilizing sequencing and clustering techniques, the analysis shows that lasting exposure to reoccurring contexts characterize neighbourhood biographies of both dense advantage and disadvantage – indicative of vastly limited life experiences. Moreover, those who experience a neighbourhood career of dense advantage are also more likely to have privileged class origins and to obtain access into the upper class in adulthood and the inverse relationship characterizes the career of neighbourhood disadvantage. Those who are surrounded by persistent advantage are, however, the most isolated geographically and thus the more likely to be socially acquaint, although all neighbourhood typologies become more clustered in the region over time. I argue that an understanding of how spatially mediated contexts unfold throughout the life course hints to processes of class structuration and thus makes for an important addition to insight into present-day inequality.

1 Introduction

Emphasis on residential segregation among the rich and affluent has been implicitly or explicitly present in the heyday of studies investigating 'the power elite' (Mills 2000 [1956]) and the upper class (Baltzell 1958). Key to these classical studies is the emphasis on the entwinement of the privileged; evident in intermar-

2

riages, intermingling, shared social club membership and equal life experiences. In effect, the neighbourly environment was sought in part to reinforce existing privilege and facilitate social integration and recognition among the affluent and thus generate social closure (Pattillo 2008). While, however, the contemporary scholarly literature on residential segregation connotes a 'poverty paradigm' (Sampson 2012) or a '"social exclusion" framework' (Cunningham and Savage 2015) – in which the spatial has figured as an important factor in explaining the persistence in social *disadvantage* (Wilson 1987) – recent findings warrant a re-visiting of a spatial dimension at the opposite end of society as well. Indeed, the affluent are discovered to be equally – and sometimes even more – segregated than the urban poor (Reardon and Bischoff 2011), even in societies that are arche-typically 'egalitarian' such as Norway (Ljunggren and Andersen 2015; Wessel 2016). Hence, concentration by affluence along with poverty should be addressed when seeking to unveil the relationship between social and spatial inequality.

This paper seeks to dig further into how the spatial figures in systematic differences between the privileged and disadvantaged. Rather than measuring how segregated social groups are from one another at one point in time, I dissect dif-ferences in contexts that people go through during pivotal phases in their life time. More specifically, I seek to trace differences in the lived environments people ex-perience starting from when they leave their parental home in late teenage-years and annually onwards into their forties – spanning a time frame of 24 years. By giving emphasis to both the utmost privileged and impoverished areas – and studying the case of Oslo – I set out to explore the following research questions; how are neighbourhood careers structured with respect to affluent and poor envi-ronments and what kinds of people experience different contexts? How are such residential patterns mapped onto physical space; do different types of neighbour-hood careers evolve in close proximity or at more distant sites? And how are neighbourhood careers segregated from each other over time? Together these questions tap into structured differences in the lived experiences that people face over time, and their corresponding trends of differential association, which may cement further separation or integration along class divisions.

That social inequality is mediated through spatial processes is recognized in a number of studies, such as in Savage and colleagues' mapping of how the ge-ographical landscape is divided into enclaves of elites (Cunningham and Savage 2015) or how the middle classes make places serve as sites for a sense of belong-ing and identity-making (Savage et al. 2005). However, not just place, but also *temporality* is an important ingredient to the structuring of socio-spatial inequality. Whereas Savage (1988) once argued for a 'missing link' between spatial mobility and social mobility (for instance due to a relationship between relocation and job advancement), Sharkey (2013) has urged for research into the relationship be-

3

tween contextual mobility – i.e. changes in residential profiles, irrespective of physical movement – and social mobility. Although contexts are shown to be persistently transmitted across generations (Sharkey 2013, 2008), the unfolding of contexts throughout the life course remains understudied. Arguably, homogeneity in neighbourhood biographies is indicative of restricted base of experience that adds to the likelihood that the spatial gives rise to a sense of belonging and the creation of social ties, which in turn may elicit processes of socio-cultural class formation (Savage 1996; Dowling 2009).

Following the Scandinavian case of the Oslo region in a time of growing income and wealth inequality coupled with political initiatives to drastically deregulate the housing market (Wessel 2016), this paper exploits the fullness of Norwegian population-wide register data to scrutinize the relationship between social and spatial inequality. Utilizing social sequence analysis, differences between neighbourhood careers in years 1989–2012 are estimated taking account of duration of exposure as well as the ordering of contextual residence. Empirical typologies of contextual mobility are then depicted by usage of cluster analysis and segregation levels between these typologies are explored through spatialized adaptations to conventional indices of segregation, measuring both spatial evenness and spatial isolation[1].

2 Theory and existing knowledge

Residential areas are not just locations in a geographical space, but may convey tacit signals of whom and what belongs within the symbolic boundaries of a site (Hauge and Kolstad 2007), and arguably so when social classes are spatially segregated. In juxtaposition to occupational identity, Savage et al. (2005) argue that 'one's residence is a crucial, possible *the* crucial, identifier of who you are' placing moving decisions 'at the heart of contemporary battles over social distinction'. As emphasized by Bourdieu (1990), a sense of one's place often necessitates a sense of other's place and a relational structure of judgements and recognition embedded in the spatial may therefore assist in cementing recognized differences between class relations insofar as the different classes live apart. Hence, an understanding of the interplay between spatial and social inequality requires not only investigating differences in residential contexts, but also exploring what *types of people* live in particular social environments – tapping into possible ways through

---

[1] The sequence analysis is performed with the R package 'TraMineR' (Gabadinho et al. 2011), the analysis of maps is performed with the ArcGIS software, while the segregation indices are estimated with the R package 'seg' (Hong et al. 2014).

4

which the spatial may give rise to collective identity-making and thus forging so-
cial class formation (Savage 1996).

Of course, residential segregation alone cannot give rise to social class
formation, which is conventionally understood to be above all contingent on
whether or not mobility between class situations is 'easy and typical' between
generations and within the life-course (Weber 1978). As argued by Giddens
(1981); such 'easy and typical' mobility closure fosters a 'homogenisation of ex-
perience' that increases the likelihood of classes becoming 'social realities'. How-
ever, denoting such mobility patterns *mediate structuration of class,* Giddens in-
troduces the concept of *proximate structuration of class* as analytically
distinguished from the former. The latter encompasses inter alia an emphasis on
homogeneity in attitudes, behaviour, consumption patterns and lifestyles[2], remi-
niscent of Bourdieu's (1984) emphasis on the space of lifestyles in fortifying class
struggles. Homogeneous groups of this kind are labelled *distributive groupings* to
which communities and neighbourhood segregation are the most significant
(Giddens 1981). Hence, both mediate and proximate class structuration – and in
particular when coinciding – promote homogenisation of experiences that adds to
the likelihood that aggregates of individuals endowed with equivalently advanta-
geous life-chances establish social ties among themselves.

However, while much effort in stratification analysis is devoted to map-
ping patterns of class mobility, Sharkey (2008) draws attention towards intergen-
erational transmissions of *context*. Through the notion of 'contextual mobility'
Sharkey (2013) analyses changes in the socio-economic environment surrounding
individuals; an individual may experience contextual mobility if the social profile
in the neighbourhood changes over time, such as becoming more affluent or more
deprived, irrespective of residential mobility. Studying the relationship between
individual neighbourhood contexts and parental contexts, he shows that social mi-
lieus are largely transmitted between generations with important implications for
reproducing and solidifying (racial) inequality in the United States (Sharkey 2013,
2008).

This multigenerational emphasis on contextual reproduction entails in-
corporating time – in addition to place – as an important aspect of inequality; *tra-
jectories* of context are acknowledged as fundamental in reproducing spatially
embedded (dis)advantage. In the heart of neighbourhood studies lies an assump-
tion that the social profile of your neighbourhood may enable or constrain individ-

---

[2] Giddens distinguishes between two aspects of Weber's Stände; i) the formation of groups due
to commonality in consumption patterns and ii) a dimension of power derived from non-
economic relationships denoting 'honor'. Giddens terms the former phenomenon *distributive
groupings* and argues that it comprises one element in processes of *proximate structuration of
class* (Giddens, 1981: 109).

ual life-chances through for instance acquiring social capital, institutional availa-
bility or cultural adaptation (Sampson 2012; Wilson 1987). However, the likeli-
hood of such 'spatial profits' (Bourdieu 1999) has a temporal component; if
neighbourhoods contribute to reinforce advantage or disadvantage, such influence
is arguably more persistent with durable, rather than temporary exposure, Sharkey
(2013) argues. Correspondingly, if neighbourly environments facilitate social
recognition and symbolic distinctions within a classed topography, the likelihood
of such boundary-making practices hinge on temporality (Bourdieu 1999). Narra-
tives of belonging and classed identity-making are thus also embedded in *trajecto-
ries* of place-types which add to negotiations between the habitus and place at-
tachment (Benson 2014). Hence, a reorientation towards *processes* of spatial
inequality urges an examination of contextual mobility 'over individual lifetimes
and across generations of family members' (Sharkey 2013).

However, while recent contributions guide attention towards the latter re-
lationships – investigating the association across generations – less is known about
the contextual mobility within individual lifetimes. Further, the empirical ap-
proach for analysing contextual mobility mirrors classical mobility tables in class
analysis, where the relationship between parental positions and children's adult
life-experiences indicates a level of social reproduction across generations
(Sharkey 2013). Whereas these designs arguably contribute to important insight
about the persistency of cross-generational transmission of inequality, they rest on
an assumption that adult position measured at one point in time or averaged across
adult-life (e.g. Sharkey 2013, 2008), is a reliable indication of mature life-
experiences. But just like mature class positions at one point in time may only be a
provisional indication of position in adulthood (Bühlmann 2010), collapsing adult
experiences into one measurement may simplify the very time-dimension empha-
sized at the offset. Indeed, the need to sufficiently investigate the unfolding of
context throughout the life course by utilizing holistic techniques, such as se-
quence analysis, is advocated in studies of residential mobility (Coulter et al.
2015).

In housing studies, the call for moving beyond snapshots in time towards
neighbourhood trajectories is incipiently investigated by implementing novel se-
quencing techniques in a slight handful of studies (Stovel and Bolan 2004; Clark
et al. 2003; Pollock 2007; Coulter and Van Ham 2013; van Ham et al. 2014). Two
of these studies operationalize residential biographies as sequences of housing
tenure (Pollock 2007; Clark et al. 2003), while one investigates moving desires
and subsequent residential (im)mobility (Coulter and Van Ham 2013). The ques-
tion of attachment to place-types is, however, explored in the two remaining stud-
ies.

6

Reminiscent of Sharkey's later concept of contextual mobility, Stovel and Bolan (2004) investigate enduring attachment to *types* of places rather than movement across geographical locations in and of itself. They find that residential careers are largely maintained within specific place-types even when geographical mobility occur – argued to indicate that geographical localities are embedded in social boundary-making. However, this study operationalizes context in terms of population size (the main opposition being sparsely populated rural areas and large metropolitan cities) and calls for further study of sequential contextualization along socioeconomic and social dimensions (Stovel and Bolan 2004). A more recent study by van Ham et al. (2014) is of particular interest in this respect, as it deals with contextual socio-economic mobility in the Stockholm region using Swedish register data (although a random sample is drawn). Following adolescents from the year of leaving their parental home and onwards for the next 18 years, the study includes a depiction of sequences[3] of exposure to disadvantaged neighbourhoods. Affirming existing research, a close relationship between parental neighbourhood and individual neighbourhood histories are detected; the overall tendency is of intergenerational transmission of contextual poverty – reassessing the importance of including measures of spatially mediated transmission of disadvantage.

While there are few studies that investigate contextual mobility by means of sequence analysis, the general literature echoes a 'poverty paradigm' in urban sociology (Sampson 2012) and insights into the spatially embedded transmission of contextual advantage along with disadvantage require further analysis (Cunningham and Savage 2015; Toft and Ljunggren 2015; Atkinson 2006); particularly so during a time in which the urban landscape becomes increasingly unequal.

2.1 The case of the Oslo region

Arguably, the Oslo region, defined as Oslo and the surrounding county Akershus, is a geographical area in significant opposition to the popular conception of social democratic redistribution and egalitarianism. While the city's contemporary features evoke the concept of a post-industrial city, it once was site for industrial production with a manufacturing sector at its peak in the 1960 with 25 per cent of the workforce. The location of industry along the river parting the city into a west-end and east-end has cemented a symbolic distinction between an affluent west and the

---

[3] This study does not, however, utilize any alignment algorithms to analyse similarities in contextual experiences.

7

deprived east-side. This has become a highly recognized distinction in the city geography. Recent growth of the service industries, such as the financial industries, has been accompanying growing economic inequalities at the high end of the income distribution making the region much more unequal than the country at large (Wessel 2016). Furthermore, along growing concentration of affluence in the economic distribution follows increasing segregation by affluence in physical space. For instance, Wessel (2016) shows how the period of 1993–2011 was characterized by intensified polarization between the top and bottom quintiles of the income distribution and he documents especially widespread segregation by affluence – at levels surpassing corresponding measurements in cities such as Amsterdam, Rotterdam and Copenhagen[4]. Spatial confinement of privilege is also the concluding result of Ljunggren and Andersen (2015) investigation of residential segregation by class.

In the context of a Scandinavian welfare regime, Norwegian housing policies stand out; not only is Norway a country of 'home-owners' but contrary to prevalent social democratic policies, housing is largely commodified and residency is greatly left to the market. As noted by Brevik (2001), the characteristic social democratic emphasis on de-commodification, i.e. governmentally initiated relief of market dependency, diverges immensely in Norwegian housing policy. Right before the observational window of the present study, large changes in housing occurred and the early 1980s marked a turn in housing policy. Price regulation on housing was removed, legislative measures facilitated property speculation, credit was more readily accessible and general tax deductions by home ownership facilitated a boost of demand in the market and a corresponding increase in housing prices. Despite a severe market crash in the late 1980s and early 1990s the period of observation is one of a housing boom; exceptional even in a European context, housing prices in the Oslo region increased by 460 per cent in the 20-year period of 1992–2012. With hardly any growth of public housing and a weak system of housing provision, finding a home became arguably a matter of 'affordability' and increasingly so as the mismatch between household income and housing prices continued to grow in the same time period (Wessel 2016). The period under study is thus a time when economic capital increasingly stratifies the ability to choose where to live in the city, and when inequalities in such capacities are intensified due to growing income inequality. It should therefore be expected that lacking economic means to navigate the housing market makes for contextual poverty to be more persistent and stationary than durable affluence and for it to be unlikely that affluent areas become sites for being 'stuck in place' (Sharkey 2013) or

---

[4] However, Wessel does not find increased segregation among the top 1 per cent in the period under study.

8

'chained to place' (Bourdieu 1999) in a matter that might be more typical for the disadvantaged.

3 Methodology and research strategy

Having access to population-wide register data allows for a unique opportunity to investigate complete sequences in a period of twenty-four years (1989–2012). I construct an analytical sub-population of three successive cohorts to account for periodic idiosyncrasies such as initiatives in the housing market, housing availability and other period effects. Also, I condition on individuals who left their parental home at ages 18–20 after the first year of observation (i.e. 1990) and who reside in the Oslo region in the observational window[5] with non-missing residential information[6] – retrieving the cohorts of 1969 (45.67 per cent), 1970 (33.69 per cent) and 1971 (20.63 per cent).

Contexts are defined as the socio-economic composition of each annual neighbourhood. Neighbourhoods are measured as basic statistical units that represent geographically continuous and stable areas that are homogeneous with respect to business base, communication, and building structure (Statistics Norway 1999). The socio-economic environment of each neighbourhood is computed as the mean of the combined annual income (earnings, capital income and self-employed income) of the adult population (ages 30-60[7]) within each tract. Residential contexts are thereafter defined as the percentile distribution of the mean aggregated income among neighbourhoods in the region using cut points at p5, p20, p50, p80 and p95.

A sequence can be understood as a list of successive states and the overarching goal of sequence analysis is to estimate the level of similarities between pairs of sequences. The most often applied technique for calculating such similarities/dissimilarities is the optimal matching algorithm that takes account of not only the duration of states, but also the specific orderings of states. For instance, having long successive states of affluent contexts in late adolescence is arguably different

---

[5] However, a robustness analysis (not shown) reveals that an analysis that only conditions on Oslo/Akershus residency in 1989 but allows for subsequent vacating finds remarkably similar typologies as those visualized in the present analysis.

[6] Out of 31,056 elements in the sequences, 192 are imputed if a gap matches two preceding or following elements.

[7] I have access to complete birth cohorts starting from 1955 and their corresponding parents. This means that contexts are based on parental information and the 1955–1959 cohorts in year 1989, while more cohorts are included in the estimation of socio-economic contexts in subsequent years.

9

from equally lengthwise exposure to affluent surroundings in one's forties. Hence, not only duration is important in estimating similarities of pairwise sequences, but also the specific ordering of states comes into play. Optimal matching relies on finding the least expensive route to changing one sequence into another by usage of three basic operations; substitution and insertion/deletion (indel). Each of these operations requires an assigned cost that makes one or the other operations more likely in the matching procedure. Setting a substitution cost matrix – i.e. the cost of each state-pair substitution – is often considered to be one of the most pivotal tasks in such analyses as it directly influences the resulting distance matrix (i.e. the level of similarity/distance between pair-wise sequences). As this cost scheme is arguably arbitrary, it is also a core feature for criticisms of sequence analysis. In this analysis I follow an often-applied procedure of using a transition probability matrix when assigning such costs, which implies that transitions between states that are less likely to occur are more costly than substituting state-pairs that are frequently successive (Cornwell 2015). In order to maintain alignment based on timing of events, the indel is set so that no substitution costs are greater than two times the indel cost (Aisenbrey and Fasang 2010). As a means to generate typologies of the resulting distance matrix from the optimal matching alignment, I employ cluster analysis using a combination of the Ward-linkage and PAM (partitioning around the means) (Studer 2013). The usage of a Bourdieu-inspired class scheme that distinguishes vertical and horizontal class divisions is utilized to investigate the differences between such typologies in respect to both parental and individual class situations (Hansen et al. 2009).

Crucially, the results presented hinge on a number of methodological choices. First, the substitution cost matrix is arguably of great importance and an alternative approach, using the percentile cut-offs as a cost matrix, has been investigated as a means to validate the results. This approach yields a very similar patterning as the one detected with the presented substitution cost matrix. Second, validation of the chosen cluster solution is also warranted, but the results do not appear vulnerable to a different linkage criterion such as an average linkage (results not shown) or different cut-offs in the neighbourhood distribution. Lastly, validation of the number of clusters is warranted and the validity of the clusters were inspected along a number of indices as suggested by Studer (2013). The chosen solution of 4 was prioritized by most indices and corresponds favourably with the substantive interest in the analysis. The ward linkage in combination with PAM proved most useful. It should, however, be noted that the cluster solution yields a fairly low AWS of about 30, and thus a fairly weak clustering of the data, although statistically significant differences in the between variance of the groups are pertinent. See the appendix for a visualisation of sequences that are distant to the gravity centre of each cluster. As seen, however, even the sequences that are

10

"poorly" represented by the clusters in a statistical sense are assigned to substantively meaningful typologies.

The usage of sequence analysis and cluster typologies enable a holistic approach to understanding what Sharkey (2013) terms 'contextual mobility', intragenerationally. While contextual mobility does not necessitate physical relocation, the extent to which sequence typologies are embedded in physical proximity or distance adds to an understanding of how social inequalities are spatially mediated. In order to address questions of physical proximity in a life-course perspective properly, a mapping of each cluster onto geographical coordinates is pursued. Distances in place are further investigated through estimates of residential segregation among the various typologies.

There are a number of indices that assist in identifying segregation among groups in a geographical space. However, the vast majority are *aspatial* in the sense that they do not account for the spatial boundaries that make up the localities to which segregation levels are estimated and solely focus on the extent to which social environments are different across people. Neglecting the construction of, for instance, census tracts may disregard tract proximity and the arbitrary boundaries that make individuals adjacent to a tract boundary erroneously more distant than individuals within the same census tract. In an effort to circumvent such issues, Reardon and O'Sullivan (2004) propose a number of spatialized versions of existing segregation indices. Distinguishing between spatial isolation (the extent that members of one group encounter members of their own group) and spatial clustering (the extent to which groups are unevenly distributed in the geographical unit), they propose a *spatial isolation index* $\tilde{p}^*$ to estimate the former, and a *spatial information theory index* $\tilde{H}$, a *spatial relative diversity index* $\tilde{R}$, and a *spatial dissimilarity index* $\tilde{D}$, to estimate the latter. These measurements utilize geographical data that incorporates a 'distance-decay' effect, namely that locations in close proximity are of larger influence to a shared social environment than distant sites. This is achieved by centring all individuals within a neighbourhood to the geographical location of its centroid. As the geographical level of neighbourhoods is computed at a very small scale (median km^2 = 0.31) this practice arguably remains satisfactory[8] and the results are robust when utilizing different spatial interpolation by assuming a uniform distribution of individuals within tracts.

---

[8] It should, however, be acknowledged that MAUP (modifiable areal unit problem) may not be completely circumvented due to the lack of accurate point-data. In addition, areas at the outskirts of the geographical region may be vulnerable to the computation due to an artificial demarcation of adjacent areas. However, as the Oslo region is of particular interests, the edges of the surrounding county Akershus are arguably of lesser analytical interest.

11

4 Results

Figure I depicts the four typologies by usage of state distribution plots denoting the relative shares of each socio-economic category within the clusters over time. The four typologies reveal an opposition between neighbourhood biographies within affluent or deprived areas, but differentiate between the extremes of the distribution and a more modest opposition along neighbourhood income. A modest opposition is detected between the first (n=420) and fourth (n=409) cluster, comprising almost 2/3 of all individuals in the analytical sub-population, while the remaining individuals are differentiated into two groups of dense affluence (n=156) and poverty (n=309). To ease interpretation these clusters are henceforth titled *dense poverty*, *modest poverty*, *dense affluence*, and *modest affluence*.
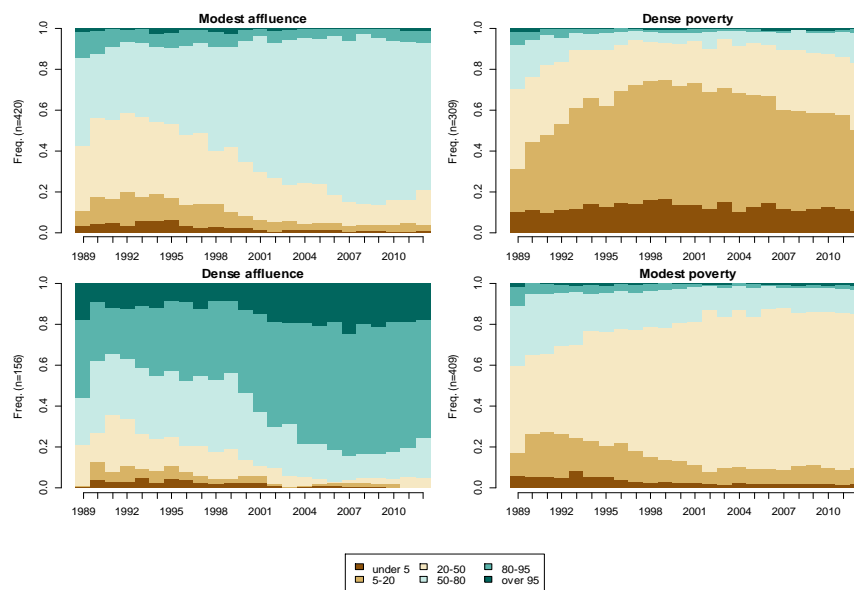
Fig. 1: Sequence distribution plot

The first entries in these plots denote the parental home environment as the research design intentionally captures individuals who leave after the first year of observation (in year 1990). Arguably, these plots unveil how neighbourhood biographies are related to parental neighbourhoods; at one extreme we find that 17 per cent of the *dense affluence* cluster grew up in the 5 per cent richest neighbourhoods in the Oslo region – 56 per cent in the top quintile. At the opposite trajectory we find 10 per cent originating in the bottom 5 per cent and 32 per cent in the bottom quintile. This points to an important addition to the scholarly literature of

12

intergenerational transmission of contextual disadvantage; equivalent associations at the high-end of the distribution appear equally persistent.

Within each cluster, leaving the parental home is associated with relative deprivation in the neighbourly context. This is especially apparent within the cluster of *dense affluence* where it may be particularly difficult to sustain the high share of affluent living and in their twenties and early thirties about 20-30 per cent within this cluster live in neighbourhoods below the median. Inspecting the temporal dimension within these clusters we see that the two clusters of *dense* and *modest affluence* are characterized by an upwardly mobile trajectory, while the opposite holds for the two latter typologies. Despite the fact that those whose contextual biography is dominated by *dense affluence* originated in very advantageous contexts, their adult-life residences are even more affluent; the later years reach levels of more than 80 per cent in the top quintiles and a max value of 25 per cent in the richest 5 per cent neighbourhoods. Among those of *dense poverty* we find less fluctuating shares of the bottom 5 per cent category – about 13 per cent within this cluster reside in such areas over time – but the combined bottom quintile is in modest decline at the end of the period. Among the two latter clusters, the categories right below and above the median become more dominated through adult life, with roughly 80 per cent of the *modest affluence*-category and the *modest poverty*-cluster living in the 30 percentile points above and below the median, respectively.

	Dense Affluence	Modest Affluence	Modest Poverty	Dense Poverty	Total
Parental class, per cent					
Upper class: culture	9.91	0.71	2.93	0.65	2.51
Upper class: symmetric	17.01	7.38	5.38	1.62	6.53
Upper class: economic	8.97	1.9	0.49	0	1.85
Upper middle class: culture	6.81	1.43	2.2	2.59	2.64
Upper middle class: symmetric	5.13	11.19	9.78	6.15	8.71
Upper middle class: economic	22.68	15.71	9.78	9.71	13.3
Lower middle class: culture	2.3	5.48	2.93	2.91	3.65
Lower middle class: symmetric	3.85	10.24	10.27	11.65	9.7
Lower middle class: economic	3.47	3.81	8.56	8.41	6.52
Skilled working class	12.18	24.05	22.98	24.27	22.33
Unskilled working class	6.41	16.67	23.47	28.16	20.32
Farmers, fishery	0	0.24	0.49	0.32	0.31
Missing	1.28	1.19	0.73	3.56	1.62
Total	100.00	100.00	100.00	100.00	100.00
Class position 2012, per cent					
Upper class: culture	3.85	1.43	1.47	0	1.39
Upper class: symmetric	3.85	2.14	0.24	0.65	1.39
Upper class: economic	5.13	2.62	0.73	0.65	1.85
Upper middle class: culture	5.77	2.14	2.44	2.27	2.70
Upper middle class: symmetric	12.18	13.1	13.69	11.97	12.91
Upper middle class: economic	17.95	15	6.85	7.12	10.90

13

Lower middle class: culture	1.92	2.38	2.69	1.62	2.24
Lower middle class: symmetric	6.41	12.38	16.87	13.59	13.37
Lower middle class: economic	10.26	11.67	11.25	8.41	10.59
Skilled working class	3.85	14.05	14.67	14.24	13.06
Unskilled working class	7.05	9.05	14.67	14.89	11.98
Welfare transference	10.26	10	8.07	19.42	11.67
Missing	11.54	4.05	6.36	5.18	5.95
Total	100.00	100.00	100.00	100.00	100.00
Income, neighbourhood 1989					
Mean	169,929	149,350	144,1955	139,249	147,7899
Income, neighbourhood 2012					
Mean	1,012,349	739,884	632,580	584,9522	701,789
Observations	3,744	10,080	9,816	7,416	31,056
Individuals	156	420	409	309	1,294
Per cent	12.06	32.46	31.61	23.88	100.00

Table I includes information about the class origins and class destinations, as well as contextual income measurements, distinguishing the four typologies. There are large differences between these typologies with respect to class origin. More than a third of everyone who experiences the *dense affluence* trajectory has origins in the upper class fractions, and this large share is especially distinctive in a country-wide comparison (in which 5.5 per cent in the 1969-1971 cohorts originates in the upper classes). The *dense poverty* trajectory is similarly disproportionately from modest origins – more than half are from the unskilled or skilled working class – but this overrepresentation is of lesser magnitude in comparison to the remaining analytical population. Also, the *modest affluence* and *poverty* typologies have origins that match their contextual trajectories.

The cluster, *dense affluence,* is also characterized by disproportionately large shares of individuals who obtain upper class position in the year 2012. As shown in table I, shares of upper class fractions amount to 13 per cent, while the corresponding share within the total analytical population adds to 4.5 per cent. Within the typology of *dense poverty* we find that a disproportionately large share belongs to the welfare transference group, and 49 per cent obtain either working class positions or transference. Again, the class character of the two remaining typologies is patterned as expected.
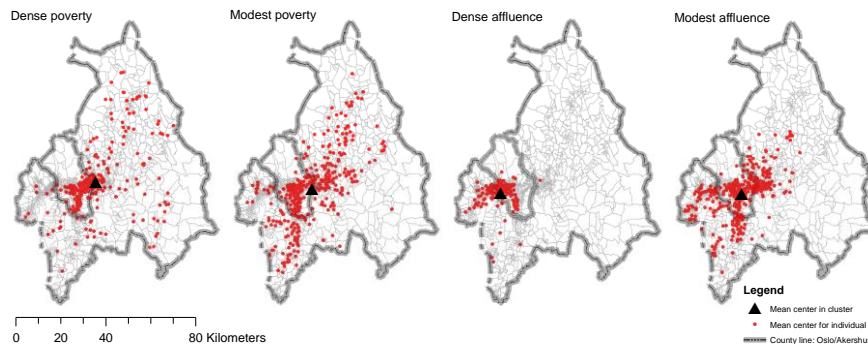
14

Fig. 2: Maps of the Oslo region. Mean center for each individual biography, by cluster typology

      In what ways are these typologies embedded in physical space? Figure III pursues this question by mapping each cluster. Each individual's typical site for dwelling is indicated by a red circle. Figure III clearly illustrates that these contextual biographies are not only different in the social profiles of the social surroundings to which people are situated, but also the ways by which such biographies are embedded in physical space. Tendencies towards an opposition between the west- and eastside of Oslo separate the two poverty-clusters from the affluence-clusters. The affluent municipality *Bærum* which borders Oslo's west side is for instance seldom site for mean localities among the two former, while almost one quarter of all localities in the cluster of *dense affluence* are located there. The mean centre in each cluster, represented by the black triangle in the map, may assist in identifying further the most typical location for each contextual typology; Oslo dominates all trajectories, but the biographies of *dense poverty* and *affluence* are centred in the east and west, respectively, while the more normal biographies of *modest affluence* or *modest poverty* are situated closer to the city centre while still connoting an east-west opposition.
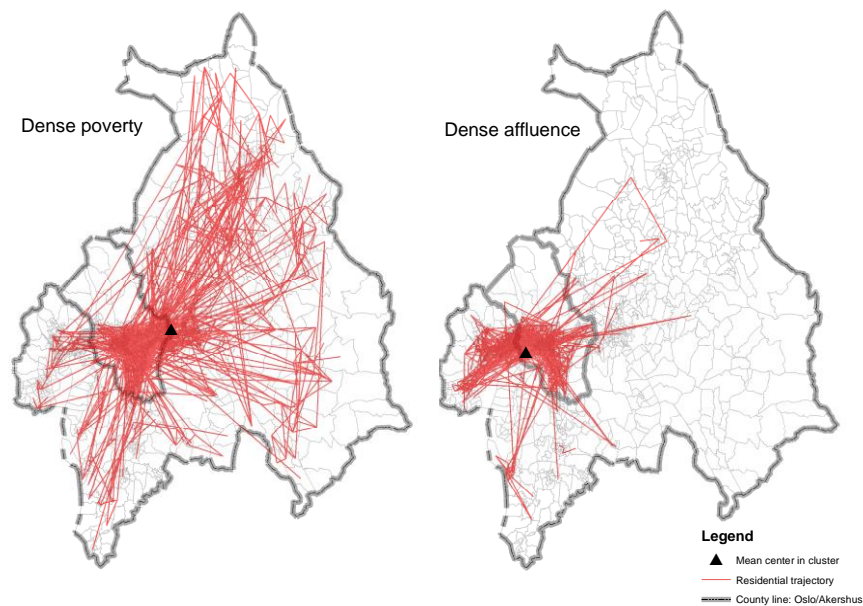
Fig. 3: Individual lines by residential trajectory. Dense poverty and dense affluence careers.

15

Differences are not only evident in *where* the clusters are located through time, but also the *scope* of geographical distance inherent in each typology. Figure III depicts all residential movements within the two clusters of *dense poverty* and *dense affluence* by individual lines per residential history. This figure greatly opposes the theoretical expectation of immobility in areas of poverty. What is observed is instead that the cluster of *dense affluence* is highly concentrated in a very small geographical region over time, while the typology of *dense poverty* is more mobile in the region.

While these maps add to our understanding of the ways in which contextual experiences are embedded in physical space, they completely conceal temporality. There is no information of whether specific localities are sites of residence in young adolescence or adult life. Figure IV therefore seeks to unveil a time dimension when assessing the level of spatial segregation between these groups. The top graph depicts three spatial segregation indices for measuring spatial evenness. These indices give an indication of how less diverse the local environments are on average in comparison to the total region. A value of 0 connotes complete evenness in respect to the region at large, while the value 1 reflects complete segregation (Reardon and O'Sullivan 2004).
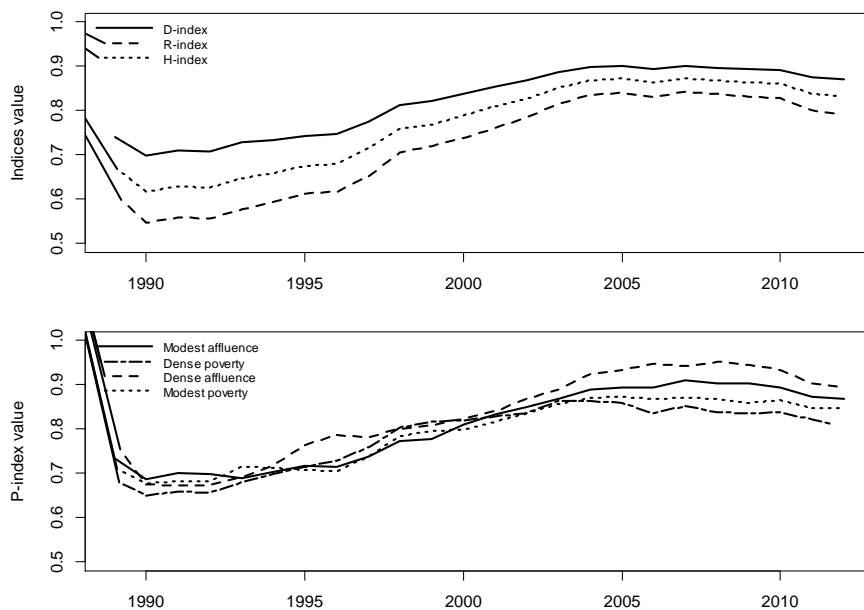
Fig. 4: Segregation indices. Top graph depicts trends of spatial evenness, while the bottom graphs estimates spatial isolation.

16

Intending to reflect the similar social phenomenon of spatial evenness or clustering, we find that all indices follow a similar shape across time: a tendency towards spatial evenness succeeds the move away from parental home, but a growing trend towards spatial clustering follows immediately. These typologies are highly clustered in the Oslo region, reaching very high segregation measures after 2005. It should be noted that the time of study is one of growing segregation by income, and that the spatial clustering of these typologies thus follow a common trend in the area. Nonetheless, the levels are arguably quite high although a slight downwards trend is detectable around year 2009.

Noting how these neighbourhood typologies are spatially clustered in the region, the bottom graph turns to a measurement of the level of isolation among these groups. Again, this index reaches a max value of 1 when a cluster is completely isolated, and has a minimum of 0. This index is based on the average composition of the local environment of each cluster and we observe that all clusters become increasingly more isolated over time although the trend is decreasing around year 2009. The typology of *dense poverty* is the least spatially isolated, while the two clusters of *modest* and *dense affluence* are the most isolated over time, and particularly high levels are detected for the group of *dense affluence* at adult age. Here the index succeeds 0.9, which is a very high index-value indicative of a very high probability for individuals who throughout life experience great affluence to be socially acquaint.

5 Discussion and conclusion

Investigating differences in the social contexts in which people live their lives adds an important dimension to understanding how structures of inequality have both a *durable* and a *spatial* manifestation within individual biographies. This analysis has shown that even in an 'egalitarian' welfare state, people go through life exposed to very different surroundings. Although the majority are differentiated by the relative prevalence of contexts in the 30 percentile points below or above the median, about one quarter live in enduring contexts of persistent poverty or dense affluence. These trajectories not only give evidence of homogeneity in intragenerational biographies, but the first-year characterization of the parental home environment hints to an additional intergenerational transmission of context. About 40 per cent of those who experience sequential environments of *dense affluence* or *dense poverty* originate in neighbourhoods in the top or bottom quintile. Hence, a relationship between parental contexts and their offspring's surroundings is not restricted to transmission of disadvantage as emphasized in existing re-

17

search, but applies in a similar vein to affluent milieus. There are further important aspects of inequality that are being neglected if one sidesteps investigations of concentrated affluence and restrict the analysis to a 'poverty paradigm' of spatial exclusion and marginalization. Arguably, incorporating concentration of wealth reorients analytical attention towards the strategic ways by which the spatial also figures in processes of social closure and class advantage.

Not only does affluence appear to be just as contextually bounded as poverty, but the spatial scope of affluent living is significantly more limited in its geographical reach than the boundedness of contextual poverty. Although all typologies are clustered from each other in the region, and increasingly so over time, trajectories in affluent environments are more isolated – especially evident at high levels in adulthood. Individuals who resemble in their persistently affluent environments are thus in close proximity in physical space, largely confined to the west-end of Oslo and its westerly bordering municipality, Bærum. This implies that those who resemble in their affluent neighbourhood career from late adolescence until their forties are likely to reside in close proximity to one another and to be socially acquainted. This typology is especially isolated in mature life, and hence the likelihood of meeting someone with similar contextual experience is very high in such areas, and this biographical affinity becomes more pivotal when also considering the class character of this group.

Indeed, systematic variations in contextual experiences are shown to be particularly *classed*. Those who follow pathways through *dense advantage* disproportionately originate in the upper classes and obtain upper class position in adulthood and this dual tendency of social and physical proximity may ease the likelihood that the spatial become sites for further acquisition of capitals (Bourdieu 1999). Conversely, individuals who face dense poverty throughout the life course are more likely to originate and partake in the working classes or having to rely on welfare transference. The relationship between class origins and class destinations is likely to facilitate 'homogeneity in experiences' (Giddens 1981) which is arguably reinforced due to prolonged experiences in comparatively disadvantaged/advantaged surroundings through life. As such, borrowing the terminology of Giddens, observed processes of 'proximate' as well as tendencies of 'mediate' structuration of class may have important implications for social class formations in the region.

That the temporal confinement of privilege is more restricted than its disadvantaged counterpart appears somewhat contrary to the theoretical expectations that the disadvantaged would be more *chained to place* (Bourdieu 1999) or *stuck in place* (Sharkey 2013). Instead, what we observe is not, exclusively, that lacking resources makes people *stuck in place*, but it appears that economic means also allow for a desired and intended spatial anchorage at the opposite end of the distri-

18

bution (Atkinson 2006). Indeed, it appears likely that the intended *immobility* of the affluent in part stimulates the mobility of the disadvantaged, as growing income inequalities follow increasing housing prices and gentrification processes in the urban centre, which may have pushed the disadvantaged towards the outskirts of the region.

An emphasis on voluntary immobility does not, of course, necessitate a conscious strategy for moving decisions. As Bourdieu (2005) reminds us, moving decisions are structurally constrained or enabled by economic means, but also one's dispositions and, importantly, the political construction of housing policies – which may influence both factors – weigh into residential patterns. Indeed, the political design of a deregulated housing market appears to offer important ways by which social distance is transmitted into physical distance, suggestively by allowing for a spatially confined area where economic means become a key criterion for access. Thus, the politically initiated housing market may serve as a closure mechanism by making housing a matter of 'affordability'.

The temporal boundaries of confined affluence may also be sustained by informal closure mechanisms. Existing research emphasises how residency may serve as an expression and identity, and this self-identifying 'sense of place' may be more strongly articulated in affluent areas, as shown elsewhere in Norway (Hauge and Kolstad 2007). Given the very limited parameters of spatially bound affluence, a similar anticipation appears reasonable in the Oslo region. Arguably, the more signs of identification are attached to place, the more likely it is that symbolic distinctions serve to maintain and reinforce symbolic boundaries, for instance through outsiders' negative selection due to risks of feeling 'out of place' (Bourdieu 1999). Hence, sentiments of "classed belonging" could serve as a useful mechanism of informal closure by cementing exclusionary practices that help enclose affluent living from the remaining public.

The holistic mapping of sites for both dense affluence and dense poverty in the present analysis leaves a depiction of the Oslo region as not only a segregated area as documented in previous research, but as a region in which distinctive life biographies unfold parallel to one another but seldom coincide. It may appear paradoxical that an otherwise 'egalitarian' country such as Norway seems to facilitate prolonged and extensive concentration of both privilege and disadvantage. Perhaps this very physical imprint of differentiated life-trajectories in the region becomes in part concealed by the very *atypical* nature of housing policies. Arguably, the more social distances are spatially embedded over time, the more likely a 'naturalization effect' comes about (Bourdieu 1999) and this may help to further reproduce and legitimize social inequalities. Given the historical context observed, it should be recognized that the present mapping of contextual mobility patterns may not apply to other life stories whether earlier or subsequently. However, as

19

economic inequality and housing prices continues to increase in the region these stories may also be suggestive for life-stories to come.

Structures of inequality have both temporal and spatial dimensions that are important in shaping the ways in which inequality is configured and reproduced. Through attentiveness to both *time* and *place*, I have shown how spatial topographies and temporal biographies are socially patterned and I have argued that both ends of the contextual distribution should be subjected to sociological research. The dominant approach of investigating spatial poverty and exclusion could as such overshadow important ways in which the spatial mediates reproduction of advantage through a homogenisation of experiences derived from similar class origin and destinations, as well as similar contextual impulses of wealth from late teenage years into one's forties. In the Oslo region, the reproduction of spatially mediated advantage is suggestively sustained through a two-sided process of closure; informally through classed sentiments of place-attachment and through politically initiated – and thus legitimized – marketization of housing. Hence, sociologists need insights into not only how the disadvantaged become spatially excluded, but also how the advantaged secure spatial withdrawal.

20

References

Aisenbrey, S., & Fasang, A. E. (2010). New life for old ideas: The "second wave" of sequence analysis bringing the "course" back into the life course. *Sociological Methods & Research, 38*(3), 420-462.

Atkinson, R. (2006). Padding the bunker: strategies of middle-class disaffiliation and colonisation in the city. *Urban Studies, 43*(4), 819-832.

Baltzell, D. E. (1958). *Philadelphia Gentlemen. The Making of a National Upper Class*. New York: The Free Press.

Benson, M. (2014). Trajectories of middle-class belonging: The dynamics of place attachment and classed identities. *Urban Studies, 51*(14), 3097-3112.

Bourdieu, P. (1984). *Distinction: A Social Critique of the Judgement of Taste*. Cambridge: Harvard University Press.

Bourdieu, P. (1990). Social Space and Symbolic Power. In *In Other Words: Essays Toward a Reflexive Sociology* (pp. 123-139). Stanford: Stanford University Press

Bourdieu, P. (1999). Site effects. In P. e. a. Bourdieu (Ed.), *The weight of the world: Social suffering in contemporary society* (pp. 123-129).

Bourdieu, P. (2005). *The Social Structures of the Economy*. London: Polity Press.

Bühlmann, F. (2010). Routes into the British service class feeder logics according to gender and occupational groups. *Sociology, 44*(2), 195–212.

Clark, W. A., Deurloo, M. C., & Dieleman, F. M. (2003). Housing careers in the United States, 1968-93: Modelling the sequencing of housing states. *Urban Studies, 40*(1), 143-160.

Cornwell, B. (2015). *Social Sequence Analysis. Methods and Applications* (Structural Analysis in the Social Sciences, Vol. 37). Cambridge: Cambridge University Press.

Coulter, R., & Van Ham, M. (2013). Following people through time: An analysis of individual residential mobility biographies. *Housing Studies, 28*(7), 1037-1055.

Coulter, R., van Ham, M., & Findlay, A. M. (2015). Re-thinking residential mobility: Linking lives through time and space. *Progress in Human Geography*, doi:10.1177/0309132515575417.

Cunningham, N., & Savage, M. (2015). The secret garden? Elite metropolitan geographies in the contemporary UK. *The Sociological Review, 63*(2), 321-348.

Dowling, R. (2009). Geographies of identity: landscapes of class. *Progress in Human Geography, 33*(6), 833-839, doi:10.1177/0309132508104998.

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

Giddens, A. (1981). *The Class Structure of the Advanced Societies* (2nd ed.). London: Hutchinson.

Hansen, M. N., Flemmen, M., & Andersen, P. L. (2009). Oslo Register Data Class Scheme (ORDC), Final report from the classification project.

21

Memorandum (Vol. 1, pp. 1-22): Department of Sociology and Human Geography, University of Oslo.

Hauge, Å. L., & Kolstad, A. (2007). Dwelling as an Expression of Identity. A Comparative Study Among Residents in High‐Priced and Low‐Priced Neighbourhoods in Norway. *Housing, Theory and Society, 24*(4), 272-292.

Hong, S.-Y., O'Sullivan, D., & Sadahiro, Y. (2014). Implementing Spatial Segregation Measures in R. *PloS one, 9*(11), e113767.

Ljunggren, J., & Andersen, P. L. (2015). Vertical and Horizontal Segregation: Spatial Class Divisions in Oslo 1970-2003. *International Journal of Urban and Regional Research, 39*(2), 305-322, doi:10.1111/1468-2427.12167.

Mills, C. W. (2000 [1956]). *The Power Elite*. London: Oxford University Press.

Pattillo, M. (2008). Race, Class, and Neighborhoods. In A. Lareau, & D. Conley (Eds.), *Social Class: How Does It Work?* (pp. 264–292). New York: Russell Sage.

Pollock, G. (2007). Holistic trajectories: a study of combined employment, housing and family careers by using multiple‐sequence analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 170*(1), 167-183.

Reardon, S. F., & Bischoff, K. (2011). Income Inequality and Income Segregation. *American Journal of Sociology, 116*(4), 1092-1153, doi:10.1086/657114.

Reardon, S. F., & O'Sullivan, D. (2004). Measures of Spatial Segregation. *Sociological Methodology, 34*(1), 121-162, doi:10.1111/j.0081-1750.2004.00150.x.

Sampson, R. J. (2012). *Great American City. Chicago and the Enduring Neighborhood Effect*. Chicago: The University of Chicago Press.

Savage, M. (1988). The missing link? The relationship between spatial mobility and social mobility. *British Journal of Sociology*, 554-577.

Savage, M. (1996). Space, networks and class formation. In N. Kirk (Ed.), *Social Class and Marxism. Defences and Challenges* (pp. 58-86).

Savage, M., Bagnall, G., & Longhurst, B. (2005). *Globalization and belonging*. London: Sage.

Sharkey, P. (2008). The Intergenerational Transmission of Context. *American Journal of Sociology, 113*(4), 931-969.

Sharkey, P. (2013). *Stuck in place: Urban neighborhoods and the end of progress toward racial equality*: University of Chicago Press.

Statistics Norway (1999). Regionale inndelinger [Regional classifications]. *Official Statistics of Norway*. Oslo - Kongsvinger: Statistics Norway.

Stovel, K., & Bolan, M. (2004). Residential Trajectories: The Use of Sequence Analysis in the Study of Residential Mobility. *Sociological Methods & Research, 32*, 559-598.

Studer, M. (2013). WeightedCluster Library Manual: A practical guide to creating typologies of trajectories in the social sciences with R. *LIVES working papers, 24*.

22

Toft, M., & Ljunggren, J. (2015). Geographies of class advantage: The influence of adolescent neighbourhoods in Oslo. *Urban Studies*, doi:10.1177/0042098015601770.

van Ham, M., Hedman, L., Manley, D., Coulter, R., & Östh, J. (2014). Intergenerational transmission of neighbourhood poverty: an analysis of neighbourhood histories of individuals. *Transactions of the Institute of British Geographers, 39*(3), 402-417.

Weber, M. (1978). *Economy and Society. An Outline of Interpretive Sociology* (Vol. 1). Berkeley: University of California Press.

Wessel, T. (2016). Economic Segregation in Oslo: Polarisation as a Contingent Outcome. In T. Tammaru, S. Marcińczak, M. van Ham, & S. Musterd (Eds.), *Socio-Economic Segregation in European Capital Cities. East meets West* (pp. 132–155): Routledge.

Wilson, W. W. (1987). *The Truly Disadvantaged: the Inner City, the Underclass and Public Policy*. Chicago: University of Chicago Press.

23

6 Appendix
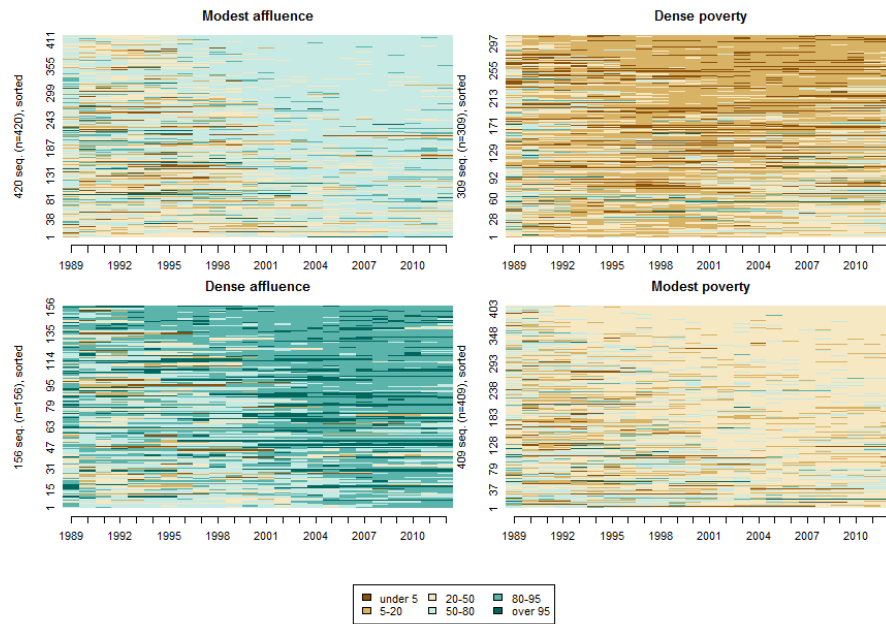
Fig. 4: Clusters ordered by silhouette values. Sequences that are the most poorly represented by the cluster solution are at the bottom of each graph (Studer 2013).

Session 10B: Multichannel

Trajectories of vulnerability: a multi-dimensional approach

How are employment, cohabitation and health related?

Matteo Antonini[1,] Felix Bühlmann[1], Jean-Luc Heeb[2],

[1] Université de Lausanne, Nccr Lives

[2] Haute école de travail social de Fribourg, Nccr Lives

Abstract Our paper explores vulnerable trajectories across three life course domains: employment activity, cohabitation and health status. We embrace a comprehensive point of view applying multichannel sequence analysis to a sample of about 1250 residents of the Canton de Vaud considering the period from age 1 to 40. Our results confirm the diffusion of the effects of vulnerable statuses across life course domains.

1 Extended abstract

1.1 Theoretical background

The idea that different life domains are interrelated is present in life course literature from the very beginning (Elder 1974). This concept is now well established in the literature (Mayer 2009; Sapin, Spini, and Widmer 2007) and suggests dense and multidirectional relations among the involved life-course domains. Nevertheless, the empirical application of this theoretical frame is often reduced to an analysis of a single life course domain. Numerous are the study focusing exclusively on either employment trajectories (Buhlmann 2008, 2010; Fuller and Martin 2012; Fuller and Stecy-Hildebrandt 2015; Manzoni, and Mooi-Reci 2011; Scherer 1999), education (Brzinsky-Fay 2016; Cooksey and Rindfuss 2001), household structure (Chaloupkovà 2010) or health condition (Cullati 2015; Cullati et al. 2014; Friedli 2009). Equally frequent are papers that take a step further and link two life course domains. Nevertheless, in many papers, one of the life course do-

main is the central topic of investigation, while the other is used as a time-variant independent variable (Charles and Harris 2007; Jacob and Kleinert 2008; Oesterle et al. 2010; Teachman 2010). Much less frequently, two life course domains are placed on the same level and analyzed together with descriptive aims (Davia and Legazpe 2014; Gauthier et al. 2010; McMunn et al. 2015; Oesterle et al. 2010). This strategy is even rarer if we consider three or more life course domains (Pollock 2007). In our paper, we follow this road and analyze three life course domains together: employment activity, cohabitation and health status. These life course domains are conceptualized in a wider sense than usually. "Activity" includes indicators on work status, education and other relate conditions (e.g. social help). "Cohabitation" contains the household structure including the changes due to family formation processes. Finally, "health" includes both mental sickness and physical impairments - from chronical condition to isolated episodes -, accidents and operations. We aim to take all these pieces of information together in order to build a comprehensive view of individuals' lives. This perspective relies on the idea that the partition of different life curse domains is a useful analytical tool, but is rather artificial. Every life course domain is interrelated with the others: each on them influences, but, at the same time, it is also influenced by the others (Elder 1974; Pollock 2007).

1.2 Objectives of our paper

The objective of our paper is twofold. First, we aim to describe individuals' life courses looking for recurrent patterns. Which configurations are more frequent and which ones are less usual? This question concerns both the trajectories in a single life course domain and the compresence of different patterns across the life course domains.

Second, we look at the connections among life course domains when vulnerable situations are present. Given the mutual relation among life course domains the question is double. On the one hand, which trajectories are more likely to be related to vulnerable situations? On the other, what are the biographical consequences of a transition through a period of vulnerability?

1.3 Data and method

Our data come from a retrospective life calendar of a sample of residents in the canton of Vaud, Switzerland. Our data oversamples the presence of households with a lower income. Starting from these data we select a subsample including all the people aged 40 years or more with valid sequence data in every life course domain: employment activity, cohabitation and health status.

3

We use two types of variables in our analyses: time-invariant variables and time-variant variables. Time-invariant variables are sociodemographic characteristics: sex, age, cohort, nationality. Time-variant variables include the measures linked to the life domains at the center of our analyses. The states describing cohabitation include: living with parents, with siblings, alone, with parent without children, with the partner and children, with children only, with friend, and a residual category gathering all the other situations. Health status has less categories: physical illness, mental illness, operation, accident, multiple events among the previous and no event. Employment activity includes: education, part-time job, full-time job, self-employment, inactivity, unemployment, social help. Employment activity categories can overlap creating statuses that mix two previous categories. As education was not explicitly addressed in the calendar questionnaire, we filled up the empty parts of the work sequences with a specific imputation procedure using the educational level.

In order to conduct a comprehensive analysis of our data, we use multichannel sequence analysis (Gauthier et al. 2010, 2013) and single-channel sequence analyses (Abbott 1990) connected using Yule's group-to-group association (Yule 1912). These techniques are applied to sequences describing individuals' life course from birth until the age of 40 years.

1.4 Results

To date, only a series of preliminary single-channel sequence analyses that shows the differences among men and women trajectories have been carried out.

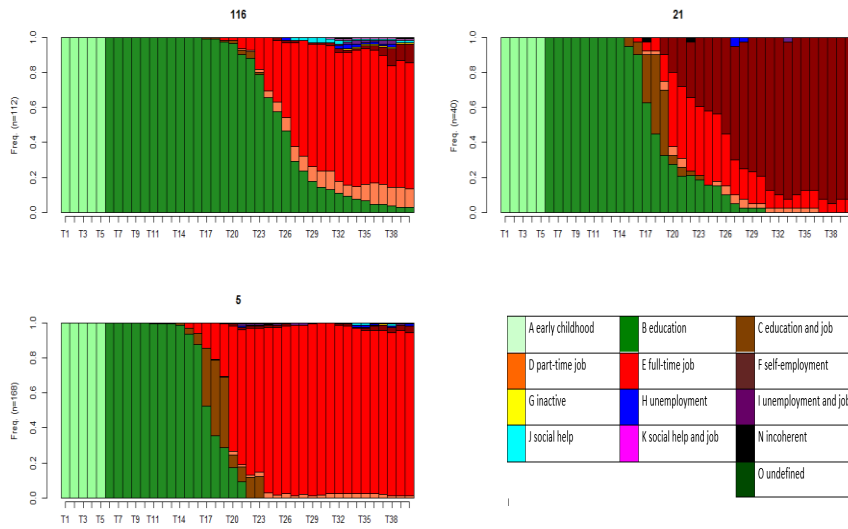**Figure 1 – Males: trajectories of employment activity**

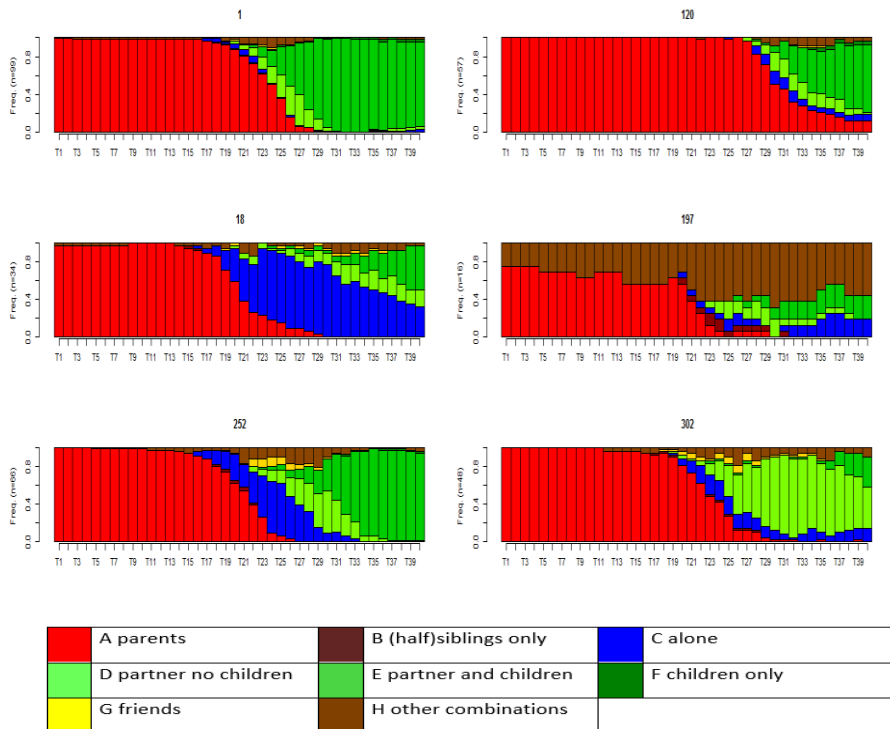**Figure 2 - Males: trajectories of cohabitation**

5

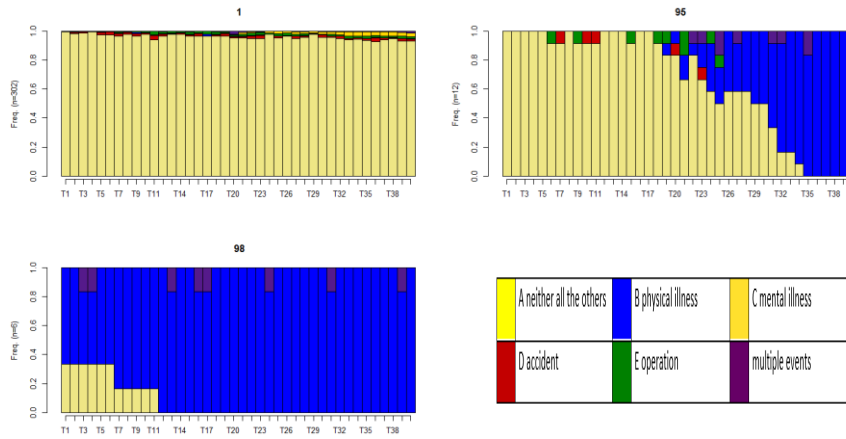## Figure 3 - Males: trajectories of health status

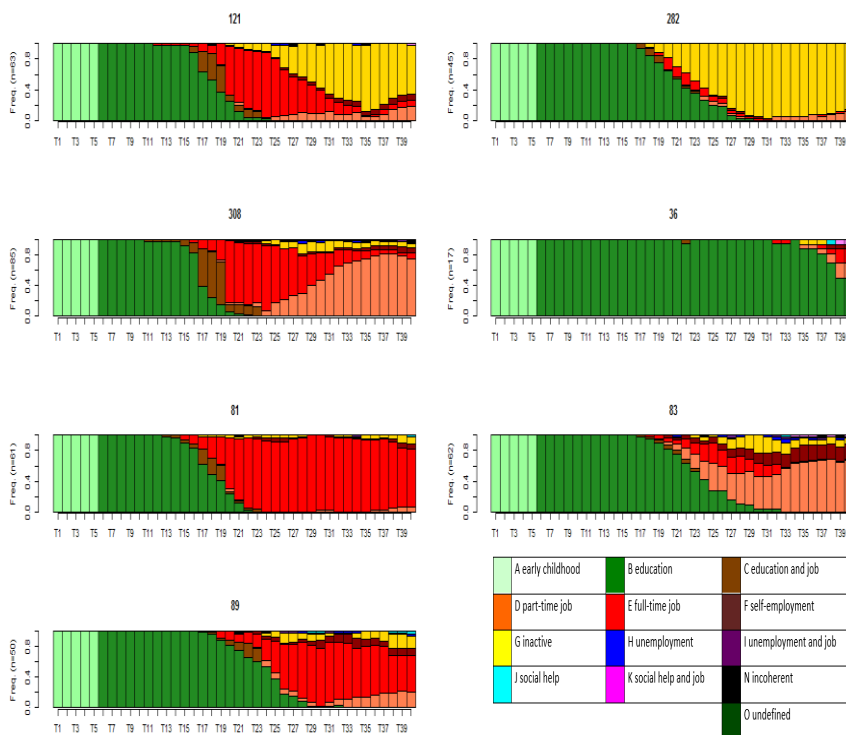## Figure 4 - Females: trajectories of employment activity

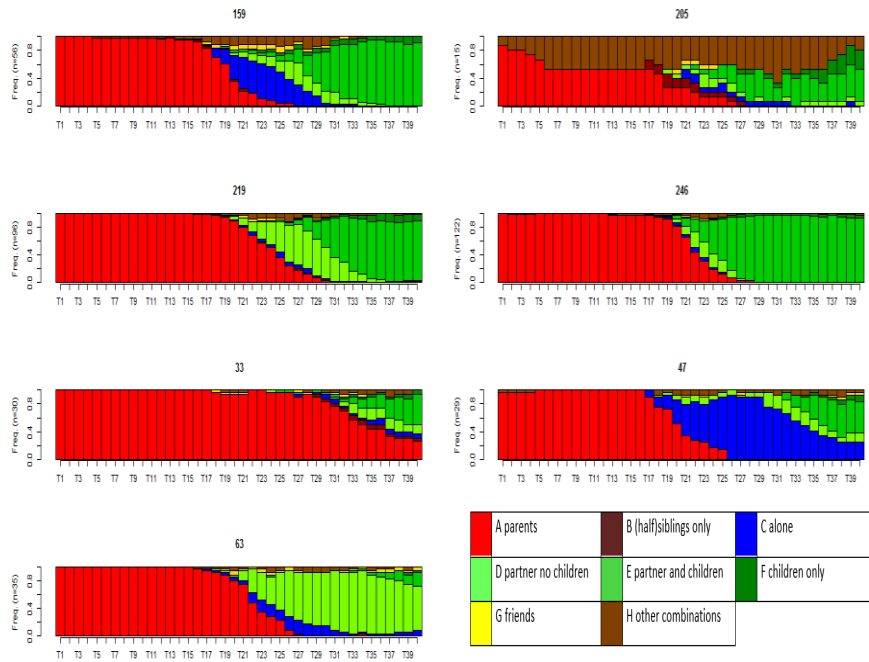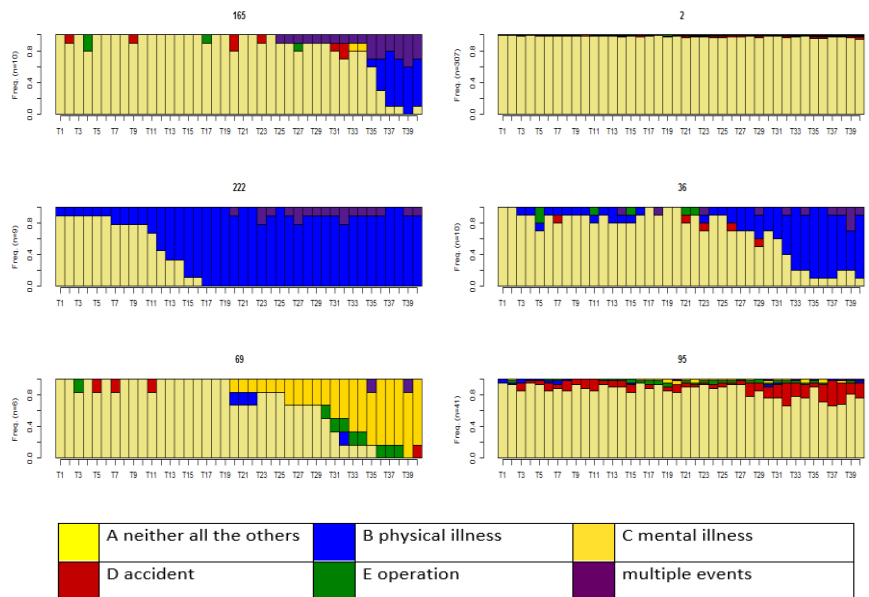**Figure 5 - Females: trajectory of cohabitation**

Figure 6 - Females: trajectories of health status

7

1.5 Conclusion

Our paper explores the link between three life course domains: employment activity, cohabitation and health status. Our results confirm that the life course domains are deeply connected and that the vulnerable statuses affect individuals' life across the life domains. This relation seems to be behind the accumulation of advantages/disadvantages.

References

Abbott, Andrew. 1990. "A Primer on Sequence Methods." *Organization Science* 1(4):375–92.

Brzinsky-Fay, Christian. 2016. "Leavers Transition? Labour Sequences of School Market." 23(4):409–22.

Buhlmann, F. 2008. "The Corrosion of Career?--Occupational Trajectories of Business Economists and Engineers in Switzerland." *European Sociological Review* 24(5):601–16. Retrieved September 29, 2014 (http://esr.oxfordjournals.org/cgi/doi/10.1093/esr/jcn019).

Buhlmann, Felix. 2010. "Routes into the British Service Class Feeder Logics according to Gender and Occupational Groups." *Sociology* 44(2):195–212.

Chaloupkovà, Jana. 2010. "The De-Standardisation of Early Family in the Czech Republic: Trajectories A Cross-Cohort Comparison." *Czech Sociological Review* 46(3):427–51.

Charles, Nickie and Chris Harris. 2007. "Continuity and Change in Work-Life Balance Choices." *The British Journal of Sociology* 58(2):277–95. Retrieved (http://www.ncbi.nlm.nih.gov/pubmed/17610623).

Cooksey, Elizabeth C. EC and RR Ronald R. Rindfuss. 2001. "Patterns of Work and Schooling in Young Adulthood." *Sociological Forum* 16(4):731–55. Retrieved (http://www.jstor.org/stable/684831\nhttp://link.springer.com/article/10.1023/A:1012842230505).

Cullati, Stéphane. 2015. "Socioeconomic Inequalities in Health Trajectories in Switzerland: Are Trajectories Diverging as People Age?" *Sociology of Health & Illness* 0(0):n/a – n/a. Retrieved (http://doi.wiley.com/10.1111/1467-9566.12232).

Cullati, Stéphane, Emmanuel Rousseaux, Alexis Gabadinho, Delphine S. Courvoisier, and Claudine Burton-Jeangros. 2014. "Factors of Change and Cumulative Factors in Self-Rated Health Trajectories: A Systematic Review." *Advances in Life Course Research* 19(1):14–27.

Davia, M. a. and N. Legazpe. 2014. "Female Employment and Fertility

Trajectories in Spain: An Optimal Matching Analysis." *Work, Employment & Society* 28(4):633–50. Retrieved (http://wes.sagepub.com/content/28/4/633?etoc).

Elder, G. .. 1974. *Children of the Great Depression. Social Change in Life Experience.* 25th anniv. Boulder: Westview Press.

Friedli, Lynne. 2009. "Mental Health , Resilience and Inequalities." *Health San Francisco* 1–64. Retrieved (http://www.mentalhealthpromotion.net/resources/mental-health-resilience-and-inequalities.pdf).

Fuller, Sylvia and Todd F. Martin. 2012. "Predicting Immigrant Employment Sequences in the First Years of Settlement." *International Migration Review* 46(1):138–90.

Fuller, Sylvia and Natasha Stecy-Hildebrandt. 2015. "Career Pathways for Temporary Workers: Exploring Heterogeneous Mobility Dynamics with Sequence Analysis." *Social Science Research* 50:76–99. Retrieved (http://linkinghub.elsevier.com/retrieve/pii/S0049089X14001999).

Gauthier, A., E. D. Widmer, P. Bucher, and C. Notredame. 2013. "Multichannel Optimal Matching: A Multidimensional Approach to Sequence Analysis." in *Gendered life courses*, edited by E. Widmer. Berlin: Lit Verlag.

Gauthier, Jacques-Antoine, Eric D. Widmer, Philipp Bucher, and Cédric Notredame. 2010. "Multichannel Sequence Analysis Applied To Social Science Data." *Sociological Methodology* 40(1):1–38. Retrieved September 29, 2014 (http://smx.sagepub.com/lookup/doi/10.1111/j.1467-9531.2010.01227.x).

Jacob, Marita and Corinna Kleinert. 2008. "Does Unemployment Help or Hinder Becoming Independent the Role of Employment Status for Leaving the Parental Home." *European Sociological Review* 24(2):141–53.

Manzoni, Anna and Irma Mooi-Reci. 2011. "Early Unemployment and Subsequent Career Complexity: A Sequence-Based Perspective." *Schmollers Jahrbuch* 131(2):339–48.

Mayer, K. .. 2009. "New Directions in Life Course Research." *Annual Review of Sociology* 35:413–33.

McMunn, Anne et al. 2015. "De-Standardization and Gender Convergence in Work-Family Life Courses in Great Britain: A Multi-Channel Sequence Analysis." *Advances in Life Course Research* 26:60–75. Retrieved (http://www.sciencedirect.com/science/article/pii/S1040260815000313).

Oesterle, Sabrina et al. 2010. "Men ' S and Women ' S to Adulthood Pathways Adolescent and Their Precursors." *Journal of Marriage and Family* 72(5):1436–53.

Pollock, Gary. 2007. "Holistic Trajectories: A Study of Combined Employment, Housing and Family Careers by Using Multiple-Sequence Analysis." *Journal of the Royal Statistical Society. Series A: Statistics in Society* 170(1):167–83.

Sapin, M., Dario Spini, and Eric Widmer. 2007. *Les Parcourse de Vie*. Lausanne:

9

Presses polytechniques et universitaires romandes.

Scherer, Stefani. 1999. *Early Career Patterns – a Comparison of Great Britain and West Germany*.

Teachman, Jay. 2010. "Family Life Course Statuses and Transitions: Relationships With Health Limitations." *Sociological Perspectives* 53(2):201–20.

Yule, G. U. 1912. "On the Methods of Measuring the Association between Two Attributes." *Journal of the Royal Statistical Society* 75:579–652.

Multiphase Optimal Matching: An Application to Careers of Participation in *Pâtissiers'* Competitions

Thomas Collas

Center for the Sociology of Organizations (Sciences Po), 19. rue Amélie, 75007 Paris
thomas.collas@sciencespo.fr

Please do not quote or circulate.

Many social sciences theories (*e.g.* Becker, 1963 ; Wilensky, 1964 ; Lang and Lang, 1984 ; for a synthesis on 'universal narratives' see Abbott, 1992) and many common notions (lifecycle, adulthood, turning point, ratchet effect, etc.) are premised upon the idea that some sets of processes follow regular patterns defined by successive phases. A phase can be defined as a moment in a sequence, a succession of episodes (time units in a sequence) that can be regarded as singular compared to other successions of episodes (other phases) in the same sequence[1].

Optimal matching analysis (OMA) is commonly used to reveal these successive phases. In this talk, I use OMA to compare sequences defined by two or more phases through what I call multiphase optimal matching (MPOM).

In a first section, I define the main properties of multiphase sequences and the main features of MPOM. In a second section, I present a simple example of MPOM to compare careers of participation in *pâtissiers*' competitions in France[2].

---

[1]Such is also the basic postulate of Qualitative Harmonic Analysis (Deville, 1974). For an application of QHA and a comparison with Optimal Matching Analysis, see Robette and Thibault (2008) and Robette and Bry (2012).

[2]I keep the french word *pâtissiers* since 'baker' refers to bread making – to *boulanger* in french –, 'pastry chef' or 'pastry cook' are relevant to describe *pâtissiers*' restaurant work and 'confectioners' refers to candies making.

2   Thomas Collas

1 Multiphase sequences and multiphase optimal matching (MPOM)

The basic principle of MPOM is to compare sequences according to equivalent phases. Phase is made a relevant unit of comparison. Sequence is both regarded as a succession of phases and as a coherent unit compared to other sequences.

1.1 Multiphase sequences: four properties

The main property of multiphase sequences is that they are defined before the analysis as composed by two or more phases that are supposed to be common to the whole set of sequences under study. For example, sequences synchronization (Blanchard, 2010; Giudici and Gauthier, 2009; Colombi and Paye, 2014) is a visualization technique suited for two-phase sequences, that is to say sequences in which one event is *postulated* as a turning point (marriage, employment, entry into an association, etc.). This event thus divides the sequence between a first phase (episodes before the turning point) and a second one (episodes occurring after the turning point). As Denis Colombi and Simon Paye (2014) have underlined, this turning point can be endogenous or exogenous. In the first case, the turning point is a transition between two states belonging to the sequence alphabet. In the second case, the turning point is defined on another channel. For example, an occupational career can be divided between two phases when married people are compared: before and after marriage. Following Andrew Abbott (1997), synchronization can be defined as a presumed division into two trajectories, *i.e.* phases characterized by distinct 'regimes of probability'. This is the second property of multiphase sequences.

Obviously, there can be more than *one* turning point in what appears to be a single sequence (*e.g.* an individual career), which is then divided into three or more phases. Synchronization can be generalized to *n* phases. Every sequence in a set can be defined as a succession of *n* phases such as $A = (\zeta_1^A, \zeta_2^A, ..., \zeta_p^A, ..., \zeta_n^A)$ and $B = (\zeta_1^B, \zeta_2^B, ..., \zeta_p^B, ..., \zeta_n^B)$, where ζ_p^A is the phase p in sequence A and ζ_p^B is the phase p in sequence B. Take tennis matches for example. To visualize matches as sequences of games (each time unit is a game, each state in the alphabet is the number of points in a game), each last point of a set can be regarded as a turning point. Thus, each match can be divided into five phases (the highest possible number of sets) to indicate that it is not only a sequence of games but, above all, a sequence of sets of games. This toy example points out two properties of multiphase sequences. First, equivalent phases often vary in length from one sequence to another. There can be six games in the first set of match A and ten games in the first set of match B. The end of the set announces in each case the second set. Even if complete sequences are of equal length, synchronization generally implies differences in phase lengths. If not, each turning point is met at the same time for the whole set of sequences. Such is the case if turning points are ages. Second, the length of a phase can be equal to zero. This phase is thus considered as an empty one. This can occur if two hypothetic idiosyncratic turning points are observed at the same time or if a supposed turning point is not yet observed, never observed or immediately observed at the beginning

of the sequence. Match A can come to an end after three sets, while match B can last five sets. Within match A, sets 4 and 5 are empty phases.

This applies to the former example: occupational careers can be divided into more than two phases with respect to marriage. A first phases can be 'being single', a second one 'being married', a third one 'being divorced or widowed', a fourth one 'being remarried' and so on. Each of these phases necessarily follows the previous ones: 'being remarried' implies celibacy, marriage and divorce or widowhood. The length of these different phases can vary largely and can obviously be equal to zero, for example for people who never get married, for people who never get divorced or for people who get remarried during the same time unit (a month for example) they get divorced (two assumed turning points are met at the same time)[3].

Let us summarize the four properties underlined:

1. Multiphase structure is not a result but a *postulate*, even if it can result from previous analysis.
2. Each phase is regarded as an *hypothetic trajectory*.
3. Equivalent phases often *differ in length*.
4. Phase length can be *equal to zero*. This does not affect the relevance of a constant multiphase definition for all sequences under study.

1.2 MPOM: a short definition

Let us now consider that we do not only want to visualize a set of *n*-phase sequences, but to compare them through optimal matching.

The main concern is to keep the multiphase structure. As Laurent Lesnard (2008, p. 463) has stressed, insertion/deletion operations 'loosen the connection of processes with their temporal scale'. Dynamic Hamming Matching (DHM), the method suggested by the author (for details, see Lesnard, 2014), rests on two conditions to avoid such a loosening. First, each time unit is regarded as incommensurable with other time units: if A's workday is compared to B's one, what A is doing (working/not working) at 7PM can only be compared to what B is doing at the exact same time. Then, only substitution operations are used and insertion/deletion operations are excluded. Second, the cost of substitution for a state to another depends on the observed rate of transition between the two states at the time unit in question. Then, a cost of substitution between each pair of states is defined for each time unit. This solution, which does not preserve only one synchronization point but *each time unit synchronization*, is highly suited for sequences defined by a limited number of states, observed in each sequence and spanning over long periods, that is to say sequences varying from one another in timing and duration.

MPOM loosens the 'connection of [sequences] with their temporal scale' (Lesnard, 2008, p. 463) only within phases. In other words, phases, not time units, are regarded as incommensurable. Three features can be underlined:

---

[3]Another simple example of multiphase sequence is a French academic career at University with turning points such as becoming *Maître de conférences* and becoming *Professeur des universités*.

4   Thomas Collas

- The three basic operations of OM (Abbott and Forrest, 1986) – substitution, insertion, deletion – are kept and used to compare equivalent phases from one sequence to another (ζ_p^A to ζ_p^B). Length variations resulting from synchronization induce several complications. While the 'core program' (Gauthier and al., 2014. p. 5) of sequence analysis 'mostly refers to sequences of equal length, with age as a time axis and year as a time unit' (Ibid.), analysts of sequences of unequal length are often mostly preoccupied, since the seminal study by Andrew Abbott and John Forrest (1986), by normalizing sequences with respect to length or by minimizing distance due to length (Abbott and Hrycak 1990; Stovel and al., 1996; Stovel and Bolan, 2004). As Philippe Blanchard (2010. p. 57) has stressed, taking length differences into account implies a cautious arbitration between substitution costs and insertion/deletion costs in the search of their most relevant relative definition regarding the sequences under study.
- Substitution costs as well as insertion/deletion cost (indel) can be defined for each phase. That involves a multiplication of cost-setting operations that may seem dubious, since many criticisms addressed to OMA focused on cost-setting operations (see the widely cited debate between Wu, 2000 and Abbott and Tsay, 2000). Transition-rates based substitution costs can be relevant in some cases and can be defined for each phase. If a pair of states can be observed in any phase, transition between these states can be defined as a combination between intra-phase transition rates and intra-sequence transition rates.
- For each pair of sequences, the aim of the operation is to compute a distance per phase and then to compute an inter-sequences distance. The simplest way to do so is to sum OM-distances per phase, the resulting distance matrix being the sum of phase-distance matrices (see example below). Thus, each phase contribution to the full distance depends on differences in states-composition and in length. Another way to compute a total distance is to standardize OM-distances per phase with respect to maximal possible distance for each phase.

2 A two-phase example: competitors' careers as ante-senior and senior phases

I used MPOM while preparing my PhD thesis on *pâtissiers*' work in France since the 1970s. I was studying careers of participation in *pâtissiers*' competitions in France, which mostly consist in sugar or chocolate sculptures competitions.

I was fishing for regular patterns, especially for cumulative patterns in terms of ranking. Cumulative models (DiPrete and Eirich, 2006), when applied to competitions, predict a close relation between length and structure (for a detailed account, see Collas, 2015): short careers' patterns should mainly follow a 'succession of failures' pattern, long careers should even follow 'successions of victories' patterns – through indirect screening of candidates (Menger, 2009) – or progressive pattern – for example through learning-by-doing due to imperfect information on the competition during the first participations.

Despite of numerous cost setting operations, OMA appeared as the best solution to search for regular patterns

- The first reason was *instability*: the sequences are characterized by a high instability from one time unit to another concerning ranking (one can rank first, then ninth, then third, etc.) ; insertion/deletion operations reduce distance due to some lags.
- The second reason was *relations between states*: a first rank is closer to a second rank than to a tenth rank. OM algorithm lies in the postulate of distance between states (through substitution costs) instead of strict difference between them.

2.1 Data

Data were gathered from 2060 rankings (120 competitions, recurring or not), mainly from trade press, but also from organizers' archives for some major competitions. Data cover the period 1953-2012. Due to source heterogeneity[4], I focused on careers beginning after 1985 and ending before 2007 and thus kept 3927 names out of 6264[5]. I used OM only to compare participation careers counting two participations or more, that is 1258 careers[6]. Career length varies from 2 participations to 21 participations.

Time unit is a participation in a competition: the time axis is thus a process one. Careers are defined with 24 states (Table 1, p. 6).

Three dimensions have been taken into account to define the states:

- Recurring competitions preceded by pre-selection contests are regarded as distinct. That is the case of *Championnat de France du dessert* (CFD, a national 'dessert on plate' competition) for apprentice and for senior competitors, of *Meilleur Apprenti de France* (Best Apprentice in France) and *Meilleur Apprentice du Monde* (Best Apprentice in the World) – these two are put in a same category, BA –, of '*Un des meilleurs ouvriers de France*' (MOF) competition[7] and, finally, of senior international competitions[8] with pre-selection contests that I put in a same category: CM. Competitions which do not verify these two conditions (recurrence and pre-selections) are categorized R2 and regarded as equivalent.
- Every state is defined by the rank occupied, in three categories: 1st rank (L), 2nd or 3rd rank (P) and 4th rank or beyond (H).

---

[4]*Journal du pâtissier* (published since 1978) mentions competitions organized in different areas in France, while the other sources mention mainly competitions taking place in Paris.

[5]The mean duration of career counting two participation or more and ending after 1980 is 5 years.

[6]1263 careers counted two participations or more but I have not kept a first rank at the '*Un des meilleurs ouvriers de France*' (MOF) competition if it was gained at the end of a career since 98 % of participation careers including this rank end with this rank. Then, it allowed me to compare MOF laureates' careers with other participation careers.

[7]This can be translated as 'one of the best craftsmen in France' but the usual category is simply the french expression for 'best craftsman in France'.

[8]*Coupe du Monde* (in Lyon), *Grand prix international de la chocolaterie*, World Chocolate Maser.

6   Thomas Collas

Table 1. States definition

State	Competition	Rank	Category
L-CFDAp	CFD	1st	Apprentice
P-CFDAp	CFD	2nd or 3rd	Apprentice
H-CFDAp	CFD	4th or beyond	Apprentice
L-BA	BA	1st	Apprentice
P-BA	BA	2nd or 3rd	Apprentice
H-BA	BA	4th or beyond	Apprentice
L-R2Ap	R2	1st	Apprentice
P-R2Ap	R2	2nd or 3rd	Apprentice
H-R2Ap	R2	4th or beyond	Apprentice
L-R2Ju	R2	1st	Junior
P-R2Ju	R2	2nd or 3rd	Junior
H-R2Ju	R2	4th or beyond	Junior
L-CFDSe	CFD	1st	Senior
P-CFDSe	CFD	2nd or 3rd	Senior
H-CFDSe	CFD	4th or beyond	Senior
L-CM	CM	1st	Senior
P-CM	CM	2nd or 3rd	Senior
H-CM	CM	4th or beyond	Senior
L-MOF	MOF	1st	Senior
H-MOF	MOF	unranked finalist	Senior
L-R2Se	R2	1st	Senior
P-R2Se	R2	2nd or 3rd	Senior
H-R2Se	R2	4th or beyond	Senior
4years	4 years void between two participations		

- Every state is defined by the age category of contestants, in three categories again: Senior (Se), Junior (Ju) or Apprentice (Ap).

A state named '4years' was added for every period lasting more than 4 years and less than 8 years between two successive participations.

These sequential data were characterized by high heterogeneity[9], high instability[10] and high length variance.

2.2 First senior competition as an assumed turning point between ante-senior phase and senior phase

In addition to that, it seemed dubious to compare these sequences from start to finish. Some competitions are for apprentices and junior competitors only. Furthermore,

---

[9]Amongst 1258 careers, we count 794 different patterns (mean occurrence is 1.56). Each sequence of length 6 or more ($N = 174$) is unique.

[10]Mean complexity (Gabadinho and al., 2010) is 0.43. This metric varies from 0 to 1 and is based on the number of transitions and the on longitudinal entropy.

a first analysis pointed out a low rotation between these competitions and senior ones : the first participation in a senior competition implies no later participation in junior or apprentice competitions, except for 6 % of senior contestants[11]. That first analysis also stressed that an important part of individual careers did not include any senior competition (44 % of the 557 sequences that include a junior or apprentice competition) and that a more important part did not include any junior or apprentice competition (69 % of the 1017 sequences that include a senior competition).

Based on that analysis and on interviews with *pâtissiers* ($N = 58$), I assumed that the first participation in a senior competition represents entry into a career of evaluation which is not based on age or scholarship, but on the single fact of being identified as a *pâtissier*.

The sequences compared are defined as two-phase sequences verifying the aforementioned properties with the first participation in a senior competition as a supposed turning point. States observed before the first participation in a senior competition were regarded as incommensurable with states observed after that participation. There are an *ante*-senior phase (defined solely by participations in apprentice and junior competitions) and a senior phase (that can include any state in Table 1). The first phase is defined by 218 distinct patterns (mean occurrence is 2.6) and the second phase by 504 patterns (mean occurrence is 2.02).

Figure 1 shows a sample of 30 phases following that synchronization[12].

2.3 MPOM as a sum of OM-distances per phase

MPOM preserves synchronization while computing distance between sequences.

In this example, I define MPOM distance between two sequences as a sum of OM-distances per phase. Let us formalize it.

Following the presentation of OM distance by Cees Elzinga (2014. p. 55), let S_1 and S_2 denote two n-phases sequences and ζ denotes the set of n postulated phases. For a phase ζ_p in $\{\zeta_1, \zeta_2, ..., \zeta_p, ..., \zeta_n\}$ over the alphabet of states $\Sigma_p = \{\lambda_p, a_p, b_p, ...\}$ in ζ_p (with λ denoting the empty state), let $t_{\zeta_p^{S_1}, \zeta_p^{S_2}} = t_1...t_l$ denotes a series of admissible phase edits. For any pair of equivalent phases from one sequence to another, many distinct series of edits that transform $\zeta_p^{S_1}$ (the phase p in sequence S_1) into $\zeta_p^{S_2}$ (the phase p in sequence S_2) may exist and I write $T(\zeta_p^{S_1}, \zeta_p^{S_2})$ to denote the set of such edit-series. Furthermore, to each edit t_i, a nonnegative cost of weight $c(t_i)$ is assigned and the cost of an edit-series $C(t_{\zeta_p^{S_1}, \zeta_p^{S_2}})$ equals the sum on the edits involved. The OM-distance per phase is defined as the minimum of the costs of the edit-series in $T(\zeta_p^{S_1}, \zeta_p^{S_2})$.

Thus, MPOM-distance $d_{MPOM}(S_1, S_2)$ between sequences S_1 and S_2 is defined as follow :

---

[11]22 % of the 311 competitors who participate in senior and apprentice or junior competitions, participate in junior or apprentice ones after a first participation in a senior one.

[12]R software (R Core Team, 2015) and TraMineR R package (Gabadinho and al., 2011) were used to visualize sequences, to compute OM-distances, to extract sets of representative sequences and to compute other metrics related to sequences (length and complexity).

8   Thomas Collas

Fig. 1. Random sample of 30 sequences
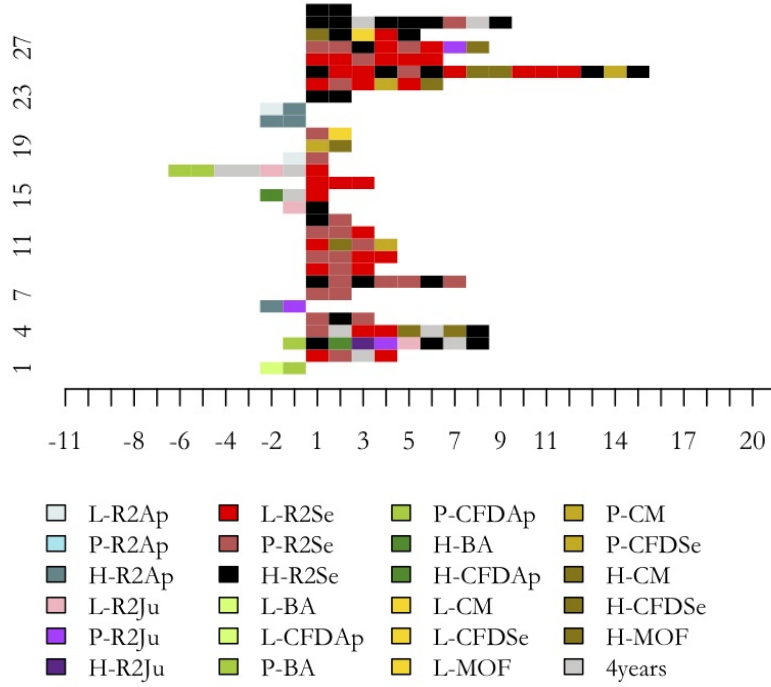
N.B.: Due to a high number of states, visualization is a bit tricky, diversity has been impoverished to improve legibility: some states, which are rare states and for which substitution costs are low, are visualized with the same colors

$$d_{MPOM}(S_1, S_2) = \sum_{p=1}^{n} min \left\{ C(t_{\zeta_p^{S_1}, \zeta_p^{S_2}}) : t_{\zeta_p^{S_1}, \zeta_p^{S_2}} \in T(\zeta_p^{S_1}, \zeta_p^{S_2}) \right\} \tag{1}$$

2.4 Definition of substitution and insertion/deletion costs

In trying to uncover cumulative patterns, substitution costs based on transition rates would be of little help. Substitution costs between states have been defined with respect to their formal closeness: competitions with pre-selection contests are closer to one another than to competitions without pre-selection contests; a first rank is closer to a second rank than to a fourth; a second or a third rank is closer to a fourth rank than to a first rank (due to the singular position of ranking first).

   Minimum substitution cost between two distinct states is 1 and maximum cost is 2. 80 % of substitution costs are superior to 1.6, which is the minimum substitution cost between two states differing by rank and by pre-selections. Due to the division into two phases, only 5 % of the costs differences between distinct states are due to a difference in age categories, 95 % are due to a difference in rankings and in types

Multiphase Optimal Matching 9

Table 2. Substitution costs for senior phase

	L-BA	P-BA	H-BA	L-R2Ap	P-R2Ap	H-R2Ap	L-R2Ju	P-R2Ju	H-R2Ju	L-CFDSe	P-CFDSe	H-CFDSe	L-CM	P-CM	H-CM	L-MOF	H-MOF	L-R2Se	P-R2Se	H-R2Se	4years
L-BA	0	1.6	1.81	1.88	1.93	1.95	1.92	1.97	1.99	1.05	1.65	1.86	1.05	1.65	1.86	1.05	1.65	1.93	1.98	2	2
P-BA	1.6	0	1.6	1.84	1.92	1.94	1.88	1.96	1.98	1.65	1.05	1.65	1.65	1.05	1.65	1.65	1.05	1.9	1.97	1.99	2
H-BA	1.81	1.6	0	1.6	1.9	1.93	1.64	1.94	1.97	1.86	1.65	1.05	1.86	1.65	1.05	1.86	1.65	1.65	1.95	1.98	2
L-R2Ap	1.88	1.84	1.6	0	1.88	1.91	1.04	1.92	1.95	1.93	1.9	1.65	1.93	1.9	1.65	1.93	1.9	1.05	1.93	1.96	2
P-R2Ap	1.93	1.92	1.9	1.88	0	1.6	1.92	1.04	1.64	1.98	1.97	1.95	1.98	1.97	1.95	1.98	1.97	1.93	1.05	1.65	2
H-R2Ap	1.95	1.94	1.93	1.91	1.6	0	1.95	1.64	1.04	2	1.99	1.98	2	1.99	1.98	2	1.99	1.96	1.65	1.05	2
L-R2Ju	1.92	1.88	1.64	1.04	1.92	1.95	0	1.88	1.91	1.92	1.88	1.64	1.92	1.88	1.64	1.92	1.88	1.04	1.92	1.95	2
P-R2Ju	1.97	1.96	1.94	1.92	1.04	1.64	1.88	0	1.6	1.97	1.96	1.94	1.97	1.96	1.94	1.97	1.96	1.92	1.04	1.64	2
H-R2Ju	1.99	1.98	1.97	1.95	1.64	1.04	1.91	1.6	0	1.99	1.98	1.97	1.99	1.98	1.97	1.99	1.98	1.95	1.64	1.04	2
L-CFDSe	1.05	1.65	1.86	1.93	1.98	2	1.92	1.97	1.99	0	1.6	1.81	1	1.6	1.81	1	1.6	1.88	1.93	1.95	2
P-CFDSe	1.65	1.05	1.65	1.9	1.97	1.99	1.88	1.96	1.98	1.6	0	1.6	1.6	1	1.6	1.6	1	1.84	1.92	1.94	2
H-CFDSe	1.86	1.65	1.05	1.65	1.95	1.98	1.64	1.94	1.97	1.81	1.6	0	1.81	1.6	1	1.81	1.6	1.6	1.9	1.93	2
L-CM	1.05	1.65	1.86	1.93	1.98	2	1.92	1.97	1.99	1	1.6	1.81	0	1.6	1.81	1	1.6	1.88	1.93	1.95	2
P-CM	1.65	1.05	1.65	1.9	1.97	1.99	1.88	1.96	1.98	1.6	1	1.6	1.6	0	1.81	1.6	1.6	1.84	1.92	1.94	2
H-CM	1.86	1.65	1.05	1.65	1.95	1.98	1.64	1.94	1.97	1.81	1.6	1	1.81	1.6	0	1.81	1.6	1.6	1.9	1.93	2
L-MOF	1.05	1.65	1.86	1.93	1.98	2	1.92	1.97	1.99	1	1.6	1.81	1	1.6	1.81	0	1.6	1.88	1.93	1.95	2
H-MOF	1.65	1.05	1.65	1.9	1.97	1.99	1.88	1.96	1.98	1.6	1	1.6	1	1.6	1.6	1.6	0	1.84	1.92	1.94	2
L-R2Se	1.93	1.9	1.65	1.05	1.93	1.96	1.04	1.92	1.95	1.88	1.84	1.6	1.88	1.84	1.6	1.88	1.84	0	1.88	1.91	2
P-R2Se	1.98	1.97	1.95	1.93	1.05	1.65	1.92	1.04	1.64	1.93	1.92	1.9	1.93	1.92	1.9	1.93	1.92	1.88	0	1.6	2
H-R2Se	2	1.99	1.98	1.96	1.65	1.05	1.95	1.64	1.04	1.95	1.94	1.93	1.95	1.94	1.93	1.95	1.94	1.91	1.6	0	2
4years	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	0

10　　　　Thomas Collas

Table 3. Substitution costs for ante-senior phase

	L-CFDAp	P-CFDAp	H-CFDAp	L-BA	P-BA	H-BA	L-R2Ap	P-R2Ap	H-R2Ap	L-R2Ju	P-R2Ju	H-R2Ju	4years
L-CFDAp	0	1.6	1.81	1	1.6	1.81	1.88	1.93	1.95	1.93	1.98	2	2
P-CFDAp	1.6	0	1.6	1.6	1	1.6	1.84	1.92	1.94	1.9	1.97	1.99	2
H-CFDAp	1.81	1.6	0	1.81	1.6	1	1.6	1.9	1.93	1.65	1.95	1.98	2
L-BA	1	1.6	1.81	0	1.6	1.81	1.88	1.93	1.95	1.93	1.98	2	2
P-BA	1.6	1	1.6	1.6	0	1.6	1.84	1.92	1.94	1.9	1.97	1.99	2
H-BA	1.81	1.6	1	1.81	1.6	0	1.6	1.9	1.93	1.65	1.95	1.98	2
L-R2Ap	1.88	1.84	1.6	1.88	1.84	1.6	0	1.88	1.91	1.05	1.93	1.96	2
P-R2Ap	1.93	1.92	1.9	1.93	1.92	1.9	1.88	0	1.6	1.93	1.05	1.65	2
H-R2Ap	1.95	1.94	1.93	1.95	1.94	1.93	1.91	1.6	0	1.96	1.65	1.05	2
L-R2Ju	1.93	1.9	1.65	1.93	1.9	1.65	1.05	1.93	1.96	0	1.88	1.91	2
P-R2Ju	1.98	1.97	1.95	1.98	1.97	1.95	1.93	1.05	1.65	1.88	0	1.6	2
H-R2Ju	2	1.99	1.98	2	1.99	1.98	1.96	1.65	1.05	1.91	1.6	0	2
4years	2	2	2	2	2	2	2	2	2	2	2	2	0

of competition. The only difference in the definition of substitution costs between the two phases (see Tables 2 and 3, pp. 9-10) is related to age categories: substitution costs between junior and apprentice competitions are higher during ante-senior phase than during senior phase.

Insertion-deletion cost (indel) is set to 1.3 for both phases, between minimum substitution cost for distinct states (1) and minimum substitution cost for states differing both in terms of ranking and pre-selection (1.6). In other words, it is less costly to turn a sequence ABC into AB than to turn ABC into ABD only if C and D are similar in ranking or in pre-selection contests organization. This indel definition takes into account the unequal length of sequences without making it the first criterion of distance between sequences.

After a comparison between several clustering methods[13], the outcome of Ward (1963) algorithm appeared as the most suitable one. The dendrogram produced indicates a 9-clusters solution nested in a 3-clusters solution, itself nested in a 2-clusters solution. I comment on the 9-clusters solution from the figures (2 to 4) and tables (4 to 7) presented below.

2.5 Two-phase structure as a key to interpreting clustering

I briefly describe each cluster. Three key elements of interpretation arose: participation in senior competitions, length and 'tonality' (that is the most often reached rank). Regarding high heterogeneity of data, clustering is quite noisy.

Cluster 1 gathers sequences characterized by a short but not empty ante-senior phase. Cluster 2 is defined by at least one 4-years void before the first participation in a senior competition. Sequences in cluster 3 share a symmetrical intensity regarding participation in ante-senior and senior phases. The senior phase is mainly characterized by 1st to 3rd ranks. Cluster 4 gathers short to medium length senior sequences mainly characterized by podium positions (2^d or 3rd ranks). Cluster 5 gathers short to medium length senior careers too, but mainly characterized by 1st ranks. Cluster 6 gathers short failure senior careers. Cluster 7 gathers senior careers characterized

---

[13] Single, Complete, UPGMA, WPGMA, Ward and Beta-flexible (Maechler and al., 2016).

Fig. 2. Dendrogram
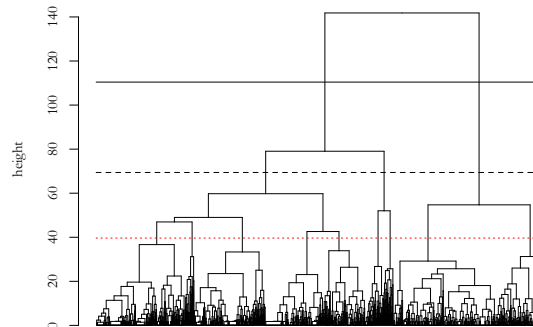
by a 4-years void after the first participation in a senior competition, that is mainly due to participation in specific competitions that are organized only every 4 years (the golden bricks on sequence index plots). Cluster 8 gathers mean to long senior careers of failures. Cluster 9 gathers long senior careers of success and is the closest to what can be called a cumulative pattern. Except for cluster 9, this clustering does not reveal a sharp connection between length and career structure.

How far does this clustering take synchronization into account ? Four points can be stressed:

- First, the 2-clusters solution separates clusters 1 and 2 from the seven other clusters. In other words, the 2-clusters solution separates careers firstly defined by ante-senior participations from careers firstly defined by senior participations.
- Second, clusters 1 and 2, both characterized by ante-senior participations, are distinct from one another with respect to participation in senior competitions.
- Third, a quarter of sequences counting one or more ante-senior participations are not clustered in clusters 1 and 2. In other words, closeness does not only rest on the (non-)emptiness of phases, but also on phases' composition (what I called tonality).
- Fourth, when, as here, synchronization is endogenous, multiphase structure simplifies greatly the interpretation. Once the phase mainly portrayed by each cluster has been highlighted, clustering's interpretation is based primary on ranking.

Conclusion

As is often the case, the structure of this paper is the exact reverse of the research story. What I have called MPOM from page 1 was first an *ad hoc* method suited to compare a set of careers of participation in *pâtissiers*' competitions. It was a kind of

12   Thomas Collas

Fig. 3. Sorted Sequence Index Plot
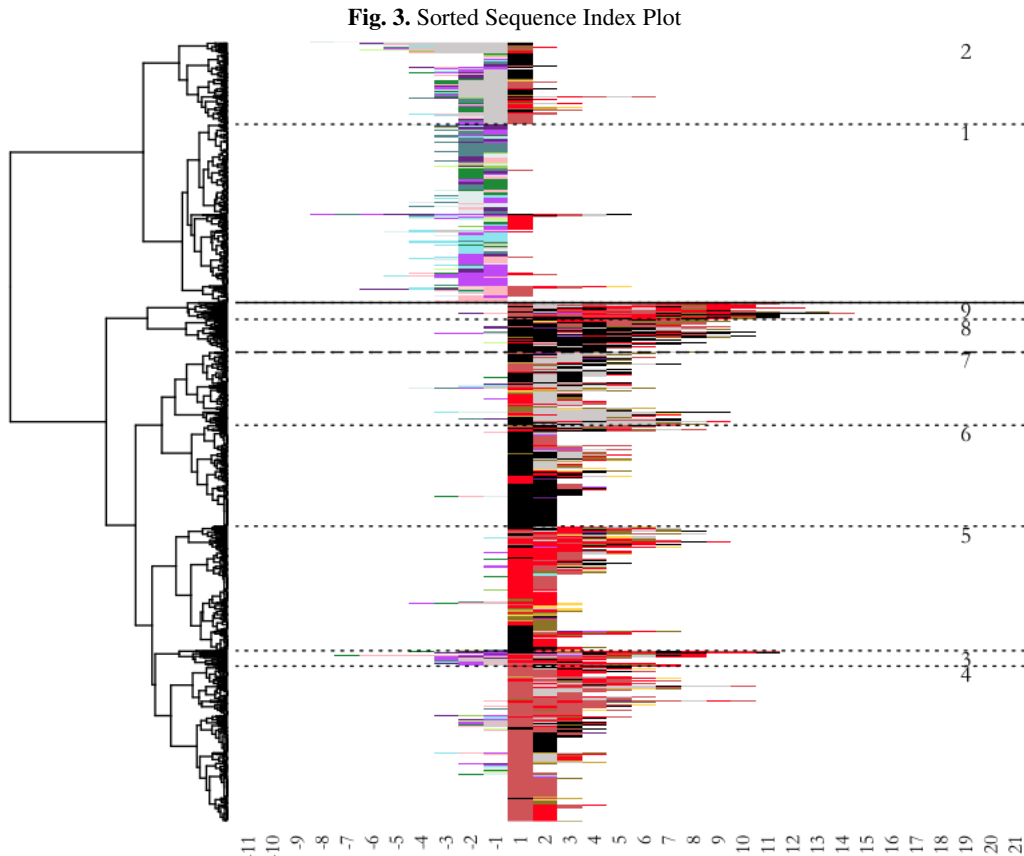
N.B. : This sorting according to the dendrogram was suggested by Joseph Larmarange (2013) and is based on R heatmap function (R Core Team, 2015). Line types refer to the three different clustering solutions kept (plain line for 2-clusters, dashed for 3-clusters and dotted for 9-clusters).

bricolage, a simple way of keeping synchronization while computing distances. That brought me to figure out what synchronization means. Let us underline two features.

First, synchronization means making phases a main character in the story. It is not at all a result, but a postulate made at the beginning of the study. Identifying what phases appear as the most important to explain differences can be regarded as a result. On this point, relations between phases and sequences could be investigated further.

Second, synchronization obviously means distorting the time axis to replace it with an odd sliced one, especially when the number of phases is superior to two. The aim of this calendar timing slaughter is to preserve an assumed social timing with phases regarded as the coherent trajectories within a career[14]. MPOM can be an

_____

[14]For sure, these are two heavy sociological postulates: something is a career – often because different events imply the same human being – and some others things are trajectories

Fig. 4. MDS Sequence Index Plot for each cluster
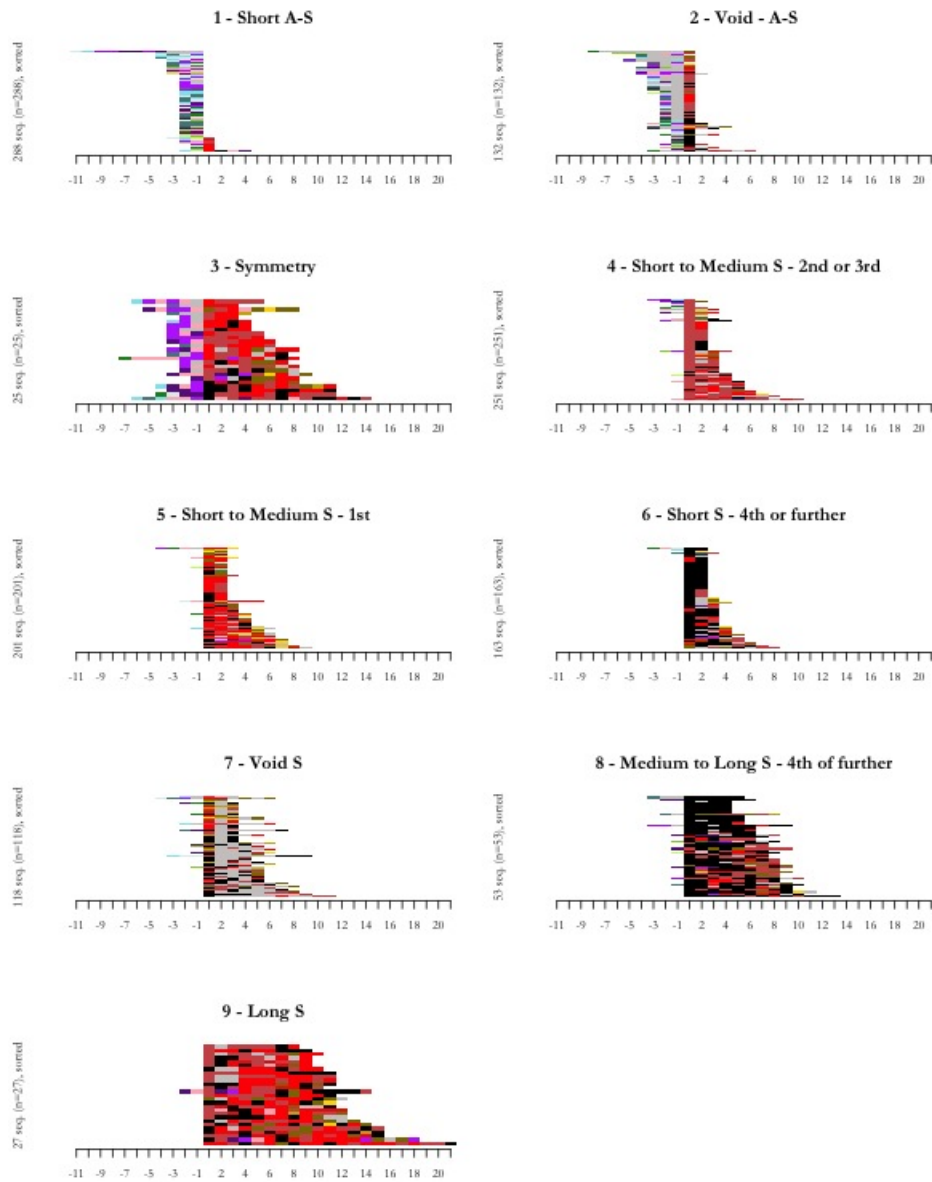
N.B : Following Piccarretta and Lior (2010), sequences are sorted according to their score on the first factor derived by applying multidimensional scaling (MDS) to the dissimilarity matrix.
A-S = Anter-Senior ; S = Senior.

14   Thomas Collas

Table 4. Mean state distribution by cluster within Ante-Senior phases for which length is superior to zero (in length percentage)

	1		2		3		4		5		6		7		8		9	
$N =$	288		132		25		53		19		9		19		11		1	
Mean/St. dev.	$\bar{X}$	σ	$\bar{X}$	σ	$\bar{X}$	σ	$\bar{X}$	σ	$\bar{X}$	σ	$\bar{X}$	σ	$\bar{X}$	σ	$\bar{X}$	σ	$\bar{X}$	σ
L-R2Ap	13	22.9	3.2	14.8	1	5	10.4	28.3	4.4	13.4	3.7	11.1	10.1	21.3				
P-R2Ap	14.2	25.7	6.4	17.5	4.3	10.5	10.7	28.1	5.3	22.9	11.1	33.3	12.3	21.4				
H-R2Ap	15.6	27.5	7.4	19.9	6.7	14.4	3.8	16.6	8.8	21.1	11.1	33.3	10.1	21.3	16.7	32.5		
L-R2Ju	13.8	26.4	9	20.5	17.6	21.8	21.1	35.7	17.1	33.4	37	48.4	2.6	11.5			50	
P-R2Ju	17.7	28	11	22.3	31.7	28.1	17.6	30.4	22.4	41.6	5.6	16.7	15.4	33.8	15.2	34.5		
H-R2Ju	9.5	21	6.8	20.2	17.8	26.6	8.5	18.1					21.1	38.4	36.4	50.5	50	
L-BA	2.3	9.7	0.3	2.9			3.8	16.6	7.9	25.1					9.1	30.2		
P-BA	3.3	12.6	0.4	3.6			4.4	17	10.5	31.5			10.5	31.5	9.1	30.2		
H-BA	8.2	19	6.9	19.6	1.6	5.7	5.7	21.2	11.8	31.6	25.9	43.4	9.6	25.6				
L-CFDAp	0.2	2.9	0.3	2.9			1.9	13.7	5.3	22.9								
P-CFDAp			1.1	9.2					5.3	22.9								
H-CFDAp	1	10.2	2.3	8.5			2.8	15.2										
4years	1.1	6.3	44.9	25.4	19.3	26.2	9.4	19.7	1.3	5.7	5.6	16.7	8.3	17.3	13.6	24.5		

Table 5. Mean state distribution by cluster within Senior phases for which length is superior to zero (in length percentage)

	1		2		3		4		5		6		7		8		9	
$N =$	49		130		25		251		201		163		118		53		27	
Mean/St. dev.	$\bar{X}$	σ	$\bar{X}$	σ	$\bar{X}$	σ	$\bar{X}$	σ	$\bar{X}$	σ	$\bar{X}$	σ	$\bar{X}$	σ	$\bar{X}$	σ	$\bar{X}$	σ
L-R2Se	51	45.3	16.7	33.8	39.3	24	13.1	18.8	44.4	26.8	7.6	15.4	11.7	16.9	5.8	8.9	32.8	11.3
P-R2Se	36.3	45	29.7	43.2	29.7	22	60.2	25.9	14.6	18.5	13	20.6	11	14.6	16.4	17.7	28	16.2
H-R2Se	3.6	11.6	41.8	47.1	13	14.5	11.9	19.4	14.4	19.6	61.9	30.1	22.9	22.1	65.7	24.6	20.7	13.6
L-CM	1.1	5.5	0.7	4.5	0.4	2.2	0.5	3.9	3.5	10.7	1.4	7.2	0.8	4	0.2	1.5	0.3	1.7
P-CM			0.9	6	1.4	5.2	1.6	7.5	1.9	8.5	0.5	4.7	2.4	10.6	0.2	1.7	0.2	1.3
H-CM			0.3	2.9	1.1	3.7	2.2	8.9	6.5	15.5	1.4	5.6	0.9	5	1.8	6.4	3	4.1
L-CFDSe			1.7	12.5	0.5	2.5	0.4	4.5	1	6.2	0.4	2.9						
P-CFDSe			0.4	4.4			0.1	2.1	0.7	5.1	0.1	1.6	1.1	4.8			0.3	1.6
H-CFDSe	2	14.3	2.3	15.1	0.5	2.5	0.7	4.8	2.6	10.5	1.7	6.5	3.8	13.4	1	5.5	0.3	1.6
L-MOF															0.2	1.2	0.3	1.4
H-MOF	0.5	3.6	3.7	17.9	5.9	10.9	0.5	3.7	3.6	10.4	1.5	6.3	4.1	10.6	1.3	3.7	4.1	5.1
L-R2Ap							0.2	3.2	0.2	3.5								
P-R2Ap	0.5	3.6					0.2	3.2	0.7	5.5	0.3	3.9						
H-R2Ap											0.4	4.1						
L-R2Ju	2.6	10.5			1.1	3.7	0.3	3.3	0.7	5.3	0.2	2	0.2	2.3	0.5	2.4	1.1	3.6
P-R2Ju	1	7.1			1.4	4	0.6	5.5	0.4	3.1	1.2	6.9	0.1	1.5	0.6	2.6	0.9	3.4
H-R2Ju	0.5	3.6			0.7	2.6	0.3	3.6	0.7	4.9	0.8	6.1	0.5	3	0.9	4.1	0.5	1.7
L-BA							0.1	1.6										
P-BA							0.1	1.6	0.1	1.4								
H-BA									0.2	2.4	0.2	2.6			0.2	1.7		
4years	0.9	4.5	1.9	7.8	5	9.6	6.9	13.4	3.8	9	7.4	12.9	40.5	13.8	5	8	7.5	8.6

alternative to multiple-sequence analysis (Pollock, 2007 ; Gauthier and al., 2010) if one channel appears to be tightly structured along phases. MPOM can be combined with this definition of multi-channel distance if more than two channels are taken into account. For example, individual sequences can be defined as a combination

---

within this career – often because we consider that some events are more challenging than others for this same human being.

Table 6. Some features of the 9 clusters

Cluster	Frequency		Mean value of three metrics		
			Distance	*Complexity*	*Length*
1 - Short Ante-Senior	288	22.9 %	4.31	0.41	2.67
2 - Void Ante-Senior	132	10.5 %	5.51	0.53	4.05
3 - Symmetry	25	2.0 %	12.66	0.61	10.20
4 - Short to Medium Senior - 2^d or 3rd	251	20.0 %	4.56	0.41	3.59
5 - Short to Medium Senior - 1st	201	16.0 %	4.96	0.46	3.56
6 - Short Senior - 4th or further	163	13.0 %	3.44	0.33	3.11
7 - Void Senior	118	9.4 %	6.30	0.50	4.97
8 - Mean to Long Senior - 4th or further	53	4.2 %	7.26	0.34	7.51
9 - Long Senior	27	2.1 %	12.84	0.54	12.07

Table 7. Sets of representative sequences for each cluster

Cluster	Ante-senior phase	Senior phase
1	(H-R2Ap,1)-(P-R2Ju,1)	
2	(P-R2Ju,1)-(4years,1)	(H-R2Se,1)
3	(P-R2Ju,1)-(4years,1)	(P-R2Se,1)-(L-R2Se,2)
	(H-R2Ap,1)-(4years,1)-(P-R2Ju,1)	(L-R2Se,1)-(P-R2Se,2)-(L-R2Se,1)-(4years,1)-(L-R2Se,1)-(H-R2Se,1)
	(P-R2Ju,1)-(L-R2Ju,1)-(H-R2Ju,1)	(L-R2Se,1)-(P-R2Se,4)
4		(P-R2Se,2)
5		(L-R2Se,2)
6		(H-R2Se,2)
7		(H-R2Se,1)-(4years,2)-(H-R2Se,1)
8		(H-R2Se,5)
9		(P-R2Se,1)-(H-R2Se,1)-(P-R2Se,1)-(L-R2Se,1)-(P-R2Se,2)-(L-R2Se,2)-(H-R2Se,1)-(P-R2Se,1)
		(P-R2Se,1)-(H-R2Se,1)-(L-R2Se,1)-(H-R2Se,1)-(P-R2Se,1)-(4years,1)-(L-R2Se,1)-(H-CM,1)-(L-R2Se,1)
		(P-R2Se,1)-(L-R2Se,1)-(P-R2Se,3)-(L-R2Se,1)-(H-R2Se,1)-(H-MOF,1)-(L-R2Se,2)-(H-R2Se,1)
		(P-R2Se,1)-(L-R2Se,2)-(H-R2Se,1)-(L-R2Se,2)-(H-R2Se,3)
		(P-R2Se,1)-(4years,2)-(L-R2Se,3)-(P-R2Se,1)-(L-R2Se,2)-(P-R2Se,1)-(L-R2Se,1)
		(P-R2Se,1)-(4years,2)-(P-R2Se,3)-(L-R2Se,3)-(P-R2Se,2)
		(P-R2Se,1)-(L-R2Se,1)-(H-R2Se,1)-(P-R2Se,2)-(L-R2Se,2)-(H-MOF,1)-(4years,2)-(H-R2Se,1)
		(P-R2Se,1)-(4years,1)-(P-R2Se,4)-(H-R2Se,1)-(L-R2Se,1)
		(H-R2Se,2)-(L-R2Se,2)-(H-R2Se,1)-(L-R2Se,1)-(P-R2Se,1)-(L-R2Se,1)-(H-R2Se,1)-(H-CM,1)-(L-CM,1)

N.B.: The centrality criterion has been used to extract these sets of representatives (Gabadinho and Ritschard, 2013). Coverage threshold is 50 % and neighborhood radius is 30 %. In other words, in each cluster, the distance of at least one sequence out of two to one of the representative sequences is inferior to 30 % of the maximum distance within each cluster. Sequences are sorted by representativeness. This presentation of sequences follows Abbott and Hrycak (1990), what has been named State-Permanence Sequence (SPS) format (Aassve and al., 2007): each state is followed by its number of successive occurrences.

of employment and residential sequences divided into phases defined by a marital sequence.

MPOM is thus a simple by-product of OMA, a by-product which is heuristic if sequences are defined by a set of phases and which takes into account the succession of states within phases. That is a major difference with other methods making the postulate of a phase division, such as QHA (Deville, 1974).

16   Thomas Collas

References

Aassve, A., Billari, F. Piccarreta, R. (2007). Strings of Adulthood: A Sequence Analysis of Young British Women's Work-Family Trajectories. *European Journal of Population*, 23-3. 369-388.

Abbott, A. D. (1992). What Do Cases Do ? Some Notes on Activity in Sociological Analysis. In Becker, H. S., & Ragin, C. C. (eds.). *What is a Case ? Exploring the Foundations of Social Inquiry* (pp. 53-82), Cambridge: Cambridge university.

Abbott, A. D. (1997). On the concept of turning point. *Comparative Social Research*,16. 85-106.

Abbott, A. D., & Forrest, J.. (1986). Optimal Matching Methods for Historical Sequences. *The Journal of Interdisciplinary History*, 16. 471-494.

Abbott, A. D., & Hrycak, A.. (1990). Measuring Resemblance in Sequence Data: An Optimal Matching Analysis of Musicians' Careers. *American Journal of Sociology*, 96-1. 144-185.

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

Becker, H. S. (1963). *Outsiders: Studies in the Sociology of Deviance*. New York: The Free Press.

Blanchard, Ph. (2010). Analyse séquentielle et carrières militantes. Research Report. http://hal.archives-ouvertes.fr/hal-0047619. Accessed 10 January 2015.

Collas, T. (2015). La pâte et le décor: Considération et formes professionnelles dans le monde des pâtissiers. PhD Thesis. Sciences Po.

Colombi, D., & Paye, S. (2014). Synchronising Sequences: An Analytic Approach to Explore Relationships Between Events and Temporal Patterns. In Blanchard, Ph., Bhlmann, F., Gauthier, J.-A. (Eds.). *Advances in Sequence Analysis: Theory, Method, Applications* (pp. 249-264), New York Heidelberg Dordrecht London: Springer.

Deville, J.-C. (1974). Méthodes statistiques et numériques de l'analyse harmonique. *Annales de l'Insee*, 15. 3-101.

DiPrete, T. A., & Eirich, G. M. (2006). Cumulative Advantage as a Mechanism for Inequality: A Review of Theoretical and Empirical Developments. *Annual Review of Sociology*, 32. 271-297.

Elzinga, C. H. (2014). Distance, Similarity and Sequence Comparison. In Blanchard, Ph., Bhlmann, F., Gauthier, J.-A. (Eds.). *Advances in Sequence Analysis: Theory, Method, Applications* (pp. 51-73), New York Heidelberg Dordrecht London: Springer.

Gabadinho, A. & Ritschard, G. (2013). Searching for typical life trajectories applied to child birth histories In R Lévy & E. Widmer (eds.), *Gendered Life Courses* (pp. 287-312). Vienna: LIT.

Gabadinho, A., Ritschard, G., Mller, N. S., Studer, M. (2011). Analyzing and Visualizing State Sequences in R with TraMineR. *Journal of Statistical Software*, 40(4). 1-37.

Gabadinho, A., Ritschard, G., Studer, M., Mller, N.S. (2010). "Indice de complexité pour le tri et la comparaison de séquences catégorielles", In *Extraction et gestion des connaissances (EGC 2010), Revue des nouvelles technologies de l'information RNTI*. E-19. 61-66.

Gauthier, J.-A., Bhlmann F., Blanchard Ph. (2014). Introduction: Sequence Analysis in 2014. In Blanchard, Ph., Bhlmann, F., Gauthier, J.-A. (Eds.). *Advances in Sequence Analysis: Theory, Method, Applications* (pp. 1-17), New York Heidelberg Dordrecht London: Springer.

Gauthier, J. A., Widmer, E. D., Bucher, P., Notredame, C. (2010). Multichannel sequence analysis applied to social science data. *Sociological Methodology*, 40. 1-38.

Giudici, F., & Gauthier, J. A. (2009). Différenciation des trajectoires professionnelles liée à la transition à la parentalité en Suisse. *Revue suisse de sociologie*, 35. 253-278.

Lang, G., & Lang, K. (1988). Recognition and Renown: the Survival of Artistic Reputation. *American journal of sociology*, 94. 79-109.

Larmarange, J. (2013). Représenter un tapis de séquences ordonnées selon un dendrogramme. http://joseph.larmarange.net/?Representer-un-tapis-de-sequences. Last access: Feb. 2nd 2015.

Lesnard, L. (2008). Off-Scheduling within Dual-Earner Couples : An Unequal and Negative Externality for Family Time. *American Journal of Sociology*, 114. 447-490.

Lesnard, L. (2014). Using Optimal Matching Analysis in Sociology: Cost Setting and Sociology of Time. In Blanchard, Ph., Bhlmann, F., Gauthier, J.-A. (Eds.). *Advances in Sequence Analysis: Theory, Method, Applications* (pp. 39-50), New York Heidelberg Dordrecht London: Springer.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., Hornik, K.(2016). cluster: Cluster Analysis Basics and Extensions. R package version 2.0.4.

Menger, P.-M. (2009). *Le travail créateur: S'accomplir dans l'incertain*. Paris: Gallimard-Seuil.

Piccaretta, R. & Lior, O. (2010). Exploring sequences: a graphical tool based on multi-dimensional scaling. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 173-1. 165-184.

Pollock, G. (2007). Holistic Trajectories: a Study of Combined Employment, Housing and Family Careers by Using Multiple-Sequence Analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 170. 167-183.

18   Thomas Collas

R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria.

Robette, N., & Bry, X. (2012). Harpoon or Bait? A Comparison of Various Metrics in Fishing for Sequence Patterns. *Bulletin de Mthodologie Sociologique*, 116. 524.

Robette, N., & Thibault, N. (2008). Analyse harmonique qualitative ou méthodes d'appariement optimal?. *Population*, 63(4). 621646.

Stovel, K., & Bolan, M. (2004). Residential Trajectories: Using Optimal Alignment to Reveal the Structure of Residential Mobility. *Sociological Methods & Research*, 4. 559-598.

Stovel, K., Savage, M., Bearman, P. (1996). Ascription into Achievement: Models of Career Systems at Lloyds Bank, 1890-1970. *American Journal of Sociology*, 102-2. 358-399.

Ward, J. H. Jr. (1963). Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*, 58. 23644.

Wilensky, H. L. (1964). The Professionalization of Everyone?. *American Journal of Sociology*, 70. 137-158.

Wu, L. L. (2000). Some Comments on Sequence Analysis and Optimal Matching Methods in Sociology: Review and Prospect. *Sociological Methods & Research*, 29. 41-64.

Session 12A: Entry into the labor market

Employment pathways forecasting

What are the future prospects for young people after three years of vocational experience? Over/under-performing

P. Rousset[1], P. Trouvé[2] and Shirley Lawes[3]

[1] and [2] : Institute of Céreq, centre d'étude et de recherche sur les qualifications

[3] University College London

This paper is an exposition of an analysis of employment entry pathways. Through a sequential analysis it selects individuals who emerged with different career pathways than might have been expected considering their working position 3 years after leaving school. It identifies young people who have developed a secure career pathway through their vocational experience. The Céreq survey *Génération* was carried out in 2011 recording the monthly "occupational calendar" over a 7 year period of 12,365 young people who left the French school system in 2004. Sequential analysis has become increasingly popular following the work of Abbott (Abbott, 1983). Its success derives from the adoption of an holistic approach which focuses on the entire career of an individual (Elzinga, Studer, 2015 ; Studer, Ritschard, 2015 ; Robette, Bry, 2012). Sequential analysis is now frequently applied to vocational pathway studies (Massoni and al., 2009 ; Grelet, 2002). The relationship between different career pathways in identifying the most successful students has already been used in a number of studies indicating a strong link between the vocational pathway and the qualification gained (Céreq, 2007). In our analysis we take the opposite perspective, that of identifying individuals who exceed expectations during the period 2007-2011 in relation to their potential, job security (high/low or increasing/decreasing level), as identified by their vocational situation at 2007. An open-ended contract or a fixed-term contract are two current employment contracts in common use that are perceived as having different potentials in relation to establishing secure employment pathways. Potential is defined in this paper as the statistical probable pathway.

The aim is to compare the career pathway achieved by an individual with the statistically probable pathway that is representative of the expected pathway. Our problem naturally relates to defining the probability, based on empirical expectations, of reaching each point B in the future, starting at point A. Transitional approaches or Markovian Models lie within this overall approach. In sequential analysis a first approach is to translate any individual pathway as a Markovian model in order to be able to compare the actual career pathway with the expected pathway. A Markovien model for any individual pathway is achieved through Drifting Markov Model DMM (Massoni & al, 2010). Such an approach requires sufficiently rich individual information and therefore very lengthy sequences. The matrices of dissimilarity, as for Optimal Matching (Halpin, 2010 ; Studer and Ritschard, 2014), only compare actual pathways. They can't define a distance between actual career pathways and probable career pathways. As an alternative approach, Euclidian metrics allow for the projection from the past to the future. Therefore for individuals defined by their affiliation to a group, the probable career pathway is the mean pathway of the group and the gap between the actual and probable pathways measures inertia. We therefore selected Euclidian metrics methods to use in our analysis. The chi-square distance, applied directly between sequences, allows for controlling frequencies effects and does not take into account the order of sequences (Robette, Thilbault, 2008). Therefore the status of employment taken at two different points in time often had the same

potential for establishing secure employment pathways but not the same *frequency*. The methodology *"emlt"* outlined in Rousset and Giret (2007) and Rousset et al (2012) relates to our present work as it established a Euclidian metric between career pathways through the transitions matrix at every step giving weighting to the short and long term. An inter-pathway matrix enables the measurement of the distance to the probable pathway. From the point of view of our study, the position of individuals reflects their potential for developing more secure employment pathways, that is the probability of maintaining or making the transition towards status of employment or unemployment.

In terms of practical application, the literature indicates broadly that entry to the labour market in France is strongly linked to qualification, measured three years after the end of education. (Céreq, 2007). Our results show however, that the probable career pathway after three years is more closely linked to a more secure pathway than an insecure pathway which suggests that beyond the achievement of a qualification individuals remain active agents in the development of their career pathways. A typology of individuals according to their achievement in terms of degree of job insecurity referring to employment status and the gap between probable pathways illustrates this point. This analysis will be developed fully in the paper as will the potential for the development of security of professional status. In the first instance, the paper will expand on the methodological analysis of career pathways. It will show, through comparison with other methodological approaches, the link with transitions and its pertinence to vocational pathways. The second part of the article will examine the prediction of pathways. The third part will illustrate the findings applied to the establishment of secure career pathways.

1 The survey

The CEREQ 'Generation 2004-7' survey (2011) involved 12,365 young people exiting the school system in 2004. Over a period of 7 years (88 months) a monthly calendar of each individual's position in the job market was produced according to five categories of employment or unemployment status : *open-ended contract[1], fixed-term employment contract, qualification contract[2], temporary work,* other subsidised employment contracts and respectively *unemployment, not actively seeking employment, training or return to study* .

2. Methodology

The *emlt* methodology employed is presented in detail and its application in sequential analysis is developed in Rousset and Giret (2007) and in Studer and Ritschard (2015). The reader can find the *seqemlt* function in Ritschard et al (2015).The following summary presents an overview of its principles. The methodology specifies in particular the metric for the Euclidean space of transitions and for those inferred for career pathways. For a more detailed technical description the reader is referred to articles by Rousset et al (2012) and to Joseph et al (2013) for an example of an application. This methodology proposes an alternative to methods such as Optimal Matching

---

[1] Open-ended contract : permanent employment contract
[2] French Public policy for young people under 26 years old

(Abbott, 1983) and chi-square distance (Grelet, 2002) taking better account of transitions and of order and sequences. Beyond this it integrates two important points : that transitions develop over time and that transitions must be considered in one and several steps. In fact exchanges on Optimal Matching (Wu 2000) ; Halpin, (2010) reveal the difficulties of this method in taking account of transitions. As far as the chi-square distance is concerned, the co-concurrence which does not integrate the sequential order replaces the notion of transition. Harmonic analysis (Robette and Thibault, 2008) is designed to better adapt the approach through an interplay of windows where adherence to a window of time depends on sequential order. The methodology presented here goes further in defining the metric on the basis of transition rates at one and several steps by modulating the long and short term.

Definition and encoding of sequences

In the formal sense, the calendar of an individual *i* gives rise to the status category that he occupies at each *t moment.* The set *E* of available outcomes of status *e* is finite. The methodology is based on the transitions to one and several steps. These transitions are time dependent, for example transitions to work are always greater in the months following exit from the school system and lessens over time. In order to take account of this, the status categories are indexed over time, and are then called 'situations'. Subsequently, the term *situation s=e*, designates the position in status category *e* at instant *t.* Probable career pathways and transitions: the realm of probable career pathways

Let us consider a situation *s=e,* and the set Ω_s of an individual who have experienced this situation. For each instant *t'>t*, we can identify empirically the probability of transition of *s* towards the set of status categories *e'* from *t'* as the rate of individuals in Ω_s who have experienced the situation *s'=e'$_{t'}$*. We can then establish the probable career pathway of *s* as the vector of transition probabilities for each *t'* instant. Finally, we can identify the set of probable career pathways for the set of *s* situations*.* We call this set 'the realm of probable career pathways'. The realm of probable career pathways is defined by the matrix ($\Phi_s^{s'}$) which general term indicates the empirical probability to reach *s'* starting at *s.* Each line corresponds to a situation *s* and provides the probable career pathway from *s.*

- $\Phi_s^{s'}$ =0, if t'<t

- $\Phi_s^{s'} = \dfrac{\sum\limits_{i=1}^{I} Y_i^s Y_i^{s'}}{\sum\limits_{i=1}^{I} Y_i^s}$, where *Y* has the value of 1 if individual *i* has experienced situation *s,*

   otherwise 0

A metric for the field of probable career pathways

In the following we asses a situation and its own probable pathway that is also considered as its potential. The distance between two situations is defined as the one between two probable career pathways. A natural metric for defining the distance between two probable career pathways is the chi-square distance, such that the gap between two transitions does not depend on situational

frequency. The methodology enables us to balance time so as to give greater weight to correlations in the short-long term. When we need to balance the distance between transitional profiles of two situations by giving more weight in the short-term than in the long-term, we replace $\Phi_s^{s'}$ by its balanced value $\beta_s^{s'} \, \Phi_s^{s'}$ where $\beta_s^{s'}$ is a parameter that diminishes with time as in this study with

$$\beta_s^{s'} = \frac{1}{t'-t+1} \ .$$

A metric for the space of sequences based on transitions

Instead of using directly this metric, Torgerson's method (Torgerson 1958, Benzécri 1973) enables us to use the traditional Euclidean quadratic distance applied to transformed coordinates of situations. These coordinates are given by the *seqenlt* function. At that time, all Euclidean methods applied to this coordinates are available, giving a powerful tool for sequence analysis. The *seqenlt* function enables us to obtain transformed coordinates for achieved career pathways by defining a sequence as the sum of successive situations. The probably career pathway coordinates are then obtained averaging real pathways. The main aim of this paper is to show how this method defines, with the same metric, both the distance between actual and probable trajectories. In this way the study of global performance (one individual compared to others) and relative performance (one individual's actual performance compared to his or her own expected performance) are based on the same criterion (the same metric).

Properties of the methodology

The fundamental property of this methodology is to draw on the transitions matrix in that the distance between the pathway trajectories is a Euclidean distance which is based on the space between transitions. It thus draws on the order of the data where others do not take it into account. In our application of this methodology to career pathways, the status of the individual is not what is written in his/her contract but his/her future potential, particularly in the short-term. The internal distance between probable pathways and the internal distance between achieved pathways as well as the distance between probable and achieved pathways is the same, and it is this which guarantees coherence of approach and results.

Comparison with other methods

In a general sense, the advantage of this method is that it brings together two status categories, not only if there are links between them but also if they have the same links in relation to other status categories and therefore they play the same role in the pathway. Rousset et al's article (2012) presents a comparative study of employment access drawing on data from *Génération* after 7 years**.** The comparison in this article aims at an examination of the absolute performance of an individual in relation to others and not the relative performance in relation to his/her own individual potential. The graph (Figure 1), a chronogram, provides juxtaposition to cumulative histograms giving monthly details of the relative share that each status category has at a particular point in time. It thus represents the development over time (the abscissa) of the contribution of each status category of

each cohort participant. This type of representation gives a good picture of the development over time of a group of individuals. On the other hand, it masks the number of individual transitions by presenting an image of systematic and regular development. Figures 2, 3 and 4 show these latter results concerning the difference between the methodology presented here, optimal matching and the chi-square distance. The results show a greater ability to draw out tendencies, for example to show employment loss or Open-ended contracts (CDIs[3]) loss. This is the only methodology that provides a category of lost CDI. We see that Optimal Matching gives a great deal of importance to frequencies extracted from CDIs, two categories are dedicated to accelerated access to CDIs. The chi-square distance gives much greater importance to rare situations (qualification contracts are more rare five years after end of study). These two methodologies each therefore reflect the effect of frequencies. For the first, the frequencies are more linked to coding than to actual reality: by separating CDIs and civil service posts or by aggregating temporary contracts with fixed-term contracts (CDD[4]), the frequencies would have been different. The adjustment of frequencies by chi-square is not necessarily adequate either in cases where public policy is evaluated in the sense that public policy aims to increase the number of beneficiaries. The comparison of trends based on transitions (only for these are frequencies adjusted) limits the effect of frequencies: thus if CDDs and temporary contracts have the same transitions to stable employment, their impact will be added. This is why, in this article, we developed a methodology based on transitions which additionally enables us to work with probable trajectories.

Key to figures 1, 2, 3 and 4

---

[3] CDI: Contrat de Durée Indéterminée
[4] CDD: Contrat à Durée déterminée

Figure 1 : Chronogram of employment status of young people over 7 years, immediately following full-time education
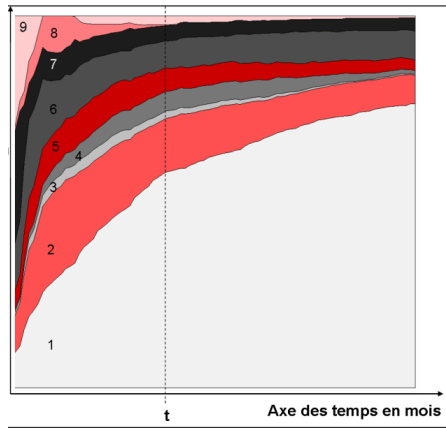
Figure 2: Typologie of career entry of young people after full-time education in 1998 over a 7 year period with the methodology presented here. Divided into 8 categories based on the seqemlt method (for key see key to figures 2, 3, 4).
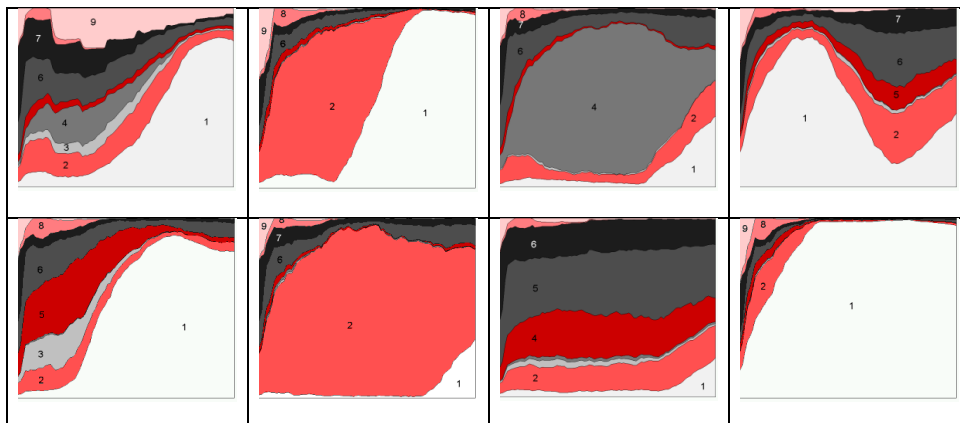
Figure 3 : Typology of career entry of young people after full-time education in 1998 over a 7 year period. Divided into 8 categories based on optimal matching (for key see key to figures 2, 3, 4).
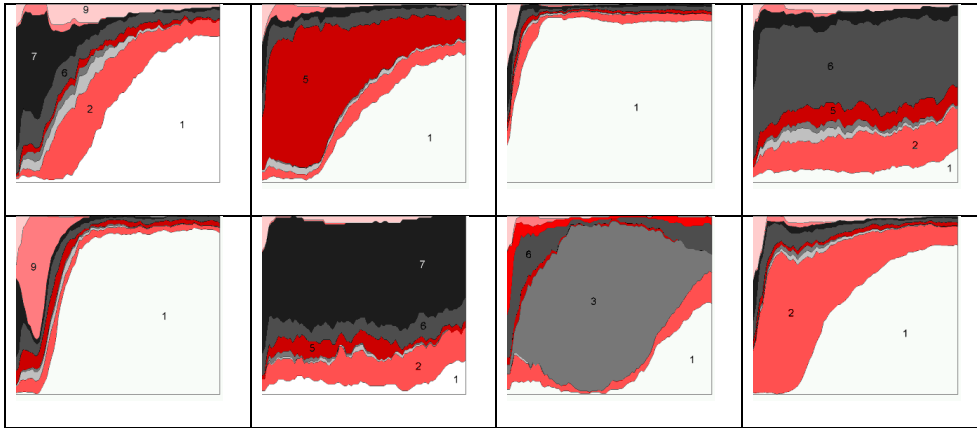
Figure 4 : Typologie of career entry of young people after full-time education in 1998 over a 7 year period. Divided into 8 categories based on harmonic analysis (for key see key to figures 2, 3, 4).
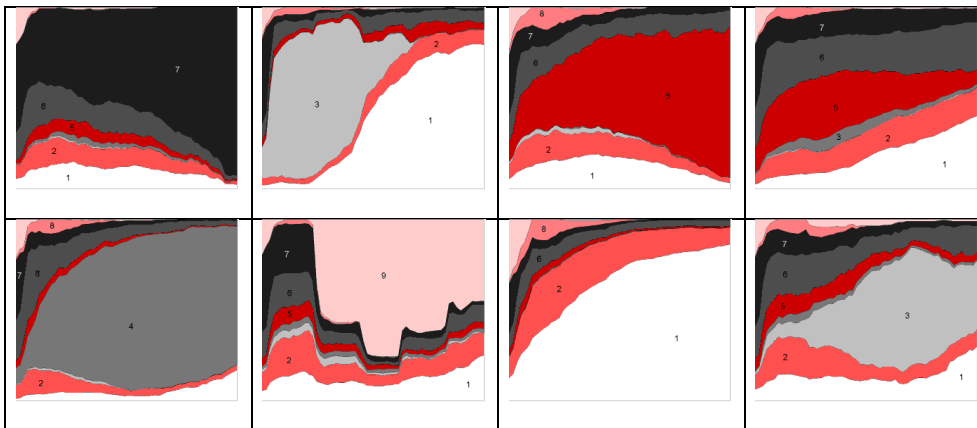
3. The Forecast

The aim of the forecast is to define the probable future career pathway of individuals based on knowledge of their present and past. Section 2 outlines the properties of this research methodology. It assimilates the position of each individual, at each *t* moment*,* with a potential for near and distant future. This potential is defined by the number of transitions as a whole. Furthermore, the distance between the achieved career pathway and the probable pathway is measured by a metric of associated space deduced from the transitions metric. The pathways are defined in Euclidian space. In this way, the mean of a group of individuals' pathway indicates the probable pathway of individuals in that group.

Definition:

Considering the referring position at t_0, the probable future pathway of a group is defined as the expectation or the mean of pathways over the period $\{t_0, ..., T\}$.

Properties:

The constitution of a group may be the total number of individuals in the same situation at t_0, referred to as prediction in relation to the present moment t_0. In this first case, the probable pathway for the group of individuals in status e at moment t_0 is specifically defined by the transitions $\Phi_s^{s'}$ in section 2, where s is the situation in status e at moment t_0 and t' each of the situations for $t>t_0$. This case is illustrated as follows. The distance between pathways is therefore exactly what enables us to define the distance between an actual pathway and a probably pathway. Remembering that this property guarantees the coherence of analysis for both actual and probably pathways.

The choice of the group may also be composed of individuals who have identical time sequences $\{t_0',$ ..., $t_0\}$, in which case individuals are described as having a common past, or else as individuals who are re-grouped in relation to a secondary variable. In the first case of re-grouping in relation to a time sequence $\{ t0', ..., t0\}$, a weighting may be used which gives more weight as t_0 gets closer in the same way as parameter β in section 2.

The mean future pathway, i.e. probable pathway, and the mean past pathway, when put together constitute a mean pathway for the whole period. This property, which gives greater coherence to the whole analysis and which broadly allows for interterpretations to be made, is not, for example, checked against the modal pathway: modal pathways over periods in the past and future do not necessarily constitute the modal pathway for the whole period.

4. Analysis of individuals who have over-performed

In our study, the forecast of a pathway enables us to identify individuals who have significantly over-performed in their expected pathway. The first result that we obtained through this method is to be able to measure the quadratic deviation between the achieved pathway and the probable pathway. In a group of individuals, the smaller the probability of deviation from the probable pathway, the lowerwill be the performance. In figure 5, employment signifies lower performance in the case of probable pathway 2 than in the case of probable pathway 1. This property is not verified either with the modal pathway which does not depend on the quadratic deviation.

Figure 5 : In case of probable pathway 2, the rate of individual in employment increase more in time than in case of probable pathway 1. Therefore, the performance of "being in employment" is smaller than in case of probable pathway1.
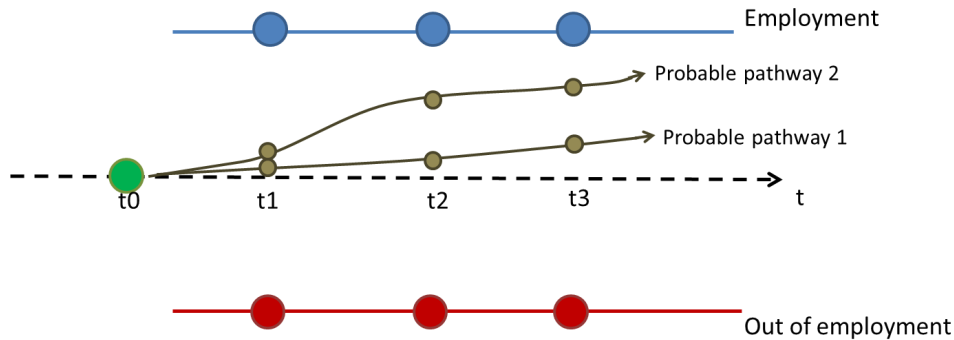
In the context of our application, we consider the position of 12,365 young people in the *Generation* study of 2004 in the labour market 3 years after the end of their studies according to a list of five employment and four unemployment status categories : *open-ended contract[5], fixed-term employment contract, qualification contract[6], temporary work, other subsidised employment contracts* and respectively *unemployment, not actively seeking employment, training or return to study.* We are seeking to find out if the employment pathway that they have achieved beyond this date corresponds to an expected pathway, an over-performing pathway, or an under-performing pathway. The expected pathway depends here on the position at a point 3 years after the end of schooling. Thus, an individual on a fixed-term contract after 3 years deviates from his/her expected pathway if the deviation D between the achieved pathway and the probably pathway is great. The level of the deviation subsequently considered is given by the variable D.

In order to know if the pathway is over or under performing, the distance between achieved pathways and the 8 patterns of pathways, namely open-ended contract, fixed-term contract, temporary contract, qualification contract, other subsidised employment contracts, unemployed, not actively seeking employment, return to study. The principle component of these 8 variables is the indicator Q, chosen to measure the quality of status. The analysis of correlations shows that the indicator Q   contrasts open-ended contracts (correlation –0.4) with other   less secure status categories (correlation between 0.4 and 0.5) and is thus an indicator of employment stability. Two classifications which emerge respectively from variables D and Q have been established. The intersection of typologies in Table 1 identifies the intersection of the two dimensions, D in lines and Q in columns.

  It shows that, in most cases in the Table, there is a gap between the future trajectory achieved and the probable trajectory with the exception of young people who had already obtained an open-ended contract. This implies that after 3 years there is still substantial uncertainty around future trajectories and that young people's efforts to find employment immediately after the end of their

---

[5] Open-ended contract : permanent employment contract
[6] French Public politics for young people under 26 years old

schooling is of great importance to the trajectory. Only young people who obtained an open-ended contract during the first three years have a secure future and this corresponds to a low gap between the probable trajectory and high performance. These young people are in a group where absolute high performance can constitute relative performance in line with expectations. This observation justifies the identification of an intersection between relative and absolute performances. Thus in cases where the future is more or less determined, as for those with open-end contracts after three years, the level of the diploma or social features have been some of the main determining factors. For all the other cases, the gap between the probable and achieved pathway is medium or high and the security of the pathway has yet to be stabilised. Where qualifications or family characteristics have not sufficed, the efforts made by the young person or the initial working environment are the determining factors. In the rest of this article we will confine our analysis to that which pertains to employment status at the point 3 years after the end of schooling. Other factors, such as the characteristics of types of companies or discrimination will be the subject of further studies.

If we study the performance for each situation at a 3 year point, as pointed out above and examined in Table 2, the open-ended contract is the most secure. Eighty-four percent of young people achieve a very successful pathway with a small gap between the probably and actual pathway. The risk for young people with open-ended contracts is thus very low. Moreover, the gap between the probable and actual pathway would need to be very large for there to be significant under performance (2.5% of open-ended contracts). For employment with temporary status, fixed-term contracts are shown to have greater security potential than temporary work, as showed respectively in Tables 3 and 4 . In effect, fixed-term contract show a medium gap in the probable pathway allowing for high and medium performances (respectively 23% and 22% of fixed-term contracts), while it allows essentially for average performance for temporary work (4% and 23% average). In the same way, with fixed-term contracts high performance is obtained with a large or medium gap (respectively 18% and 23%) as opposed to a high gap for temporary work (4% medium and 18% high). Nevertheless, for temporary as for fixed-term contracts, under-performance constitutes a large gap which shows that young people have a greater tendency to make their pathway secure by their own efforts to find employment rather than to fall into insecurity. What is more, underperformance is generally a result of insecure employment pathways where there is a low incidence of unemployment. In respect of subsidised work, we find a more uncertain configuration where most pathways show a large gap with probably pathways (77% of young people in the line 'high gap'), which reveals a situation of greater uncertainty that young people find themselves in. Nevertheless, 18% are able to gain more secure employment by obtaining an Open-ended contract which is evidence of high performance. Average performance correlates with a reduction in job insecurity, through a belated Open-ended contract (17%) or for 22% a less insecure status (resumption of studies, temporary work). The 37% of under-performing pathways correspond to pathways where the incidence of unemployment is at approximately 50% on average. Young people who are unemployed at the 3 year point, are equally in a very uncertain position since their future trajectories represent a large gap with the probable pathway (80% of young people are in the 'high gap' line). Furthermore, the population is divided into two more or less equal parts: 46% achieve medium performance often with an Open-ended contract at the end of the period, while 54% experience a pathway where employment is often temporary and unemployment accounts for more than half of the period.

Key to tables 1to 6

9	initial training
8	return to training
7	not actively seeking employment
6	unemployment
5	temporary work
4	other subsidised contracts
3	qualification contract
2	fixed-term contract
1	open-ended contract

Table 1: Who is over/under-performing 3 years after living school? Chronogram of the cohort of school leavers in 2004 . Employment pathways in terms of contracts : 1- Open-ended, 2- Fixed-term, 3- Qualification, 4- Subsidised work, 5- Temping, 6- Unemployment, 7- Inactivity, 8- Training, 9- Studies
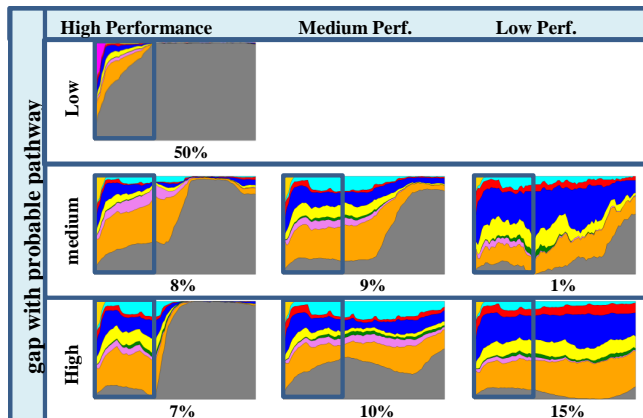
Table 2: Who is over/under-performing 3 years after living school when the situation at 3 years is a **Open-ended contract**? Chronogram of the cohort of school leavers in 2004 . Employment pathways in terms of contracts : 1- Open-ended, 2- Fixed-term, 3- Qualification, 4- Subsidised work, 5- Temping, 6- Unemployment, 7- Inactivity, 8- Training, 9- Studies
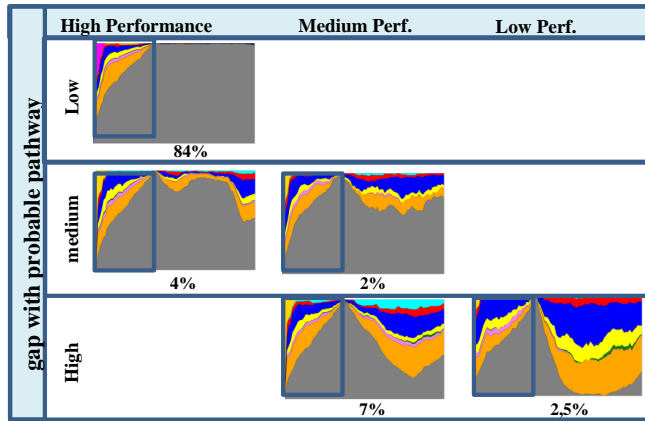
Table 3: Who is over/under-performing 3 years after living school when the situation at 3 years is a **Fixed-term contract**? Chronogram of the cohort of school leavers in 2004 . Employment pathways in terms of contracts : 1- Open-ended, 2- Fixed-term, 3- Qualification, 4- Subsidised work, 5- Temping, 6- Unemployment, 7- Inactivity, 8- Training, 9- Studies
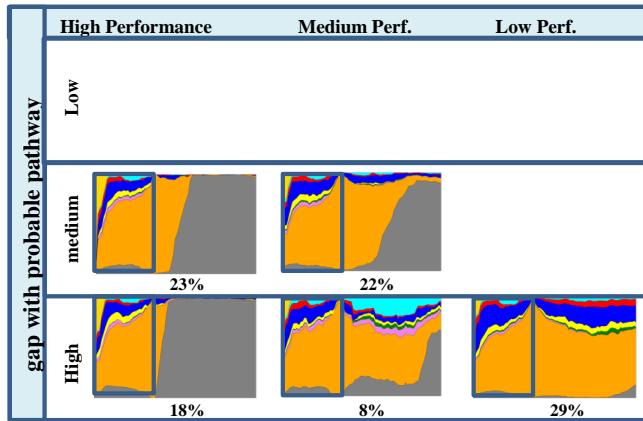
Table 4: Who is over/under-performing 3 years after living school when the situation at 3 years is a **Temporary work**? Chronogram of the cohort of school leavers in 2004 . Employment pathways in terms of contracts : 1- Open-ended, 2- Fixed-term, 3- Qualification, 4- Subsidised work, 5- Temping, 6- Unemployment, 7- Inactivity, 8- Training, 9- Studies
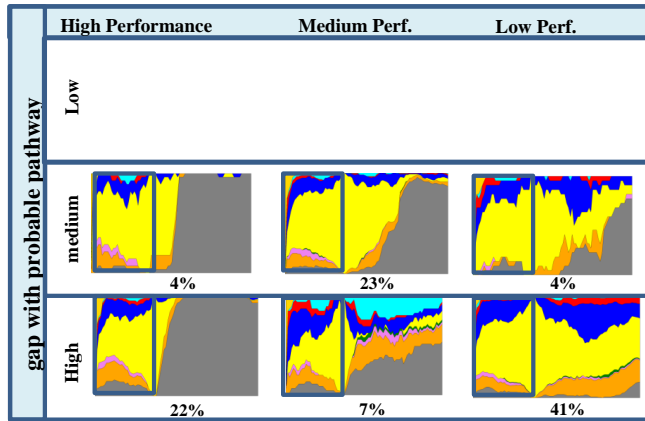
Table 5: Who is over/under-performing 3 years after living school when the situation at 3 years is **Subsidised work**? Chronogram of the cohort of school leavers in 2004 . Employment pathways in terms of contracts : 1- Open-ended, 2- Fixed-term, 3- Qualification, 4- Subsidised work, 5- Temping, 6- Unemployment, 7- Inactivity, 8- Training, 9- Studies
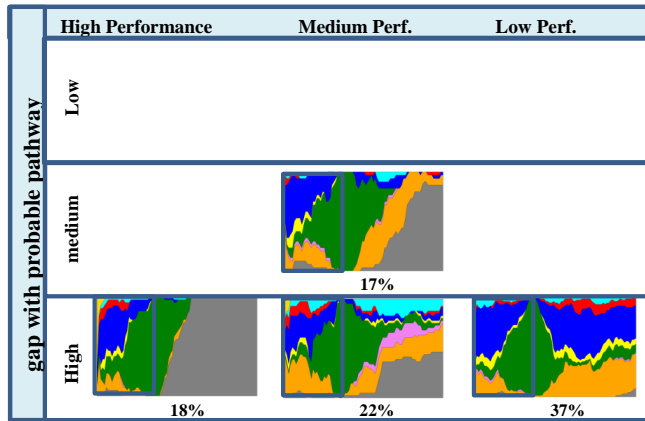
Table 6: Who is over/under-performing 3 years after living school when the situation at 3 years is **Unemployment**? Chronogram of the cohort of school leavers in 2004 . Employment pathways in terms of contracts : 1- Open-ended, 2- Fixed-term, 3- Qualification, 4- Subsidised work, 5- Temping, 6- Unemployment, 7- Inactivity, 8- Training, 9- Studies
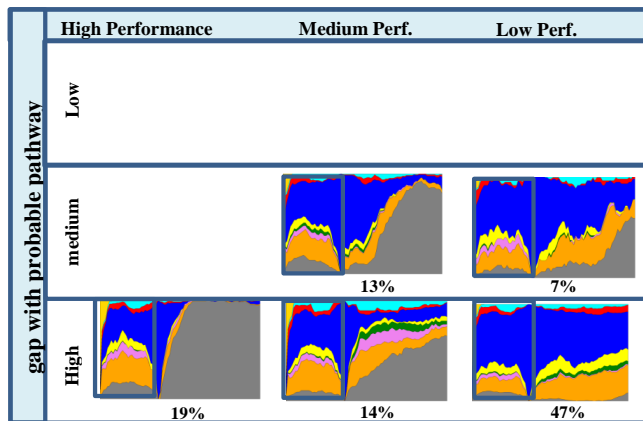
Conclusion

A comparison of the pathway achieved by an individual with the probable pathway enables the potential that an individual achieves in his/her pathway to be taken into account. The method used in this article enables both an examination of the absolute performance of an individual in relation to others and also the relative performance in relation to his/her own individual potential. The essential contribution of this article is to explain how our method takes into account the potential of an individual and can draw simultaneously on the two criteria that are the absolute performance and gap with probable pathway. This contribution is derived directly from the method and the metric presented in the methodology section.

In terms of application, we have been able to classify young people entering the labour market in relation to their ability to improve their situation by capitalising on their activity in employment. The methodology, significantly, enabled us to distinguish performance in absolute terms, due to qualifications and social categories (as mentioned in several studies), and also to relative performance linked to young peoples' proactive approach to job seeking.. In this article we have highlighted the difference in potential between different employment status groups. Moreover, the methodology we employed offers numerous possibilities of further exploration beyond status categories. We intend to use it to analyse the impact of companies on the provision of secure jobs; in particular to verify if large companies, companies belonging to networks or family-owned companies have the same perspectives. The methodology should enable us to identify cases where a company is able to affect the potential of an individual, from a company that recruits young people who already show good potential. In the same way, through an analysis of perceived discrimination, it could enable us to resolve the traditional problem of distinguishing outcomes resulting from the negative impact of discrimination from outcomes that result from a personal feeling of failure.

References

- Abbott, A. (1983) Sequences of social events: concepts and methods for the analysis of order in social processes. Hist. Meth., 16, 129–147.

- Benzecri J.P., 1973, L'Analyse des Données : TIIB n°2 : représentation Euclidienne d'un Ensemble fini de masses et de distances, Paris, Dunod, p. 65-95.

- Céreq – Centre d'études et de recherches sur les qualifications, 2007. *Quand la carrière commence… Les sept premières années de vie active de la Génération 98,* Marseille.

- Elzinga, C. H. and Studer, M. (2015) Spell sequences, state proximities and distance metrics. Sociol. Meth. Res.,44, 3–47.

- Grelet, Y. (2002) Des typologies de parcours: méthodes et usages. Notes de Travail Génération 92. Céreq, Paris.

- Joseph O., Lemière S., Lizé L., Rousset P., (2013), The Feeling of Dis-crimination and Job-Market Entry in France, Brussels Economic Review , Editions du DULBEA, 2013, 56 , pp.5-42

- Halpin, B. (2010) Optimal matching analysis and life-course data: the importance of duration. Sociol. Meth. Res.,38, 365–388.

- Massoni, Olteanu, Rousset (2009) Career-Path Analysis Using Optimal Matching and Self-Organizing Maps, Advances in self Organizing maps, Lecture Notes in computer science, vol 5629, 2009, Springer Berlin / Heidelberg.

- Massoni, Olteanu, Rousset. Career-path analysis using drifting Markov models (DMM) and self-organizing maps. MASHS, 2010, Lille, France. 2010.

- Robette,N. and Bry, X. (2012) Harpoon or bait?: a comparison of various metrics in fishing for sequence patterns.Bull. Sociol. Methodol., 116, 5–24.

- Robette N., Thilbault, N., (2008), « Analyse harmonique qualitative ou méthode d'appariement optimal? Une analyse exploratoire de trajectoires profes-sionnelles », Population, 63(4), 621-646.

- Rousset P., Giret J.F., (2007), « Classifying qualitative time series with SOM: the typology of career paths in France », in Sandoval F., Prieto A., Cabes-tani J., Grana M. (ed.), Computation and Ambiant Intelligent, Iwann 2007 proceeding, Lecture Note in Computer Science, Berlin, Springer, p. 757-764

- Rousset, P., Giret, J.-F. and Grelet, Y. (2012) Typologies de parcours et dynamique longitudinale. Bull. Sociol. Methodol., 114, 5–34.

- Studer, M. and Ritschard, G. (2014) A comparative review of sequence dissimilarity measures.Working Paper 33. Swiss National Center of Competence in Research LIVES, Geneva.

-       Studer, M. and G. Ritschard (2015), What matters in differences between life trajectories: A comparative review of sequence dissimilarity measures,Journal of the Royal Statistical Society, Series A. (Early View, Open Access), doi: 10.1111/rssa.12125.

-       Ritschard, G., M. Studer, R. Buergin (2015), Package `TraMineRextras', CRAN

-       Torgerson W.S., (1958), Theory and methods of scaling, Wiley, New-York.

-       WU L., (2000), Some comments on "Sequence analysis and optimal matching methods in sociology, review and prospect", Sociological methods and research, vol. 29(1), p.41-64

Job access and the spatial mobility trajectories of higher education graduates in the Netherlands

Marten M. Middeldorp, Arjen A. Edzes, Jouke van Dijk

Department of Economic Geography, Faculty of Spatial Sciences, University of Groningen

Abstract   The successfulness of the transition from education into working life is closely related to further career success. Graduates with good access to jobs earn higher wages and have lower chances of being unemployed. Access to jobs at the start of the career is therefore an important determinant of early career success and of importance for the whole career. In this paper, we study the effect of job access on the mobility patterns of recent higher education graduates. We use a GIS to calculate a job accessibility index based on driving time and use sequence analysis to create spatial mobility histories for 13,621 recent graduates of higher education. We subsequently relate job access at the start of the career to spatial mobility histories to analyze whether a suboptimal starting location in terms of job access leads to differing spatial mobility trajectories. Finally, we analyze how job access and spatial mobility influence labor market outcomes.

1. Introduction

This paper studies the relation between access to jobs at the start of the career and spatial mobility and their combined effect on early career labor market success. Success of the transition from education into working life is of interest to graduates, employers and policymakers at both national and local levels of government. Generally, graduates will want to find a job matching their skill set in order to reap the full benefits of their education, whilst employers are interested in attracting the best graduates for their firms. Optimal allocation of human capital enables countries to enjoy sustained competitiveness in a globalizing world economy (OECD, 2012), whilst at a more local level attracting and retaining higher education graduates is associated with higher levels of economic growth (Berry & Glaeser, 2005; Faggian & McCann, 2006).

When local opportunities are insufficient, spatial flexibility (migrating for employment reasons or accepting a daily commute to a job at a distance) can benefit

graduates entering the labor market. Jobseekers are able to achieve better matches by extending their search radius beyond the local labor market (Van Ham, 2001). Migration and commuting are thus not only important for individual outcomes, but also for the functioning of the labor market as a whole (Haas & Osland, 2014; Zabel, 2012). Indeed, young workers and especially higher education graduates have long been known to be more mobile, both spatially (Venhorst, Van Dijk, & Van Wissen, 2011) and in terms of employment (Topel & Ward, 1992).

The decision to migrate or commute is tied to both regional and personal factors. Different theoretical frameworks and empirical approaches have been used to explain spatial mobility (Herzog Jr, Schlottmann, & Boehm, 1993; Venhorst & Cörvers, 2015). However, the general focus is on the effects of periods in and moves between certain locations on labor market outcomes (e.g. Ahlin, Andersson, & Thulin, 2014; van Ham, 2003; Venhorst et al., 2011). This approach ignores that mobility and local labor market circumstances may have different effects on labor market outcomes depending on their timing within the career start and their relation to other mobility. This paper uses sequence analysis to create ideal-typical spatial mobility histories and demonstrates how these typologies can uncover patterns that account for all states, and their relation to each other, in the period under study. This provides further insight into the effect of job access and spatial mobility on early career labor market outcomes. Moreover, it enables us to analyze timing within and simultaneousness of spatial mobility processes. The paper analyses whether job access at the start of the career influences how subsequent spatial mobility takes shape and to what extent effects of job access and spatial and job mobility on labor market outcomes differ, depending on spatial mobility trajectory.

The remainder of this paper is organized as follows. In the next section, the paper summarizes the relevant literature. Then, in Section 3, it presents the data and discusses the empirical strategy. Section 4 presents spatial mobility trajectories, followed by estimates of the influence of job access and various forms of mobility on wages. The final section discusses and concludes.

3

2. Literature

Jobs are increasingly more spatially concentrated than people, a phenomenon also known as spatial mismatch (Holzer, 1991; Kain, 1968). Access to jobs at the beginning of the career is an important determinant of early career success and career advancement throughout the life course (Van Ham, 2001; van Ham, 2003). Workers that live in areas that are spatially mismatched search less intensively for jobs, have longer unemployment spells, or are employed in lower quality jobs or jobs not matching their education (Détang-Dessendre & Gaigné, 2009; Gobillon, Selod, & Zenou, 2007; Hensen, de Vries, & Corvers, 2009). Finding and keeping suitable employment early in the career is important, because unemployment job mismatch and non-standard working arrangements hinder the accumulation of specific human capital, wage growth and are associated with lower levels of job security (Arulampalam, 2001; Kunze, 2002; Light & Ureta, 1995).

In neoclassical labor market theory, worker mobility balances regional labor markets as deficits in one region are supplemented by surpluses from another. However, in practice worker mobility is limited, thus limiting the effective size of labor markets and the extent to which imbalances can be equilibrated (Blau & Duncan, 1967; Phelps, 1969). For workers, the cost of covering distances between regional labor markets can be significant, both in monetary and psychological terms. In this sense, spatial mobility is a means for jobseekers to extend their reach onto other labor markets. Mobile workers are often compensated or rewarded for their effort by earning higher wages and tend to migrate to labor markets with better opportunities, higher economic growth and lower levels of unemployment (Herzog Jr et al., 1993). This implies that the decision to be mobile is a personal consideration of costs and benefits. Hence it is not surprising that young individuals with high levels of human capital (for instance, recent graduates of higher education) are known to be especially mobile (Faggian & Mccann, 2009; Venhorst & Cörvers, 2015; Venhorst et al., 2011). Both the costs of staying in an inferior location and benefits of moving toward a more opportunity-rich location are higher, and young individuals have more time to change the costs of a move into the benefits of a better job (Sjaastad, 1962). Furthermore, recent graduates have relatively weak ties to the place where they have studied, which makes them more prone to be mobile (Fischer & Malmberg, 2001).

Several studies have linked job access, spatial mobility and (early) career success; we discuss a number of studies regarding The Netherlands. Van Ham (2003) finds that job access at the start of the career is related to higher occupational status and that the effect of job access increases with age. He hypothesizes that access to jobs at the start of the career gives jobseekers a head start over other workers, so that they accumulate human capital more rapidly through job mobility. Accepting a job at a distance from the residence is also related to higher occupational achievement, indicating that spatial mobility is beneficial for careers. In a study among Dutch graduates, Hensen et al. (2009) find that spatial mobility leads

to a better matching and higher quality job. Venhorst & Cörvers (2015) find positive returns of spatial mobility on wages, but note that this effect disappears after controlling for self-selection. This indicates that it is the higher human capital individuals who are spatially mobile and that this is driven by necessity (no suitable job opportunities nearby).

An alternative literature stresses the importance of considering migration and commuting as alternatives or substitutes (e.g. Eliasson, Lindgren, & Westerlund, 2003; Reitsma & Vergoossen, 1988). For instance, women often earn lower wages or have to balance work and family roles, making them generally less likely to commute long distances (Clark, Huang, & Withers, 2003). Home-owners on the other hand lack the spatial flexibility to migrate and therefore may have longer daily commutes (van Ham & Hooimeijer, 2009) and dual-earner households show a preference for rural regions that provide access to multiple urban labor markets within commuting distance (Green, 1997). Furthermore, for some commuting may precede migration, whilst for others this may be the other way around (Haas & Osland, 2014). The reciprocal relation between migration and commuting has long been acknowledged (e.g. Hanson & Pratt, 1988; van Ommeren, Rietveld, & Nijkamp, 1997), whilst from the previous it is apparent that timing and order are important characteristics of spatial mobility decision-making processes.

Currently, the literature lacks an approach to spatial mobility that accounts for order, simultaneousness and timing of migration and commuting. Spatial mobility is often included as an "ever mobile" variable in models, conceptualized as mobility probability or simply included as the distance between two locations, for instance place of residence and workplace (e.g. Faggian, Corcoran, & McCann, 2013). However, the decision to be spatially mobile and the form it takes – migration or commuting– will be related with personal factors and local opportunities, but also with previous mobility choices and changes in other life course domains, for instance housing and family careers (cf. Mulder & Hooimeijer, 1999). Hence, we conceptualize spatial mobility as a process that unfolds in time. Sequence analysis, a combination of methods that allows to study trajectories as wholes instead of focusing on durations, risks and transitions, has been proposed as a way to extend our knowledge of such processes (Aisenbrey & Fasang, 2010).

Studies using sequence analysis to uncover career patterns are manifold (see Dlouhy & Biemann (2015) for an extensive chronological overview). Most early studies were descriptive – in the sense that creating and analyzing the typology was the main goal of the study (e.g. Abbott & Hrycak, 1990; Halpin & Chan, 1998). Later studies had a more comparative approach and related the trajectories to hypotheses stemming from the literature. For instance, Brzinsky-Fay (2007) attempts to find grounds for a theoretical typology by comparing school-to-work transition trajectories in European countries and Schoon et al. (2001) compare two birth cohorts to identify differences in the extent and direction of changes in school-to-work transitions. A number of studies use the trajectories in further analysis, in order to identify how trajectories are related to other factors. Anyadike-Danes & McVicar (2005) relate observable background characteristics of

5

young men at birth, age 10 and age 16 to the likelihood of following a certain career trajectory in order to identify factors that predict negative career pathways. They find educational achievement and school disciplinary record at age 16, health and learning progress at age 10 and region of birth to be the strongest predictors of career paths. Biemann et al. (2012) use a panel, spanning 20 years of employment data, to distinguish six career patterns that deviate from the traditional career path of employment within a single firm. They then use multinomial logistic regression to relate individual characteristics and occupational sector at the start of the career to the probability of having one of the 'new' career patterns. They find that women, young, singles and higher educated more often have career patterns that deviate from the traditional path. Kovalenko & Mortelmans (2014) confront two juxtaposing theories about the effect of 'transitional' career patterns on objective and subjective career success. After constructing a career typology through sequence analysis, they relate the career trajectories to objective (wage and home-ownership) and subjective (satisfaction and disappointment) measures of career success. They find that neither of the competing theories is able to completely explain career outcomes, but that a synthesis of the two perspectives would provide an understanding that better matches the outcomes observed in their study.

Application of the methodology on socio-spatial phenomena has been limited and to our knowledge, sequence analysis has not been used to create and analyze spatial mobility trajectories. Coulter & Van Ham (2013) analyze sequences of moving desire and behavior and distinguish between eight types of mobility biographies. The use of sequence data highlights the importance of heterogeneity in experiences. Although for some respondents moves are followed by (desire for) more mobility, for others it does not. Furthermore, it stresses the importance of adopting a life course approach when studying mobility biographies as the results suggest that the impact of states on personal well-being are better understood in a broader context.

Taken together, the previous indicates that a sequence approach to spatial mobility during labor market entry can be of value for our understanding of the manner in which job access and various forms of spatial mobility interact to influence early career labor market success.

Data & Methods

Data and sample

The study draws on longitudinally linked registry micro-data, provided by Centraal Bureau voor de Statistiek (CBS; Statistics Netherlands). The micro-data files contain information on labor market states, jobs (size and type of contract, sector, location, wage), education and personal and household characteristics (such as place of residence) of all inhabitants of the Netherlands for the period 2006-2013. Job accessibility was calculated at level of five digit postal codes (PC5) using the LISA dataset, a database of business establishments. As there are 32,000 PC5 areas in The Netherlands, this is a highly detailed spatial resolution. ESRI's 2008 StreetMap Premium road network dataset was used to calculate travel times for the GIS network analysis.

Our sample consists of all graduates of tertiary education and was refined to ensure that the selected persons have comparable career experience and did not leave education only for a very brief time. First, we selected all individuals between 20 and 30 years old that obtained a tertiary degree in the period May – August 2006 and were registered in the data as having the state 'Student' for at least five months between January and September 2006. The selection was then refined by excluding all graduates that were registered as 'Student' anytime between October 2006 and January 2007. We also exclude graduates from the sample who have missing values for labor market states, home locations or job information (when employed) during the period under study. It was unfeasible to exclude all graduates with missing job locations at any moment in the period under study, due to the way the job location is registered[1], as this would reduce our sample by 25 per cent. The final selection thus includes all graduates for whom we have complete information on labor market states, home locations and educational achievement, resulting in a sample of 13,621 graduates.

---

[1] Job locations are only available for jobs that exist in the month December of a certain year. Originally, this meant that only about 17 per cent of all job-months had locational information. We imputed the locations of jobs based on a number of criteria. First, we checked whether a job had a known location in the previous year. If so, that location was used as the location of the job. Then, we checked if more than 80 per cent of all workers in a firm worked in one location in a given year. If so, that location was used as the location of the job. This raised the number of job-months with known locational information to 73 per cent.

7

Analytical approach

Our analysis consists of three parts; network analysis using a GIS to calculate job access at the career start, sequence analysis to define ideal-typical spatial mobility histories, and regression modeling to estimate the effect of job access and personal factors on spatial mobility trajectories and of mobility and access on labor market success (measured as hourly wage).

First, the number of full-time jobs and working age population were aggregated at the PC5 level and then geocoded. We calculate job access as an index that takes the following basic form:

$$A_i = \sum_j \frac{\varphi J_{ji}}{\sum_k \varphi P_{kj}} \qquad (1)$$

Where A_i is the job access index of a location i, J_{ji} the number of full-time jobs in locations j within reach of location i and P_{kj} the working age population in locations k that is able to reach job location j. The term φ is a factor that controls for declining commuting tolerance as distance increases (i.e. as commuting time increases, less people are willing to travel to a certain location) by only counting the number of jobs and population within reach for a certain percentage, see table 1. In a final step, we normalize the resulting access indexes by the (working population) weighted average of job access in The Netherlands, thus centering mean job access around 1.

Table 1. Commuting tolerance boundaries, reflecting the percentage of working persons traveling a certain amount of time to reach their job. Source: SCP (2007), own calculations.

Travel time (minutes)	%
< 15	100
15 – 30	71
30 – 45	38
45 – 60	18
60 – 75	10
75 – 90	5

Figure 1, below, shows that job access is not spread evenly across the country. Although peripheral regions have, generally, the lowest access to jobs, there are also areas close to central locations such as Amsterdam that have low levels of access. Corridors of higher access can be discerned around important highways, whilst peripheral locations have low levels of access. We have no access to detailed employment information in neighboring areas but did minimize border effects by also including the road network in Germany and Belgium. Since cross-

border mobility is still limited, we are certain that our results are not distorted too much by this limitation. Although the colleges and universities at which they studied are mostly located in the larger cities of the Netherlands which are often located in areas with high levels of access, 65% of all graduates start their careers in areas that have an access index in the range of 0.75–1.25; this range corresponds to approximately 1 standard deviation around the mean (see table A1 in the appendix for more sample statistics by level of access).
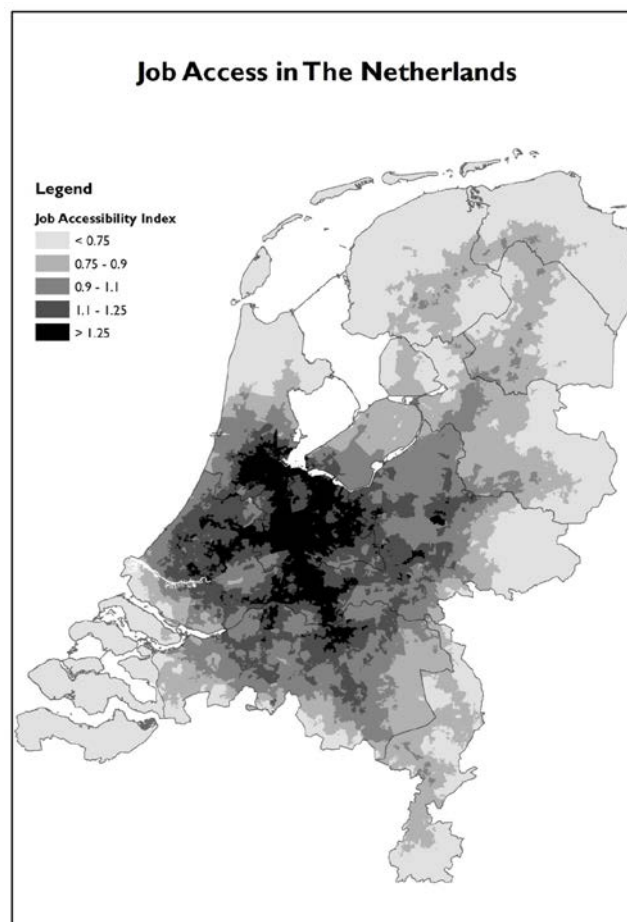
Fig. 1: Job access in The Netherlands, 2006. Source: own calculations.

We then use sequence analysis to create spatial mobility trajectories. Sequence analysis roughly consists of three steps: defining the sequence, measuring dissimilarity between sequences and grouping similar sequences together. For our analy-

9

sis of spatial mobility histories, we follow the graduates during the first five years of their careers, plus an additional half year (from April – September 2006) to account for potential pre-graduation mobility. For migration, we consider the number of moves across provincial borders (max. 6 during the period under study). For commuting, we use two levels of commuting (Short <= 30 min. and Long: > 30 min.), and two types of missing (Not employed and No data).

Both sequences were analyzed in R using TraMineR and WeightedCluster (Gabadinho, Ritschard, Müller, & Studer, 2011; Studer, 2013). We use OM_future as proposed in (Studer & Ritschard, 2016) to calculate the dissimilarity matrices, because it weighs the dissimilarity by the probability of ending up in a certain state. To test the sensitivity of our results to the chosen distance costs and algorithm, we also calculate dissimilarities using the optimal matching algorithm and both weighted (based on theoretical similarity of states, indel costs of 1.5) and unweighted costs (substitution costs of 2, indel costs of 1.5). The dissimilarities were clustered using Ward's (1963) clustering algorithm, which was shown to produce the best results for dissimilarity matrices (Dlouhy & Biemann, 2015). Although, according to the average silhouette width (Kaufman & Rousseeuw, 2005), a solution of two to three clusters would be preferred. However, these solutions only distinguish between the number of migrations across provincial borders (no migration, migration and, in the three cluster solution, multiple migration). We decide on using the six-cluster solution, as it also explains the dissimilarity matrix quite well and shows a trajectory that is defined by commuting.
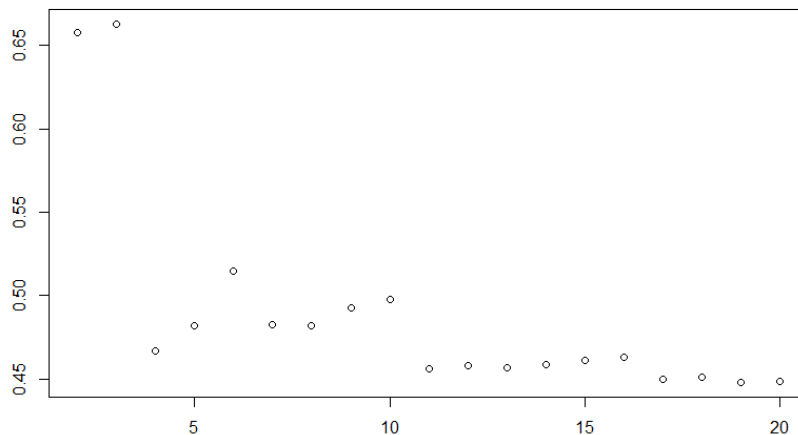
Fig. 2: Average silhouette width, by number of clusters. Source: CBS, own calculations.

Results

Spatial mobility histories of higher education graduates

The trajectories resulting from our combined analysis of migration and commuting histories are presented below in figure 3a-f. The left panel of each graph describes the number of moves across provincial borders in the first five years after graduating, the right panel the commuting decisions during the same period.

a.

b.

c.

11
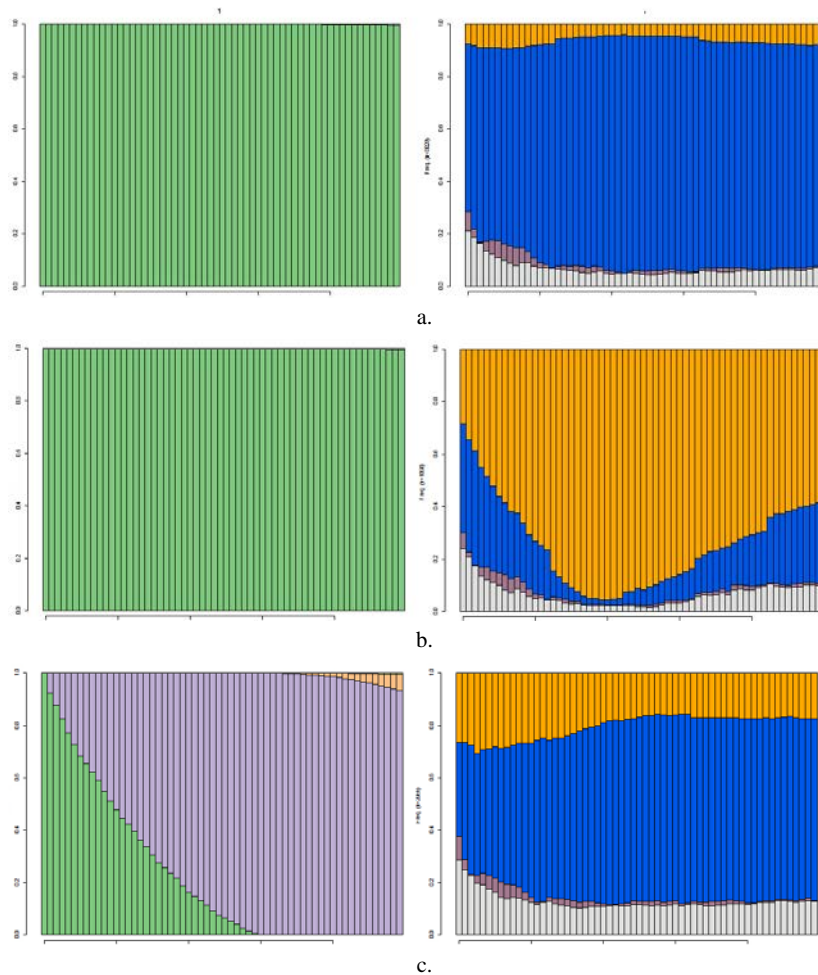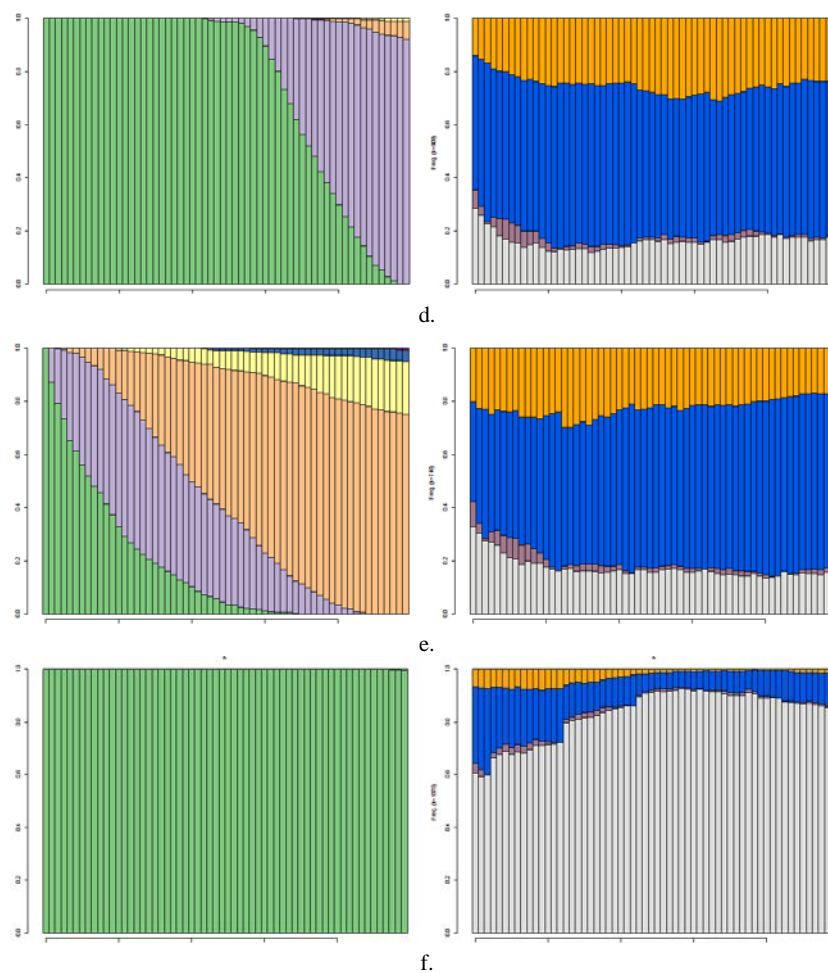
Fig. 3 a-f: Proportional transversal state distributions (x = time in months, y = cumulative state proportion), by trajectory. Left: migration across provincial borders; right: daily commuting distance. Source: CBS, own calculations.

The trajectory in figure 3a corresponds to the spatial mobility histories of approximately 60% of all graduates and is characterized by immobility. At any point in time, at most ten percent of all graduates are making long distance commutes to their job and moves across provincial borders are very rare and happen only in the last year, if at all. Graduates in the second trajectory (fig. 3b) are willing to commute long distances between home and the workplace, whilst moves across provincial borders happen rarely, if at all. At any moment in time, from one year after

graduating onwards, 75% of all currently working graduates in this trajectory travel more than thirty minutes during a single commute. The trajectories in figure 3c and 3d are best characterized by their differences in timing. In both trajectories, commuting is only a temporary phenomenon, and a single move is made, sometimes followed by a very late second move. In the third trajectory (fig. 3c), the move happens earlier (within the first two years after graduating) than in the fourth trajectory (fig. 3d). Graduates in the fifth trajectory (fig. 3e) make several moves, and the second move follows the first move very quickly (usually within 18 months). Finally, graduates in the sixth trajectory (fig. 3f) are distinguished more by their labor market states than their spatial mobility choices. This group, a little under ten percent of the total sample, is not employed during the most part of the period under study. This does not mean that the group is homogeneous or that their careers are by definition unsuccessful (e.g. many of the graduates in this trajectory become self-employed). Since this trajectory is best defined by our lack of information on the graduates in it, we will not further discuss the graduates in this trajectory in the remainder of this paper.

Table 2. Percentage of graduates per spatial mobility trajectory, by education level and field and level of job access, October 2006. Source: CBS, LISA; own calculations.

	Immobile	Commuter	Mobile Early	Mobile Late	Mobile Multiple	Other
Education level						
> College (BSc)	61.9	9.6	7.5	6.9	3.3	10.8
> University (BSc)	50.6	6.3	10.2	6.6	6.3	20.1
> University (MSc)	53.8	10.9	13.1	10.6	5.6	6.0
Education field						
> Teaching	75.2	4.8	5.7	6.5	2.2	5.6
> Agriculture	46.0	15.3	13.7	10.9	5.9	8.2
> Natural sciences	59.1	12.3	8.0	8.0	5.2	7.4
> Engineering	56.5	13.1	9.7	7.3	3.6	9.8
> Healthcare	62.1	9.4	9.8	8.9	4.1	5.7
> Economics	58.9	11.6	9.4	8.4	5.1	6.6
> Law	55.6	6.9	16.4	11.1	6.2	3.7
> Behavioral & social	64.9	9.8	8.7	7.6	3.1	5.8
> Language & arts	39.5	6.4	10.0	8.0	4.6	31.4
Access						
> Low (< 0.75)	57.8	13.1	10.3	7.3	4.6	7.0
> Medium (0.75-1.25)	56.9	10.5	10.5	9.3	4.5	8.4
> High (>= 1.25)	60.7	7.9	8.4	7.9	3.7	11.5
Total	**59.0**	**9.9**	**9.4**	**8.1**	**4.1**	**9.5**

13

The propensity to have a particular trajectory differs by job accessibility and both level and field of education, as is shown in Table 2 (above). For instance, the percentage of university graduates that have an immobility trajectory is lower than the percentage of college graduates with the same trajectory. Sector of studies also is related to the propensity to be mobile and the specific type of mobility. Those with teaching backgrounds are the least mobile, the commuter trajectory is more common among those with backgrounds in agriculture and those with law degrees are more likely to move quickly after graduation. Interestingly, low levels of access seem related to higher percentages of those with a commuting trajectory, which may seem counterintuitive. However, this can at least partly be explained by the fact that jobs that are more than 30 minutes away are discounted for a large part in the accessibility measure. Long commutes are necessary to reach these jobs, which is why lower levels of job access may be related to a higher propensity to commute. Figure 4 confirms that commuters have job access levels slightly below the average. It also depicts that, on average, graduates who move across provincial borders do so in the direction of locations with higher levels of job accessibility. Interestingly, it seems that this is only a characteristic of the first move; further moves do not further increase average levels of job access, as depicted in the graph for the fifth trajectory (the multiple movers).
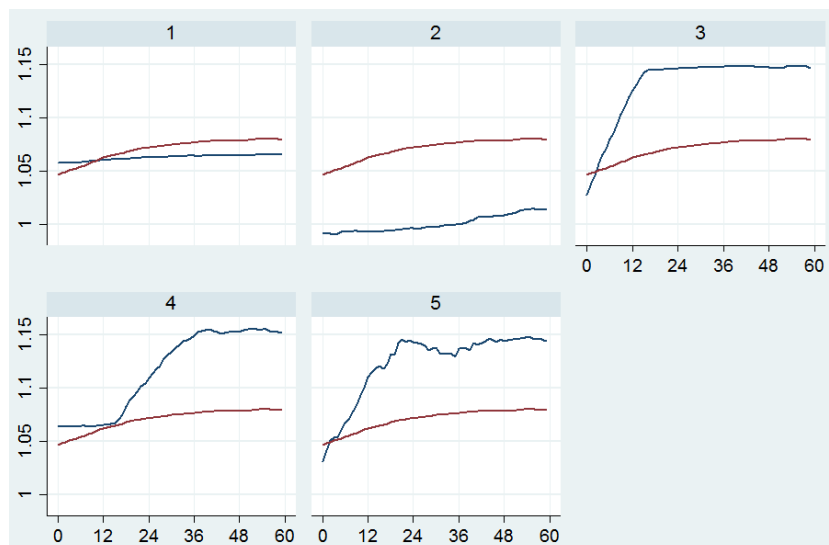
Fig. 4: Average versus trajectory levels of job access, in time (x = time in months, y = job access index), by trajectory. Red: average level of access in sample; Blue: level of access for graduates with trajectory. Source: CBS, LISA; own calculations.

Job access, spatial mobility and early career success

We use multinomial logistic regression to analyze how, net of other factors, level of job access and individual and education characteristics relate to the probability of following one of the spatial mobility trajectories. We test whether the trajectories can be seen as independent from each other, by executing the Small-Hsiao test of the IIA assumption and Wald tests for combining alternatives. Both indicate that the trajectories are suitable to be used as a dependent variable in multinomial logit analysis. Table 3 presents the average marginal effects derived from this model. Marginal effects have the benefit of allowing more straightforward interpretation of the effect of covariates on the probability of having a particular spatial mobility trajectory. In the case of categorical covariates, they are interpreted as the effect of a discrete change with respect to the base level. For job access and age, the only continuous variables in our model, the interpretation is as the effect of a one unit increase on the probability of belonging to a trajectory. We stress that since we cannot control for many factors, among which ability and ambition, the results should be interpreted carefully and in terms of association, not causation.

Although, in fact, a one unit increase ($\pm$ 4 standard deviations) in job accessibility is very improbable, a ten percentage point higher level of access is related to a 2 percent higher probability of belonging to the immobility trajectory. This is reflected in particular in a negative correlation with the probability to become a commuter. University graduates are, indeed, more mobile than college graduates, although this relationship may be moderated by field of study as the differences between fields of study in their associations with trajectories are sometimes stronger than those between levels of education. For instance, a behavioral and social sciences college graduate has a higher probability to become a commuter than a university teaching graduate.

Although previous literature found higher propensities for women to be spatially mobile (e.g. Venhorst et al., 2011), we find a negative association with the probability of having the commuter trajectory and no statistically significant relationship with any of the three mobile trajectories. As expected, singles have a lower probability to be immobile, compared to graduates in other household situations. However, graduates with partners or still living with their parents seem more willing to accept a commute in order to bridge the distance between home and the workplace. The associations with other personal factors are small, if significant at all.

15

Table 3. Multinomial logistic regression results, average marginal effects;
dependent variable: spatial mobility trajectory.

			Mobile		
	Immobile	Commuter	Early	Late	Multiple
Job access	.197***	-.101***	-.072***	.014	-.038***
Education level					
> College (BSc)	ref.	ref.	ref.	ref.	ref.
> University (BSc)	-.017	-.026	.021	-.006	.028*
> University (MSc)	-.101***	.022***	.032***	.031***	.015***
Education field					
> Teaching	.06***	-.057***	-.012	.009	.000
> Agriculture	-.171***	.050***	.057***	.039**	.025**
> Natural sciences	-.009	.013	-.022	-.003	.021*
> Engineering	-.05***	.02*	.017*	.005	.008
> Healthcare	-.059***	-.003	.023*	.024**	.015*
> Economics	-.074***	.009	.02**	.02**	.027***
> Law	-.047**	-.038***	.046***	.015	.024**
> Behavioral & social	ref.	ref.	ref.	ref.	ref.
> Language & arts	-.097***	-.001	.034***	.037***	.026***
Gender					
> female	.04***	-.033***	.001	-.004	-.004
Age	-.001	-.001	.006***	-.001	-.002**
Origin					
> Dutch	ref.	ref.	ref.	ref.	ref.
> Western	-.032*	.014	.005	-.004	.017*
> non-Western	.026	.004	-.023**	.002	-.009
Household status					
> single	ref.	ref.	ref.	ref.	ref.
> couple	.132***	.027***	-.091***	-.009	-.028***
> with parents	.123***	.021***	-.075***	-.015**	-.053***
> other	.076***	.017	-.064***	.007	-.036***
High earning parents	-.027***	-.002	.009	.011**	.008*
Chi² (df = 120)	1592.6***				
N (individuals)	11,839				

Figures 5 and 6 provide a naive idea of the effect of mobility trajectories on labor market outcomes. In figure 5, the percentage of graduates that are employed in 'normal' working arrangements (full-time or part-time jobs with fixed weekly hours) are depicted as deviations from the average. Overall, commuters have the highest levels of employment, but this may well be due to how commuters are defined (i.e. to be a commuter, one has to be employed). For the early movers, this percentage is decreasing over the course of the study period. Interestingly, the decreaese seems to set in just before or around the time of the first move, twelve months into the study period. In figure 6, the same is done for wages. The steepest slope can be found among the graduates who move multiple times during the study period. Over the course of five years, they go from earning around or slightly below the average towards around € 200,- more than average. The wages of those in the immobility trajectory lag in comparison.
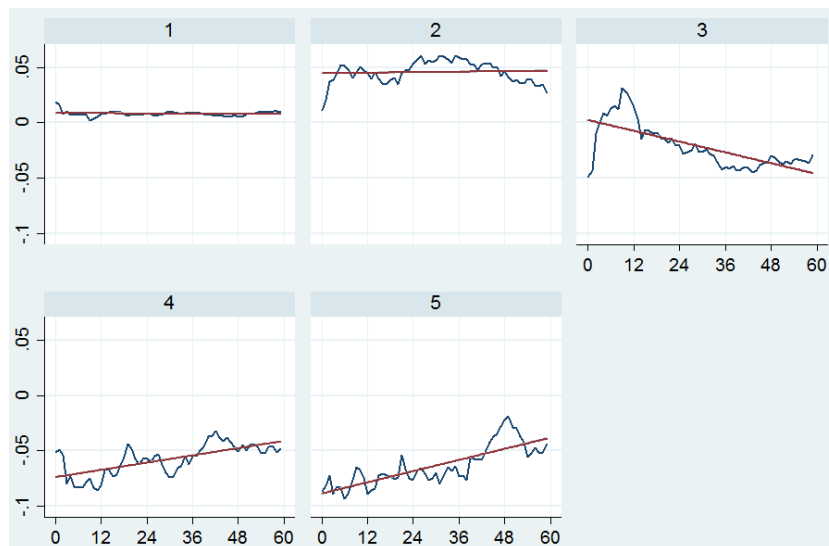
Fig. 5: Percentage in employment as deviations from the average, over time (x = time in months, y = deviation in percentage employed), by trajectory. Blue: deviation; Red: trend line. Source: CBS, own calculations.

17
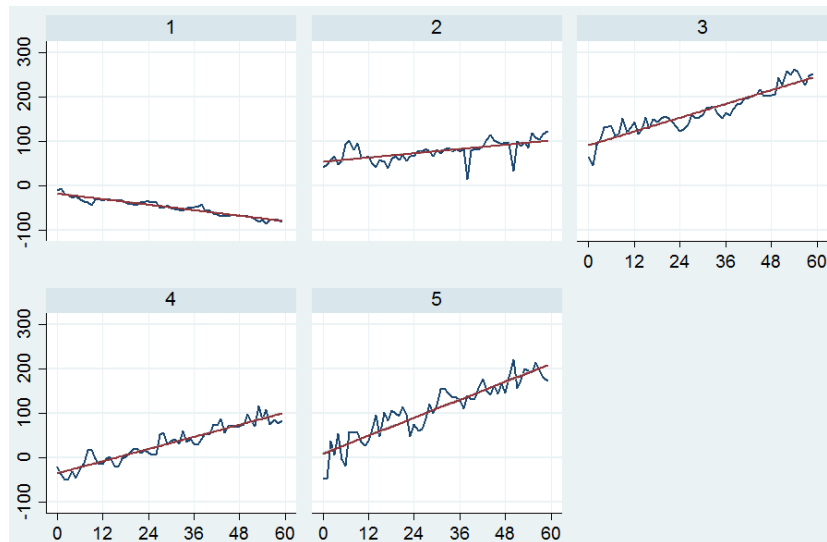
Fig. 6: Wage levels as deviations from the average, over time (x = time in months, y = deviation in hourly wage), by trajectory. Blue: deviation; Red: trend line. Source: CBS, own calculations.

Of course, these effects may be as much or more due to personal factors explaining the selection into specific trajectories as to the trajectory self. To further probe the effect of job access and spatial mobility on labor market outcomes, we employ a fixed effects regression model. A fixed effects regression model eliminates estimate bias due to time invariant factors at the individual level by only explaining within-person differences in the dependent variable by within-person changes in independent variables. In this model, we regress (log of) hourly wage in October 2006-2013 on job access, spatial and job mobility, tenure in full-time (> 35 hours), part-time (20-35 hours) and non-standard (< 20 or flexible hours) employment. We also control for job and firm characteristics[2] and include regional dummies. Table 4, on the next page, presents the results.

---

[2] Job characteristics: type of contract (permanent, temporary, other), job size (full-time, part-time, small)

Firm characteristics: firm size (very small, small, medium, large, very large), firm broad sector, dummy: firm location unknown

Table 4. Fixed effects regression results; dependent variable: (log of) hourly wage.

	All	Immobile	Mobile			
			Commuter	Early	Late	Multiple
Job access	.059***	.044**	.023	.113***	.04	.132***
Mobility						
> move	.032***	.009	.015	.052***	.045***	.048***
> + job change	.013*	.003	.008	.053***	.02	.011
> commute	.015***	-.004	.06***	.016***	.019**	.008
Job change	.028***	.032***	.016***	.012***	.027***	.021***
Tenure						
> full-time	.076***	.074***	.082***	.082***	.072***	.081***
> part-time	.051***	.051***	.061***	.048***	.043***	.034***
> non-standard	.053***	.054***	.06***	.049***	.038***	.034***
Controls						
> region	Y	Y	Y	Y	Y	Y
> job	Y	Y	Y	Y	Y	Y
> firm	Y	Y	Y	Y	Y	Y
R^2 (within)	.455	.518	.431	.518	.546	.474
N (obs)	88247	58460	9876	8738	7442	3731
N (individuals)	12291	8032	1355	1261	1090	553

In the first column, the results for all graduates are presented. Job access, spatial and employment mobility are all positively associated with wage level. The results show that living in areas with a ten percentage points higher job access level is associated with a 0.6% higher wage. The wage premium of a move across provincial borders (3.2%) is generally higher than that of long distance commuting (1.5%). Switching jobs can also be instrumental to career advancement (there is an associated wage gain of 2.8 %) and full-time jobs have higher hourly wages than part-time or non-standard employment.

However, we also find that the size of these effects can be quite different depending on spatial mobility trajectory. For those in the immobile trajectory, tenure effects are generally lower than for those in mobile trajectories, suggesting that they acquire less specific human capital in their jobs than graduates that are spatially mobile. However, the effects of job switching are more pronounced, which implies that some are able to achieve upward mobility by switching to other firms in their vicinity. For commuters, the effect of commuting is much higher than it is for other graduates and they also have high tenure premiums, whilst the benefit of switching jobs is relatively low. This may indicate that commuters are able to achieve good matches on the labor market relatively quickly and are aware of the value of their current position. Early movers have a high wage premium associated with a long distance move, especially when it is combined with taking up or switching jobs. The direction of the move is of importance too as there is a rela-

tively strong association with job accessibility. On average, early movers gain little over .15 points on the job access index during their first move, this leads to a wage premium between 1.5 and 2 %. For many late movers, the move may in the end be related to other life factors (e.g. family formation), since they are in many respects more similar to the immobile than to early movers. For late movers, there is no wage premium associated with higher levels of job access and the effect of tenure is also low. Finally, the association between job access and wage is strongest for multiple movers.

Conclusion

This paper analyses how access to jobs at the start of the career influences the spatial mobility choices of graduates of higher education in The Netherlands. Although the reciprocal relationship between migration and commuting has been noted in previous research, this study is among the first to construct comprehensive trajectories based on migration and commuting histories. This paper focuses on a homogeneous group that is often considered highly mobile. Graduates of higher education embody high levels of human capital and suitable job opportunities are generally spread thin across the country. Graduates in areas with lower job accessibility thus can use spatial mobility as an instrument toward gaining a better job or accessing more advantageous labor markets.

Our study highlights a number of issues. First, although higher education graduates are usually depicted as highly mobile, only 35% is mobile during the first 5 years after graduating from college and only 15% is highly mobile (migrating several times or commuting long distances). This may not seem surprising and related to the study setting, as the Netherlands is a dense country and distances are relatively short. However, this is also reflected in our states: the average province is smaller than 3,000 km^2 and a thirty minute commute is considered short in many countries.

Second, access to jobs is negatively associated with spatial mobility and positively associated with early career success. Graduates in areas with better access to jobs are especially less likely to commute or migrate early. Graduates that live in areas with better access earn higher wages. This effect is stronger for graduates that move early or often than for immobile graduates. A statistically significant effect of access to jobs could not be found for late movers and commuters.

Finally, our study indicates that outcomes of mobility (in terms of wages) are heterogeneous toward the type and timing of mobility. Failure to account for different types of mobility and their timing may underestimate the effect of mobility on labor market outcomes. These results show that distinguishing between type and timing of mobility can be helpful in determining the value of mobility for graduates of higher education entering the labor market.

References

Abbott, A., & Hrycak, A. (1990). Measuring Resemblances in Sequence Data: An Optimal Matching ANalysis of Musicians' Careers. *American Journal of Sociology*, *96*(1), 144–185. http://doi.org/10.1086/229495

Ahlin, L., Andersson, M., & Thulin, P. (2014). Market Thickness and the Early Labour Market Career of University Graduates: An Urban Advantage? *Spatial Economic Analysis*, *9*(4), 396–419. http://doi.org/10.1080/17421772.2014.961534

Aisenbrey, S., & Fasang, A. E. (2010). New Life for Old Ideas: The "Second Wave" of Sequence Analysis Bringing the "Course" Back Into the Life Course. *Sociological Methods & Research*, *38*(3), 420–462. http://doi.org/10.1177/0049124109357532

Anyadike-Danes, M., & McVicar, D. (2005). You'll never walk alone: Childhood influences and male career path clusters. *European Association of Labour Economists 16th Annual Conference, Universidade Nova de Lisboa, Lisbon, 9th – 11th Sepetember, 2004*, *12*(4), 511–530. http://doi.org/http://dx.doi.org.proxy-ub.rug.nl/10.1016/j.labeco.2005.05.008

Arulampalam, W. (2001). Is Unemployment Really Scarring? Effects of Unemployment Experiences on Wages. *The Economic Journal*, *111*(475), 585–606. http://doi.org/10.1111/1468-0297.00664

Berry, C. R., & Glaeser, E. L. (2005). The divergence of human capital levels across cities. *Papers in Regional Science*, *84*(3), 407–444. http://doi.org/10.1111/j.1435-5957.2005.00047.x

Biemann, T., Zacher, H., & Feldman, D. C. (2012). Career patterns: A twenty-year panel study. *Journal of Vocational Behavior*, *81*(2), 159–170. http://doi.org/10.1016/j.jvb.2012.06.003

Blau, P. M., & Duncan, O. D. (1967). *The American Occupational Structure. The American Occupational Structure*. http://doi.org/10.2307/2092400

Brzinsky-Fay, C. (2007). Lost in Transition? Labour Market Entry Sequences of School Leavers in Europe. *European Sociological Review*, *23*(4), 409–422. http://doi.org/10.1093/esr/jcm011

Clark, W. A. V, Huang, Y. Q., & Withers, S. (2003). Does commuting distance matter? Commuting tolerance and residential change. *Regional Science and Urban Economics*, *33*(2), 199–221. http://doi.org/10.1016/S0166-0462(02)00012-1

Coulter, R., & van Ham, M. (2013). Following People Through Time: An Analysis of Individual Residential Mobility Biographies. *Housing Studies*, *28*(7), 1037–1055. http://doi.org/10.1080/02673037.2013.783903

Détang-Dessendre, C., & Gaigné, C. (2009). Unemployment duration, city size, and the tightness of the labor market. *Regional Science and Urban Economics*, *39*(3), 266–276. http://doi.org/http://dx.doi.org/10.1016/j.regsciurbeco.2009.01.003

21

Dlouhy, K., & Biemann, T. (2015). Optimal Matching Analysis in Career Research: A Review and Some Best-Practice Recommendations. *Journal of Vocational Behavior*, *90*(0), 163–173. http://doi.org/10.1016/j.jvb.2015.04.005

Eliasson, K., Lindgren, U., & Westerlund, O. (2003). Geographical Labour Mobility: Migration or Commuting? *Regional Studies*, *37*(8), 827–837. http://doi.org/10.1080/0034340032000128749

Faggian, A., Corcoran, J., & McCann, P. (2013). Modelling geographical graduate job search using circular statistics. *Papers in Regional Science*, *92*(2), 329–343. http://doi.org/10.1111/pirs.12026

Faggian, A., & McCann, P. (2006). Human capital flows and regional knowledge assets: A simultaneous equation approach. *Oxford Economic Papers*, *58*(3), 475–500. http://doi.org/10.1093/oep/gpl010

Faggian, A., & Mccann, P. (2009). Universities, agglomerations and graduate human capital mobility. *Tijdschrift Voor Economische En Sociale Geografie*. http://doi.org/10.1111/j.1467-9663.2009.00530.x

Fischer, P. A., & Malmberg, G. (2001). Settled people don't move: On life course and (Im-) mobility in Sweden. *International Journal of Population Geography*, *7*(5), 357–371. http://doi.org/10.1002/ijpg.230

Gabadinho, A., Ritschard, G., Müller, N. S., & Studer, M. (2011). Analyzing and Visualizing State Sequences in R with TraMineR. *Journal of Statistical Software*, *40*(4), 1–37. http://doi.org/10.18637/jss.v040.i04

Gobillon, L., Selod, H., & Zenou, Y. (2007). The Mechanisms of Spatial Mismatch. *Urban Studies*, *44*(12), 2401–2427. http://doi.org/10.1080/00420980701540937

Green, A. E. (1997). A question of compromise? Case study evidence on the location and mobility strategies of dual career households. *Regional Studies*, *31*(7), 641–657. http://doi.org/10.1080/00343409750130731

Haas, A., & Osland, L. (2014). Commuting, Migration, Housing and Labour Markets: Complex Interactions. *Urban Studies*, *51*(3), 463–476. http://doi.org/10.1177/0042098013498285

Halpin, B., & Chan, T. W. (1998). Class careers as sequences: An optimal matching analysis of work-life histories. *European Sociological Review*, *14*(2), 111–130.

Hanson, S., & Pratt, G. (1988). Reconceptualizing the Links between Home and Work in Urban Geography. *Economic Geography*, *64*(4), 299–321.

Hensen, M. M., de Vries, M. R., & Corvers, F. (2009). The role of geographic mobility in reducing education-job mismatches in the Netherlands. *Papers in Regional Science*, *88*(3), 667–682. http://doi.org/10.1111/j.1435-5957.2008.00189.x

Herzog Jr, H., Schlottmann, A., & Boehm, T. (1993). Migration as spatial job–search: a survey of empirical findings. *Regional Studies*, *27*(4), 327–340. http://doi.org/10.1080/00343409312331347595

Holzer, H. J. (1991). The Spatial Mismatch Hypothesis: What Has the Evidence

Shown? *Urban Studies*, *28*(1), 105–122. http://doi.org/10.1080/00420989120080071

Kain, J. F. (1968). Housing Segregation, Negro Employment, and Metropolitan Decentralization. *The Quarterly Journal of Economics*, *82*(2), 175. http://doi.org/10.2307/1885893

Kaufman, L., & Rousseeuw, P. J. (2005). *Finding Groups in Data: An Introduction to Cluster Analysis (Wiley Series in Probability and Statistics)*. *eepe.ethz.ch*.

Kovalenko, M., & Mortelmans, D. (2014). Does career type matter? Outcomes in traditional and transitional career patterns. *Journal of Vocational Behavior*, *85*(2), 238–249. http://doi.org/10.1016/j.jvb.2014.07.003

Kunze, A. (2002). *The Timing of Careers and Human Capital Depreciation. IZA Discussion Paper Series*. Retrieved from http://hdl.handle.net/10419/21432\nNutzungsbedingungen:

Light, A., & Ureta, M. (1995). Early-Career Work Experience and Gender Wage Differentials. *Journal of Labor Economics*, *13*(1), 121. http://doi.org/10.1086/298370

Mulder, C., & Hooimeijer, P. (1999). Residential Relocations in the Life Course. *Population Issues*, 159–186. http://doi.org/10.1007/978-94-011-4389-9_6

OECD. (2012). *OECD Economic Surveys: Netherlands 2012*. Paris.

Phelps, E. S. (1969). Introduction: The new microeconomics in employment and inflation theory. In E. S. Phelps (Ed.), *Microeconomic foundations of employment and inflation theory* (pp. 1–26). New York: Norton.

Reitsma, R. F., & Vergoossen, D. (1988). A causal typology of migration: the role of commuting. *Regional Studies*, *22*(4), 331–340. http://doi.org/10.1080/00343408812331345010

Schoon, I., McCulloch, A., Joshi, H. E., Wiggins, R. D., & Bynner, J. (2001). Transitions from school to work in a changing social context. *Young*, *9*(1), 4–22. http://doi.org/10.1177/110330880100900102

Sjaastad, L. A. (1962). The Costs and Returns of Human Migration. *Journal of Political EconomyJournal of Political Economy*, *70*(5), 80–93.

Studer, M. (2013). *WeightedCluster Library Manual: A Practical Guide to Creating Typologies of Trajectories in the Social Sciences with R* (LIVES Working Papers No. 24).

Studer, M., & Ritschard, G. (2016). What matters in differences between life trajectories: a comparative review of sequence dissimilarity measures. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *179*(2), 481–511. http://doi.org/10.1111/rssa.12125

Topel, R. H., & Ward, M. P. (1992). Job Mobility and the Careers of Young Men. *The Quarterly Journal of Economics*, *107*(2), 439–479. http://doi.org/10.2307/2118478

Van Ham, M. (2001). Workplace mobility and occupational achievement. *International Journal of Population Geography*, *7*(4), 295–306. http://doi.org/10.1002/ijpg.225

23

van Ham, M. (2003). Job access at labour market entry and occupational achievement in the life course. *International Journal of Population Geography*, *9*(5), 387–398. http://doi.org/10.1002/ijpg.295

van Ham, M., & Hooimeijer, P. (2009). Regional differences in spatial flexibility: Long commutes and job related migration intentions in The Netherlands. *Applied Spatial Analysis and Policy*, *2*, 129–146. http://doi.org/10.1007/s12061-008-9016-2

van Ommeren, J., Rietveld, P., & Nijkamp, P. (1997). Commuting: In Search of Jobs and Residences. *Journal of Urban Economics*, *42*(3), 402–421. http://doi.org/http://dx.doi.org/10.1006/juec.1996.2029

Venhorst, V. A., & Cörvers, F. (2015). Entry into working life: Spatial mobility and the job match quality of higher-educated graduates. Retrieved from http://pub.maastrichtuniversity.nl/974bb512-71c2-4a5a-a897-d1e265d50bd5

Venhorst, V. A., Van Dijk, J., & Van Wissen, L. (2011). An analysis of trends in spatial mobility of Dutch graduates. *Spatial Economic Analysis*, *6*(1), 57–82. http://doi.org/10.1080/17421772.2010.540033

Ward, J. H. J. (1963). Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*, *58*(301), 236–244.

Zabel, J. E. (2012). Migration, Housing Market, and Labor Market Responses to Employment Shocks. *Journal of Urban Economics*, *72*(2-3), 267–284. http://doi.org/10.1016/j.jue.2012.05.006

Sample statistics

Table A1. Sample statistics, October 2006. Source: CBS, own calculations.

			Access	
	Total	Low (< 0.75)	Medium	High (>= 1.25)
Education level				
> College (BSc)	0.65	0.74	0.67	0.52
> University (BSc)	0.03	0.02	0.02	0.04
> University (MSc)	0.33	0.24	0.31	0.44
Education field				
> Teaching	0.11	0.16	0.11	0.06
> Agriculture	0.04	0.04	0.05	0.02
> Natural sciences	0.02	0.01	0.02	0.05
> Engineering	0.15	0.19	0.15	0.10
> Healthcare	0.07	0.07	0.07	0.07
> Economics	0.25	0.27	0.25	0.22
> Law	0.04	0.03	0.04	0.06
> Behavioral & social	0.21	0.17	0.20	0.24
> Language & arts	0.11	0.05	0.10	0.19
Gender				
> female	0.60	0.55	0.59	0.63
Age	23.99	23.67	23.92	24.44
Origin				
> Dutch	0.83	0.90	0.83	0.79
> Western	0.09	0.06	0.08	0.11
> non-Western	0.08	0.04	0.09	0.10
Household status				
> single	0.30	0.18	0.28	0.44
> couple	0.26	0.23	0.26	0.26
> with parents	0.40	0.56	0.42	0.23
> other	0.04	0.02	0.04	0.07
High earning parents	0.31	0.27	0.31	0.35
N	**13,621**	**1,916**	**8,859**	**2,846**
%	100	0.14	0.65	0.21

Extended abstract LACOSA II

Sibling Similarity in Entry into the Labor Market

Aleksi Karhula, Jani Erola, Marcel Raab and Anette Fasang

Introduction

The effect of family background on labor market outcomes is the founding question in the research on intergenerational mobility. If social origin plays a significant role on labor market outcomes, we have failed to achieve one of the most universally accepted goals of every society: the equality of opportunity. Indeed the association between social origin and occupational destination often enforced through education has been one of the most studied in all social sciences.

However, the complete pathways of labor market entry have been researched surprisingly little with a holistic approach. Most of the existing literature concentrates on the final outcomes, i.e. the education or labor market outcomes at a certain age at which it is reasonable to assume that they have stabilized. This gives us an accurate picture of how the family background affects e.g. education (Branigan et al. 2013; Breen and Jonsson 2005; Sieben et al. 2001), labor market outcomes (Björklund et al. 2002; Erikson and Goldthorpe 2010; Solon 1992) and in more elaborate cases labor market outcomes through education (Blau and Duncan 1967; Hauser and Mossel 1985). In the first two cases these studies tell us very little, if anything at all, of the pathways to these final outcomes. Even in the case of education mediating the effect of family background, the description of trajectories is often very rough. The studies analyzing the educational and labor market trajectories holistically as a single entity are all together rare although there are some notable exceptions (e.g. Brzinsky-Fay 2007; Pollock 2007).

2

In this paper we return to this founding question of intergenerational mobility research analyzing the effect of family background on combined educational and labor market trajectories. Our aim is to analyze how much family background influences these trajectories in the early adulthood (from the age of 16 to 35). Furthermore we study how well the observed family background characteristics account for the total family background effect and what kind of trajectories are most strongly linked to family background. Last we claim that the use of labor market trajectories can broaden our view on intergenerational inequality and further show how much of the intergenerational transmission of inequality through trajectories would remain unseen analyzing only the outcomes at certain age as is often done. We utilize high quality yearly register data and the latest methodology of sequence analysis.

Research questions and design

To identify the family background effect we use a sibling comparison design as is often done in the research on social mobility (e.g. Björklund et al. 2002; Conley and Glauber 2008; Erola 2009; Mazumder 2008; Solon 1992). Most of the studies employing sibling methods use sibling correlations, but as this is not possible when considering sequences we employ dyadic regression approach we have previously applied on family formation (Raab et al. 2014).

We aim to expand the classical origin-education-destination (OED) -framework into holistic analysis of whole educational and occupational sequences. This way we can see the family background effect on the whole career pathways in young adulthood instead of concentrating on the outcomes at certain age. We are interested in four research questions:

(1) Are entry into labor market trajectories of siblings more similar?
(2) Can observed parental background characteristics account for sibling similarity?
(3) In which way is siblings' labor market entry more similar?

3

    (4) How much of the sibling similarity in trajectories would we miss by just looking at the outcomes at age 35?

In our research design we first compare the differences in siblings' trajectories to differences in trajectories between one of the siblings and randomly assigned unrelated person (Figure 1). If the distances to siblings are smaller compared to distances to randomly assigned unrelated persons, we can conclude that family background has an effect on early career trajectories (RQ 1). Furthermore by matching the randomly assigned persons conditionally on family background characteristics in our quasi experimental design, we can see whether or not the measurable family background accounts for the smaller amount of differences (RQ 2). After analyzing the first two research questions we turn to third one by clustering the sequences and analyzing whether certain clusters are strongly influenced by family background, i.e. analyze whether the siblings are more likely to reside in same clusters in certain cases. Last but not least we match the sibling dyads with randomly assigned dyads with similar outcomes. This gives us an estimate of the family background effect on trajectories even, if the outcomes are identical between the dyads and thus the association with family background would be unseen without sequence measure.

We define the career trajectories as sequences with seven different states: studying, unemployed, otherwise outside workforce and income in four categories. The study category contains people who are studying fulltime according to registers. The unemployment is defined as being unemployed at the last work week of the year. People considered to be outside the workforce are those with no clear employment, unemployment or studies. These people consist mainly of parents staying home with children and in case of men people serving the mandatory military or civilian service (6 to 12 months). Four income groups are defined relative to the income quantiles at the age of 35 in our cohorts.

4

Siblings are linked through the mother and all the variables and information on family relations is based on register data from Statistics Finland.

Results

In the preliminary analysis we have used optimal matching with theory based substitution costs and indel cost to define the distances between persons. The distributions of distances in sibling and unrelated dyads according to sex can be seen in Figure 2. We can clearly see that the siblings resemble each other more compared to unrelated dyads (RQ 1). The sibling dyads have smaller distances especially when comparing the same sex dyads. This would imply that some of the family background effects are sex-specific. The results show clearly that family background has a strong effect on the career trajectories in early adulthood.

5
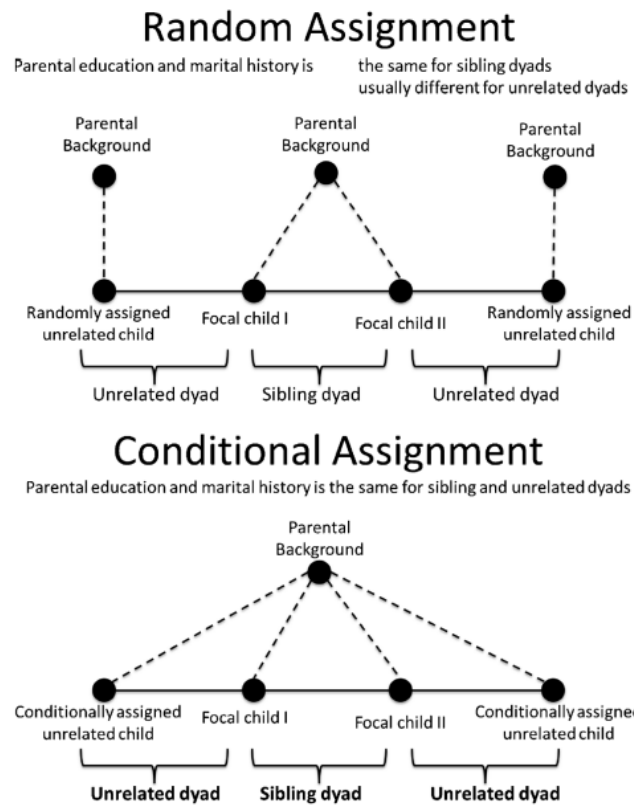
Figure 1. Random and conditional assignment of dyads (from Raab et al. 2014)*.*

6
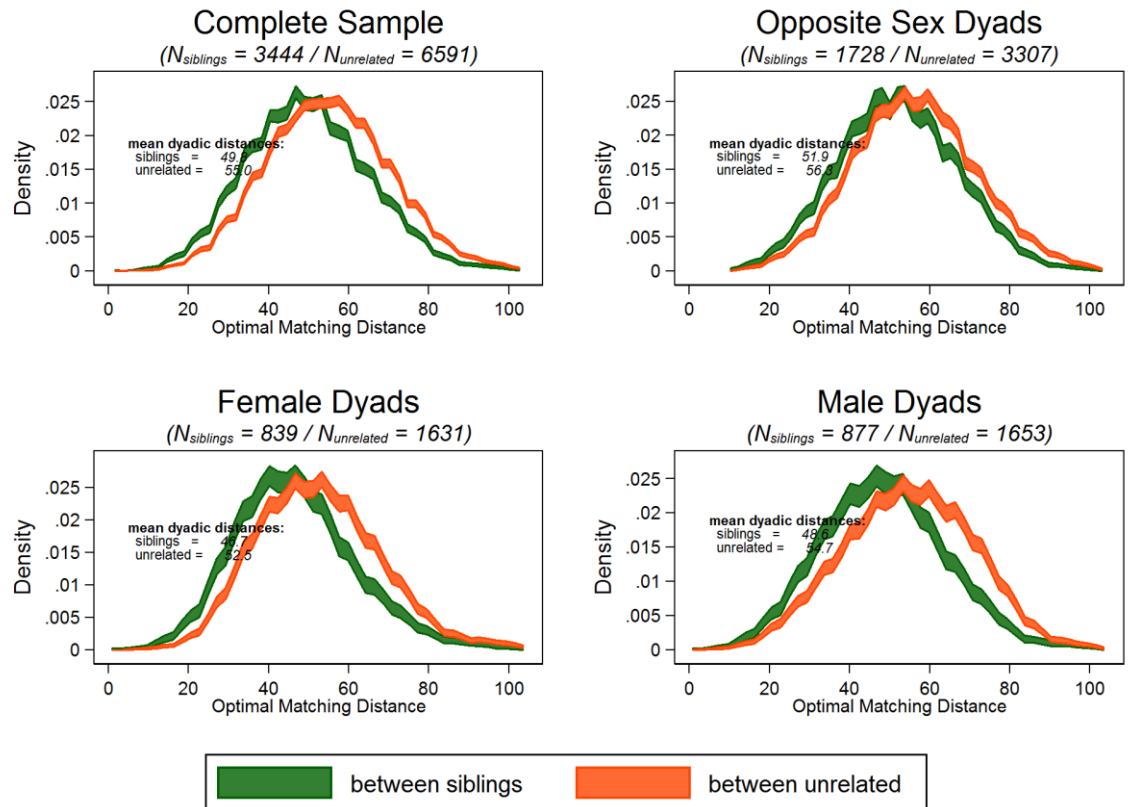
Figure 2. Distributions of distances for sibling and unrelated dyads for all, opposite sex dyads, women and men (standardized between 0 and 100).

In the case of family formation surprisingly little of the similarity between siblings was explained by the measurable family background (Raab et al. 2014). In the case of career trajectories we expect stronger effect of measurable family background. However, when looking at the sibling similarity in our models (M1 and M2 in table 1), we can see that upon conditional assignment on family background the sibling similarity compared to unrelated dyads decreased surprisingly little although significantly (19 percent).

7

Table 1. Similarity of sibling dyads compared to unrelated dyads in in different matching models.

	M1 Random Assignment	M2 Conditional Assignment on Family Background	M3 Conditional Assignment on earnings (at age 35)	M4 Conditional Assignment on education (at age 35)	M5 Conditional Assignment on both outcomes (at age 35)
Sibling dyad	-5.52***	-4.45***	-4,27***	-3,59***	-2,72***
	(0.17)	(0.17)	(0,15)	(0,16)	(0,15)
Observations	34,724	33,299	34712	34712	34712

In the last models (M3-M5) we can observe that matching the sibling dyad with a dyad with similar outcome at the end of our observation period at age 35, did not result into non-existing sibling effect (RQ4). It is noteworthy that in case of income (M3), education (M4) and both (M5) quite much of the sibling similarity remains in the sequences although no sibling similarity would be observed looking at the outcomes. This shows that a lot of sibling similarity would be missed when only looking at sibling similarity of the outcomes. This further implies that we do not capture intergenerational inequalities fully with only the end outcomes.

For the lack of space we do not present the results to our RQ3 (In which way is siblings' labor market entry more similar?) in detail here. It suffices to say that sibling similarity seems to be strongest when it comes to disadvantageous trajectories and lessened in academic trajectories.

If we return to the equality of opportunity mentioned in the beginning, our results of sibling similarity indeed imply lack of it. It would seem that when it comes to trajectories we cannot identify clearly the reasons behind the similarities although we can identify many family background factors associated with the similarities. Further the similarity of trajectories is stronger in disadvantaged trajectories and is likely linked to other life events and accumulating disadvantage. Last but not least the fact that much of sibling similarity in trajectories would remain hidden when look-

8

ing at only the end outcomes underlines that low sibling similarity in the end outcomes does not necessarily imply equality of opportunity as is too often hastily assumed. Similarities in outcomes might result from seriously different and unequal trajectories leading to these outcomes.

9

References

Björklund, A., Eriksson, T., Jäntti, M., Raaum, O., & Österbacka, E. (2002).
Brother correlations in earnings in Denmark, Finland, Norway and
Sweden compared to the United States. *Journal of Population
Economics*, *15*(4), 757–772. doi:10.1007/s001480100095

Blau, P. M., & Duncan, O. D. (1967). The American Occupational Structure.

Branigan, A. R., McCallum, K. J., & Freese, J. (2013). Variation in the Herit-
ability of Educational Attainment: An International Meta-Analysis.
Social Forces, *92*(1), 109–140. doi:10.1093/sf/sot076

Breen, R., & Jonsson, J. O. (2005). Inequality of Opportunity in Compara-
tive Perspective: Recent Research on Educational Attainment and
Social Mobility. *Annual Review of Sociology*, *31*, 223–243.
doi:10.2307/29737718

Brzinsky-Fay, C. (2007). Lost in transition? Labour market entry sequences
of school leavers in Europe. *European Sociological Review*, *23*(4),
409–422.

Conley, D., & Glauber, R. (2008). All in the family?: Family composition,
resources, and sibling similarity in socioeconomic status. *Research*

10

in Social Stratification and Mobility, *26*(4), 297–306. doi:10.1016/j.rssm.2008.08.003

Erikson, R., & Goldthorpe, J. H. (2010). Has social mobility in Britain decreased? Reconciling divergent findings on income and class mobility. *The British Journal of Sociology*, *61*(2), 211–230. doi:10.1111/j.1468-4446.2010.01310.x

Erola, J. (2009). Social Mobility and Education of Finnish Cohorts Born 1936—75 Succeeding While Failing in Equality of Opportunity? *Acta Sociologica*, *52*(4), 307–327.

Hauser, R. M., & Mossel, P. A. (1985). Fraternal Resemblance in Educational Attainment and Occupational Status. *American Journal of Sociology*, *91*(3), 650–673. doi:10.2307/2780205

Mazumder, B. (2008). Sibling similarities and economic inequality in the US. *Journal of Population Economics*, *21*(3), 685–701.

Pollock, G. (2007). Holistic trajectories: a study of combined employment, housing and family careers by using multiple-sequence analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *170*(1), 167–183. doi:10.1111/j.1467-985X.2006.00450.x

11

Raab, M., Fasang, A. E., Karhula, A., & Erola, J. (2014). Sibling Similarity in

Family Formation. *Demography*, *51*(6), 2127–2154.

doi:10.1007/s13524-014-0341-6

Sieben, I., Huinink, J., & De Graaf, P. M. (2001). Family background and

sibling resemblance in educational attainment. Trends in the for-

mer FRG, the former GDR, and the Netherlands. *European Socio-

logical Review*, *17*(4), 401–430.

Solon, G. (1992). Intergenerational income mobility in the United States.

The American Economic Review, 393–408.

Session 12B: Gender inequalities

The Cohorts of Convergence?

Danish Women and the Changing Paradigm of Women's Labour Market Participation

Luize Ratniece[a] and Lisbeth Palmhoej Nielsen[b.]

[a]Universitat Pompeu Fabra, [b]SFI - The Danish National Centre for Social Research

Abstract: Cohorts born during the 1940s have been the ones identified as breaking the middle-class ideal of separation of gendered work spheres and the following cohorts have been gradually converging towards their male peers. This article explores labour market trajectories of the cohorts of women who spearheaded the switch towards an uninterrupted presence in the Danish labour market. The extent and conditions of gender convergence in the Danish labour market for the 1941-1980 cohorts is assessed via sequence and regression analysis on Danish register data. The novelty of this article lies in a longitudinal approach that includes the intensity of the labour market attachment as a key aspect of analysis. The results indicate that the gender convergence when it comes to work regime is still incomplete. The divide between public and private sector work is an essential aspect that genders the labour force during the time span examined.

Key words: Gender revolution, gender gap, female labour market participation, Denmark.

1 Introduction

The interest in the gendered patterns of labour market attachment and success has been stirring the social science research for decades. The explanations for women's attachment to the labour market are more complex than those used for a masculinized workforce (Goldin 1990; Blau et al 1998). We now have the opportunity to map out the breathtaking advancement that women have made (Blau et al 1998; Goldin 2006; Coontz 2011) which is a continuous puzzle about the absence of full gender convergence in labour market and income trajectories (Gerson 2009, 2011; England 2010, 2011; Goldin 2014). An overview of the current literature reveals complex interactions that link personal preferences and structural constraints.

This paper explores the gender differences in labour market trajectories that are still present even in one of the pioneer countries of gender revolution. Denmark has undoubtedly been at the forefront of the gender revolution, encompassing profound changes both at the labour market and at the reproductive work (Esping-Andersen 2009 [2013], Esping-Andersen et al 2013), which makes it a case of great interest for anybody wishing to explore the process of women becoming intensely attached to the labour market and men starting to pick up their fair share of house and care work. According to the most of the common indicators of formal gender equality (such as the Gender Equality Index (UNDP 2015) or labour market participation rates (OECD Statistics 2016)), the Danish society seems very close to a complete gender convergence.

Yet despite the high participation rates and an early attitudinal shift towards gender egalitarianism (Jensen and Rathlev 2010), the Danish labour market is highly segregated by gender (Esping-Andersen 1993, Leth-Sorensen and Rohwer 2001, Statistics Denmark 2015). Women typically congregate in the public sector as care workers. Hence, we explore the extent of convergence of the labour market trajectories and closure of the gender gap in Danish labour market across a set of key cohorts. Examining the years during which care work passed from the private to the public sphere via increased gender egalitarianism and the advances of the Welfare State (Hernes 1987, Jensen and Rathlev 2010), we expect to map the gender revolution in the Danish labour market by comparing older Danish cohorts to younger cohorts.

There are three aspects of the methodical design of this analysis that contributes to the existing literature on how women construct their careers. The first one is choosing to compare women with their male peers. While throughout the years the labour market trajectories of men have served as the benchmark for women's achievements (Moen & Roehling 2005), in only a few cases men are explicitly assumed to be the counterfactuals of women. More often women have been compared among themselves, emphasizing the differences between the married and unmarried, and women with and without children. The strategy of comparing women to their male peers has been applied for the highly educated (for example, Goldin 2014), but not that often across all social classes. Taking into account

that during the time period of interest also male labour market trajectories have changed towards less labour market attachment (mostly due to education expansion and earlier retirement (Leth-Sorensen and Rohwer 2001, Jensen and Rathlev 2010), this approach permits to avoid comparison with an ideal-typical male trajectory and do it with actual data instead.

The second contribution is that the use of longitudinal data. Such design ensures the capacity to analyze the working lives as a dynamic phenomenon, capturing the key moments of labour market trajectories with greater precision than any cross-sectional data ever could. The female labour market revolution is a relatively recent phenomenon, even in the countries, such as Denmark, that were among first to shift from women holding jobs to women having careers, also known as the quiet revolution (Goldin 2006). Due to the short time spell, longitudinal research on full labour market trajectories is becoming possible just now as the cohorts who pioneered the massive incorporation in the formal labour market reach the retirement age (see, for example, García-Manglano 2014).

The third contribution stems from the comparison of key cohorts, hence following the gradual change of milieu in the population and convergence in labour market outcomes for men and women. This is only possible, because we have access to the best available dataset containing very detailed, longitudinal data on labour market experiences of Danes on an individual level. The Danish registry data cover the period from 1986 till 2011. A large representative sample is drawn from these population data and sequence analysis followed by regression analysis is applied to assess the extent and conditions of gender convergence in labour market trajectories of Danish women throughout the years observed.

The rest of this paper is structured in the following way. Section two offers an overview of the state of art regarding longitudinal analysis of women's labour market participation with a special focus of their relevance for Denmark. In section three we describe the data and the analytical methods we use. Section four contains all the empirical results, while section five discusses them. Conclusions and ideas for future development of this research question can be found in section six.

2 Background and relevance

In the West it has been taken as a given that women have some labour market experience, at least throughout the second half of 20th century (Goldin 1990, Blau et al 1998, Cooke 2011, Coontz 2011). Women are expected to enter the labour market before even considering the possibilities of opting out of it. The great majority of people with no labour market experience are among the youth who are yet to enter in the labour market.

The framing of paid work done by women has evolved. Participation rates have been rising up till the 1990s (Goldin 2006, 2014), especially for married women and women with young children (Cooke 2011). Also the significance of women's paid work has changed. The idea of jobs that would sustain one until marriage, or help family budget in case of occasional need (Warren & Warren Tyagi 2004) has been replaced with that of careers,

bringing about more purposeful investment in human capital and establishment of a professional identity (Goldin 2006). And, while this change has been more obvious for highly educated women, Rubin (1994) observe a very similar discourse among the working class women who do paid work primarily because of the economic need. Therefore, the notion of having a career is used in its widest meaning. It implies an ongoing attachment to the labour market and the very idea that doing paid work is most likely not a transitory phase in a woman's life (Goldin 2006).

This shift from jobs to career has been named *the quiet revolution* by Claudia Goldin (2006). For her the quietness of this change stems from the fact that it was not primarily driven by a discourse of labour market participation as an emancipatory practice in the battle for gender equality. Instead, it was a response to changes in patterns in union formation and fertility (such as reliable contraceptives and high divorce rates) and labour market characteristics (such as growth in administrative, service or institutionalized care jobs, and improvements in working hours and conditions). The trend towards convergence in women's labour market trajectories that Goldin observes in the US starting from the 1970s is a change due to exogenous reasons. "As opposed to the noisy revolution [Civil Rights movement, anti-war protests, women's liberation], the quiet revolution was accomplished by many who were unaware that they were part of a grand transformation. They were the unwitting foot soldiers of an upheaval that would alter women's employment, education, and family" (Goldin 2006: 32-33). Therefore asking if one has ever worked for pay is not that relevant anymore.

After observing this shift from framing the female employment as jobs to that of careers (Goldin 2006), the question is about the what extent that attachment women have to the labour market are now equaling that of men. The research interest now lies in the characteristics of women's careers. It is the entry and exit points, interruptions, length of the trajectories, intensity of dedication, professional growth during the years of attachment to the labour market that are the variables of interest. In some contexts the change towards the notion of career has effectively been labeled as the masculinization of women's life course (Esping-Andersen 2009 [2013]).

It has been recognized that this revolution has its own steps that depend on the labour market context: participation that goes from (1) working at certain life stages to working throughout life, especially the reproductive years, (2) part time to full time work, (3) work in feminized sectors to dissipation of the notion of feminized/masculinized sectors. Each country that has undergone the transition towards a more gender equal labour force has had its own unique mixture of these aspects.

Depending on the socio-cultural context and historical contingencies, the patterns of participation vary a lot across countries. While most women in Scandinavia are attached to the labour market throughout their lives), the women in the US are obliged to decide between full attachment or exit from the labour market (Pfau-Effinger 2004, Goldin 2006, Esping-Andersen 2009 [2013], Cooke 2011). At the same time women in the UK

experience a bimodal distribution of attitudes towards labour market participation and attachment to it (Hakim 1996). In Germany women tend to opt for part-time work instead of full time attachment once they form a family, and in the Netherlands women seem to be moving away from housewifery followed by part-time pattern towards their own quiet revolution where the new ideal might be dual earner couples where both partners work part-time (Pfau-Effinger 2004).

Even among the vanguard countries of female employment a complete convergence with men is closer in some countries than in others. The female labour force participation in Finland was 0.96 of that of male participation rate in 2013. Norway, Iceland, Sweden and Denmark all had ratios above 0.9. It was 0.85 in the US, and 0.78 was the average for all the OECD countries (OECD 2016). These numbers signifies the fast pace of female convergence since the quiet revolution that changed the way how women relate to the labour market (Goldin 2006). For example, the ratio of the Civilian labour force participation rates in the US labour force for those over 16 was 0.39 in 1950, 0.54 in 1970, 0.75 in 1990 and 0.82 in 2010 (US Department of Labour 2015). At the same time an inquiry about the failure to achieve a complete convergence is also possible (Williams 1999; Stone 2007; Gerson 2009, 2010; Cha 2010, 2013; Goldin 2014; Pedulla & Thébaud 2015). The pay gap – the other star indicator of women's progress in paid work – is experiencing exactly the same dynamics. The gap has been closing consistently if looking at the nation-wide averages while maintaining pockets of yawning gaps in certain professions (Goldin 2014).

Recognizing that the change in women's labour market attachment was the first step in the revolution, we focus on that. Figure 1 shows the labour market participation rates of women and men in different age groups from 1940 to 2010's. The graph clearly shows that women in all age groups have increased their labour market participation from below 40 per cent to more than 80 per cent. Yet part-time work is still rather prominent for Danish women, despite the fact that hour-wise these are typically "long part-time hours" (Blossfeld and Drobnič 2001b). It is not the almost half of all working women that it was in the 1970s (Bonke 1997) - and still is in Germany and Austria, and to a greater extent in Netherlands and Switzerland (Eurostat 2016) – but it ought to have an impact on the overall labour market outcomes for Danish women. While the impact should not be as severe as for typical British, Dutch or German women (Hakim 1996, Esping-Andersen 2009 [2013]), it is still a prominent gendered difference in the Danish labour market.

The time of the great advancement in women's labour market attachment is meshed with several other phenomena: the overheating of the Danish economy in the early 70's (Esping-Andersen 1990 [1991]: 170-173, Koch-Nielsen 1998) and the preference to incentivize the labour market participation of women instead of inviting foreign guest-workers (Esping-Andersen 1990 [1991]: 173, Lewis 1992, Koch-Nielsen 1998, Jensen and Rathlev 2010), as well as the rise of the consumer culture and subsequent demand for 1.5 or 2 salaries per household (Koch-Nielsen 1998, Warren and Warren Tyagi 2004).
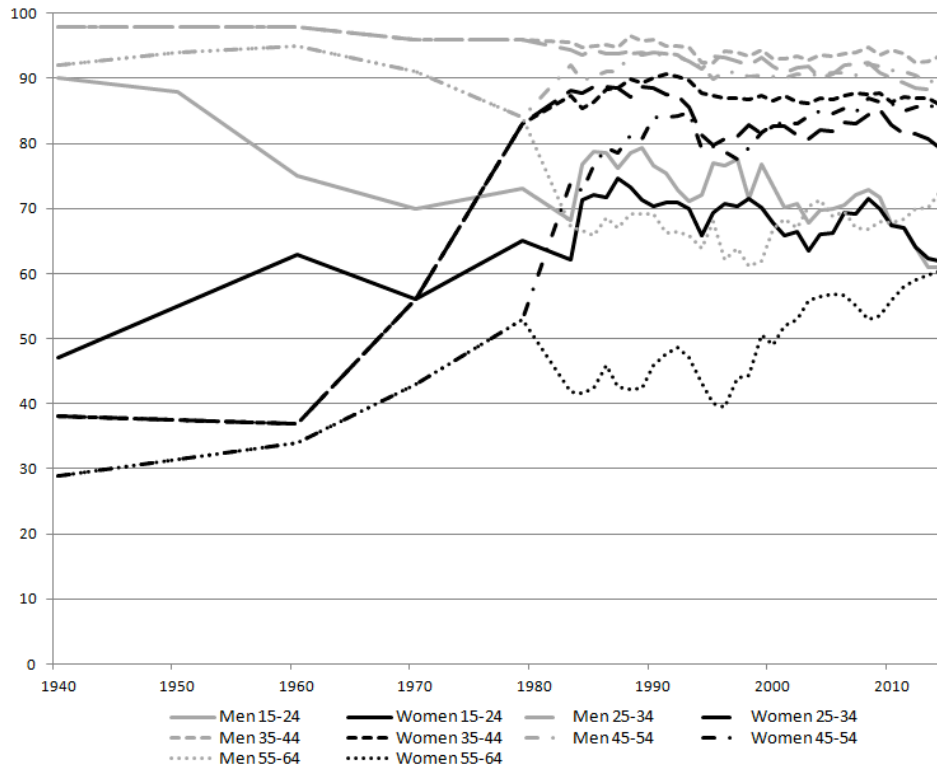
Figure 1: Changes in the rates of active population in Denmark between 1940 and 2014 between ages 15 and 64 by gender. Source for data until 1983: Bonke (1997). The values for women in 1950 are imputed due to missing data and the age groups used there are wider than in the OECD data. Source for data starting from 1983: OECD Statistics (2016).

At the same time, the understanding of what constitutes a standard work week has changed. Since the early 1990s, normal weekly hours of 37 and overtime premiums have been established through collective bargaining (OECD 1998, Lee 2004), a change that started with the Metal Working Industry Agreement and later spread to other sectors. While this change has made shorter hours the mode of the duration of a full-time work-week (Bishop 2004, Lee 2004), in early 2000s almost a third of the men and 10 per cent of the women employed reported having a work-week over 40 hours (Bishop 2004).

The care work was crowded out by the welfare state and is now taken care of mainly by women employed by the public administration (Jensen and Rathlev 2010) in a very formalized way (Jensen et al 2010, Jensen and Rathlev 2010, Pfau-Effinger et al 2010), hence there was no obvious need for changing the care work patterns inside the families (Blossfeld and Drobnič 2001b). The ideal-typical model of the Danish family is having two breadwinners and the state serves as the major care-provider (Esping-Andersen 1999, Jensen et al 2010, Jensen and Rathlev 2010, Pfau-Effinger et al 2010). An example of this ideal is reflected in the parental-leave policies. In comparison with Norway and Sweden that have had generous maternity-leave arrangements since the early 1960s, Denmark has

6

opted for using public childcare instead. The relatively short maternity leave coupled with a high coverage by public day-care institutions "helps to preserve women's close contact with the labour market and their human capital investments during maternity" (Jensen and Rathlev 2010: 46).

Aligned to this is also the tax system. There is no fiscal incentive for the breadwinner-homemaker model from the highly individualized Danish taxing system (Blossfeld and Drobnič 2001b, Leth-Sorensen and Rohwer 2001), as having two incomes does not undermine the benefits and tax returns, and high income earners pay a larger share of their income than low income earners. The expansion of the Danish welfare state has created a virtuous loop of public employment for women, boosting the participation rates. In the late 1990s, 30 per cent of all employees worked in the public sector, and almost 45 per cent of all employed women (Leth-Sorensen and Rohwer 2001, Statistics Denmark 2015).

While Danish women have partly overcome the part-time hurdle and the country is characterized by an early attitudinal shift towards gender egalitarianism (Jensen and Rathlev 2010), Denmark at the same time has a highly gender segregated labour market, both sectorial and occupational (Esping-Andersen 1993). Even more, at least back in the 1990s some would confirm that,

> Although [Denmark and Sweden] have proactively tried to integrate married women with children in the labour force via full-time work, reduced full-time work, and part-time work, 'the wife's role as a supplementary worker has hardly changed' (Bernhardt 1993). It is generally acknowledged that Scandinavian women are less dependent on their husbands or partners in financial terms than women in many other Western European countries, but they, by and large, still suppress their own long-term job opportunities, earning profiles, and other job-related interests when they raise young children. (Blossfeld and Drobnič 2001a: 7)

The more recent work suggests that Denmark has undergone a full gender revolution in intra-couple behaviour since then (Esping-Andersen [2009] 2013, Esping-Andersen et al 2013), converging in the amount of housework done. Yet the tensions between careers and family life described are at recent if not present and have been a reality for many of the women who are still active in the labour market.

We set up three hypotheses that we test empirically:

Hypothesis 1: The labour market trajectories are not gender-neutral, but a gradual convergence is observable across the birth cohorts. It is expected that the regression analysis will confirm the hypothesis that the extent to which gender is a strong predictor of a labor market attachment-based cluster membership is fading out with subsequent birth cohorts (Han and Moen 2001).

Hypothesis 2: Women with higher education credentials are more likely to have trajectories closer to those typical of men. It has been established that one of the most important predictors of labour market attachment is larger investment in education (Han and Moen 2001, Goldin 2014). In previous Swedish research the authors found that higher education for women reduce risks of labour market exits, increase the likelihood of re-entries and speeds up the return (Henz and Sundstrom 2001). For Denmark it has been

found that less education is linked to higher risk of experiencing unemployment and exiting form the labour market (Leth-Sorensen and Rohwer 2001). Despite the educational expansion that has taken place across the cohorts of our interest, we expect that there will be differentiation by education level in all birth cohorts.

Hypothesis 3: For women the presence of young children in the family has a much stronger impact on work trajectories than for men and are more likely to be towards less intense labor market attachment (Blossfeld and Drobnič 2001a). A major hindrance to career success for women is children, both for those that exit the labour market and those that return (Stone 2007, Cooke 2014). Previous results suggest that – while the institutional framework mediates the intensity of the effect – even in the most family-friendly national contexts there are significant effects (Aisenbrey et al 2009). In Sweden, having small children increases the likelihood of part-time work and labour market exits of women (Henz and Sundstrom 2001). Also for Denmark, having children have been found to decrease the likelihood of uninterrupted presence in the labour market for women (Leth-Sorensen and Rohwer 2001). For older cohorts in the US, Han and Moen (2001) found that strong labour market attachment for women even precluded stable unions and vice versa, while it did not impact the labour market trajectories of men (2001).

In light of the move towards more gender-egalitarian patterns of sharing the care work at home (Esping-Andersen et al 2013), we expect to observe a gradual change across the cohorts. We expect that the labour market trajectories of younger Danish women will be less negatively affected by parenthood, while men's will begin to experience a negative impact similar to that of women. Among older cohorts we foresee a bimodal tendency. Only women with significant human capital, or in urgent need of resources, will adopt a fully attached working life whereas men's labour market trajectories are not expected to be sorted by union or fertility events.

3 Data and Methods

The optimal data for an analysis that is able to assess a full labour market trajectory are those covering the whole life course and at the same time containing a wealth of additional information. Cross-sectional data would offer only a glimpse at an individual's connection to the labour market instead, as proposed, to examine the complete development of women's careers. Previous research has found women's labour market trajectories to be dynamic to an extent that preferences stated early in life, or even behavior at some point, hold little prediction value over the later life course (Gerson 1985; Hakim 2002; García-Manglano 2014).

The Danish register data maintained by Statistics Denmark permits to carry out a longitudinal research design. The unique datasets follows the entire population since the mid-1980s and contain individual, yearly, greatly detailed information on, among other variables, labour market participation (For a more detailed description see, for example,

Leth-Sorensen and Rohwer 2001). In comparison to the second-best data which would be longitudinal panels, register data offer more precision due to the fact that the data are not self-reported, why we avoid attrition problems. Because the data are not retrospective, we also avoid recall bias. The drawback from the register data is that there is no information on attitudes or interpretations of the behaviors, such as the reasons behind retreat from the labour market or choice of work sector. Furthermore, the fact that the work with these sensitive data has to occur on the Statistics Denmark servers, the computational capacity required for the chosen methodology obliges us to select a smaller random sample to work with instead of the entire population of birth cohorts of interest. The final sample is still a large sample though, of 10.000 observations per subsample.

Aiming at analyzing only the complete labour market trajectories of women born in 1940s and after, there are very few cohorts available for such research design. The literature agrees that the quiet revolution in the US started with women born during the 1940s (Gerson 1985; Goldin 2006; Coontz 2011; García-Manglano 2014), which was also true for Denmark (Koch-Nielsen 1998, Blossfeld and Drobnič 2001b, Henz and Sundstrom. 2001, Pfau-Effinger 2005). Therefore, it is reasonable to take these cohorts as the baseline for the evolution of the women's labour market trajectories in the last decades of the 20th century, and then follow the subsequent cohorts.

Due to data limitations, we are unable to work with full labour market trajectories, as the 1940's birth cohorts entered in the labour market in the late 1950s and early 1960s when the registers were not yet set up. Additionally, some of the younger cohorts that are currently still in the labour market are only half-way through, hence their full labour market trajectories are not available either. The register data available allows us to observe 26 consecutive years between 1986 and 2011. In order to work with sequences meaningful to our research question the final sample is restricted to individuals who immigrated in Denmark no later than 1998 and individuals who died *after* 1999, assuring that they have been "present" for at least half of the observed years. Individuals with missing data on labour market attachment (most probably due to having spent time abroad) have been excluded. Individuals are included in analysis from the moment that the individuals are 17 years old, due to the fact that late teens are the years when divergence among labour market trajectories start.

Due to data limitations, until 1994 it's impossible to distinguish the type of paid work that the person observed is engaged in. Only from then on we are able to distinguish between employment in public or private sector, part-time work and full-time work. Despite the fact that until then the only labour market dynamic observable is being in paid work versus different reasons for being out of it, we opt to include these years in our analysis. For the oldest cohort – the most affected by this limitation – it is still analytically interesting to look at the dichotomy between paid work and absence from the labour market, even if there are no other details available.

Recognizing that we work with a complex dataset where the variables of interest reflect both cohort and period effects, we work with four separate segments of the sample. We end up with four subsamples of the birth cohorts grouped as follows: (1) those born in 1941-1950, (2) 1951-1960, (3) 1961-1970, and (4) 1971-1980. For the sake of simplicity from now on we will be referring to these groups as "birth cohorts".

For each of the cohorts there is a common chunk of 17 (fifteen for the youngest cohort) years observed for all subsamples, and we choose to expand it including all the available data (see table 1). While accepting that the results will be based on fewer cohorts on the tails of the sequences, this approach allows us to make the most of the available observed years and achieve an almost full vision of the adult lives of the three older birth cohorts.

Table 1: The age ranges observed for each of the birth cohorts.

	Total age range observed	The common age range	Number of common observed years for all observations	Number of observed years for each observation
1941-1950	36-70	45-61	17	26
1951-1960	26-60	35-51	17	26
1961-1970	17-50	25-41	17	25
1971-1980	17-40	17-31	15	15

As the literature has shown (Gerson 1985; Hakim 1996, 2002; Stone 2007; García-Manglano 2014), there is a lot of change during the life course and a lot of heterogeneity among women. To carry out this research design, the method of choice has to be one that maintains the complexity of trajectories in a longitudinal analysis (Widmer & Ritschard 2009; Barban 2011). As much of the sociological thinking explains phenomena as sequences and interactions, it makes sense to work with a method that addresses that directly (Abbott 1995). It also has to be a method that is able to capture the patterns behind the sequences (Abbott 1995). Sequence analysis permits all of this (Abbott 1995; Abbott & Tsy 2000; Billari 2001; Aassve et al 2007; Widmer & Ritschard 2009; Barban 2011; Barban & Billari). Sequence analysis is especially relevant for analyzing life courses, as it captures the complex lives of people as heterogeneous sequences in different states (Billari 2001). Aassve, Billari & Piccarreta (2007) have carried out a similar design to ours for the trajectories of women in the UK, finding the method fruitful and the results relatively easy to interpret.

Sequence analysis serves as a purely descriptive technique, permitting the construction of whole sequence of diverse events without any assumption about them (Abbott & Tsay 2000). We use it to create two longitudinal trajectories (R Development Core Team 2011, Gabadinho et al 2009, 2011), one of labour market related activity and the other of household composition. The Labour market sequence includes the relationship with the labour market, the intensity of that attachment if the individual is working and, for the specific case we are analyzing, if the paid work is carried out in the private or public sector.

There are 12 possible sequence states: (1) working full-time in the public sector, (2) working full-time in the private sector, (3) working part-time in the public sector, (4) working part-time in the private sector, (5) working but with further details unknown (this is the state for everybody working until 1994 due to data limitations, see above for explanation), (6) being on a leave (sick, maternity, sabbatical, etc), (7) being unemployed, (8) being in school, (9) being out of the labor market (retired or otherwise not active), (10) dead, (11) not yet in Denmark (for those who immigrate during the observed period), and (12) missing due to data limitations regarding the years observed.

The Household sequence captures partnership status and number of children under 18 in the household. The register data permits to include all couples that are sharing a dwelling, be they married or cohabiting. The importance of cohabiting in Scandinavia (Bracher and Santow 1998, Henz and Sundstrom 2001, Leth-Sorensen and Rohwer 2001) makes sharing a dwelling more appropriate than using civil status. The only care aspect included is the number of children. Taking into account the fact that Danish families have a relatively small care burden due to the high levels of formalized care by the welfare state (Jensen and Rathlev 2010), the number of children raised in the household is the most straightforward and common stratifier of amount of care demanded from a person. There are 11 sequence states of household composition: (1) living with parents, (2) single, (3) living with a partner, (4) single with a child, (5) in a union with a child, (6) single with two children, (7) in a union with two children, (8) single with three or more children, (9) in a union with three or more children, (10) unknown (implies exit from the sample mostly due to death), (11) missing due to data limitations regarding the years observed.

First we provide a description of the typical activity in the labour market and household composition trajectories at the observed ages for each of the birth cohorts. We cluster each of our subsamples according to the most common trajectories (Studer 2013). Across the four cohorts, the optimal number of clusters for the Labour market sequence is three, though there are slight differences due to the different age spans we are observing for each cohort. For the same reason the number of clusters for Household sequence varies for each cohort. Once the clusters are created, we explore the correlation between the two sets of clusters for each cohort descriptively and apply regression analysis to examine the variables that predict cluster membership.

In the following we first provide a description of the typical work and family trajectories at the observed ages for each of the birth cohorts, and the way those two trajectories correlate. Second, we assess to what extent the Labour market cluster membership is predicted by gender and Household cluster membership.

The key results of the regression analysis (See tables 6-9 in the Annex) is the predicted margins of the likelihood of the Household cluster membership depending on the effects of the three-way interaction among gender, education level, and Household cluster membership. These are presented in the graphical form in the Results section below (Figures 18-29).

4 Results

4.1 The 1941-1950 Birth Cohorts and Labour Market

The years between 1986 and 2011 for the 1941-1950 birth cohorts, observed between ages 36 and 70, are dominated by paid market work and gradual move into retirement. The patterns observed in the figure 2 confirm the gendered division between private and public sectors. It is an index plot summarizing all 10,000 sequences for this cohort. Five blue hues represent different types of paid work while a range of other colors represent the rest of states. It can be observed that the trajectories of men (left side of figure 2) are dominated by private sector jobs while those of women (right side of figure 2), especially of the older cohorts (upper part of the figure), are dominated by work in the public sector. Part time work is not randomly distributed either. It is more prevalent among women. We do not distinguish between different types of exits from the labour market, but we can observe that trajectories of women become dominated by being out of the labour market sooner than those of men.

The most typical trajectories confirm the gendered differences in exit from the labour market and in the likelihood of working in either the private or public sector. These are the 20 most common trajectories, covering 14% of all male trajectories and 10% of all female. There is a total 6608 unique sequences in this birth cohort. The elevated number is partly due to the unequal length of the observed sequences, partly due to the heterogeneity of the population. Yet it is clear that for a Danish man born in the 1940s the typical trajectory has been to work continuously and mostly (while not exclusively) in the private sector, followed by retirement from the labour market in their sixties or later. The most typical women's trajectories range between continuous paid work and permanent absence from the labour market. Work in the public sector dominates the trajectories of women participating in the labour market. Retirement among those is experienced earlier than among their male counterparts working in the private sector, and is common already in their fifties.

The mean years spent in each state is another way of observing the same divisions in the labour market. The great heterogeneity among the observations is confirmed by the fact that in no state, any of the genders spend more than ten years on average. Paid work in all its different modalities dominates the trajectories, followed by having exited from the labour market. Men spend more years in paid work (in all work categories together) than women, while women are the ones spending more years out of the labour market. Also evident is the gendered differences between public and private sectors and part-time and full-time work. While all categories are present in the sample, it is clear that the auxiliary categories of being on a leave, unemployed, in education, already dead and not yet in Denmark, only unemployment has a significant appearance among average years spent in different states.

As specified above, deaths and arrivals to Denmark have been censored, while being in education and on leave are not states relevant to this birth cohort.

Division in clusters reveal divisions created along the lines of trajectories dominated by private or public sector work for clusters 1 and 2, and between continuous labour market attachment (clusters 1 and 2) and a less typical cluster that gathers trajectories marked by absence from the labour market and early retirement, unemployment and part-time work. 38 per cent of the sample is sorted in the first cluster, 32 in the second one, and 30 per cent in the third. Taking into account that this is the birth cohort that we observe entering into retirement, the elevated "population" of the third, marginal, cluster is not surprising.
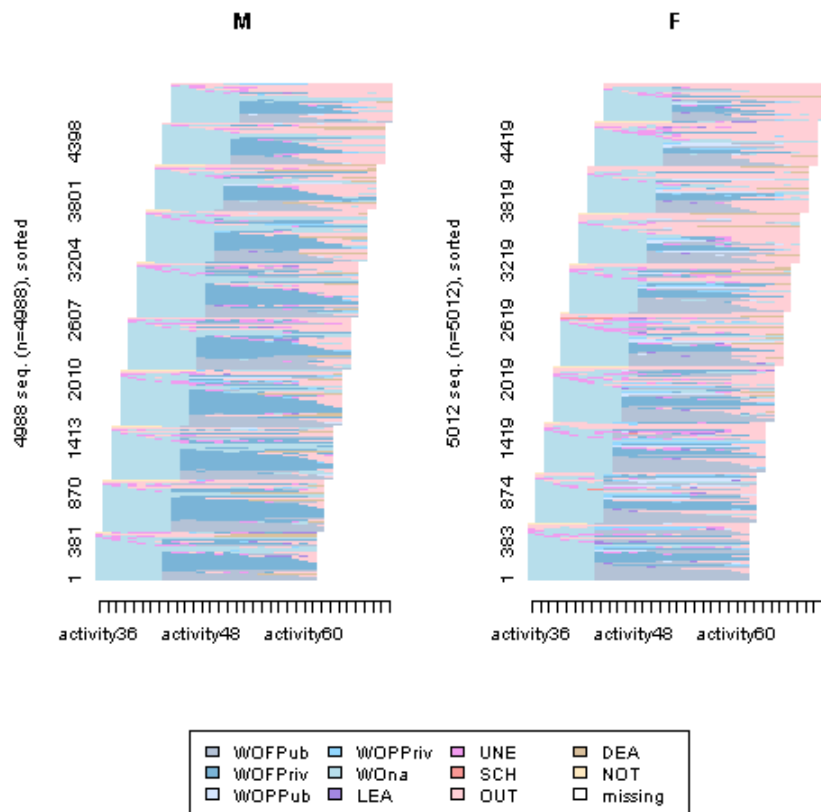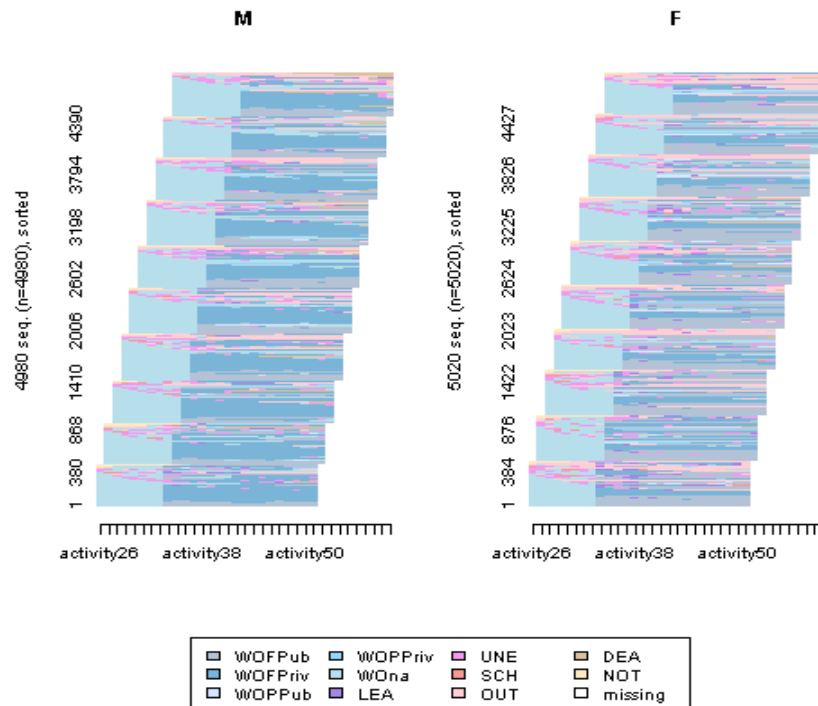

Fig. 2: Labour market sequence full-sequence index plots for the 1941-1950 birth cohorts.
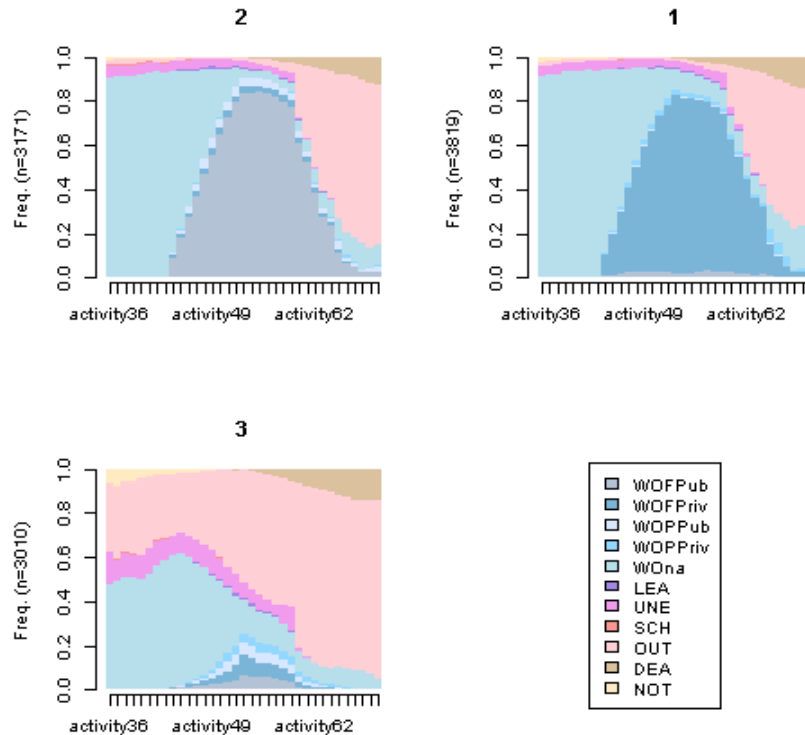
Fig.3: State distribution plots by Labour market cluster for the 1941-1950 birth cohort.

4.2 The 1941-1950 Birth Cohorts and Household Composition

Though the heterogeneity of the trajectories is high (7460 unique sequences, the twenty most common cover 9 per cent of male trajectories and 8 per cent of female), when it comes to the household composition at these ages, we mainly observe the process of "emptying the nest" (see figure 4). Most households start with having children under 18 living with them and then move towards living with their partner or alone.

According to the mean years at each state, men and women follow similar dynamics of mostly spending these years with their partners. One of the visible gendered differences is that women on average spend more time being unpartnered but with a child, confirmation that single mothers are more common than single fathers. Yet men on average spend more time with a partnered household maintaining two or more minors.

Clustering divides the sample in two (see figure 5). Cluster 1 is dominated by singlehood and cluster 2 – by being in a partnership. The great majority of the population (78 per cent) is sorted into the cluster of being in a partnership.

14
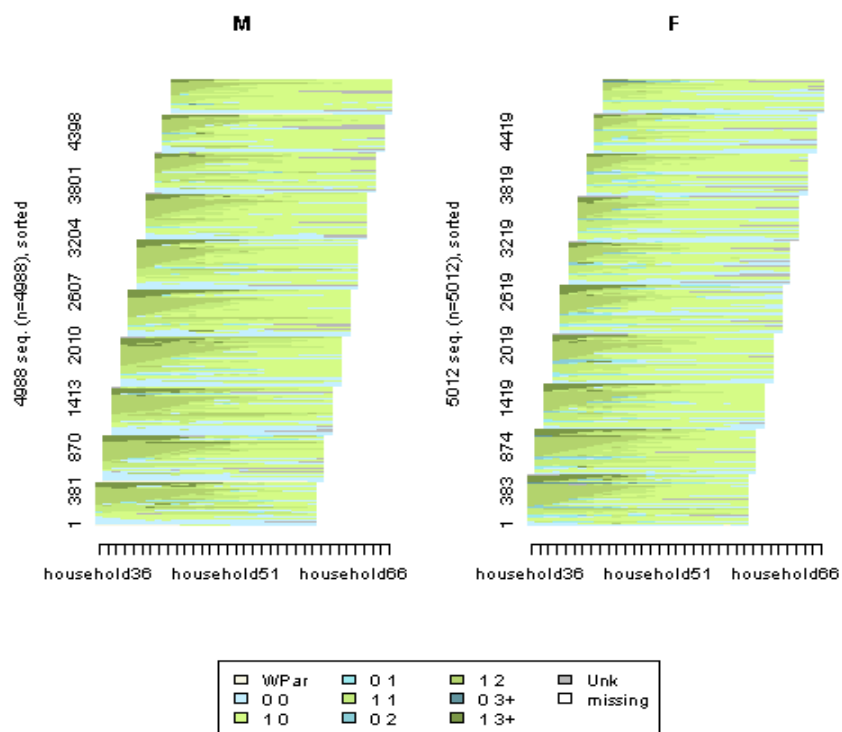
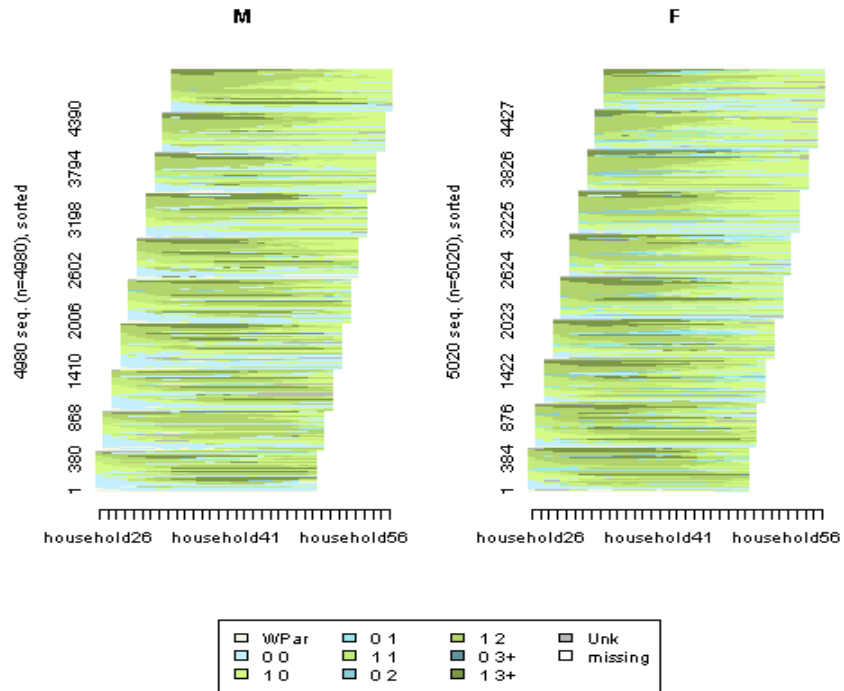**Fig. 4:** Household sequence full-sequence index plots for the 1941-1950 birth cohort.
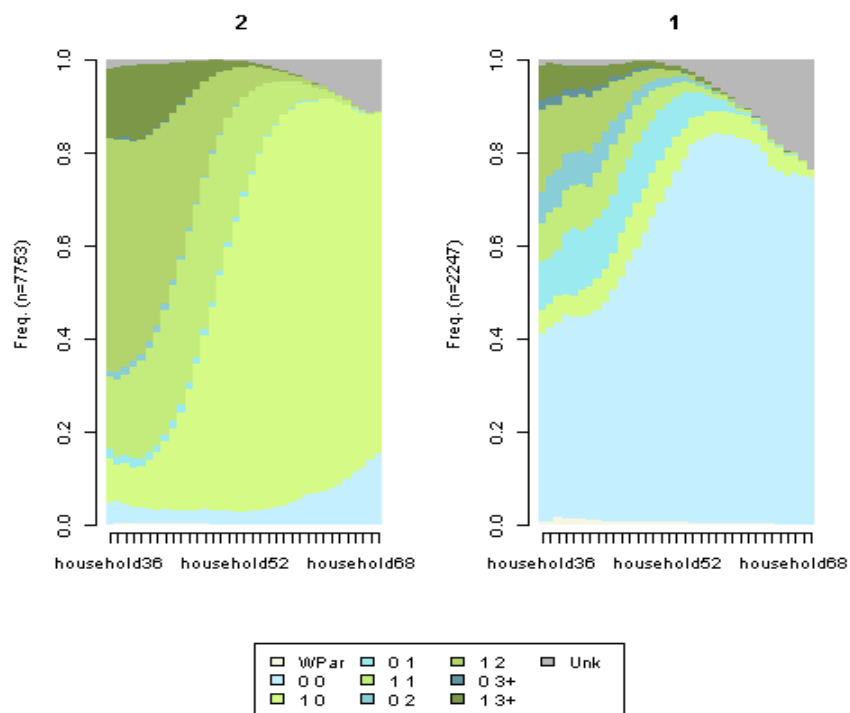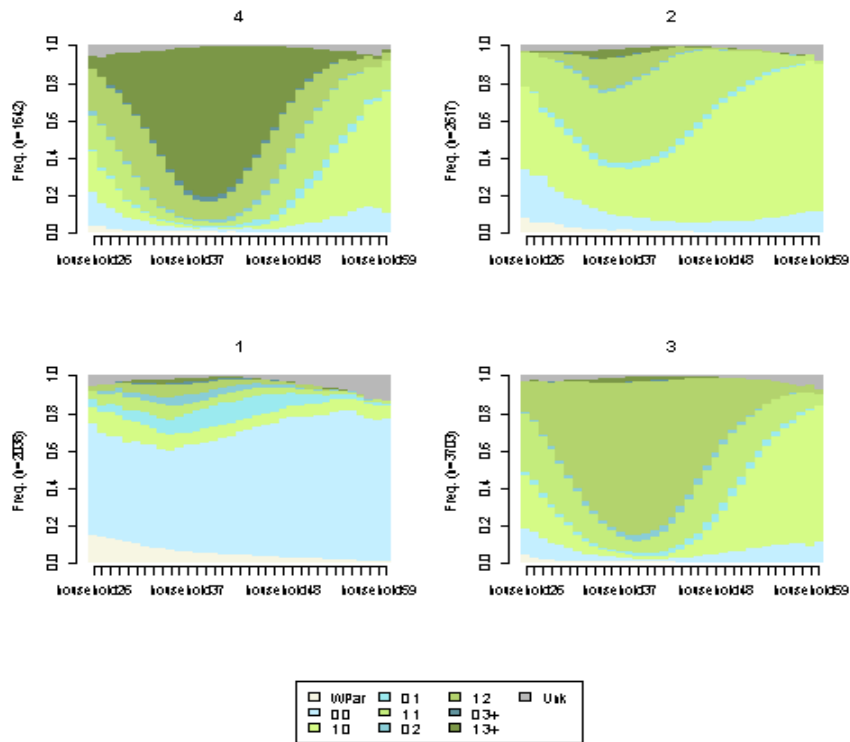

Fig.5: State distribution plots by Household cluster for the 1941-1950 birth cohort.

15

4.3 The 1941-1950 Birth Cohorts' Labour Market Activity and Households

The two cluster memberships are not homogeneous across each other (See Table 2). The majority of the sample is people whose trajectories are dominated by being in a union during the time span observed, 78 per cent of men and 77 per cent of women are sorted in the Union cluster. The Labour market cluster membership is not gender neutral, though. Most men (52 per cent) belong to the cluster characterized by full-time private sector employment, and the rest is evenly divided between full-time public sector work and absent-intermittent trajectories. The proportion of private sector cluster members is even bigger among the men from the Union cluster (56 per cent). Yet among the members of the household cluster characterized by singlehood, the most numerous group is that of Absence-intermittence – grouping together labour market absentees and retirees – members (44 per cent).

Women of this birth cohort are mostly members of Public sector and Absence-intermittence cluster (39 and 37 per cent respectively), while only a quarter belong to the Private sector cluster, the most common among men. The distribution of Labour market cluster membership among women from the Union cluster mirrors that of the total sample. But women with trajectories dominated by singlehood follow a similar pattern than the men from that cluster. Those single tend to be also the ones belonging to the Absence-intermittence cluster (44 per cent), followed by the full time cluster characteristics to one's gender (38 per cent of women in the Public sector and 37 per cent of men in the Private sector). This initially counterintuitive finding is in line with previous findings of Leth-Sorensen and Rohwer (2001) linking lack of success at the labour market for men with decreased likelihood of having experienced fatherhood.

Table 2: A contingency table between Labour market and Household clusters for the 1941-1950 birth cohort.

		Household clusters					
		C1 Single Men (22)	C2 Union Men (78)	Total Men (100)	C1 Single Women (23)	C2 Union Women (77)	Total Women (100)
Labour market clusters	C1 Private	37	56	52	18	26	25
	C2 Public	19	26	24	38	39	39
	C3 Absence-Intermittence	44	18	24	44	34	37
		100	100	100	100	100	100

Items may not sum exact 100 due to rounding.

4.4 The 1951-1960 Birth Cohorts and Labour Market

The age range observed of those born between years 1951-1960 is 26-50. While their trajectories are peppered with part-timing, schooling, leaves, unemployment and exits, for both men and women they are dominated by full-time work (figure 6).

An inspection of the most common trajectories confirms the predominance of uninterrupted work. Covering 28 per cent of men and 21 per cent of women in the sample (from a total of 6687 unique sequences), it also reiterates the gendered division between public and private employment. While the division is not exclusive, of course, it's much more common for women to work in the public sector than it is for men. The mean years spent in each state illustrate the dominance of paid work. The gendered differences are largely the same as in the oldest birth cohort. Women do more public sector work and more part time work. They are also spending more time out of the labour market and on leaves.

The clustering of this birth cohort follows the same logic as with those born between 1941-1950, yet the proportions of cluster membership are different (figure 7). Cluster one – dominated by private sector employment – include 47 per cent of the sample. Cluster two - dominated by the public sector employment – covers 39 per cent. The Absence-intermittence cluster becomes a residual cluster (14 per cent) of trajectories vaguely attached to the labour market and marked part-time work, leaves, unemployment and absence from the labour market.
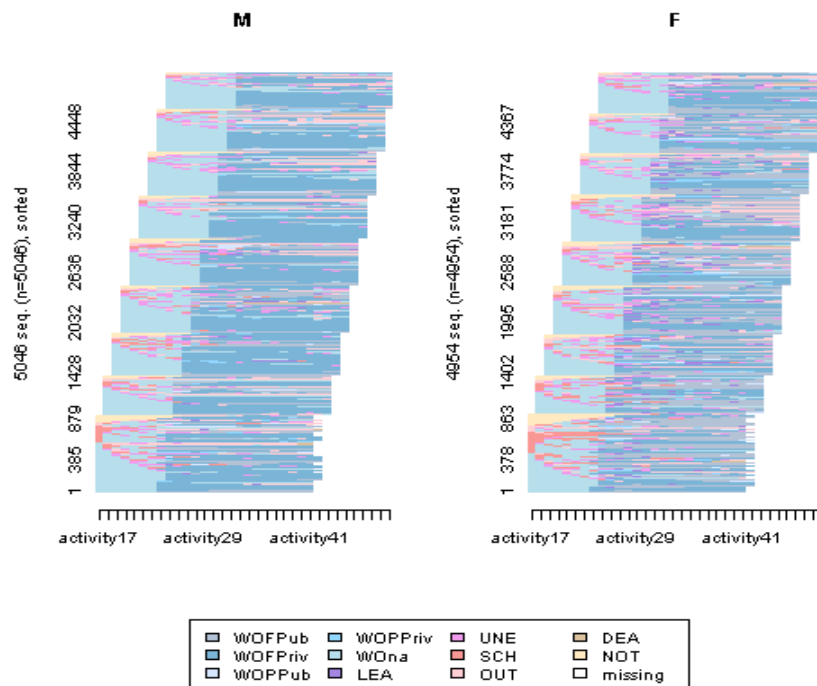

Fig. 6: Labour market sequence full-sequence index plots for the 1951-1960 birth cohort.

17
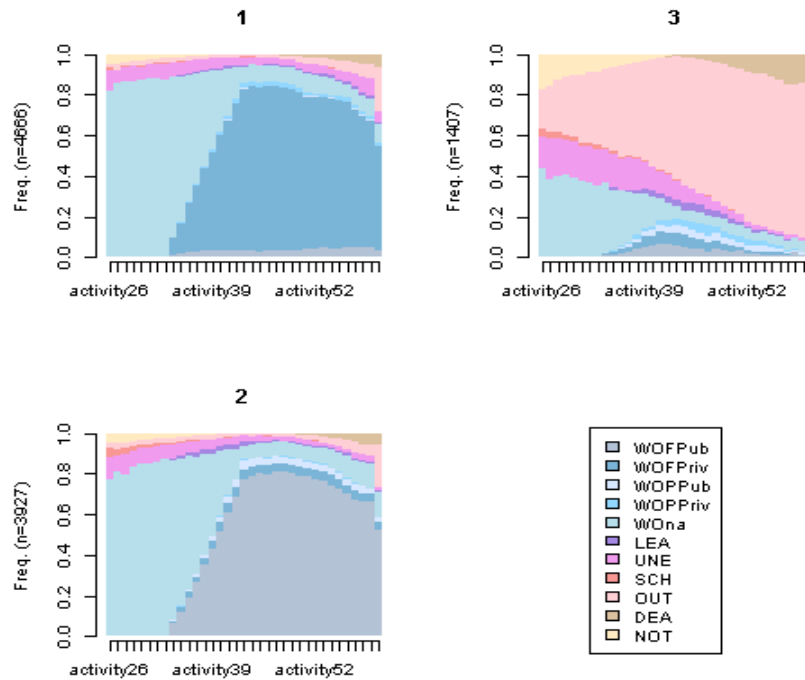
Fig.7: State distribution plots by Labour market cluster for the 1951-1960 birth cohort.

4.5 The 1951-1960 Birth Cohorts and Household Composition

Between their late twenties and early fifties those born between 1951 and 1960 have lived through a great variety of family arrangements (figure 8).. There are 8726 unique sequences, and the 20 most common cover only 7 per cent of sequences for men and 4 per cent for women. While most of the sequences are dominated by living with a partner, there are periods of singlehood both at the beginning and the end of the trajectories.

The mean number of years spent in each state reveals the diversity of states. None average above seven years yet it is distributed among having had a differing number of children inside a partnership. The observable gendered differences are that men have been somewhat more likely to live with their parents while among women there is greater prevalence of single parenthood.

The clustering of the family trajectories is done based on the same logic as that of the oldest birth cohort. Cluster 1 separates the trajectories dominated by singlehood (20 per cent of the sample), and cluster 2 to 4 sort people according to the number of children in the household during most of the time span observed (figure 9). Cluster 2 gathers the childless and those with one child (26 per cent). Cluster 3 gathers those who have mostly had two children (37 per cent), and Cluster 4 is for partnerships that have been living with three or more minors under their roof (16 per cent).

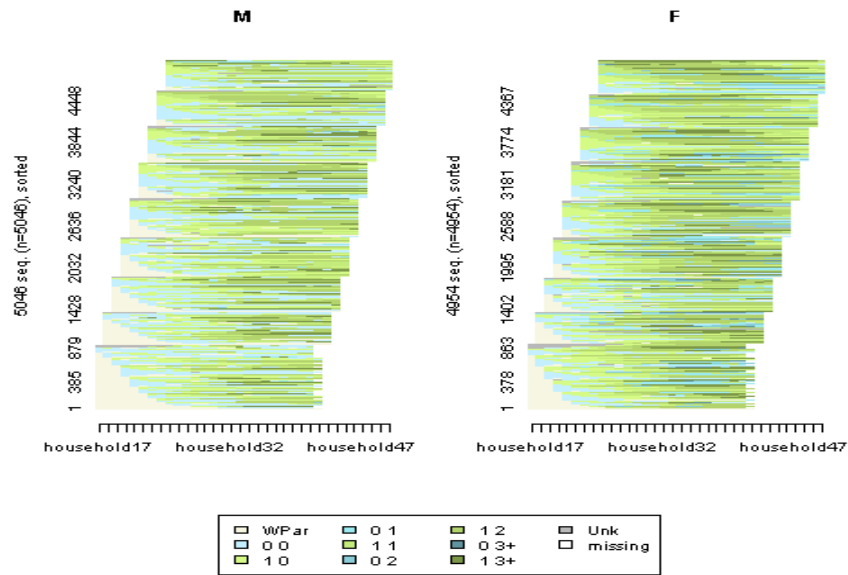**Fig. 8:** Household sequence full-sequence index plots for the 1951-1960 birth cohort.

Fig.9: State distribution plots by Household cluster for the 1951-1960 birth cohort.

4.6 The 1951-1960 Birth Cohorts' Labour Market Activity and Households

When it comes to the burden of care work, most of the people during this period have had a household with two or more minors (See Table 3). 36 per cent of men belong to the cluster 3, characterized by households that have had two children, and additional 16 per cent belong to the cluster 4, implying having raised 3 or more children. For women the respective percentages are 38 and 17. Yet Household clusters seem to be loosely connected with the Labour market cluster membership. Across all Household clusters, the most common Labour market cluster for men is the one dominated by full-time employment in the private sector (ranging from 47 to 66 per cent) and the most common Labour market cluster for women is the one dominated by full-time public sector work (40 to 57 per cent of all women in the sample). It is true, though, that public-sector cluster membership is more common among women with the most care burden, and so it is also among men. 30 per cent of the Household cluster 4 members form part of the Public-sector Labour market cluster.

Table 3: A contingency table between Labour market and Household clusters for the 1951-1960 birth cohort.

	Household clusters				
	C1	C2	C3	C4	
	Single	Union 1	Union 2	Union 3	Total
	Men	Men	Men	Men	Men
	(25)	(23)	(36)	(16)	(100)
C1 Private	47	67	66	63	61
C2 Public	22	26	30	30	27
C3 Absence-Intermittence	30	7	4	7	12
	100	100	100	100	100

	C1	C2	C3	C4	
	Single	Union 1	Union 2	Union 3	Total
	Women	Women	Women	Women	Women
	(16)	(30)	(38)	(17)	(100)
C1 Private	24	35	37	26	33
C2 Public	40	50	55	57	51
C3 Retired, marginal	36	15	7	18	16
	100	100	100	100	100

Items may not sum exact 100 due to rounding.

4.7 The 1961-1970 Birth Cohorts and Labour Market

The Labour market sequences for those born between 1961 and 1970 (observed between 17 and 50 years) reveal entrance in the labour market and strong attachment to it (figure 10). Yet there is a lot of variance (8089 unique sequences, the twenty most common covering 19 per cent of men and 9 per cent of women).

Analysis of the most common sequences show that for both men and women the most common has been to work full time in the private sector. Public sector work has been, however, more common among women than men also during the observed years of this birth cohort. The so far marginal categories of education, unemployment of leaves appear for the first time in the most common trajectories. Yet leaves – most probably maternity – appears only for women and unemployment only for men.

Nevertheless, the mean amount of years spent in each state show that men and women have spent on average the same years in unemployment while women have had more time on leaves. The gendered differences by sector show that women on average have spent only slightly more years in private sector than in the public while men have spent significantly more in the private sector.

The clustering logic for the Labour market sequences is the same as with the previous birth cohorts (See figure 11). Cluster 1 includes trajectories dominated by private sector work and covers 57 per cent of the sample. Cluster 2 is driven by trajectories dominated by public sector work (32 per cent). And Cluster 3 – of trajectories marked by part-time work, unemployment and absence – is of 11 per cent of the sample.

Fig. 10: Labour market sequence full-sequence index plots for the 1961-1970 birth cohort.
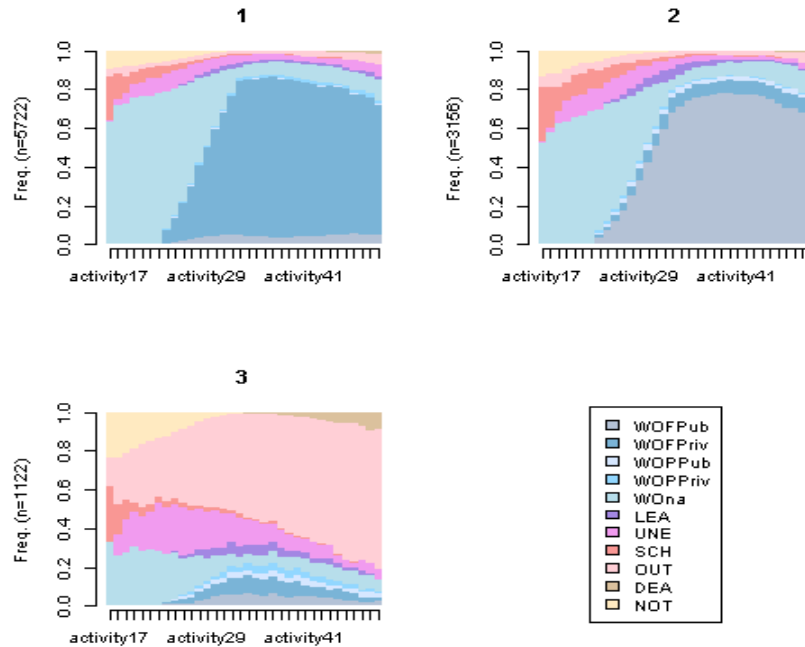
Fig. 11: State distribution plots by Labour market cluster for the 1961-1970 birth cohort.

4.8 The 1961-1970 Birth Cohorts and Household Composition

The 1961-1970 birth cohorts are observed during the process of emancipation and family formation (ages 17-50) (figure 12). Most start out from the household of their parents, and, following a spell of singlehood, form unions, have children, and see even their own children leaving the nest. The sequences observed are very heterogeneous. There are 9553 unique sequences, and the twenty most common sequences cover only 3 per cent for men and 2 per cent of the women in sample.

The diversity of household arrangements for this birth cohort is clear when analyzing the number of mean years spent in each state. On average, men have spent most of the time being single, followed by being in a partnership and raising a child. The state that the average woman has spent most years is being in a partnership and raising a child. Most significant gendered differences are observed when it comes to emancipation and single parenthood. On average, men have spent more time living with their parents while women have spent more time being a single parent.

The clustering of the family trajectories follows the logic of the family trajectories of the birth cohorts described above (Figure 13). Cluster 1 encompasses trajectories dominated by singlehood and, differently from the previous ones, late emancipation from the parental home (30 per cent of the sample). Cluster 2 gathers trajectories dominated by having one or

two children in a partnered household (47 per cent). Cluster 3 covers households that have had three or more children (23 per cent).

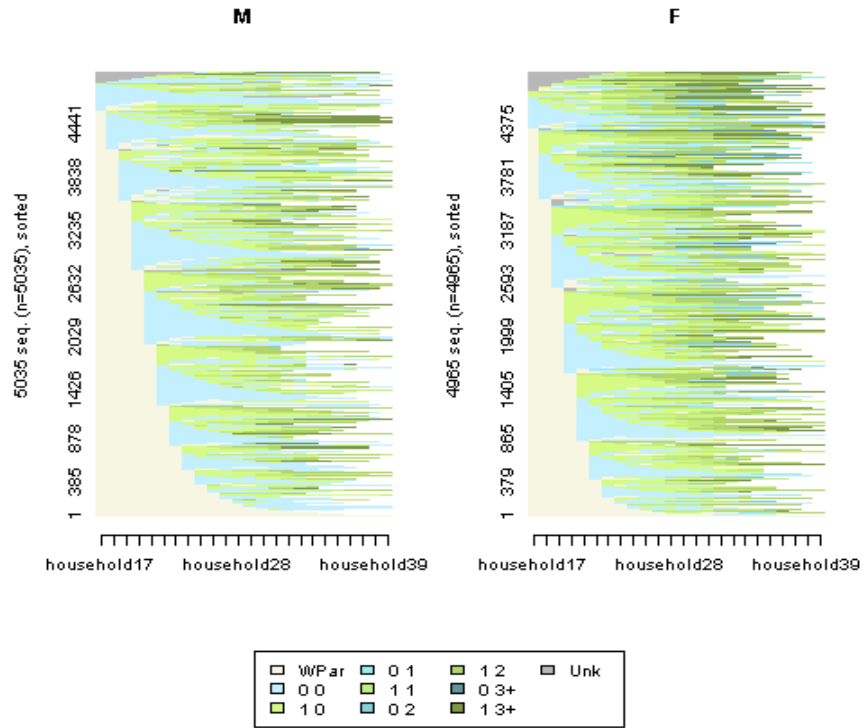**Fig. 12:** Household sequence full-sequence index plots for the 1961-1970 birth cohort.
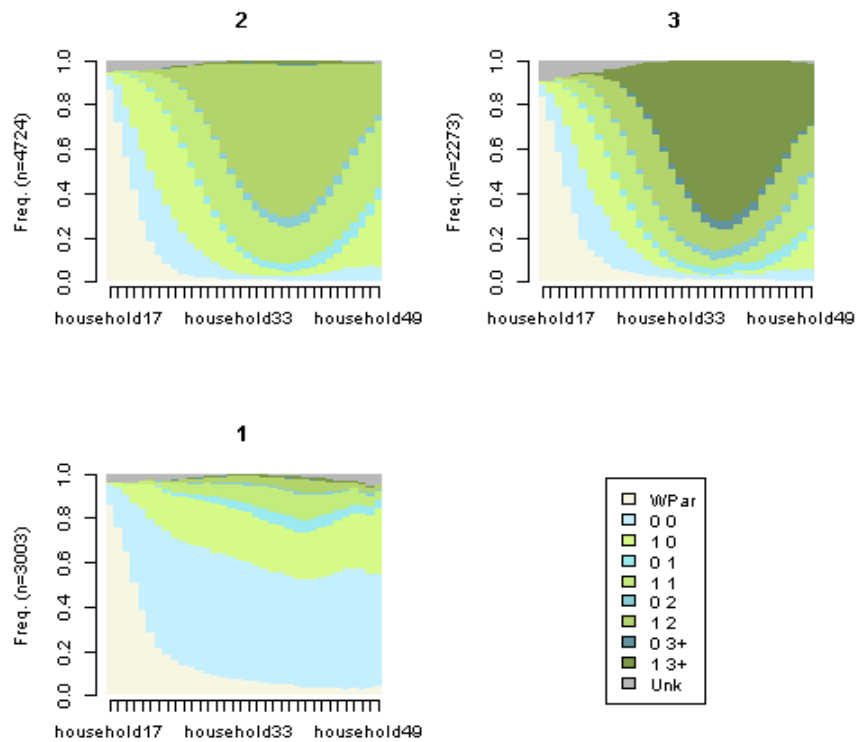
Fig. 13: State distribution plots by Household cluster for the 1961-1970 birth cohort.

4.9 The 1961-1970 Birth Cohorts' Labour Market Activity and Households

In this birth cohort the care burden of women does differentiate the Labour market cluster membership (See Table 4). Among all women the most common (46 per cent) is Labour market Public sector cluster membership, followed by the Private sector cluster membership (42 per cent). When analyzed by the care burden, only those with the least care burden (Single Household cluster) are more likely to find themselves in the Private sector Labour market cluster (44 per cent), followed by the Public sector cluster membership (36 per cent). This dynamic is also observed on the other end of the care burden spectrum. Most women (52 per cent) that have raised three or more children form part of the Public sector Labour market cluster.

Among men Private sector Labour market cluster membership is the case for most of the sample, going up to 80 per cent for up to two children households. Absence-intermittence Labour market cluster membership is negligible for men with care burden.

Table 4: A contingency table between Labour market and Household clusters for the 1961-1970 birth cohort.

		Household clusters			
		C1 Single Men (39)	C2 Union 1 Men (41)	C3 Union 2 Men (20)	Total Men (100)
Labour market clusters	C1 Private	62	80	77	72
	C2 Public	19	17	17	18
	C3 Absence-Intermittence	20	2	7	10
		100	100	100	100

		C1 Single Women (21)	C2 Union 1 Women (54)	C3 Union 2 Women (25)	Total Women (100)
Labour market clusters	C1 Private	44	46	32	42
	C2 Public	36	47	52	46
	C3 Absence-Intermittence	21	8	15	13
		100	100	100	100

Items may not sum exact 100 due to rounding.

4.10 The 1971-1980 Birth Cohorts and Labour Market

The youngest cohort (1971-1980) – observed between the ages of 17 and 40 – are seen during their formative years and entrance in the labour market. As seen in figure 14, their

trajectories start by either work or being in education and then develop into attachment to the labour market.

The twenty most common sequences (out of 9083 unique sequences) cover 9 per cent of men and 3 per cent of women. The most typical ones for both men and women is continuous work in the private sector. Nevertheless, due to the fact that for most of this birth cohort data on full or part time work is available from the very beginning of the trajectory, we can appreciate the importance of part-time work in the private sector as a strategy to enter in the labour market. Also education appears in the most typical sequences and – only for women – full time work in public sector.

Only in this birth cohort the mean number of years spent working full-time in the private sector during the observed period is higher than work in the public sector for both men and women. Yet the difference between the two genders is striking, because for women, as observed above, work in the public sector is still much more common.

Also for the youngest birth cohort the Labour market cluster solution is one of three clusters (See Figure 15). Private sector cluster (57 per cent of the sample) gathers trajectories dominated by full- and part-time work in the private sector. Public sector cluster (37 per cent of the sample) covers trajectories of full-time public sector works, part-time work in both sectors and somewhat more education. The Absence-intermittence cluster (11 per cent) is dominated by absence from the labour market and unemployment while none of the paid work modalities is salient.
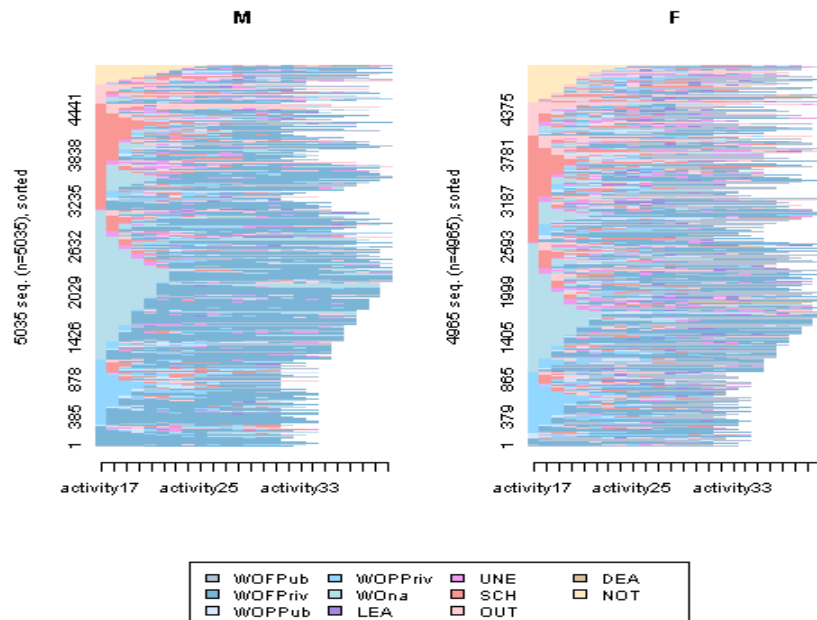
Fig. 14: Labour market sequence full-sequence index plots for the 1971-1980 birth cohort.

25
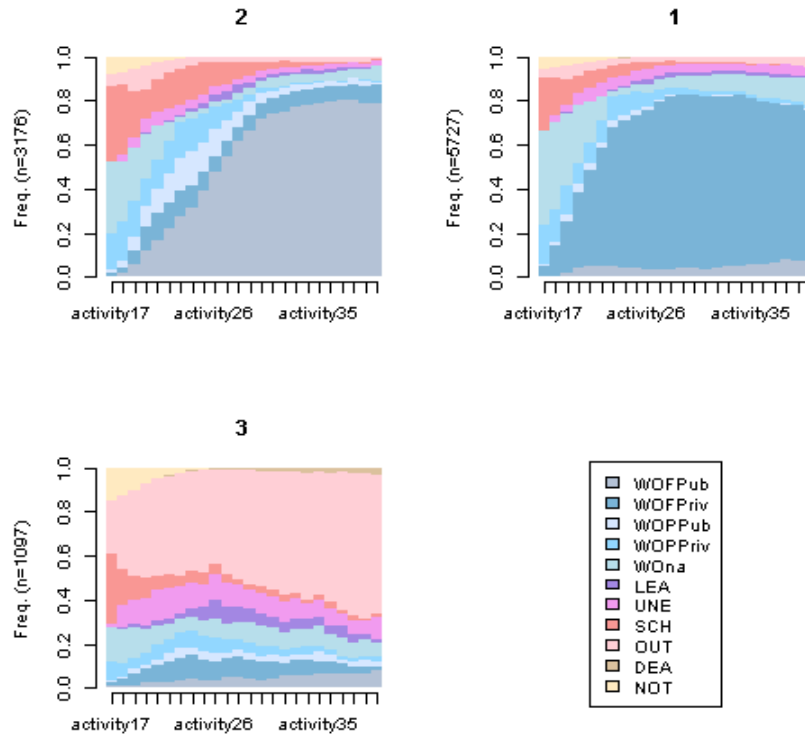
Fig. 15: State distribution plots by Labour market cluster for the 1971-1980 birth cohort.

4.11 The 1971-1980 Birth Cohorts and Household Composition

When it comes to household composition sequences, the 1971-1980 birth cohorts so far have emancipated from the parental households and, mostly after spells of singlehood, have formed unions and had children (figure 16). This transition has not been uniform, though. There are 9328 distinct sequences and the twenty most common cover 3 per cent of male trajectories and 2 per cent of female trajectories.

This diversity is reflected by the mean years spent in each state. Only being single for men surpass an average of five years. For men, the following largest on average spells have been spent at the parental home and in union without children. For women, the three states with largest on average spells are the same but in the following order: being single, in union with no children, and in the parental home.

As with the other birth cohorts, the clustering of household sequences separates singles and those in union (figure 17). Here the Singlehood cluster is dominated by trajectories with long spells of singlehood after emancipation and late, if any, union formation. It covers 36 per cent of the sample. The Early unions cluster gathers individuals whose emancipation from the parental home was very soon followed by union formation and childbearing (64 per cent of the sample).

Fig. 16: Household sequence full-sequence index plots for the 1971-1980 birth cohort.
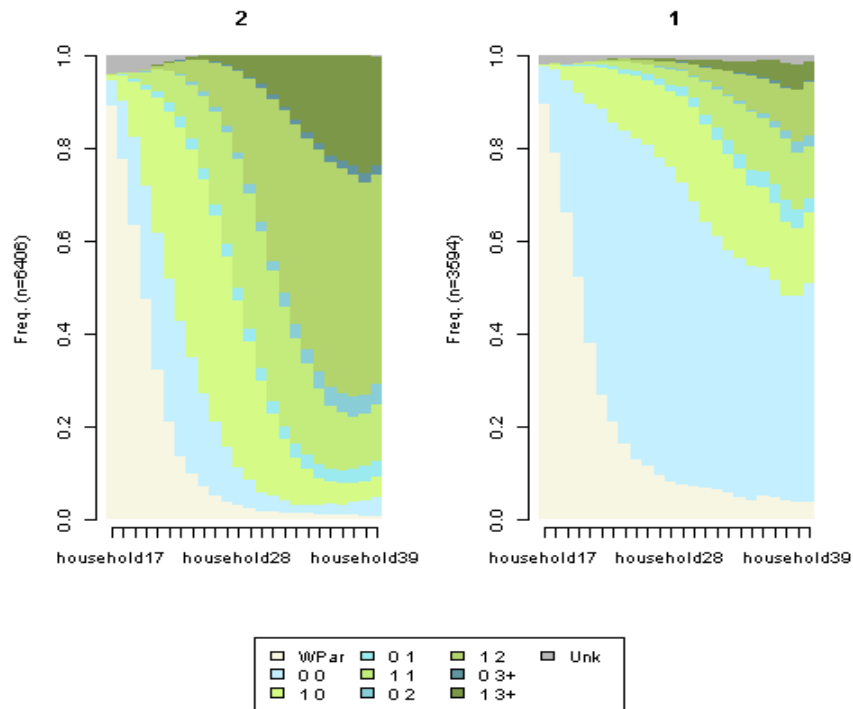
Fig. 17: State distribution plots by Household cluster for the 1971-1980 birth cohort.

4.12 The 1971-1980 Birth Cohorts' Labour Market Activities and Households

Among the youngest birth cohorts there are no big differences in Labour market cluster membership patterns between those entering in unions early (Household cluster 2) or delaying them (Household cluster 1) (Table 5). Most men belong to the Private sector Labour market cluster, although more so those that have established early unions than those who have enjoyed a longer spell of singlehood (79 and 64 per cent respectively). Women are divided between the Private sector and Public sector Labour market clusters, with slightly more being members of the Labour market cluster dominated by work in the public sector than in the private (43 vs. 39 per cent of the Singles and 46 vs. 43 per cent of the Early unions). As with the other birth cohorts, the Absence-intermittence cluster membership is correlated with being in the Household cluster dominated by singlehood.

Table 5: A contingency table between Labour market and Household clusters for the 1971-1980 birth cohort.

			Household clusters			
	C1 Single Men (46)	C2 Union Men (54)	Total Men (100)	C1 Single Women (26)	C2 Union Women (74)	Total Women (100)
C1 Private	64	79	73	39	43	42
C2 Public	19	17	18	43	46	45
C3 Absence-Intermittence	17	3	9	18	11	13
	100	100	100	100	100	100

Items may not sum exact 100 due to rounding.

4.13 Regression Outcomes for the 1941-1950 Birth Cohorts

Among those in the Single Household cluster, men of all levels of education are more likely than women to be in the Private sector Labour market cluster (figure 18). Among those in the Union Household cluster, men with only secondary education or less are the ones most likely to find themselves in the Private sector Labour market cluster, followed by men with post-secondary education and then men with tertiary education. Women with post-secondary and tertiary education who are in union are the ones least likely to be in the Private sector Labour market cluster. This likelihood is not significantly different between

these two groups of women, though it is significantly smaller than that of women with less education.
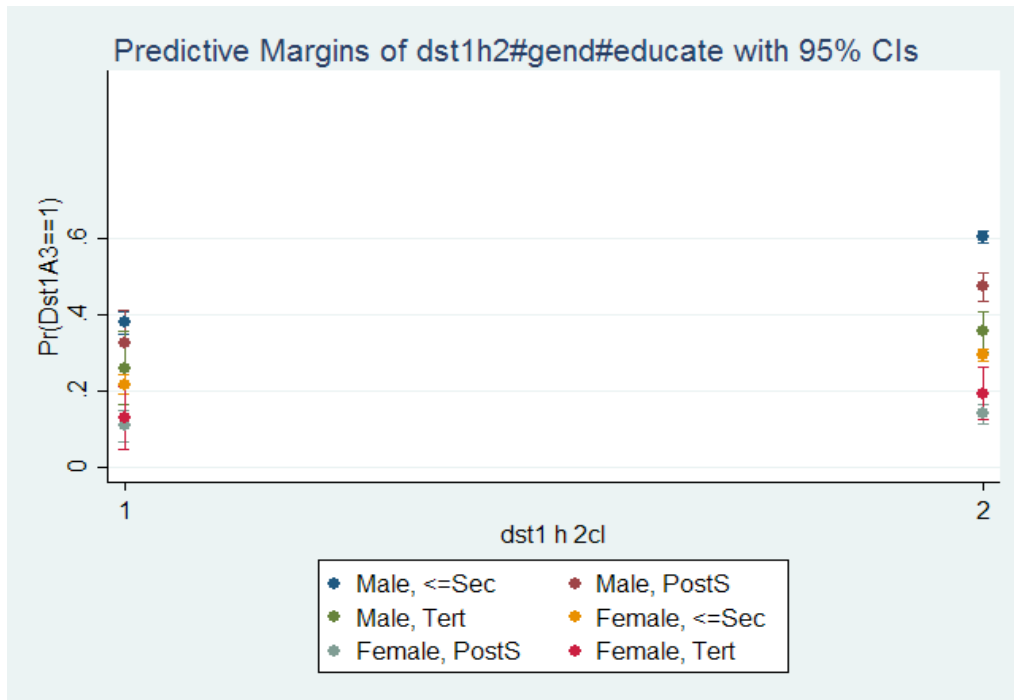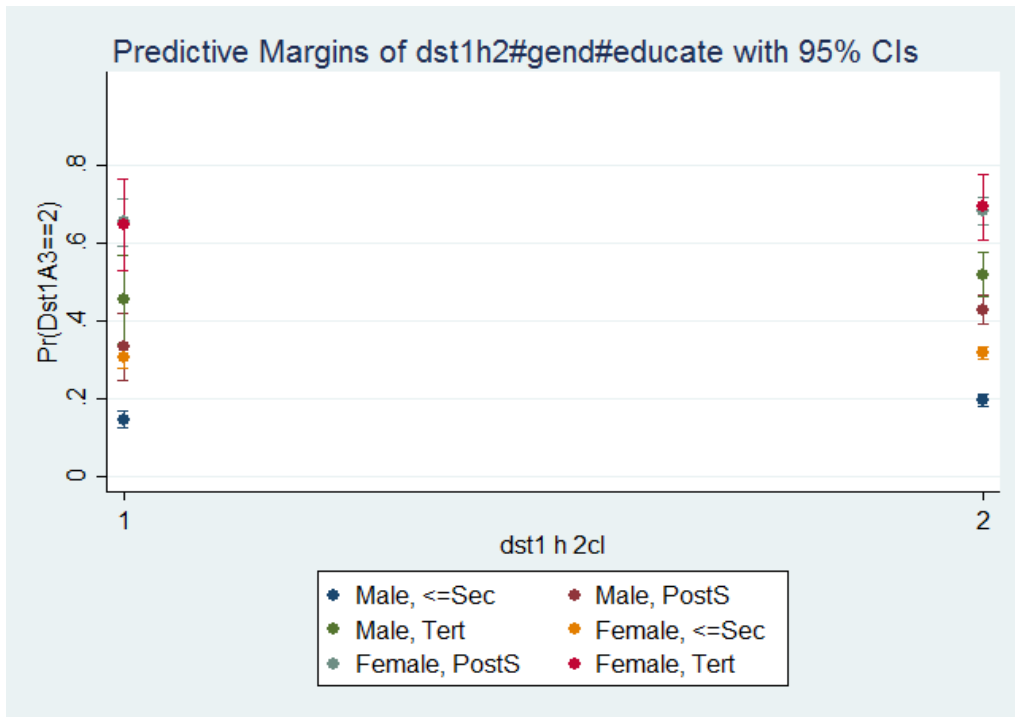

Fig. 18: Predictive margins of the Household cluster, gender and education level interaction for the Private sector Labour market cluster, birth cohort 1941-1950.
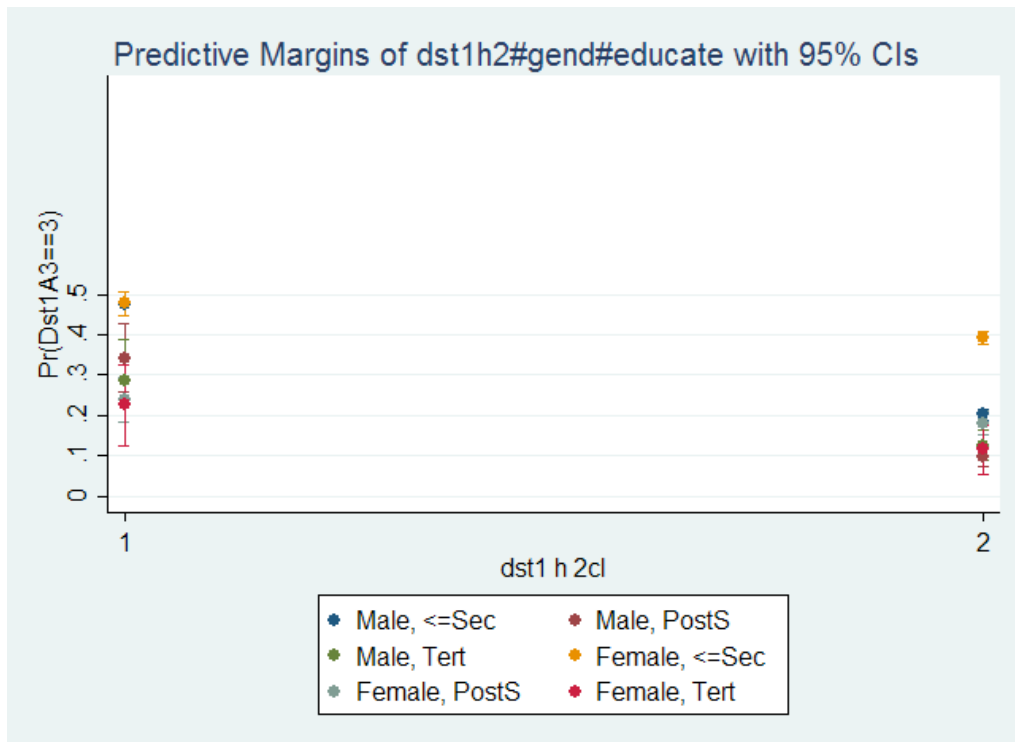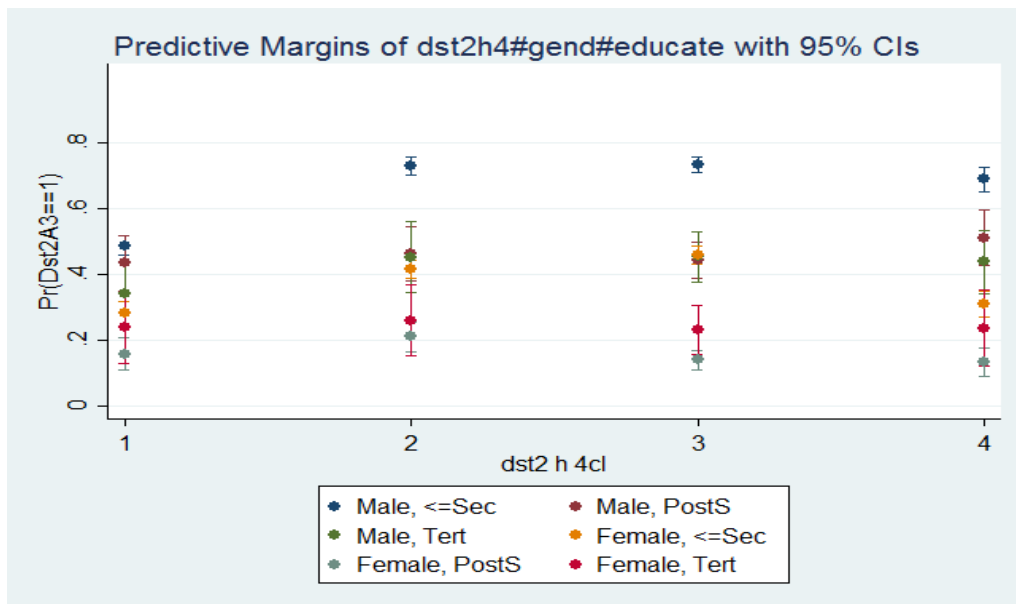
The "profile" that is the most likely to find themselves in the Public sector Labour market cluster almost mirrors that of the Private sector Labour market cluster. Those most likely to be in the Public sector cluster are women with post-secondary and tertiary education, both single and in union (figure 19). The difference is that they are followed by men with post-secondary and tertiary education. Those least likely to be in this Labour market cluster are men and women with least education, both single and in unions.

Altogether those in the Single Household cluster are more likely to form part of the Retired-Absence Labour market cluster (figure 20). Among the singles, it's men and women with the least education that are the ones most likely to find themselves in this Labour market cluster. In the Union Household cluster, it's only the least educated women who are the ones clearly most likely to be in the Retirement-Absence cluster. There are no significant differences among the rest of gender/education groups.
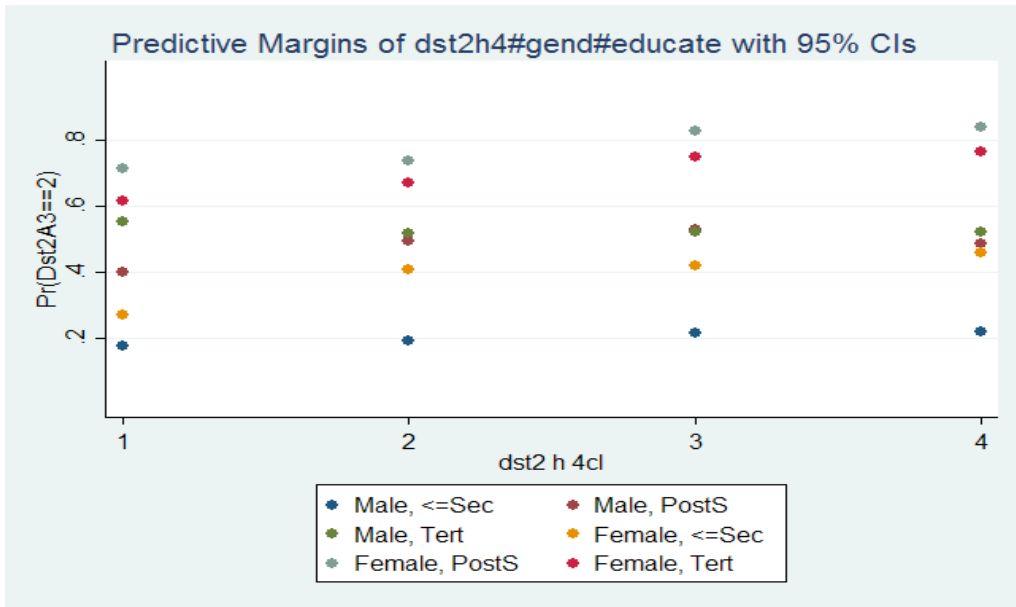
Fig. 19: Predictive margins of the Household cluster, gender and education level interaction for the Public sector Labour market cluster, birth cohort 1941-1950.
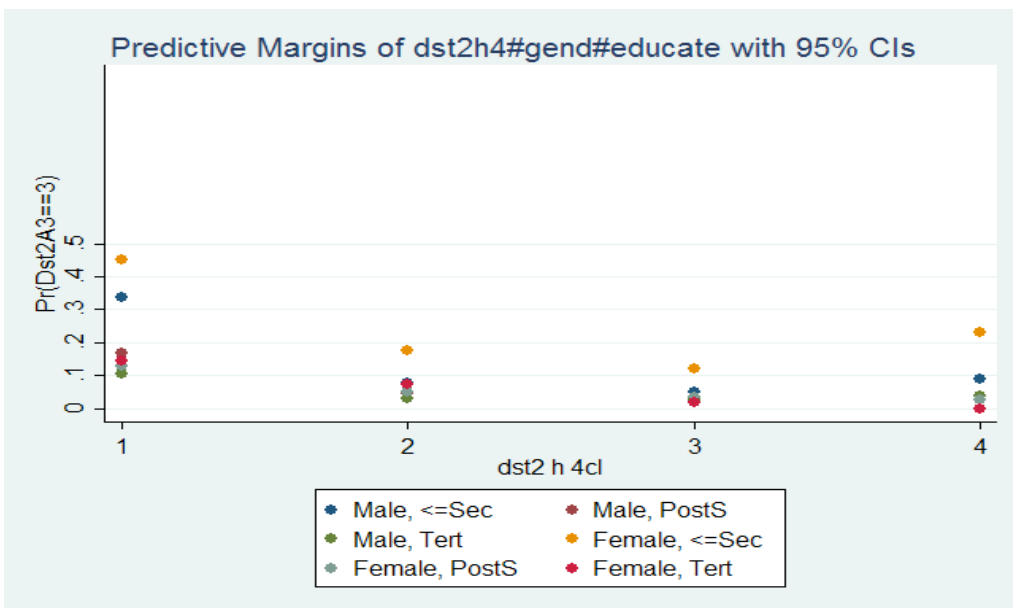

Fig. 20: Predictive margins of the Household cluster, gender and education level interaction for the Absence-Intermittence Labour market cluster, birth cohort 1941-1950

30

4.14 Regression Outcomes for the 1951-1960 Birth Cohorts

The characteristics that make one likely to be in the Private sector Labour market cluster are similar to that of the previous birth cohort (figure 21). There is some variance, though, in the correlation between the unions and care burden experienced during the observed years. The differences are insignificantly small among the Single Household cluster members. However, in the Union clusters men with least education are the ones most likely to find themselves in the Private sector Labour market cluster. Women with little education have similar likelihood than men with post-secondary and tertiary education. Irrespective of the care burden, women with more education are the ones least likely to form part of the Private sector Labour market cluster.

Fig. 21: Predictive margins of the Household cluster, gender and education level interaction for the Private sector Labour market cluster, birth cohort 1951-1960.

Women with post-secondary and tertiary education are the ones most likely to find themselves in the Public sector Labour market cluster, especially so the ones with post-secondary education (figure 22). And here increase in care burden at home raises the likelihood of membership in this cluster. This dynamic is also true for women with less education, but the probabilities are significantly lower, lower than those of post-secondary and tertiary educated men. The gender/education group least likely to be part of the Public sector Labour market cluster is men with secondary education or less. However, for men there is no clear pattern linking care burden and Labour market cluster.

The pattern for the likely Retirement-Absence Labour market cluster membership in these cohorts is very similar – if only clearer – to that of the previous one (figure 23). Women with little education across the different household structures are the ones most

likely to find themselves in this cluster. However, it is especially the single ones – as are single little-educated men and the rest of the gender/education groups, but to a much smaller extent – that are the ones most likely to be in the Retirement-Absence Labour market cluster.
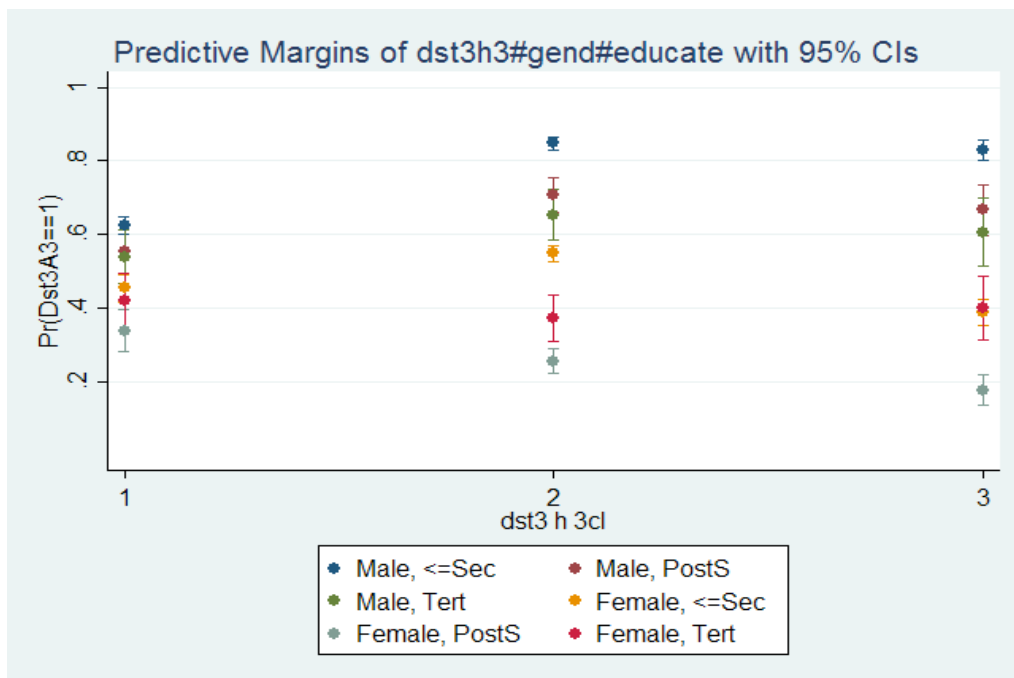

Fig. 22: Predictive margins of the Household cluster, gender and education level interaction for the Public sector Labour market cluster, birth cohort 1951-1960.

Fig. 23: Predictive margins of the Household cluster, gender and education level interaction for the Absence-Intermittence Labour market cluster, birth cohort 1951-1960.

4.15 Regression Outcomes for the 1961-1970 Birth Cohorts

The predicted probabilities for the Private sector Labour market cluster membership in the 1961-1970 birth cohorts reveal differential patterns by care burden at the household (figure 24). While ordering of the Single Household cluster members according to gender and education is rather unclear due to overlapping confidence intervals, there is a clear dispersion of probabilities for those in unions. Increased care burden clearly diminishes the likelihood that a woman with post-secondary credentials finds herself in the Private sector Labour market cluster. For other women the pattern is not that clear, while for men it's the reverse. Men of all educational levels, and especially the less educated, are more likely to be in this cluster when they have a partner and children.

As observed in the other cohorts, increased care burden for women increase the likelihood of them being in the Public sector Labour market cluster (figure 25). These increased probabilities are the most clear for women with post-secondary education or less while women with post-secondary education or more are the ones most likely to find themselves in this cluster across the Household clusters. Men with Secondary education or less are the ones least likely to be in this cluster. Among men there is little correlation between Household care burden and Labour market clusters.

As with the previous cohorts, the least educated women are the ones most likely to find themselves in the Retirement-Absence Labour market cluster (figure 26). In case of singles, it's also the case for the least educated men.

Fig. 24: Predictive margins of the Household cluster, gender and education level interaction for the Private sector Labour market cluster, birth cohort 1961-1970.
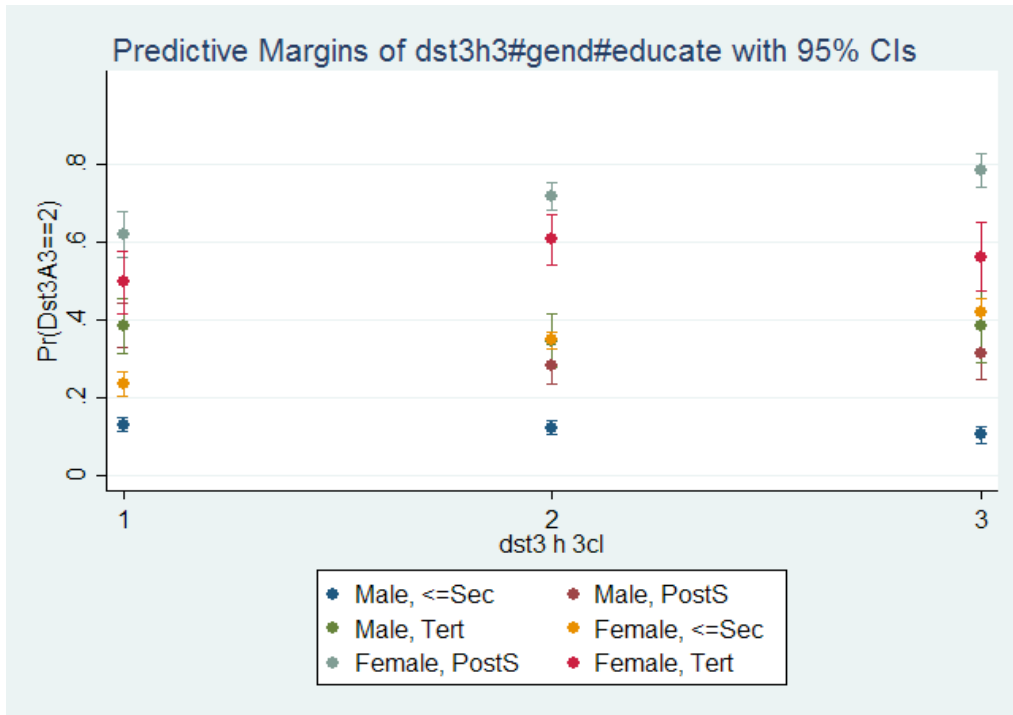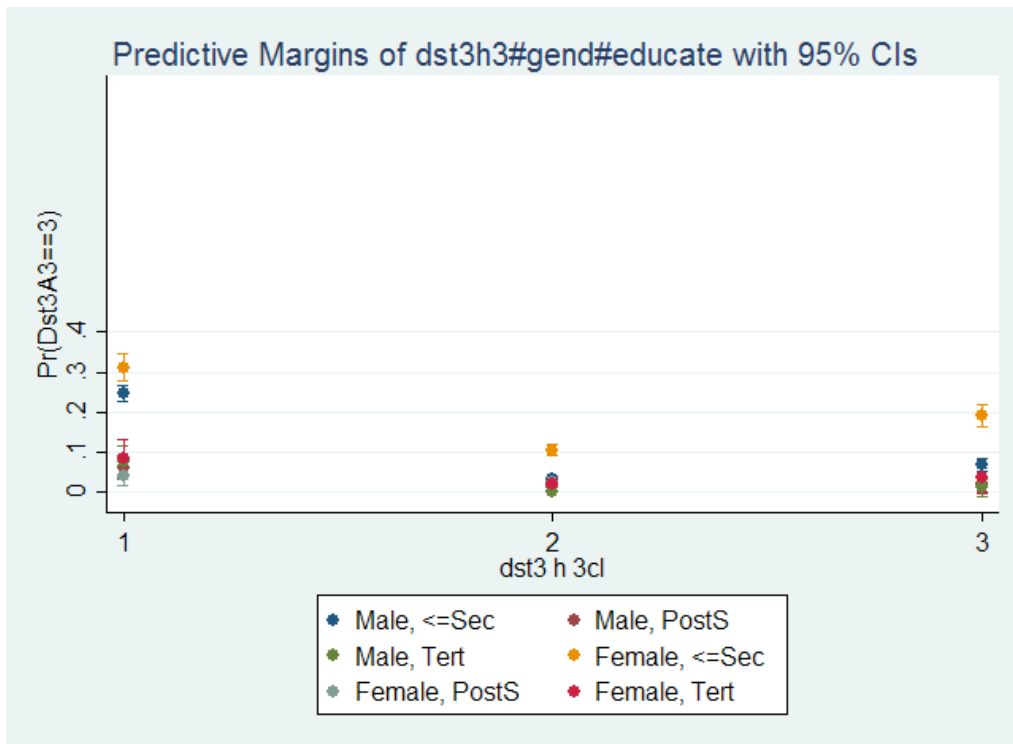
33

Fig. 25: Predictive margins of the Household cluster, gender and education level interaction for the Public Sector Labour market cluster, birth cohort 1961-1970.

Fig. 26: Predictive margins of the Household cluster, gender and education level interaction for the Absence-Intermittence Labour market cluster, birth cohort 1961-1970.

4.16 Regression Outcomes for the 1971-1980 Birth Cohorts

For the youngest cohort, there are few differences between the two Household clusters when it comes to likelihood of Labour market cluster membership. For the Private sector Labour market cluster, early entrance in union disperses the likelihood of membership in this Labour market cluster (figure 27). Partnered low educated men are more likely to be in the Private sector cluster than their single counterparts. And partnered women with post-secondary or tertiary education – especially the ones with post-secondary credentials - are clearly less likely to be in this cluster than less educated women or men of any educational level.

These women are much more likely to be in the Public sector Labour market cluster, the ones with post-secondary education more than tertiary educated, especially among the ones in unions (figure 28). For women of least education, being in union from early on increases the likelihood of being in this cluster, but it still below that of higher educated men. As for all other cohorts, the least educated men are the ones least likely to form part of this cluster.

Replicating the pattern described above, the least educated women are the ones most likely to find themselves in the Retirement-Absence Labour market cluster (figure 29). In case of singles, it's also the case for the least educated men. For the rest of the gender/education groups, the likelihood of being in this cluster is negligible.
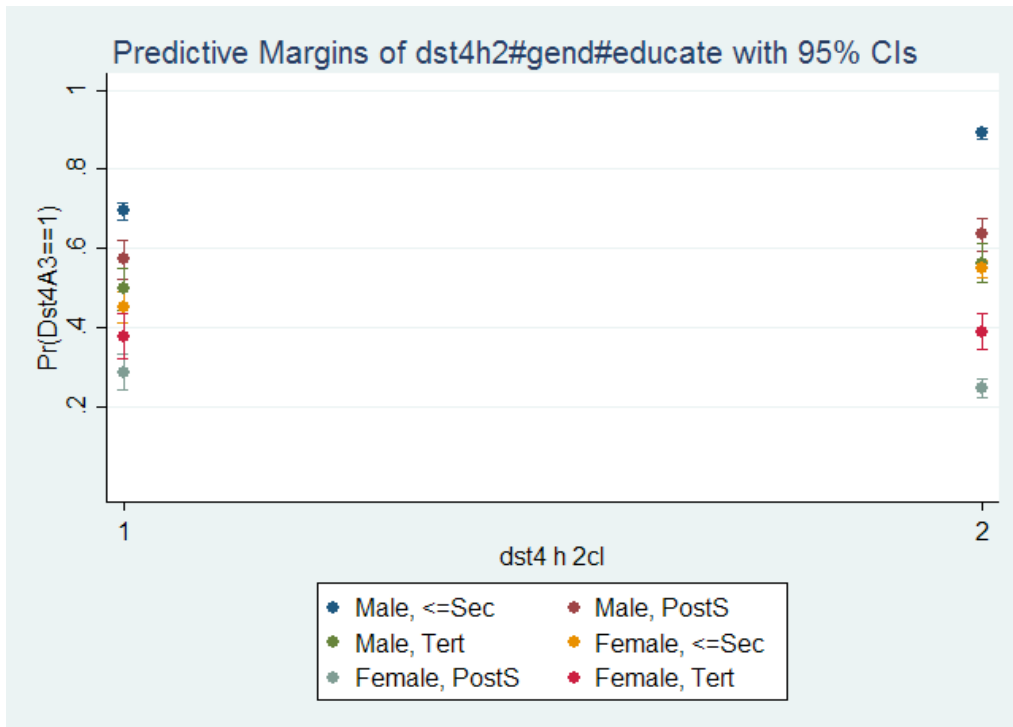

Fig. 27: Predictive margins of the Household cluster, gender and education level interaction for the Private sector Labour market cluster, birth cohort 1971-1980.
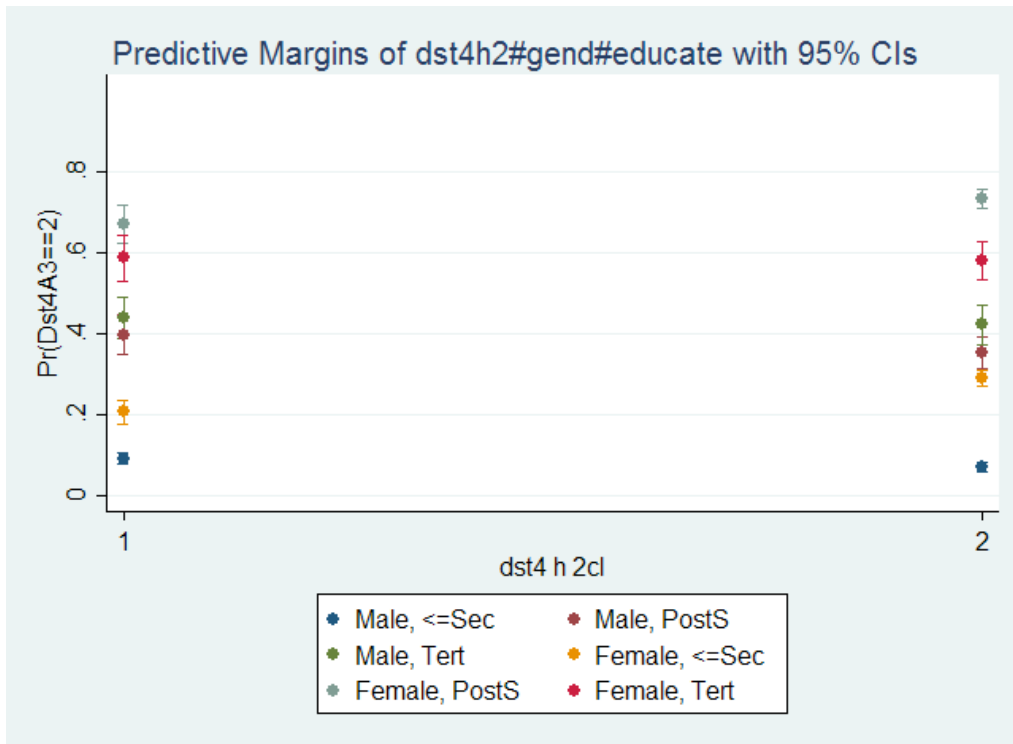
35

Fig. 28: Predictive margins of the Household cluster, gender and education level interaction for the Public sector Labour market cluster, birth cohort 1971-1980.
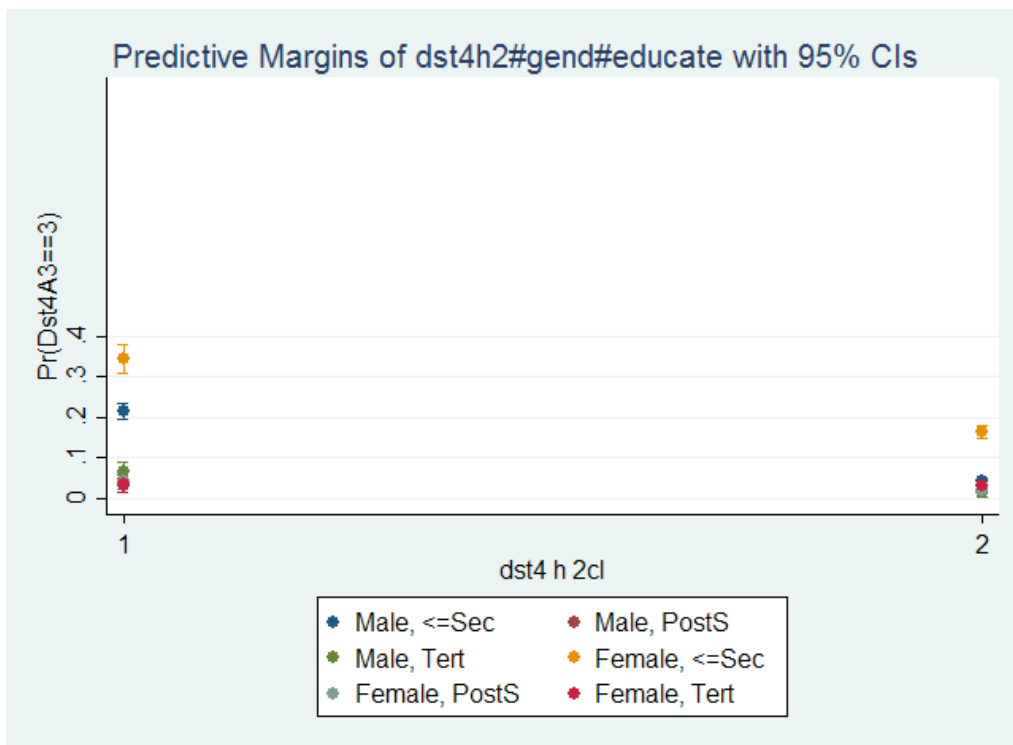

Fig. 29: Predictive margins of the Household cluster, gender and education level interaction for the Absence-Intermittence Labour market cluster, birth cohort 1971-1980.

36

5 Discussion

According to our Hypothesis 1, we expected to find evidence of both gendered labour market participation patterns and gradual dissipation of those in the most recent birth cohorts. We do find gender to be of great importance in predicting the likelihood of being in one Labour market cluster or other. However, nothing in the analysis we have done suggests disappearance of the gendered patterns captured in this paper. For all four birth cohorts women are much more likely to be in the Labour market clusters dominated by public sector work and absence/loose attachment/early retirement from the labour market. We observe that gender interacts with other variables, namely, the level of education achieved and the composition of the household. Hypothesis 1 is partly confirmed: there are important gender differences, but no sign of closing the gendered sectorial divide.

Our Hypothesis 2 predicted that higher education would increase the likelihood of more men-typical trajectories for women. In the operationalized terms of our research it would mean higher likelihood of being in the Private sector Labour market cluster. While it is true that the post-secondary educated women are the ones least likely to form part of the Private sector Labour market cluster, under none of the conditions that we have specified tertiary educated women are more likely to form part of that Labour market cluster than women with secondary education or less. Hence Hypothesis 2 is rejected while admitting that there is a non-linear pattern of linking Private sector paid work with women's education, namely that the trajectories of women with post-secondary education are the most different ones from male trajectories.

Hypothesis 3 was formulated to capture the impact of the care burden of the family life. We expected to find that having been in a partnership and having raised more children during the observed years would decrease the likelihood of women to have a male-typical career, i.e. form part of the Private sector Labour market cluster. However, across the four birth cohorts we find no evidence that the least care burden – singlehood – increases the likelihood of being in the Private sector Labour market cluster or that having raised more children increases the likelihood of Absence-intermittence Labour market cluster membership. It is true, though, that Public sector Labour market cluster seem to be the refuge for women with higher care burden, especially among the birth cohorts of 1951-1960 and 1961-1970.

Only among the women with least education (and in the three older cohorts only) we actually observe that partnership and children tend to increase the likelihood of being in the Private sector Labour market cluster.

Only in one instance we find such dynamic for men. Among those born in 1951-1960, increased care burden does raise the likelihood of Public sector Labour market cluster membership for men with post-secondary education. For the rest of cohort/education groups of men a pattern of increased likelihood of Private sector Labour market cluster membership is observed. Hence Hypothesis 3 is partly confirmed: women's Labour market

cluster membership is more impacted by the household care burden than men's yet there is no sign of a generational change of this pattern across the cohorts observed.

6 Conclusions and future research ideas

These findings allow us to conclude that the four cohorts we are observing have gone through similar dynamics during this time window. Women with post-secondary or higher education credentials tend to reconciliate their care burden with paid work via public sector employment while for men unions and children have been a correlate for more private sector work. Both men and women with secondary education or less rely heavily on the private sector work that offers more opportunities for people for scarce educational credentials.

Despite the methodological limitations for overarching conclusions, the analysis carried out confirms the partiality of the apparent gender egalitarianism of the Danish labour market. While the strong labour-market attachment of Danish women is very clear, so is the fact that at least so far they have tended to build their careers in the more sheltered public sector. We find that more education – especially the post-secondary education that is not yet tertiary – make it more likely for them to find a position in this more family-friendly part of the economy. The most important finding is that across the years we observe all four cohorts follow the same pattern of clustering around the same three modes of relating with the labor market. Curiously enough, only one of the two key stratifiers of the work regime introduced in the analysis (the part-time/full-time division and public sector/private sector division) drives the clustering. While the distribution of part-time work is clearly gendered, it is not a salient part of the trajectories except for the youngest cohort where it is common to use part-time to reconciliate work with studies. The preliminary conclusion with the data available is that, while a significant part of women are working part-time at any given moment in time, it serves as a transitory phase and women do not tend to stay in it. Our analysis so far permits us to conclude that gender convergence in the Danish labor market is at the point where both men and women work full-time for most of their lives and the major stratifier by gender is the sector of the economy. At least so far this difference does not show clear signs of disappearing with the most recent cohorts.

These data offer a rich material for further exploration and analysis. At the moment we are considering several options. One would be more detailed analysis of the characteristics of the gender-outliers, i.e. women that pursue a private-sector career, men that opt for public-sector work, and the extent to which the two profiles are similar.

Onother path could be a reduced research design option that would allow to disentangle cohort and period effects would be to analyze only (a) trajectories between ages 25-34 comparing cohorts 1961-1977, (b) trajectories between ages 35-44 comparing cohorts 1951-1967, and (c) trajectories between ages 45-54 comparing cohorts 1941-1957.

Nevertheless, that would imply working with significantly reduced sequences, passing from the minimum sequence length of 15 years and maximum of 26 to a uniform trajectory length of 10 years. This design would also allow to use the transversal entropy (Billari 2001, Widmer and Ritschard 2009) as an alternative dependent variable able to measure the dissolving of the male breadwinner pattern into a range of diverse arrangements of reconciling paid work and family.

An alternative clustering of Household trajectories, forming them around theoretically reasoned ideal-typical paths (Elzinga and Liefbroer 2007) could enrich our understanding of household composition. Such approach would permit a clearer interpretation of the Household cluster membership and probably a more intuitive division too.

A way to add precision to the Labour market sequence could be a multichannel sequence analysis design that would add income to the Labour market-intensity sequence (Gauthier et al 2010). There is little doubt that income potential is a powerful stratifier of labour market trajectories (Henz and Sundstrom 2001), hence there are probably additional layer of complexity – except for the partial effect that we are capturing via education credentials – which we are not currently taking into account.

Another design change could be one that analyzes together the parallel trajectories of couples. An analysis of the intertwined trajectories of the unions would allow a more precise assessment of the complex interdependencies of the three careers of each couple (Han and Moen 2001), their "linked lives" (Elder): the two individual ones and the common (Hertz 1986, Reichart et al 2007, Gerson 2011). This is a feat that few data sets allow (Han and Moen 2001) to do but the Danish Register data offer an opportunity to fully track all the partners that a person has shared a household in Denmark with (Leth-Sorensen and Rohwer 2001).

7 References

Aassve, Arnstein, Francesco C. Billari, and Raffaella Piccarreta. 2007. "Strings of Adulthood: A Sequence Analysis of Young British Women's Work- Family Trajectories." *European Journal of Population*, 23 (3-4): 369–88.

Abbott, Andrew. 1995. "Sequence Analysis: New Methods for Old Ideas." *Annual Review of Sociology*, 21: 93–113.

Abbott, Andrew, and Angela Tsay. 2000. "Sequence Analysis and Optimal Matching Methods in Sociology: Review and Prospect." *Sociological Methods & Research*, 29 (1): 3–33.

Aisenbrey, Silke and Anette E. Fasang. 2014. "Work-Family Trajectories in Germany and the United States." Paper presented at the 2014 Annual Meeting of the Population Association of America.

Barban, Nicola. 2011. "Family Trajectories and Health: A Life Course Perspective." Milan: Dondena working papers, 39.

Barban, Nicola and Francesco C. Billari. 2012. "Classifying Life Course Trajectories: A Comparison of Latent Class and Sequence Analysis." *Journal of the Royal Statistical Society*, 61 (5): 765–84.

Bernhardt, Eva M. 1993. "Fertility and Employment." *European Sociological Review*, 9(1): 25-42.

Billari, Francesco C. 2001. "Sequence Analysis in Demographic Research." *Canadian Studies in Population*, 28 (2): 439-458.

Bishop, Kate. 2004. "Working Time Patterns in the UK, France, Denmark and Sweden." *Labor Market Trends*, March 2004, 113-122. London: Office for National Statistics.

Blau, Francine D, Marianne A Ferber and Anne. E. Winkler. 1998. *The Economics of Women, Men, and Work*. Upper Saddle River, NJ: Prentice Hall.

Blossfeld, Hans-Peter, and Sonja Drobnič. 2001a. "A Cross National Comparative Approach to Couples' Careers." In Blossfeld, Hans-Peter, and Sonja Drobnič, (eds.). 2001. *Careers of Couples in Contemporary Society: From Male Breadwinner to Dual-Earner Families*. Oxford: Oxford University Press.

Blossfeld, Hans-Peter, and Sonja Drobnič. 2001b. "Theoretical Perspectives on Couples' Careers." In Blossfeld, Hans-Peter, and Sonja Drobnič, (eds.). 2001. *Careers of Couples in Contemporary Society: From Male Breadwinner to Dual-Earner Families*. Oxford: Oxford University Press.

Bonke, Jens (ed.). 1997. *Levevilkår i Danmark: Statistik Oversigt 1997*. Copenhagen: Danmarks Statistik and Socialforksningsinstituttet.

Bracher, Michael, and Gigi Santow. 1998. "Economic Independence and Union Formation in Sweden." *Population Studies*, 52(3): 275-294.

Cha, Youngjoo. 2010. "Reinforcing Separate Spheres: The Effect of Spousal Overwork on the Employment of Men and Women in Dual-Earner Households." *American Sociological Review*, 75(2): 303–329.

Cha, Youngjoo. 2013. "Overwork and the Persistence of Gender Segregation in Occupations." *Gender & Society*, 27(2): 158–184.

Cooke, Lynn Prince. 2014. "Gendered Parenthood Penalties and Premiums across the Earnings Distribution in Australia, the United Kingdom, and the United States." *European Sociological Review*, 30 (3): 360–372.

Coontz, Stephanie. 2011. *A Strange Stirring: The Feminine Mystique and American Women at the Dawn of the 1960's*. New York: Basic Books.

Elder, Glen H. 1994 "Time, human agency, and social change: Perspectives on the life course." *Social Psychology Quarterly*, 1994: 4-15.

Elzinga, Cees H., and Aart C. Liefbroer. 2007. "De-Standardization of Family-Life Trajectories of Young Adults: A Cross-National Comparison Using Sequence Analysis." European Journal of Population 23 (3-4): 225–50.

England, Paula. 2010. "The Gender Revolution: Uneven and Stalled." *Gender & Society*, 24, Vol. 2: 149–166.

England, Paula. 2011. "Reassessing the uneven gender revolution and its slowdown." *Gender & Society*, 25, Vol. 1: 113–123.

Esping-Andersen, Gøsta. 1990 [1991]. *The Three Worlds of Welfare Capitalism*. Cambridge: Polity Press.

Esping-Andersen, Gøsta. 1993. *Changing Classes: Stratification and Mobility in Post-Industrial Societies*. London: Sage.

Esping-Andersen, Gøsta. 1999. *Social Foundations of Postindustrial Economies*. Oxford: Oxford University Press.

Esping-Andersen, Gøsta. 2009 [2013]. *The Incomplete Revolution: Adapting to Women's New Roles*. Cambridge: Polity Press.

Esping-Andersen, Gøsta, Diederik Boertien, Jens Bonke, and Pablo Gracia. 2013. "Couple Specialization in Multiple Equilibria." *European Sociological Review* 29 (6): 1280–94.

Eurostat. 2016. Labour Force Survey: Persons Employed Part-time. Retrieved on April 18, 2016 from http://ec.europa.eu/eurostat/web/lfs/data/main-tables.

Gabadinho, Alexis, Gilbert Ritschard, Matthias Studer and Nicholas S. Muller. 2009. "Mining Sequence Data in R with the TraMineR Package: A User's Guide." Technical report, Department of Econometrics and Laboratory of Demography, University of Geneva, Geneva.

Gabadinho, Alexis, Gilbert Ritschard, Nicolas S. Muller, and Matthias Studer. 2011. "Analyzing and Visualizing State Sequences in R with TraMineR." Journal of Statistical Software 40 (4): 1–37.

García-Manglano, Javier. 2014. "The Life-Course Employment Profiles of Early Baby-Boom Women: A Group-Based Trajectory Analysis." Paper presented at the Population Association of America 2014 Annual Meeting in Boston, MA (May 1-3).

Gauthier, Jacques-Antoine, Eric D. Widmer, Philipp Bucher and Cédric Notredame. 2010. "Multichannel sequence analysis applied to social science data." *Sociological Methodology*, 40(1): 1–38.

Gerson, Kathleen. 1985. *Hard Choices: How Women decide about Work, Career, and Motherhood*. Berkeley/Los Angeles: University of California Press.

Gerson, Kathleen. 2009. "Falling Back on Plan B: The Children of the Gender Revolution Face Uncharted Territory." In B. J. Risman (Ed.), *Families as They Really Are* (pp. 379–392). New York: Norton.

Gerson, Kathleen. 2011. *The Unfinished Revolution: How a New Generation Is Reshaping Family, Work, and Gender in America*. New York: Oxford University Press.

Goldin, Claudia. 1990. *Understanding the Gender Gap: An Economic History of American Women*. New York / Oxford: Oxford University Press.

Goldin, Claudia. 2006. "The Quiet Revolution that Transformed Women's Employment, Education, and Family." *American Economic Review*, 96(2): 1–21.

Goldin, Claudia. 2014. "A Grand Gender Convergence: Its Last Chapter." *American Economic Review*, 104(4): 1091–1119.

Hakim, Catherine. 1996. *Key Issues in Women's Work: Female heterogeneity and the polarization of women's employment*. London: The Athlone Press.

Hakim, Catherine. 2002. "Lifestyle Preferences as Determinants of Women's Differentiated Labor Market Careers". *Work and Occupations*, Vol. 29, No. 4: 428-459.

Han, Shin-Kap, and Phyllis Moen. 2001. "Coupled Careers: Pathways Through Work and Marriage in the United States." In Blossfeld, Hans-Peter, and Sonja Drobnič, (eds.). 2001. *Careers of Couples in Contemporary Society: From Male Breadwinner to Dual-Earner Families*. Oxford: Oxford University Press.

Henz, Ursula, and Marianne Sundstrom. 2001. "Earnings as a Force of Attraction and Specialization in Sweden." In Blossfeld, Hans-Peter, and Sonja Drobnič, (eds.). 2001. *Careers of Couples in Contemporary Society: From Male Breadwinner to Dual-Earner Families*. Oxford: Oxford University Press.

Hernes, Helga Maria. 1987. *Welfare State and Woman Power: Essays in State Feminism*. Oslo: Norwegian University Press.

Hertz, Rosanna. 1986. More *Equal than Others: Women and Men in Dual-Career Marriages*. Berkeley / Los Angeles: University of California Press.

Jensen, Per H., Birgit Pfau-Effinger, and Lluís Flaquer. 2010. "The Development of Informal Work-Welfare Arrangements of European Societies." In Pfau-Effinger, Birgit, Lluis Flaquer and Per H. Jensen (eds.) *Formal and Informal Work: The Hidden Work Regime in Europe*. New York / London: Routledge.

Jensen, Per H. and Jakob Rathlev. 2010. "Formal and Informal Work in the Danish Social Democratic Welfare State." In Pfau-Effinger, Birgit, Lluis Flaquer and Per H. Jensen (eds.) *Formal and Informal Work: The Hidden Work Regime in Europe*. New York / London: Routledge.

Koch-Nielsen, Inger. 1998. "The Roots of Equal Opportunity in the Family." In Ludvigsen, Peter and Lone Palm Larsen (eds.) *The People's Century*. Copenhagen: Arbejdermuseet and ABA.

Lee, Sangheon. 2004. "Working-hour Gaps: Trends and Issues." 29-59. In Messenger, Jon C. (ed.). 2004. *Working Time and Workers' Preferences in Industrialized Countries: Finding the Balance*. New York / London: Routledge.

Leth-Sorensen, Soren, and Gotz Rohwer. 2001. "Work Careers of Married Women in Denmark." In Blossfeld, Hans-Peter, and Sonja Drobnič, (eds.). 2001. *Careers of Couples in Contemporary Society: From Male Breadwinner to Dual-Earner Families*. Oxford: Oxford University Press.

Lewis, Jane. 1992. "Gender and the Development of Welfare Regimes." *Journal of European Social Policy*, 2 (3): 159–73.

Moen, Phyllis and Patricia Roehling. 2005. *The Career Mystique: Cracks in the American Dream*. Lanham, MD: Rowman & Littlefield.

OECD. 1998. OECD *Employment Outlook* 1998. Paris / Danver, MA: OECD.

OECD Statistics. 2016. *Labour Force Statistics: Labour Force Participation Rate*. Retrieved on April 14, 2016 from http://stats.oecd.org.

Pedulla, David S. and Sarah Thébaud. 2015. "Can We Finish the Revolution? Gender, Work-Family Ideals, and Institutional Constraint." *American Sociological Review*, 80, No 1, 116–139.

Pfau-Effinger, Brigit. 2004. *Development of Culture, Welfare States and Women's Employment in Europe.* Hants, England: Ashgate Publishing.

Pfau-Effinger, Birgit. 2005. "Development paths of care arrangements in the framework of family values and welfare values." In Pfau-Effinger, Birgit and Birgit Geissler (eds.) *Care and Social Integration in European Societies*. Bristol: The Policy Press.

Pfau-Effinger, Birgit, Per H. Jensen, and Lluís Flaquer. 2010. "Formal and Informal Work in European Societies: A Comparative PersepctiveNo Title." In Pfau-Effinger, Birgit, Lluis Flaquer and Per H. Jensen (eds.) *Formal and Informal Work: The Hidden Work Regime in Europe*. New York / London: Routledge.

R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Reichart, Elisabeth, Noelle Chesley, and Phyllis Moen. 2007. "The End of the Career Mystique?" *Zeitschrift für Familienforschung*, 19 Jahrh, Heft 3/2007: 337-370.

Rubin, Lillian. 1994. Families on the fault line. New York: Harper Collins.

Statistics Denmark. 2015. Statistical Yearbook 2015. Copenhagen: Statistics Denmark. Retrieved on May 24, 2016 from www.dst.dk/yearbook.

Stone, Pamela. 2007. *Opting Out? Why Women Really Quit Careers and Head Home*. Berkeley: University of California Press.

Studer, Matthias. 2013. "WeightedCluster Library Manual: A practical guide to creating typologies of trajectories in the social sciences with R". LIVES Working Papers, 24.

UNDP. 2015. Gender Inequality Index (GII). Retrieved on May 24, 2016 from http://hdr.undp.org/en/content/gender-inequality-index-gii

US Department of Labor 2016. "Labor Force Statistics from the Current Population Survey: Access to historical data series by subject." Retrieved on April 22, 2016 from http://www.bls.gov/cps/cpsdbtabs.html.

Warren, Elizabeth, and Amelia Warren Tyagi. 2004. *The Two-Income Trap: Why Middle-Class Parents are Going Broke*. New York: Basic Books.

Widmer, Eric. D., and Gilbert Ritschard. 2009. "The de-Standardization of the Life Course: Are Men and Women Equal?" *Advances in Life Course Research*, 14 (1-2): 28–39.

Williams, Joan C. 1999. *Unbending Gender: Why Family and Work Conflict and What to Do about It*. Oxford: Oxford University Press.

8 Annex

Table 6: Multinomial regression output for 1941-1950 birth cohorts

(DV reference category: Private sector Labour market cluster).

	Specification without interaction		Specification with interaction	
	Public sector Labour market cluster	Absence-Intermittence Labour market cluster	Public sector Labour market cluster	Absence-Intermittence Labour market cluster
Gender: Woman	1.344***	1.325***	0.910***	0.544
(Ref: Man)	(0.05)	(0.06)	(0.27)	(0.43)
Year of birth	-0.0575***	-0.317***	-0.0587***	-0.324***
	(0.01)	(0.01)	(0.01)	(0.01)
Immigration background: 1st or 2nd gen.	0.382**	1.332***	0.384**	1.366***
(Ref: None)	(0.15)	(0.13)	(0.15)	(0.13)
Household cluster: Single	0.356***	1.089***	0.214	1.557***
(Ref: Union)	(0.07)	(0.07)	(0.11)	(0.09)
Education: Post-secondary	1.216***	-0.397***	1.022***	-0.546***
(Ref: Secondary or less)	(0.07)	(0.09)	(0.10)	(0.15)
Education: Tertiary	1.477***	-0.251	1.511***	0.0428
(Ref: Secondary or less)	(0.11)	(0.15)	(0.14)	(0.21)
Single#Woman#Post-secondary			0.815*	0.447
			(0.37)	(0.53)
Single#Woman#Tertiary			0.125	-0.402
			(0.47)	(0.60)
Constant	110.8***	615.9***	113.0***	629.4***
	(18.90)	(20.34)	(18.99)	(20.55)
Number of observations		10000		10000
pseudo R2		0.144		0.149

Standard errors indicated in parentheses.

* p<0.05 ** p<0.01 *** p<0.001

.

Table 7: Multinomial regression output for 1951-1960 birth cohorts
(DV reference category: Private sector Labour market cluster).

	Specification without interaction		Specification with interaction	
	Public sector Labour market cluster	Absence-Intermittence Labour market cluster	Public sector Labour market cluster	Absence-Intermittence Labour market cluster
Gender: Woman	1.289***	1.259***	0.961*	-14.14
(Ref: Man)	(0.05)	(0.07)	(0.39)	(1004.7)

45

Year of birth	-0.0727***	-0.0673***	-0.0735***	-0.0707***
	(0.01)	(0.01)	(0.01)	(0.01)
Immigration background: 1st or 2nd gen.	0.151	1.573***	0.151	1.575***
(Ref: None)	(0.11)	(0.11)	(0.12)	(0.11)
Education: Post-secondary	1.449***	-0.639***	1.423***	-0.123
(Ref: Secondary or less)	(0.06)	(0.12)	(0.13)	(0.39)
Education: Tertiary	1.306***	-0.831***	1.379***	-0.216
(Ref: Secondary or less)	(0.09)	(0.20)	(0.18)	(0.53)
Household cluster: Single	0.140*	2.111***	0.234*	2.380***
(Ref: Union 2)	(0.07)	(0.09)	(0.11)	(0.15)
Household cluster: Union 1	-0.0771	0.429***	-0.108	0.419*
(Ref: Union 2)	(0.06)	(0.10)	(0.11)	(0.18)
Household cluster: Union 3+	0.193**	0.728***	0.0951	0.647***
(Ref: Union 2)	(0.07)	(0.11)	(0.12)	(0.20)
Single#Woman#Post-secondary			0.152	14.39
			(0.46)	(1004.7)
Single#Woman#Tertiary			-0.361	14.21
			(0.48)	(1004.7)
Union 1#Woman#Post-secondary			0.244	15.12
			(0.43)	(1004.7)
Union 1#Woman#Tertiary			-0.0252	15.35
			(0.46)	(1004.7)
Union 3+#Woman#Post-secondary			0.626	14.73
			(0.43)	(1004.7)
Union 3+#Woman#Tertiary			Omitted	Omitted
			(.)	(.)
Constant	141.0***	128.9***	142.5***	135.4***
	(16.62)	(22.87)	(16.72)	(23.01)
Number of observations		10000		10000
pseudo R-sq		0.149		0.153

Standard errors indicated in parentheses.
* p<0.05 ** p<0.01 *** p<0.001

Table 8: Multinomial regression output for 1961-1970 birth cohorts
(DV reference category: Private sector Labour market cluster).

	Specification without interaction		Specification with interaction	
	Public sector Labour market cluster	Absence-Intermittence Labour market cluster	Public sector Labour market cluster	Absence-Intermittence Labour market cluster
Gender: Woman	1.482***	1.182***	1.137***	13.96
(Ref: Man)	(0.05)	(0.07)	(0.28)	(502.1)
Year of birth	-0.0589***	-0.0444***	-0.0584***	-0.0434***
	(0.01)	(0.01)	(0.01)	(0.01)
Immigration background: 1st or 2nd gen.	0.427***	1.872***	0.478***	1.925***
(Ref: None)	(0.11)	(0.11)	(0.11)	(0.11)
Education: Post-secondary	1.289***	-1.195***	1.018***	-0.703
(Ref: Secondary or less)	(0.06)	(0.14)	(0.14)	(0.53)
Education: Tertiary	1.006***	-1.296***	1.308***	-13.78
(Ref: Secondary or less)	(0.08)	(0.19)	(0.17)	(502.1)
Household cluster: Single	0.145*	1.857***	0.389***	2.518***
(Ref: Union 1)	(0.06)	(0.09)	(0.11)	(0.17)
Household cluster: Union 2	0.267***	0.807***	-0.146	0.835***
(Ref: Union 1)	(0.06)	(0.10)	(0.15)	(0.21)
1.dst3h3#1.gend#1.educate			0.0212	-14.48
			(0.34)	(502.1)
1.dst3h3#1.gend#2.educate			-0.702*	-0.886
			(0.29)	(0.64)
3.dst3h3#1.gend#1.educate			1.444***	-12.27
			(0.31)	(502.1)
3.dst3h3#1.gend#2.educate			Omitted	Omitted
			(.)	(.)
Constant	113.8***	84.14***	112.8***	81.76***
	(17.35)	(24.47)	(17.47)	(24.53)
N		10000		10000
pseudo R-sq		0.153		0.161

Standard errors indicated in parentheses.
* p<0.05 ** p<0.01 *** p<0.001

47

Table 9: Multinomial regression output for 1971-1980 birth cohorts
(DV reference category: Private sector Labour market cluster).

	Specification without interaction		Specification with interaction	
	Public sector Labour market cluster	Absence-Intermittence Labour market cluster	Public sector Labour market cluster	Absence-Intermittence Labour market cluster
Gender: Woman	1.414***	1.275***	0.691***	1.088*
(Ref: Man)	(0.0526)	(0.0751)	(0.143)	(0.543)
Year of birth	0.0246**	-0.00229	0.0244**	-0.00241
	(0.00884)	(0.0122)	(0.00885)	(0.0122)
Immigration background: 1st or 2nd gen.	0.312**	1.694***	0.328**	1.706***
(Ref: None)	(0.117)	(0.105)	(0.119)	(0.108)
Education: Post-secondary	1.736***	-1.537***	1.986***	-0.862*
(Ref: Secondary or less)	(0.0576)	(0.150)	(0.130)	(0.430)
Education: Tertiary	1.567***	-1.322***	2.288***	-0.558
(Ref: Secondary or less)	(0.0682)	(0.159)	(0.140)	(0.469)
Household cluster: Single	0.102	1.356***	0.543***	1.991***
(Ref: Union)	(0.0555)	(0.0760)	(0.129)	(0.136)
Single#Woman#Post-secondary			0.215	-1.046
			(0.241)	(0.751)
Single#Woman#Post-secondary			-0.501*	-1.810***
			(0.204)	(0.479)
Constant	-50.90**	1.707	-50.74**	1.462
	(17.46)	(24.13)	(17.49)	(24.19)
N		10000		10000
pseudo R-sq		0.183		0.190

Standard errors indicated in parentheses.
* p<0.05 ** p<0.01 *** p<0.001

Gender inequality regarding retirement benefits in Switzerland

Olga Ganjour*, Jacques-Antoine Gauthier** & Jean-Marie Le Goff**

*University of Geneva, **University of Lausanne

Introduction

The aim of this paper is to compare two methods for constructing typologies of life trajectories based on sequential data. The focus is put on the forms of causality that links together ascribed, antecedent positional factors, life trajectories and some outcome variable measured afterwards. By using the term "forms of causality", we refer to the classification proposed by Godard and de Coninck (1990) about temporal forms of causality that are mobilized by life course researchers in order to give sense to the life trajectories they study. These authors distinguish three forms of causality, that they call respectively the *archeological* model, the *processual* model and the *structural* model. In the first model, the causal relation posits that trajectories ensue almost deterministically from the social position(s) individuals held at their starting point, for example at birth. This model promotes the historical or historicist explanation of individual life (Stinchcombe, 1968) or structuralist view of social life (Abbott, 2000) which is determined by social background of individuals. The brightest example relates to homogamous marriages as reproduction of social structures. In the second model, causal relations are based on the ways in which events and phases of the life courses are chained to one another, showing for example an underlying logic of cumulative advantages or disadvantages during the life course (Dannefer, 2009). The trajectories or "master narrative" result overarching of social process that has the character of coercing processes within it, and indeed of preventing those processes from creat-

ing combinations that disrupt it (Abbott, 2000). It is this coercive characteristic that makes trajectories master narratives (Abbott, 2000). The institutionalisation process exemplifies this model, as it begins with commitment of powerful individuals and ends with the domination of state institutions (Stinchcombe, 1968). In this perspective, individual trajectories, that is particular time varying participation profile (Levy & Widmer, 2013), may or may not change the initial situation.

The third model proposed by Godard and de Coninck (1990) makes links between life trajectories and the social context in which they occur and to question how social structures and institutions shape life trajectories over time. The social institutions as causal forces intervene in the process of self-generation of social structures (Stinchcombe, 1968). For example, industrial development increases the demands on the professional in the new industries and favours the education. The "routinization of innovation" and rationalisation (in Weber's sense) are the traits of modernisation, which in turn promotes institutionalisation. The most audacious hypothesis in this purpose is that the process of institutionalisation (educational and occupational trajectories as the evidence) decreases the social inequality determined by social background.

In this article we empirically address the issue of the time related link existing between initial positional factors (e.g. gender), individual occupational trajectories and retirement benefits.

In this perspective, several studies realized in occidental countries demonstrated that retirement insurances are "life course sensitive" (Leisering, 2003)., Life course sensitivity means that the level of retirement benefits is related to the occupational trajectories of individuals. In the case of Switzerland for example, those who had an occupational trajectories characterized by full time employment without interruptions get a higher level of retirement benefits than those that had trajectories characterized by discontinuous and part time employment. However, a rich body of research has also shown that the level of retirement benefits is related to social background of individual. For instance, the inequality of retirement benefits in Switzerland is reinforced by the gendered nature of individual life course. The gendered dimension of life course is theoretically framed among others with

the notion of master status (Levy & Kruger, 2001; Levy et al, 2006). It postulates that the institutions of the welfare state - contribute to primary assign to women the role of caregivers and to men that of breadwinners. Therefore, female employment is legitimated only if it is subsidiary to domestic and cares responsibilities, while paid work remains the prerogative of men. Their involvement in family tasks is possible as long as their role as breadwinners is fulfilled (Levy, Widmer & Kellerhals, 2002; Levy & Widmer, 2013).

Research results are however ambiguous concerning the relations between positional factors, occupational trajectories and retirement benefits. Schematically, in the theoretical perspective linking welfare state institutions and the gendered master status, the predominant form of causality is the archaeological model. Indeed, gender differences lead to sex specific occupational trajectories from which depends the level of retirement benefit. In this case, the occupational career plays some sort of mediating role between gender and benefits. On the other hand, focusing on the life course sensitivity of retirement schemes calls for a processual model of cumulative disadvantages and advantages. In this perspective, an important share of women, but not all of them, have discontinuous occupational career mainly for family reasons with the consequence that eventually they will be disadvantaged in terms of retirement benefits in comparison to men who have a continuous full time career.

This raises the issue of which of the archeological and the processual models is most adequate to explain the links between positional factors, occupational trajectories, and retirement benefits? To answer to this question, we will compare results obtained using two different methodologies that are combined with sequence analysis in order to describe types of trajectories. The first one is the cluster-based approach which is broadly used in the literature on life course research (Gauthier et al. 2009) and second, the discrepancy analysis that was recently proposed by Studer et al (2011).

The cluster-based approach is a two-steps analysis. In a first step, a cluster analysis is performed on the matrix gathering all pairwise distances between individual trajectories stemming from sequence analysis in order to create homogene-

ous types of trajectories. In the second step the relation of clusters with initial individual positional characteristics is analyzed using for instance logistic regression or factor analysis. Initial characteristics are conceived as covariates that characterize social positions of persons before the starting time of the considered trajectory (gender, birth cohort, social class of origin, etc.). The discrepancy analysis is an ANOVA related method that also aims at delineating types of trajectories according to initial characteristics of individuals (Studer at al, 2011). This regression tree method that reveal in an ordered way the individual characteristics that best discriminate the trajectories at hand. Both methods create groups of trajectories, either based solely on their structural proximity (cluster analysis) or on individual characteristics (discrepancy analysis). Studer et al (2011) insist on the fact that discrepancy analysis allows making a direct link between individual characteristics and trajectories. In this sense such methodology supports the archaeological model presented above. Similarly, focusing uppermost on the comparison between trajectories prior to linking them with the initial characteristics of individuals, the clustering based approach reflects the processual model of causality underlying life trajectories proposed by Godard and de Cononck (1990).

In the following, we will address the comparison of these two methods in relation to retirement benefits. We make the assumption that the most adequate method along with the underlying form of causality, will be the one producing the lowest internal variability and the highest intergroup variability on an external variable as in this case the amount of retirement benefits. Before to develop tools we used and to present our result, we will succinctly present in the following section the retirement scheme of Switzerland. This presentation is necessary in the sense that the construction of sequences that we used for our analysis depends strongly of this scheme.

The retirement scheme in Switzerland

The life course sensitivity of retirement insurance implies that the institutional context determines opportunity of low or high retirement benefits. In order to pre-

scribe retirement benefits we relate Swiss regime of retirement insurance with occupational trajectories and social characteristics of individuals. The regime of social protection in Switzerland during the retirement includes three pillars: the oldness and survivor insurance or *Assurance-vieillesse et survivants* (AVS), the occupational benefit or *Prévoyance professionnelle* (PP) and personal private insurance.

The oldness and survivor insurance is universal and offers state allowance for all individuals achieving the mandatory age fixed in 2015 to 64 years for women and 65 years for men (www.avs-ai.ch.). Although this insurance remains income dependent, it's mandatory for unemployed persons. Particularly, a husband's reallowance completes the contributions for unemployed wife. This insurance aims at covering the basic economic needs of individuals; its minimal amount is fixed 1'175 CHF and maximum is 2'350 CHF per month (www.avs-ai.ch.). In the case of divorce and widowhood never employed women keep the right for the part of allowance, which is calculated due to period and amount of contribution of their husbands.

Occupational insurance is compulsory in Switzerland since 1985 for the individuals earning more than 21'150 CHF per year (for 2015, cf. www.avs-ai.ch.). These contributions depend on the individual income during the work life and allow increasing of retirement benefits. Therefore, the prevalence of part-time work among married women, associated with caring commitment may exclude many women from occupational pension schemes. Particularly in Switzerland in 2008 56,8% women contribute to an occupational pension scheme (LPP) compared to 81,7% of men (Actualités OFS, 2011; SESAM data).

The capacity of women to obtain an adequate occupational pension is constrained by the unpaid domestic provisioning and caring they undertake for the benefit of others (Ginn & Arber, 1993). Wives who do not contribute to an occupational insurance scheme depend financially on their husband. Although married women may share the benefit of their husband's occupational pension when they retire, widows receive about 60% their deceased husband's pension depending on age at death and the matrimonial union regime (http://www.bsv.admin.ch/). Ac-

cording to law (article 122 al. 1 CCS) divorced women may claim the half of their ex-husband's occupational pension. Hence, pension schemes of all types depend to some extent on duration and level of earnings, making women more likely to have lower retirement benefits than men.

Self-employed individuals have not access to the LPP. They have to contribute to private pension funds in order to secure an adequate income in later life.

The final level of retirement benefits depends on the level and continuity of earnings obtained during ones occupational career. Mostly the level of retirement benefits is related to the occupational insurance or the second pillar. This dependence presents a disadvantage for those, especially women, who have low wages and/or do not fit the profile of full-time continuous employment. Moreover, we consider that social characteristics of individuals, particularly the type of sexual division of labor influence their occupational careers and contribute to increase inequalities between the retired workers who fully contributed to the social welfare system and those who did less so or not at all.

Data and methods

We use the data from the Survey of Health, Ageing and Retirement in Europe (SHARE), particularly the retrospective biographic data SHARELIFE gathered during its third wave in 2009. The initial Swiss sample of SHARELIFE counts 1296 respondents aged of 50 years and more. We selected individuals born before 1949 and therefore were at least 60 years old at the end of the data collection. The resulting sample includes 833 respondents. This allows to precisely observing the ways in which individuals leave the labour market, possibly before the mandatory age of retirement (currently 65 for men and 64 for women). The trajectories allow observing simultaneously timing of transition between the states, duration and reversibility of states.

We first built the occupational trajectories of individuals, as a succession of unambiguous yearly states from age 45 to 70. We distinguished six possible occupational states. The first three states are linked with employment, namely: 1. "Full time employment with LPP contributions", 2. "Full time employment without LPP contribution" and 3. "Part-time employment with and without LPP contribution". The next states are linked with unemployment or situations in which individuals are out of the labour market, namely 4. "Insurances benefits including disability insurances and unemployment insurances", 5. "At home or occupationally inactive due to personal reasons" and 6. "Retirement".

Optimal matching analysis allows quantifying the level of dissimilarity (called distance) between a pair of individual sequences and hence eventually producing a matrix of all interindividual distances (Gauthier, 2013).

Secondly we developed the first method of classification of occupational trajectories - cluster analysis. Cluster analysis using the Ward method (Ward, 1963) performed on the distance matrix allows grouping the homologous sequences in to homogeneous groups in order to build sociologically meaningful types of individual occupational trajectories. Based on a standard clustering quality criterion (silhouette index =0.69) (Rousseeuw, 1987) and face value relevance, we retain a four-group typology (Figure 1).
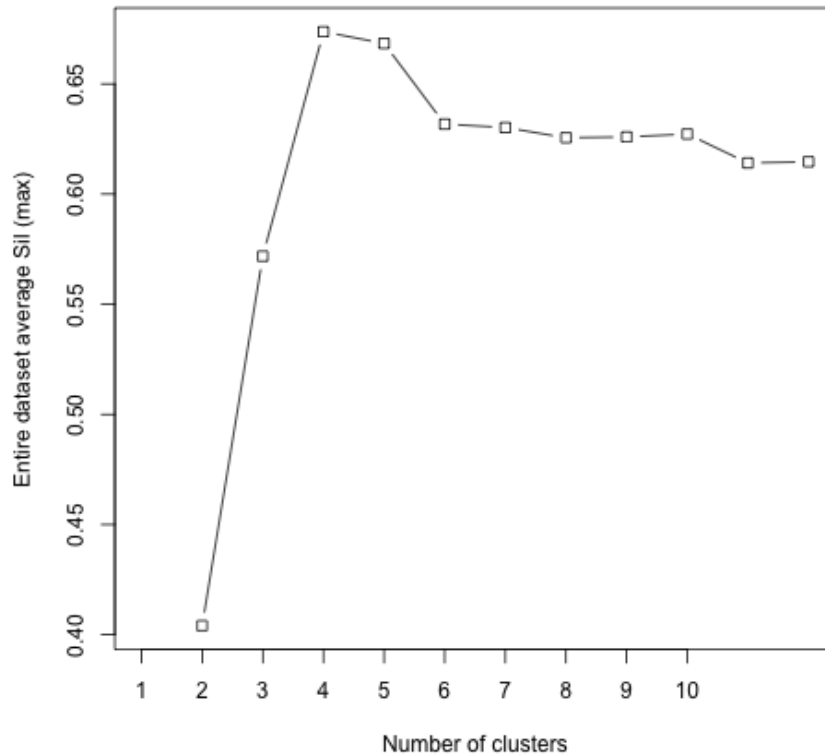
Fig. 1: Silhouette index

Unsurprisingly, the resulting patterns of individual occupational trajectories are mainly structured by the various levels of occupational activity and by the fact that employment is associated or not with compulsory contributions to a pension fund.

Eventually, we realized regression analyses in order to link the types of occupational trajectories with a selection of social characteristics of individuals, namely sex, birth cohort, occupational status, educational status, marital status and nationality. Unfortunately, due to data separation regarding the distribution of sex and age between the groups, logistic regression provides aberrant regression coefficients. This situation occurs often, in particular when comparing trajectories that

are highly sensitive to certain variables, as it is the case when studying gender inequalities. For this reason we applied multiple correspondent analysis (MCA).

Thirdly, we develop the second method of classification of occupational trajectories - ANOVA discrepancy analysis (Studer et al, 2011). The discrepancy of sequences is directly measured from their pairwise dissimilarities computed using optimal matching analysis (OMA) (Gauthier, 2013). The discrepancy analysis is based on the effect of each new covariate, which is evaluated by discrepancy between this covariates and others covariates when this covariate is included in the model. The algorithm is based on the maximisation of the between groups discrepancy and is sensible to order in which the covariates are added in the model (Studer et al, 2011). We used the algorithm when all variables are included in the model and then each variable by turn is excluded from the model. The effect of each variable is evaluated by the minimisation of the between groups discrepancy, and calculated by ratio of the explained discrepancy once all covariates were included in the model to the residual discrepancy if the covariate removes from the model. The significant reduction of this ratio when the covariate is excluded from the model signifies it's high influence on the individual's trajectories.

Finally, we compare the logarithmic score of retirement benefits in the groups of individuals produced with cluster analysis and ANOVA discrepancy analysis. We evaluate these groups with criterium of within/between group homogeneity (median, mean, variance, skewness, kurtosis). Heterogeneity of scores of retirement benefits between groups and homogeneity of scores within the groups signifies the quality of groups.

All computations are made with the R environment for statistical computing (R core team, 2014). In particular, multiple correspondence analysis are done using the FactoMineR package (Husson, Lê & Pagès, 2009), sequence, discrepancy and regression tree analyses are based on the TraMineR package (Gabadinho, Ritschard, Müller & Studer, 2011).

Results

1. Cluster analysis

We can read from Figure 2, that patterns of individual occupational trajectories (or cluster types) are related to employment states with or without contributions to a pension fund or states of occupational inactivity
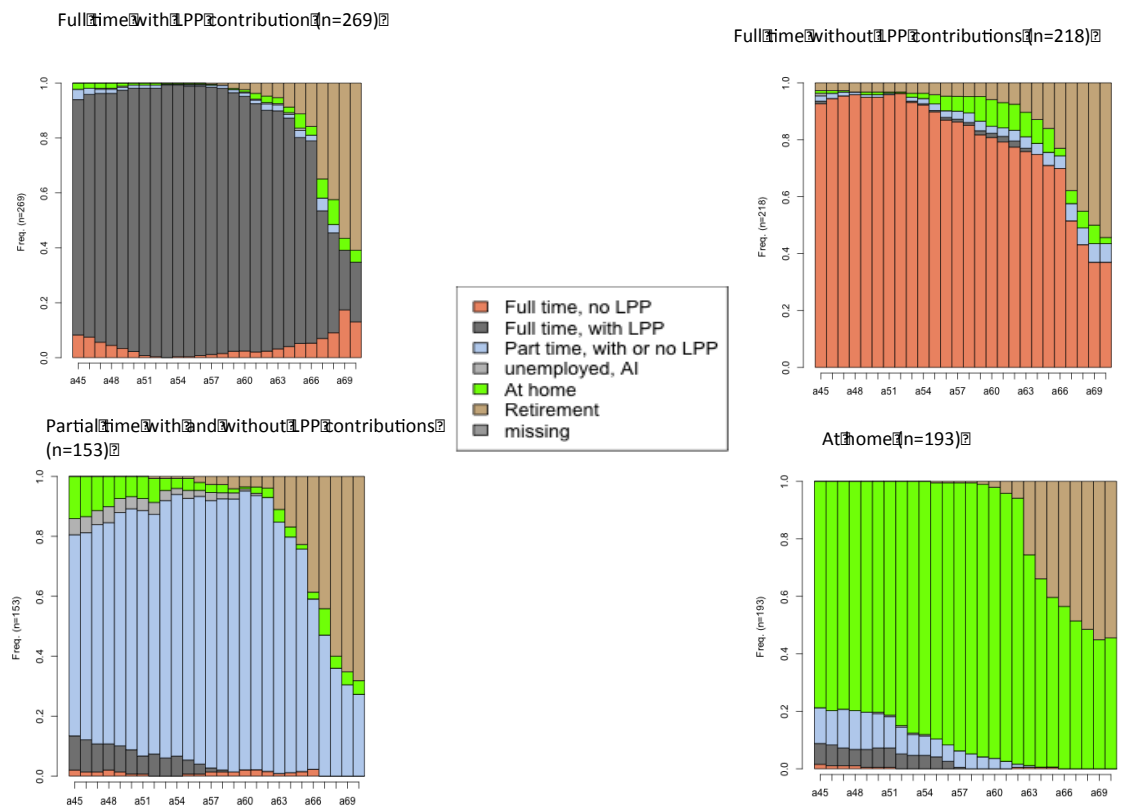
Fig. 2: Cluster types of retirement transition

The first pattern "Full time employment with LPP contribution" (32 % of the selected sample) includes individuals who are continuously full time employed and contribute to an occupational pension fund. This pattern represents highly standardized trajectories of both occupational careers and transitions to retirement. Individuals who follow this type of trajectories did not (or rarely) experience unemployment or disability during their career; only a few of them anticipated their retirement, but only at the age of 57 for the most precocious of them.

The second pattern "Full time employment without LPP contribution" (26% of the selected sample) includes individuals who are continuously full time employed, but who do not contribute to any occupational pension fund. This pattern presents less standardized trajectories of occupational careers and retirement. For them, the process of the transition to seniority starts as early as 45 years and one third of individuals belonging to this group left the labor market before the legal AVS age, sometimes progressively through part time employment and/or occupational inactivity.

The third pattern "Part time employment with and without LPP contribution" (19% of the selected sample) includes individuals who are part time employed, whatever they contribute or not to an occupational pension fund. This pattern presents also less standardized trajectories of occupational carrier and retirement. The transition to occupational inactivity due to unemployment, disability and/or occupational inactivity occurring more often around age 60, i.e. short before retirement age.

The fourth type that we named "At home" (23% of the sample) includes individuals who are occupationally inactive on the long run. This pattern presents destandardized trajectories of occupational carrier and retirement. At the legal AVS age only one half of individuals declare their retirement, while the other half continues to consider themselves as occupationally inactive after that age. This underlines the high level of institutional exclusion these individuals are facing.

At the next step by application of MCA we link the patterns of occupational trajectories with social characteristics of individuals.
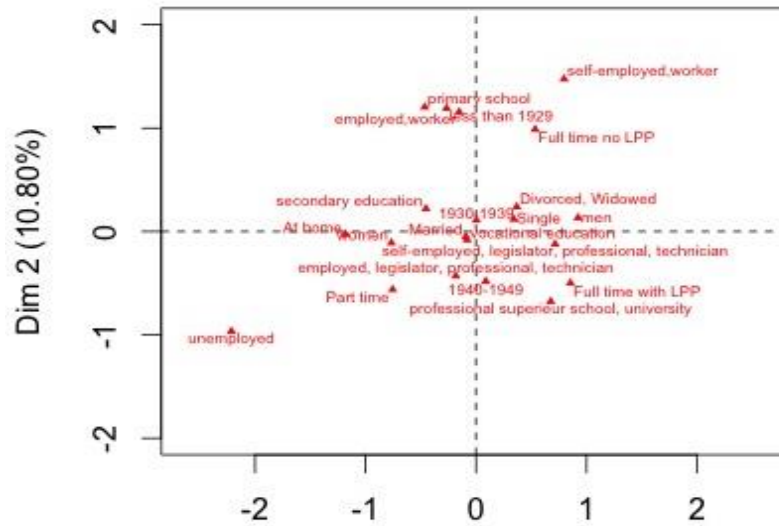
Fig. 3: Multiple correspondent analysis of categories of cluster types and categories of social characteristics of individuals

According to Figure 3 the first axe contrasts the patterns of full time activity with or without LPP contributions, and the patterns of part time activity and occupational inactivity. This opposition is also that of men and women trajectories Men are involved in the trajectory characterized by a transition to retirement from a full time job with or without LPP contributions, while women follow trajectories characterized by a transition to retirement after an occupational inactivity or part time employment. The second axe marks an opposition between the pattern "Full time employment with LPP contributions" and the pattern "Full time activity without LPP contribution". This opposition is strongly associated with the birth cohort of respondents. The individuals born before 1929 are more likely to belong to the pattern "Full time employment without LPP contributions" while the individuals born between 1940 and 1949 are more frequent in the pattern "Full time employment with LPP contributions". The second axe reveals also the opposition between self-employed who did not contribute to LPP during their career and employed who contributed to the LPP. The level of education influences also on

probability to belong to the type . A low level of education is more often associated with the type "Full time activity without LPP contributions", while a high level of education is frequent in the type "Full time activity with LPP contributions".

2. ANOVA discrepancy analysis

As alternative way for classifying occupational trajectories we used ANOVA discrepancy analysis (Studer et al, 2011).

Table 1 shows the discrepancy of the trajectories due to the covariates or variables of social characteristics.

Table 1. Effect of the covariates on the occupational trajectories according to Multifactor Discrepancy Analysis.

Full Model			
Variable	F	Pseudo-R^2	p-value
Sex	55.78	0.119	0.000
birth cohort	5.35	0.023	0.000
occupational status	1.89	0.008	0.038
level of education	0.93	0.006	0.507
marital status	4.23	0.018	0.000
nationality	2.19	0.005	0.043
Total	9.19	0.216	0.000

Note. Inconclusive intervals: 0.00724 <0.01 <0.0128; 0.04396 <0.05<0.0560.

The model with all included covariates explains 21,6% of total discrepancy (R^2=.216). The variable sex is the most influent covariate on the individual occupational trajectories (F=55.78 and p-value =0.000). If we remove this variable, the R^2 value of the model decreases by 0.119, and the total discrepancy of the trajectories decrease by 11,9%. In opposite, other covariates are not influent or not significant, beside the birth cohort, which explains 0.02% of the total discrepancy of the trajectories (R^2=.023). Although the multifactor analysis allows the recognition of the most influent social characteristics on the individual's trajectories, it does not identifies what the effects are or how the trajectories change under influence of the

covariates. In order to visualize the link between the modifications of the occupational trajectories under influence of the covariates we applied the regression tree analysis. In order to make the number of regression tree groups equal to the number of cluster groups we limited the deep of tree to four branches.
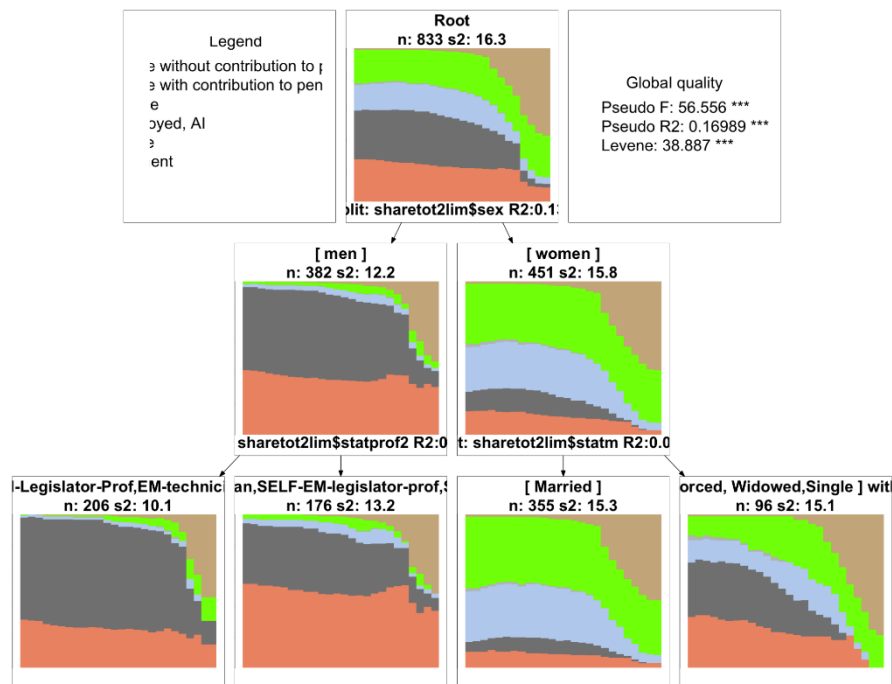
Fig. 4: The regression tree of the occupational trajectories

The model of the regression tree explained 16, 9% of total discrepancy (R^2=.169). We observed that sex is the most influential variable to discriminate occupational trajectories before the transition to retirement. Then, the occupational status is the most influential social characteristic to distinguish men trajectories; while it is the marital status for women. It confirms to a certain extent master status hypothesis. First branch (25,0%) includes employed men with advanced occupational status as legislator, occupational, technician. Second branch (21,0%) includes employed men as workers, and self-employed men, with all occupational status as legislator, occupational, technician, worker and unemployed. The third branch (43,0%) in-

cludes married women., The fourth branch (11,0%) includes divorced, widowed and single women,.

3. Repartition individuals from the classified groups

We crossed the cluster groups and the regressions tree groups in order to examine if the individuals developing the equal occupational trajectories are included in the same groups (Table 2).

Table 2. Repartition of individuals in cluster and regression tree group

	Cluster 1 Full time with LLP, %	Cluster 2 Full time without LPP, %	Cluster 3 Part time with or without LPP, %	Cluster 4 At home, %	Total
1. Employed legislator, professional men	68.9	26.7	1.9	2.4	100 N=206
2. Employed workers and self-employed all prof. status men	38.6	50.0	7.4	4.0	100 N=176
3. Married women	9.0	12.4	34.1	44.5	100 N=355
4. Divorced, separated and single women	28.4	32.6	14.7	24.2	100 N=95

According to Table 2 the individuals developing the same occupation trajectories are mostly included in correlated clusters and regression tree groups. Hoverer, the regression tree groups include the individuals who don't belong to the same clusters according to occupational trajectories.. The first regression tree group "Employed legislator, professional men" includes about 70% of men belonging to the cluster "Full time employment with LPP contribution". About 27% of men of this group belong to the cluster "Full time employment without LPP contributions" and the insignificant proportions of men of this group belong to the clusters: "Part time employment with or without LPP contribution" and "At home". The second regression tree group "Employed workers and self-employed of all professional status men" includes 50% of men belonging to the cluster "Full time employment without LPP contributions". However about 40% of men of this group belong to the cluster "Full time employment with LPP contributions". About 7% of men of this group developed less typical occupational strategies and belong to the cluster "Part time employment with or without LPP contribution". The third regression tree group "Married women" predominantly includes women who belong to the cluster "At home". This group includes also 34% of women be-

17

longing to the cluster "Part time employment with or without LPP contribution". However about 20% of women of this group develop the strategies of full time employment and belong to the clusters "Full time employment without LPP contributions" and "Full time employment with LPP contribution". The fourth regression tree group "Divorced, separated and single women" includes women who belong to different clusters, mostly to the clusters "Full time employment without LPP contributions", "Full time employment with LPP contribution" and "At home". Minority of women of this group belong to the cluster "Part time employment with or without LPP contribution".

On the next step we compared the median personal income across the regression tree and cluster groups (Table 3).

4. Median personal income across the clusters and regression tree groups

We would have expected a strongest correlation of personal income inside of regression tree groups in case of a predominant archaeological model of the life course. And oppositely, we expected a strongest correlation of personal income inside of clusters groups in case of a predominant processual model of life course. The personal income derives from household income divided by number of household members (SHARE wave 2).

Table 3. Median personal income or median level of the individual life

	Cluster 1 Full time with LLP, %	Cluster 2 Full time without LPP, %	Cluster 3 Part time with or without LPP, %	Cluster 4 At home, %	Median group (all clusters)
1. Employed legislator, professional men	3500	3050	2050 (N=4)	4000 (N=5)	3500
2.Employed workers and self-employed all prof. status men	2675	2175	4500 (N=11)	2396 (N=4)	2500
3.Married women	3600 (N=24)	2117 (N=30)	2700	2500	2700
4.Divorced, separated and single women	4500 (N=21)	2450 (N=20)	2533 (N=13)	2750 (N=17)	2750
Median cluster (all groups)	3400	2500	2633	2650	2850

Note: we noted the weak size of category between the brackets.

The variation of median personal income among the regressions tree groups shows that the income is weakly correlated with social characteristics of individuals as we would have expected from an archaeological model of the life course. In the regression tree group "Employed legislator, professional men" median personal income of employed full time with LPP contribution men is equal to median income of this group. However, there is heterogeneity of income distribution in the group "Employed workers and self-employed all professional status men". Particularly, median income of men who are employed part time is differed from median income of this group. Again in the groups "Married women" and "Divorced, separated and single women" the income of employed full time women doesn't correspond to median income of these groups. In opposite, in the clusters, especially in the two first clusters which unit the individuals developing the trajectories

19

of full time employment with and without LPP, median income correspond to occupational trajectories and predominantly corresponds to median income of these clusters. Such result promotes the processual model of causality rather than the archaeological model (Godard & de Coninck, 1990).

We also explored the variability of income inside of the groups and the clusters. For this purpose we computed Gini index of inequalities of income inside of the regression tree groups and the clusters (table 5).

Table 5. Index of Gini

	Cluster 1 Full time with LLP, %	Cluster 2 Full time without LPP, %	Cluster 3 Part time with or without LPP, %	Cluster 4 At home, %	Median group (all clusters)
1. Employed legislator, professional men	27.81	33.97	(51.10)	(12.51)	30.36
2. Employed workers and self-employed all prof. status men	26.17	37.39	(38.43)	(6.29)	34.19
3. Married women,	22.81	38.63	22.08	34.56	30.57
4. Divorced, separated, single women	26.75	35.27	16.41	24.39	29.12
Median cluster (all groups)	27.69	37.40	27.38	32.30	31.80

Note: index of Gini of the category of weak size is noted between the brackets

According to table 5 there are more income distribution inequalities inside of the regression tree groups than inside of the clusters. This result shows that the degree of inequalities is more determined by social characteristics of individuals than by their occupational trajectories. As in the case of the median income variation among the clusters and the regression tree groups (table 4), these results rather

promote a processual interpretation of the equal classification than an archeological interpretation.

In the case of the trajectory of full time employment with LPP, inequality is weakest than in the case of the trajectories of full time employment without LPP, part time or staying at home. We can explain this difference by higher degree of institutionalisation of professional trajectories with LPP contributions. The law of pension fund promulgated in 1985 had for consequence the decrease of inequalities between workers and high status employees. The low level of institutionalisation in the case of part time employment or unemployment explains the high variance of personal income and inequality in the clusters "Part time with or without LPP" and "At home".

Provisory conclusion

The aim of our paper was to investigate how to link the social background of individuals, their life trajectories (occupation) ,and an outcome, continuous variable (the level of retirement benefits). In order to investigate the causal relationship between these factors and personal income two methods were tested: the first, is the most diffused in actual social science research community, namely cluster analysis combined with sequence analysis; the second is the discrepancy analysis (Studer et al., 2011).

The application of these methods is based on the different methodology. The cluster analysis was applied to a matrix of distances between occupational trajectories, which was previously computed with a sequence analysis. The cluster analysis provides the groups of individuals due to proximity of their occupational trajectories. In a life course perspective the clusters result the influence of institutionalisation on the individuals trajectories. The social institutions such as employment, social insurances or unemployment influence on the individual trajectories together with background. The essential point of this processual model is

to consider that social institutions produce an impact on the individual trajectories. This model doesn't favour the dominance of background and supposes the existence of social mobility in the society. For example, two persons with different background can have similar occupational trajectories through participation in the labour market and benefice similar income after retirement. In opposite, the application of the discrepancy analysis supposes the link between the individual background and the occupational trajectories. As the cluster analysis the discrepancy analysis was applied on the matrix of distance between occupational trajectories, but it allows splitting the groups of individuals based on the proximity of their occupational trajectories due to background characteristics. This archeological model supposes influence of the individual background on the occupational trajectories. This regard to the classification enhances the influence of background characteristics on the occupational trajectories and reduces the impact of institutionalisation.

We applied two methods of classification of occupational trajectories in order to compare the homogeneity of the groups, particularly, homogeneity of distribution of personal income among the individuals of the clusters and regression tree groups. Relating these methods to the forms of causality proposed by Godard & de Coninck (1990) we show that processual model analysing the relationship between occupational trajectories and personal income is centred on the more homogeneous groups. According to the results, the distribution of median personal income among cluster types is more homogeneous than the distribution of median personal income among the regression tree groups. Hence, the processual model promotes the more uniform groups in relation to income distribution.

The variation of personal median income, the index Gini, appears to be more equal inside the clusters than inside the regression tree groups. This result confirms the hypothesis that institutionalisation reduces the social inequalities determined by individual background.

Our results are limited by the application of the methods of classification of occupational trajectories before the retirement transition; the inclusion of other stages of life course in the analysis is also relevant.

References

Abbott, A. (2000) Time Matters on theory and method, Chicago: The University of Chicago Press.

Actualités of OFS (2011) Protection sociale : Indicateurs de la prévoyance vieillesse. Résultats des indicateurs clés, Neuchâtel.

Dannefer D. (2009). Stability, homogeneity, agency: Cumulative dis/advantage and problems of theory. Swiss Journal of Sociology, 35, 193–210.

De Coninck F. & Godard F. (1990), L'approche biographique à l'épreuve de l'interprétation. Les formes temporelles de la causalité. Revue française de sociologie, 31-1, 23-53.

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

Gauthier, J.-A., Widmer, E. D., Bucher, P., & Notredame, C. (2009). How much does it cost? Optimization of costs in sequence analysis of social science data. Sociological Methods & Research, 38(1).

Gauthier J.-A., (2013) Optimal matching, a tool for comparing life-course sequences. In: Levy R., Widmer E. D. (eds.) Gendered life courses between standardization and individualization : a European approach applied to Switzerland. LIT, Zürich; Berlin, pp. 37-52.

Ginn, J. & Arber, S.(1992) Towards Women's independence: pension systems in three contrasting European Welfare States. Journal of European social Policy, 2(4), 255-277.

Ginn, J. & Arber, S. (1993) Pension penalties: The Gendered Division of Occupational Welfare. Work, Employment and Society, vol.7, n1, 47-70.

Groves, D. (1987) Occupational pension provision and women's poverty in old age. In Glendinning C. & Millar, J. (Eds) Women and Poverty, Brighton Wheatsheaf.

Hanappi , D., Bernardi, L. & Spini, D. (2015) Vulnerability as a heuristic concept for interdisciplinary research: assessing the thematic and methodological structure of empirical lifecourse studies, Longitudinal and Life Course Studies, Vol. 6 Issue 1, 59-87.

Husson F., Lê S. & Pagès J. (2009) Analyse de données avec R. Press Universitaires de Rennes.

Krüger H., & Levy R. (2001) Linking Life Courses, Work, and the Family: Theorising a not so Visible Nexus between Women and Men. Canadian Journal of Sociology, 26(2), 145-166.

Leisering, L. (2003) Gouvernement and the life course. in Mortimer J.T, Shanahan M. T (Eds) Handbook of the Life Course. Springer.

Levy R., Gauthier J.-A., Widmer E. (2006) Entre contraintes institutionnelle et domestique : les parcours de vie masculins et féminins en Suisse. Canadian Journal of Sociology/Cahiers canadiens de sociologie 31(4), 461-489.

23

Levy R., Widmer E & Kellerhals J. (2002) Modern Families or Modernized Family Traditionalism ? Master Status and the Gender Order in Switzerland. Electronic Journal of Sociology, vol. 6, n 4.

Levy R., Widmer E. D. (2013) (Eds.) Gendered life courses between standardization and individualization : a European approach applied to Switzerland. LIT, Zürich; Berlin.

R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL   http://www.R-project.org/.

Rousseeuw, (1987) Silhouettes: A graphical aid to the interpretation and

validation of cluster analysis. Journal of Computational and Applied Mathematics, vol 20, 53 − 65.

Stinchcombe, A. (1968). Constructing Social theories, Harcourt, Brace & World, Inc.

Studer, Ritschard, G., Gabadinho, A., Müller, N. S. (2011) Discrepancy Analysis of State Sequences, Sociological Methods & Research, 40(3), 741-510.

Glass Ceilings, Escalators and Revolving Doors: Comparing Gendered Occupational Trajectories and the Upward Mobility of Men and Women in West Germany

Lydia Malin and Ramsey Wise

Abstract Drawing from the literature on glass ceiling" and glass escalator" effects, we analyze gender differences in career advancement across gender-typical occupations in West Germany. More specifically, we argue that gender-typical occupations provide different opportunity structures for upward mobility, in part due to their location in different labor market segments – which imply deviating work organization logics. As West Germany is typically characterized by strong occupational sex segregation that is reinforced through educational credentialism and traditional gender role expectations, we contribute to previous empirical research, which has thus far primarily focused on the American or Scandinavian context. In addition, we focus not only on the likelihood of entering a leadership position, as the glass ceiling and glass escalator literature discusses, but we also consider the likelihood of staying in a leadership position, applying the logic behind revolving doors" to discuss why women may be more likely to leave an occupation or leadership position during the career.

Lydia Malin

University of Cologne, Research Training Group SOCLIFE e-mail: malin@wiso.uni-koeln.de

Ramsey Wise

University of Bremen, Bremen International Graduate School of Social Sciences (BIGSSS), e-mail: rwise@bigsss.uni-bremen.de

Session 14A: Education

Research of Students' Performance in Higher Education through Sequence Analysis

Giampiero D'Alessandro[1] and Alessandra Decataldo[2]

[1]PhD in Methodology of Social Science, Department of Communication and Social Research, Sapienza University of Rome

[2]Assistant Professor, Department of Sociology and Social Research, University of Milano Bicocca

DO NOT QUOTE WITHOUT AUTHORS' PERMISSION

Abstract

The paper aims to prove how suitable the Sequence Analysis' tools are for studying the complexity of Italian Higher Education System. Through longitudinal administrative data of a cohort (AY 2001-2002) of enrolled for the first time at Sapienza University of Roma, this approach permits to: (1) describe the phenomena of late graduate and late performance (retention); (2) identify different type of students' dropping out (attrition); (3) evaluate others particular phenomena (that are frequent in Italy) that can delay the students' career (as such as intra or extra faculty mobility).

Introduction

This paper aims to prove that the Sequence Analysis (SA) tools are suitable for studying the complexity of Italian higher education students' careers (Blanchard, Bühlmann and Gauthier 2014; D'Alessandro 2015). The SA is an efficient method that involves examination of the ordered social process (such as university career, life course trajectory, etc.). Actually, SA is commonly used to study professional careers, occupation histories, school-to-work transitions, etc.

    Italian university system is confronted by the three main problems (Benvenuto et al. 2012; Moscati and Vaira 2008): (1) low number of graduates; (2) excessively long university careers; (3) very high number of drop outs. In the course of the last fifteen years in Italy, "Bologna process" has introduced radical

2

reforms with the principal aim to control the phenomena of dropping out, perpetual students and low number of graduates. More specifically, the AY 2001-02 represents the transition from a university education based on a single level degree (4 or 5 year-courses) to a dual level degree: bachelor - BA (3 years)/master - MA (2 years) or other types of master (5 or 6 year-course, such as Architecture or Medicine).

For the purpose of our research, Sapienza University of Rome was identified as a representative case due to its dimensions (i.e. in AY 2015-16 there are 101,492 students - 6,1% of Italian university students and 3,555 professors - 6,9% of Italian university professors) and its variety, in terms of scientific and educational areas of academic training.

1 Research aims

This paper shows a secondary analysis of longitudinal administrative data. We use SA in order to (1) describe in detail the phenomena of late graduation and late performance (retention); (2) identify different kinds of students who are dropping out from the university (attrition); (3) evaluate other particular phenomena that could delay the careers (such as the mobility within and between faculties). Our interest is focused on the linearity of the university career. The non-linearity of career (mobility is a relevant aspect of retention), from a student's point of view, can have a double value. On one hand, it might be a prelude to a successful outcome, in case that a student who is taking up a new university career obtains a degree; on the other, it might cause delay or even dropping out in a student's career. From the university point of view, however, student mobility could determine important impacts, such as increasing the phenomena of late performance and dropping out, but also contributing to the creation of new graduates. The hypothesis is that different types of mobility could cause different outcomes. Mobility between faculties represents a greater educational transition than mobility within faculty.

The case study shows the construction of a longitudinal dataset of information extracted from the administrative archives of the Sapienza University, which refer to the registered students.

SA allows us to identify useful longitudinal measures for describing and analysing the students' careers (Gabadinho et al. 2009; 2011). In particular, the main interests of this study are to 1) identify pattern of the entire sequences; 2) analyse their diffusion; 3) identify groups of similar careers; 4) highlight relations between career outcomes and students' characteristics (micro-level variables); 5) characterize relation between career outcomes and university context characteristics (meso-level variables).

The SA approach has proven to be valuable for pointing out the relevance of phenomena in temporal order (Blanchard, Bühlmann and Gauthier 2014). The

main hypothesis is that the timing (i.e. the observation point when an event occurs) is the most relevant aspect which determinates the career outcome. For instance, a change of study course during the first part of career could be more advantageous than the one performed in the central part of studies (McCormick and Carroll 1997). Indeed, an early change could be an admission of a choice of study course that does not correspond to a student's interest and aspiration (career reorientation); whereas, a late change could be a proof of a student's difficulties in reaching the attainment in the first enrolment study course.

2 Data

Due to the relevance of the AY 2001-02, since it represents the turning point for the passing from the old to the BA/MA university system, the analysis focuses only on this cohort (AY 2001-02) (23,854 students). Information concerns the students enrolled for the first time and without previous university career[1]. The cohort was monitored until March 2014 (the end of AY 2013-14). Therefore, each student is monitored up to 25 semesters. It means that, even if a student graduated at the sixth semester, all his/her following states would be "graduate".

Such information are collected for administrative purposes, like, for example, monitoring fee payments and exam registrations. However, the information are collected in different archives. These archives of Sapienza are related to four main areas: socio-economic profile, high school education, higher education career (first enrolment course, graduations, possible second enrolment course, etc.), and university performance indicators (number of exams, credits, etc.). This kind of (de)structuration is not suitable for the purpose of temporal dynamics studying.

Furthermore, the information are registered in a synchronic way (each information refers to a single moment) and this represents a further critical issue for temporal dynamics studying (Menard 2007; Scherer 2013). Indeed, the phenomena characterized by continuous process of change, such as university careers, can be studied in a more appropriate way through the use of longitudinal data (Blossfeld et al. 1989).

Starting from these archives, a unique dataset has been re-constructed in a longitudinal way, using the key codes (such as the identification numbers of the study course and of the students) in order to joint the data. Through the recording of the events related to each student's career (payment of every semester fees from the first enrolment to graduation or dropping out, passing the exams, etc.), a longitudinal data vector for each student was created. By positioning the

---

[1] Since the AY 2001-2002 was the first year of the "3+2 system" (in this way it is known the Anglo-Saxon university education system in Italy), in this cohort there are no students that enrol for the first time at Sapienza in a MA study course.

4

events on a chronologically ordered string, it is possible to control the dynamics of each academic career and to highlight specific phenomena (dropping out, stopping out, mobility, degree attainment, etc.). In fact, these data allow us to follow students' careers from the enrolment until graduation (or dropping out).

The data query was very complicated and it required a large quantity of Structured Query Language (SQL) code writing in order to obtain a longitudinal data vector for each student from the first semester of AY 2001-02 to the first semester of AY 2013-14. Firstly, the dataset consisted of several vectors for each student. Every vector matched every degree course that student had taken up at Sapienza University (i.e. one for BA and one for MA). A unique data vector for each student was created with the use of the key codes. In this way, it was possible to reconstruct the entire career for each student, including the temporary stopping outs and mobility (between the different and within the same degree courses).

In the final dataset, the data referring to the socio-demographic characteristics and high school training are synchronic (registered at students' first enrolment), whereas those referring to university career and performance indicators are diachronic (repeated for each semester of students' university careers).

The used data encoding format is the STS (State Sequence). It is «a natural way of representing list is a (row) vector of k elements» (Ritschard et al. 2009, p. 159). In the dataset, the row vector has variables as moments of observation for each student's career. Therefore, the column vector corresponds to a predetermined unit of time (the semester).

The structure of the database implies the persistence of the last observed state (dropping out or graduation) for the student throughout the following observation period. Indeed, the aim to observe how different types of careers within the University would lead to different outcomes is well accessible through the University monitoring viewpoint, which suggests the non-use of missing values.

3 Methodology

The basic idea is to list each student's academic trajectories at Sapienza as an ordered list of administrative states, which are defined on a time axis. The time interval being used is the semester.

For the scope of this analysis, we consider ten different administrative "states", that define the alphabet for the SA. They are: (1-4) four types of enrolment (due to the course length: 2, 3, 5 or 6 year-course); (5) continuation (after the previous semester); (6) graduation; (7) formal dropping out (with a formal request of study interruption, probably due to a transition to another university); (8) informal dropping out (due to no six-month fee payment); (9) change of

5

faculty (between two faculties); (10) change of the course (within the same faculty).

A distance matrix between each pair of careers was calculated by Optimal Matching (OM) algorithm (MacIndoe and Abbott 2004; Studer and Ritschard 2014; Blanchard, Bühlmann and Gauthier 2014). The OM produced a huge matrix[2], used for Wald cluster analysis (Studer 2013). The purpose is to achieve well-defined clusters and to characterise them with reference to the information regarding the students (i.e. personal data, high school information – micro-level variables) and the first year performance indicator (i.e. number of exams, number of university credits – micro-level variables), the university context (i.e. length of first enrolment study course and limited enrolment course – meso-level variables).

4 Results

The first step is to give an overview that takes into account the homogeneity/heterogeneity of the students' careers within the University.

Students act within a regulated framework and their chances of action are limited in various ways. In a chart of ideal careers (each student enrolls and continues his/her career until graduation, Graph 1), immediately after the different types of enrollment (three different tones of green: soft green for BA, medium green for 5-year course, such as Architecture, and dark green for 6 year-course, such as Medicine), can be observed only continuation (yellow) and graduation (blue) states. In addition, there can be observed the graduation states in different points of students' careers, coherently to study course length. In particular, the possible ideal careers for students enrolled in BA are two: the first one stops with graduation in the second semester of the third year (sixth moment of observation); the second one continues in MA from the seventh point of observation (first semester of the fourth year, very soft green in the middle of the plot) until the tenth point (second semester of the fifth year).
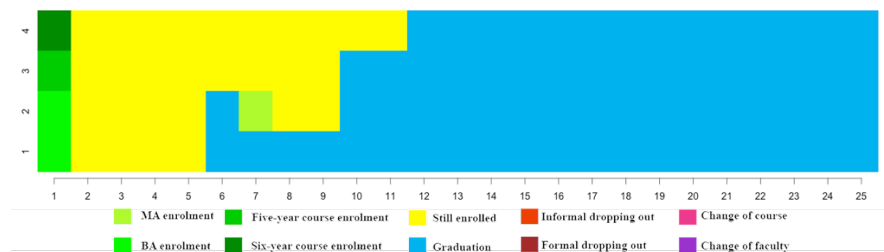
Graph. 1. Ideal index-plot

---

[2] The matrix had about 566 million of cells.

6

The Graph 2 represents the index-plot of the entire 2001-2002 cohort. Contrary to the ideal index plot, there is a preponderance of red color, which indicates dropping out in its two connotations (formal and informal). This outcome concerns particularly the BA students' careers. Although the majority of dropouts are registered in the transition between the first and second year of the study course, many students experience dropping out later (even after many years).
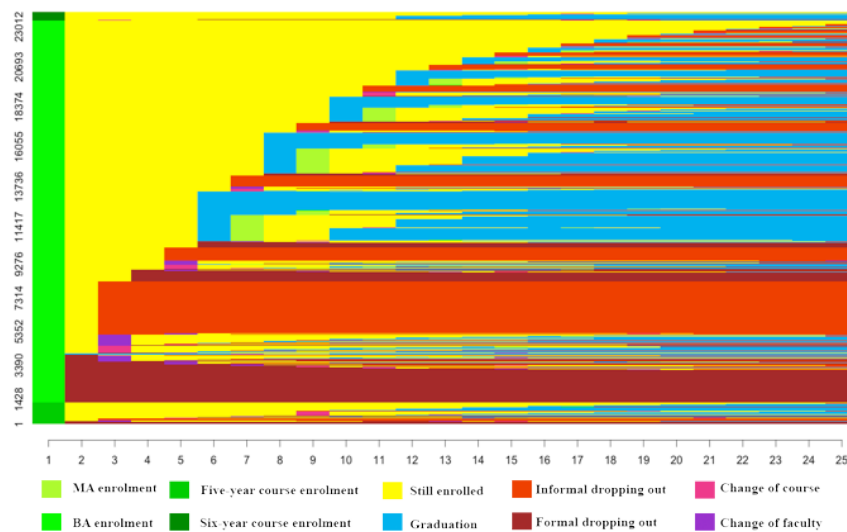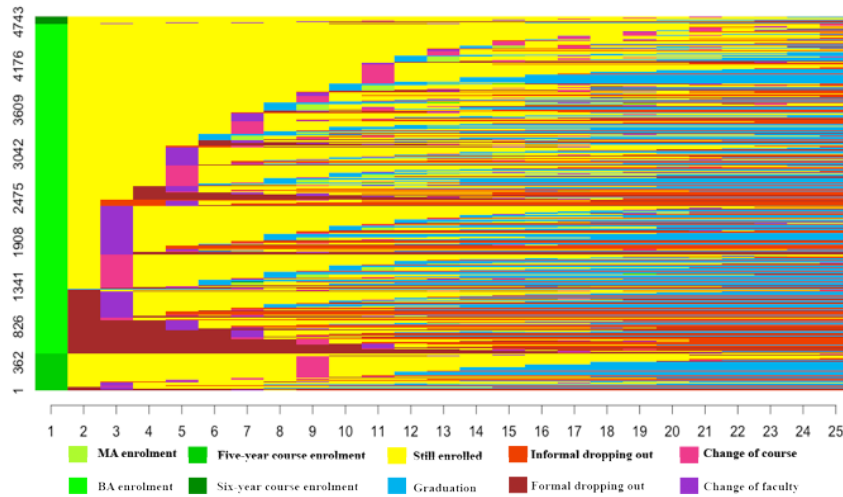

Graph. 2. Index-plot of the entire 2001-02 cohort*
* In the Index-plot each horizontal line corresponds to one career; the events that form the career are represented through different colours (as showed in the legend)

The index-plot of the careers (4,804 students, 20.1% of them enrolled in 2001-02 cohort) with one or more mobility events (Graph 3), shows a much more complex situation. The careers look even more partitioned and it is difficult to highlight trends. Anyway, the states of mobility (change of faculty, marked by purple colour, and change of the course, marked by fuchsia) frequently involve students' careers in the period immediately after enrolment.

The complexity in the right side of the graph, which refers to the end of the observation period and, therefore, to the outcomes of careers, could suggest an interesting association between the time of the change of the faculty or study course and the career outcome. Students who take a reallocation of study course in the years immediately following the enrolment are those that may stand more chances to complete successfully their career.

7

Graph. 3. Index-plot of the mobility students' careers

By splitting the entire cohort into groups according to the study course length (Graph 4), we can observe how the heterogeneity is different among the groups. Whereas in the entire cohort 42.5% of the career is represented by the 10 most frequent sequences (45.8% of the BA careers), for the six-year course students the homogeneity is greater: almost 70% of the careers can be expressed in terms of the 10 most frequent sequences. However, for the five-year course students, the share is close to 50% and thus the homogeneity is more in line with that registered for the BA students.
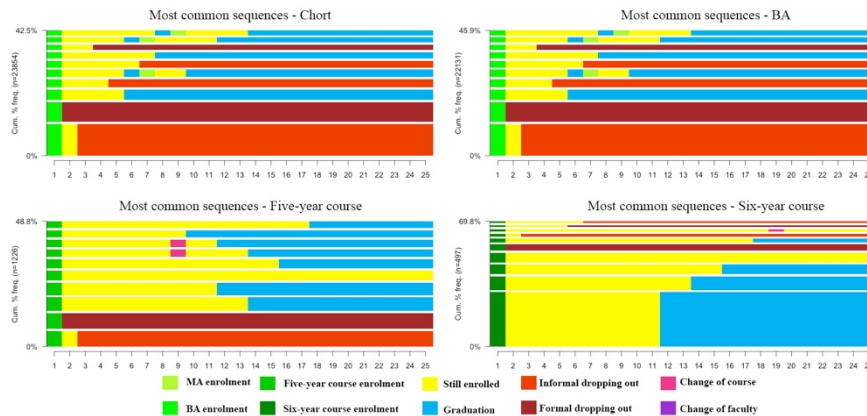
Graph. 4. Most frequent careers of the entire 2001-02 cohort

Among the 10 most frequent sequences (Graph 5) of students' careers with mobility events, the fifth is a discontinuous career (0.77%, 37 cases): this is the case of a BA student who leaves (formally) the university for one semester,

8

changes faculty, there remains registered only for one semester and then drops out permanently (in an informal way). This is a very complex career that might suggest the attempts implemented by the students to pass the limited enrolment course test. Furthermore, similarity of this career with the second one is rather interesting. These pathways differ only for the administrative state during the second half of observation: continuation in one case; formal dropping out in the other.

Formal dropping out is relevant in BA careers: it becomes a definitive state as a result of a change of faculty (ninth most common sequence, 26 cases, 0.61%). The direct comparison among the three subgroups allows us to highlight how the BA students have the worst performance with regard to the retention. Indeed, for these students dropping out state appears in 4 out of 10 most common sequences, while for five-year students it appears only once, and not even once for the six-year students.

The homogeneity among subgroups is radically different. The most frequent sequence for BA students includes only 7.9% of the students; the first one of five-year course students 45.6%; the six-year course students 31.2%. This index plot shows that the courses with limited enrolment test (all the six-year and the majority of the five-year courses) could ensure a better chance for success in university career. Indeed, the students enrolled in longer study courses achieve graduation more frequently than BA students. The only exception among the 10 most frequent sequences is that the sixth career refers to five-year students: the outcome of the study course change is the informal dropping out.

Another interesting element is the consistency of delay (perpetual students) in study courses, especially for the students enrolled in the six-year study courses. Indeed, 3 out of the 10 most common sequences relate to students that are still enrolled after 25 semesters (15 cases, 15.6%). The phenomenon is more restrained for five-year course students (24 cases, 5.1%).
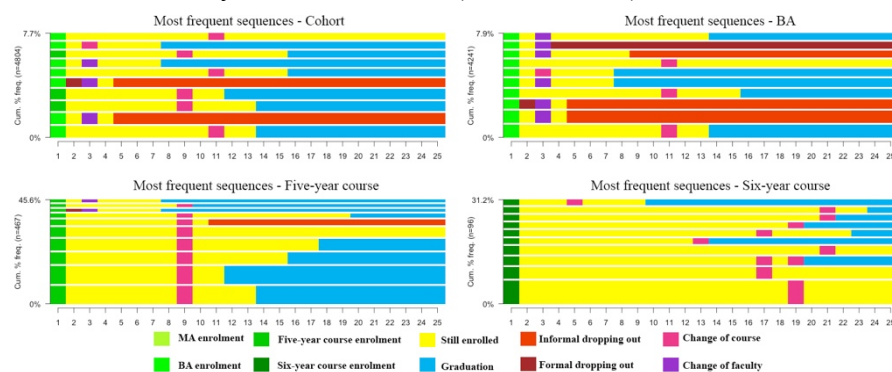
Graph. 5. Most frequent careers of the mobility students

Through cluster analysis (Graph 6), performed on the OM distance matrix, six well-defined groups of students were identified. The majority (35,5%) of the

9

first group are the students that are still enrolled at the end of the observation period (25 semesters); part of them are drop outs (informal, 32,3%, and formal 8,3%) or graduates after 10 years since the enrolment (22,2%). The second and the fourth groups differ for the dropping out timing: the students of the second group keep postponing the dropping out decision until the 9th semester, whereas for the fourth group the dropping out occurs in the first half of the observation period. The second type of dropping out (the formal one) characterises the sixth group. In other words, the majority of these students drops out at the beginning of their career. The other two groups (the third and the fifth) are composed of students that end their career with graduation. Although the students of the third group graduate at the regular course length, the timing of graduation in the fifth group is slower.
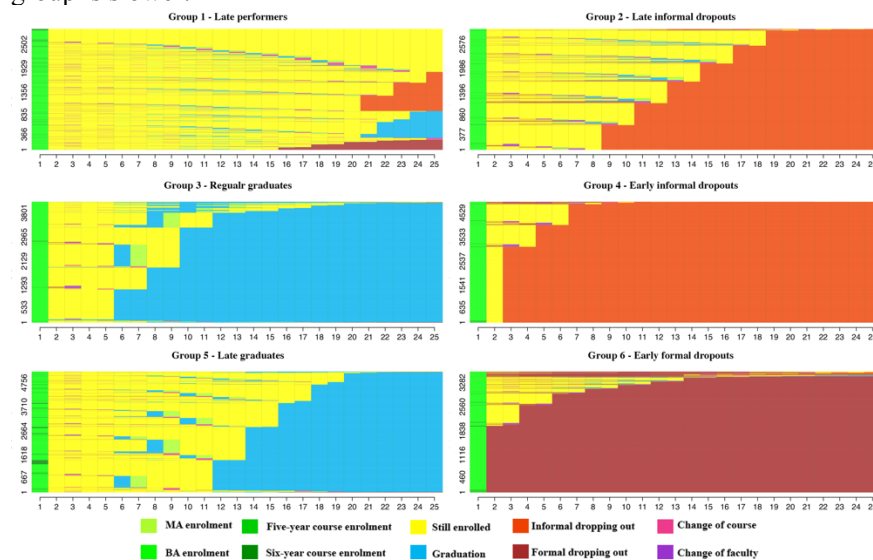
Graph. 6. Index-plot for the six groups, cohort 2001-02

These six groups are used as dependent variables in a multinomial logistic regression model (ML) (Table 1). The ML includes two types of independent variables: micro-level variables [social characteristics of students (gender, age at enrolment, residence, family income), high school information (high school type and high school final mark) and first year performance indicators (credits and average mark)]; meso-level variables [university context information (length of first enrolment study course and limited enrolment course)].
The main result shows the importance of the warming-up in determining the development of each career. The first year performance indicators are the best predictors of success. Furthermore, the university context information, especially the length of first enrolment study course, are important: students enrolled in a 5 or 6 year-course stand more chances to graduate than BA students.

Table 1. Odds ratio of ML for entire cohort 2001-2002

		Early informal dropouts		Late performers		Regular graduates		Late graduates		Early formal dropouts		Late informal dropouts	
		OR	sig.	OR	sig.	OR	sig.	OR	sig.	OR	sig.	OR	sig.
	Constant	0.673	***	0.048	***	0.014	***	0.051	***	0.497	***	0.077	***
Gender	Male	1.245	***			0.823	***					1.097	.
Age at enrolment	20-24 years	1.814	***	0.826	***	1.424	***	0.672	***	0.751	***	1.403	***
	25 year-courses or more	2.681	***	0.735	***			0.352	***	0.596	***		
High school type	Technical-Professional	1.499	***	0.845	**	1.224	***	0.773	***	0.909	.		
	Other	1.500	***			1.357	***	0.732	***				
High school final mark	Medium	0.889	*			0.895	.	1.182	***	1.183	***	0.870	*
	High	0.713	***	0.852	*	1.157	*	1.158	*	1.335	***	0.655	***
Residence	Regional	0.756	***			1.376	***						
	Out of region	0.685	***							1.627	***	0.771	***
Family income	1-10000 euro			1.305	**							1.195	.
	10001-20000 euro			1.337	***								
	20001-30000 euro			1.313	***								
	30001-40000 euro			1.189	.								
	40001 euro or more	0.836	.					0.864	.	0.852	.		
Length of study course	5 years	0.758	*	1.643	***	0.160	***	2.620	***	0.711	***		
	6 years	0.295	***	1.854	***	0.013	***	5.081	***	0.571	*	0.306	***
Limited enrolment course	Yes	0.788	***			1.239	***	0.836	***	1.255	***		
First year credits	1-15 Ects	0.705	.	3.184	***					0.570	*	2.588	***
	16-30 Ects	0.234	***	3.920	***	5.292	*	4.238	***	0.206	***	2.949	***
	31-45 Ects	0.067	***	2.426	***	15.837	***	6.204	***	0.083	***	1.769	.
	46 Ects or more	0.025	***			36.346	***	4.836	***	0.028	***		
First year average mark	Low	0.648	*					2.244	.				
	Medium	0.582	*					2.447	.				
	High	0.586	*										

*Sig. Code: 0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1*

Through the use of the same clustering algorithm, performed on the OM distance matrix related to careers with one or more states of mobility, we obtain similar groups (Graph 7). Obviously, a greater complexity of careers, which has already been highlighted in the Graph 3, is also present in these subgroups. The discontinuity of career of the sixth group (early formal dropout) is very blatant. The index-plot for this group (Graph 7) highlights how the formal dropping out is very widespread. This phenomenon affects the students' career in a different way respect to the informal dropping out: many students, after formally dropping out, enroll again at Sapienza, maybe in a different faculty or study course.
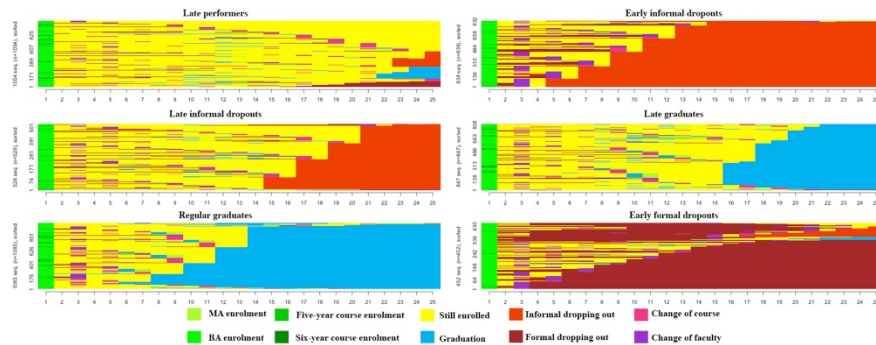
Graph. 7. Index-plot for the six groups, mobility students' careers

Through the studying of the mobility among faculties/courses as a population migration and thinking of the institution (Sapienza) as a multitude of elements, variously connected among each other (Mohr and White 2008), the Social Network Analysis (SNA) tools are used to identify structure characteristics of the mobility (Graph 8).

In order to study the relevance of the connections among the events of each trajectory (such as within and between mobility), the OM analysis was performed on the careers with one or more mobility events. With the use of the previous clustering algorithm, we obtained similar groups. The SNA was used to identify attraction and repulsion nodes among the Sapienza study courses.
The network among study courses allows us to visually observe how the core (stronger connections) is mainly occupied by the BA courses (in purple). The MA courses (in red) are arranged radially around the central core. This position is due to the links that some of these courses have with other different duration courses, particularly five-year courses. For instance, a well-defined group is composed by the six-year courses. These nodes are only the degree programmes in Medicine, which has a six-year term. The proximity of these courses to the rest of the network is given by the relations with some Pharmacy courses and some Mathematical, Physical and Natural Sciences courses (MMPPNNSS).

12

This analysis allowed us to identify specific study courses (such as Life Sciences, BA course in MMPPNNSS faculty, and Pharmacy, five-year course) used by students to circumvent the admission tests of medicine courses.

Five-year courses (15 in total) do not resemble among each other, but they appear to be divided into small groups, and they are more similar to other BA courses, which partially share educational curricula: for example, Construction Engineering (BA) is similar to Architecture course (five-year course).
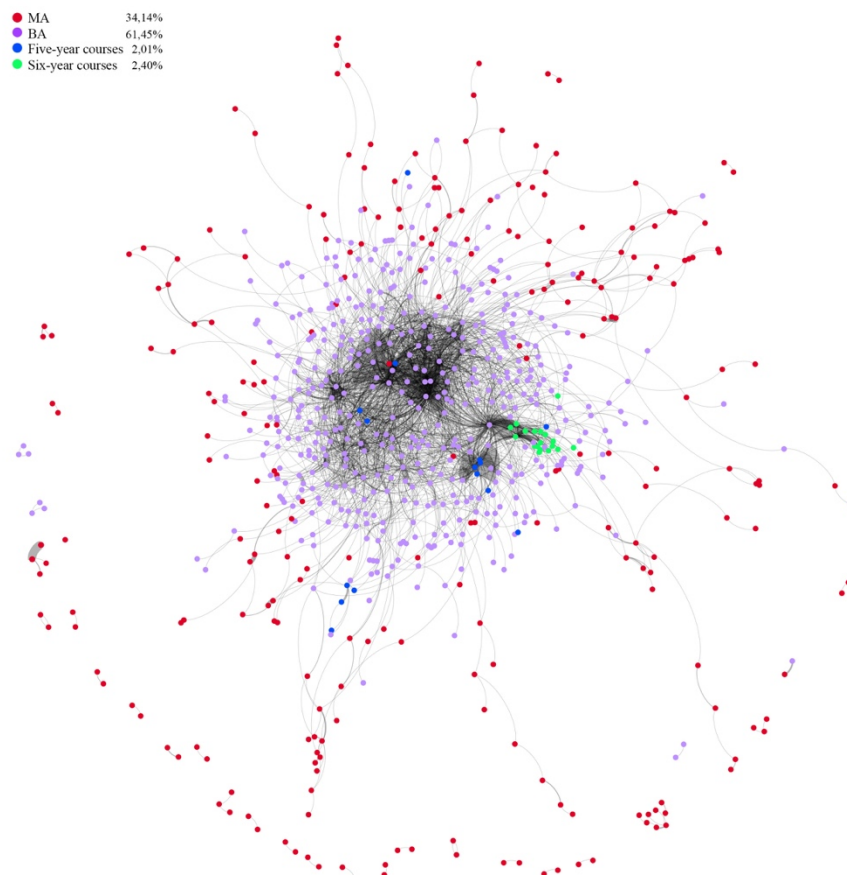
Graph. 8. Change of study course according to legal course length, entire cohort 2001-2002

A second ML aims to explain the outcome through temporal and spatial information. The dependent variable in the analysis is the career outcome: still enrolled (reference category), graduate and dropout. The dependent variables are all features related to the ongoing changes: time related (timing of change; difference in course length) and context related (type of change; difference in course scientific area). The timing of change is identified as the most relevant

13

success predictor, confirming that the warming-up period is fundamental for the career outcome.

Table 2. ML for mobile students' careers of BA*

Outcome[a]			T	Std. Err.	Wald	gl	Sign.	Exp(B)	Exp(B)-1*100
		Intercept	0.3	0.566	0.28	1	0.596		
Degree	Timing of change	2th year	3.308	0.202	266.900	1	.000	27.323	2632.303
		3th year	2.753	0.219	158.568	1	.000	15.683	1468.343
		4th year	2.487	0.269	85.537	1	.000	12.026	1102.604
		5th year	2.523	0.319	62.521	1	.000	12.468	1146.782
		6th year	2.194	0.236	86.208	1	.000	8.975	797.509
		Beyond the 6th year	0[b]			0			
	Type of change	Faculty change	-0.183	0.203	0.812	1	0.368	0.833	-16.724
		Study course change	0[b]			0			
	Difference in course length	Equal length	-0.740	0.520	2.021	1	0.155	0.477	-52.271
		Upper	-1.535	0.536	8.203	1	0.004	0.216	-78.448
		Lower	0[b]			0			
	Difference in course scientific area	Same sc. area	0.387	0.21	3.401	1	0.065	1.473	47.3
		Different sc. area	0[b]			0			
Dropout		Intercept	-0.241	0.566	0.182	1	0.67		
	Timing of change	2th year	1.813	0.194	87.591	1	.000	6.128	512.799
		3th year	1.518	0.210	52.063	1	.000	4.561	356.082
		4th year	1.764	0.260	46.060	1	.000	5.834	483.403
		5th year	1.882	0.316	35.561	1	.000	6.569	556.886
		6th year	1.683	0.243	48.163	1	.000	5.382	438.178
		Beyond the 6th year	0[b]			0			
	Type of change	Faculty change	1.059	0.199	28.178	1	.000	2.883	188.349
		Study course change	0[b]			0			
	Difference in course length	Equal length	0.007	0,521	,000	1	0,989	1.007	0.735
		Upper	-1.674	0.541	9.558	1	0.002	0.188	-81.244
		Lower	0[b]			0			
	Difference in course scientific area	Same sc. area	0.509	0.205	6.156	1	0.013	1.664	66.354
		Different sc. area	0[b]			0			

a. Reference category: Still enrolled.
b. This parameter is zero due to redundancy
* the model refers only to the BA students and at the first career.
Cox e Snell pseudo R^2 0.187; Nagelkerke pseudo R^2 0.222; -2log final model 412.917 sig. <0.001

14

Firstly, the model parameters show that the timing of change is crucial: the chance to graduate rather than to remain registered is higher as soon as the change occurs.

The type of change (course or faculty) is not significant with reference to the probability of graduation or to the fact of still being enrolled, but it is related to the probability of dropping out: a change of faculty nearly doubles the probability of dropping out rather than being still enrolled with respect to a change of study course (Exp (B) -1 * 100 = 188.3). Therefore, we can confirm the assumption that a strong reorientation (which determines the faculty change) requires a higher investment compared to a weak reorientation (such as a change of study course within the same faculty), exposing more students at risk of dropping out.

Mobility between equal-length courses seems not to have a significant influence neither on the probability of graduating nor dropping out, compared to the probability of persistence inside the system. On the contrary, the "upper" exchange, i.e. the course change from a shorter to a longer one, reduces both the probabilities of graduation and dropping out, rather than the continuation (respectively with Exp (B) -1 * 100 = -78.4 for the graduation and -81.2 for the dropping out. This effect could be determined simply by the increased duration of the second course of study. However, with reference to the lower probability of dropping out, this effect could hypothetically also be determined by a strong motivation of these students.

Any difference between the scientific area of the original course and the scientific area of the second one has a double effect on the outcome: students who change course or faculty within the same scientific area are more likely both to graduate or to drop out rather than to continue (Exp (B) -1 * 100 47.3 for the graduation and 66.3 for the dropping out). However, only the dropping out effect is statistically significant. This evidence shows that the mobility between different scientific areas represents a higher chance of still being enrolled than the changes within the same area.

Summary and conclusions

This paper shows the value of transforming synchronic administrative information into diachronic data vectors. Throughout this transformation, we were able to obtain a longitudinal dataset of students' careers and to investigate their dynamics by highlighting specific phenomena (dropping out, stopping out, mobility, degree attainment, etc.) and the connections among them.

The longitudinal structure allows us to study all the processes characterised by continuous changes, such as university careers. The students' careers complexity was examined in detail with the use of SA. The main results show the importance of the warming-up period (in terms of study course reorientation and first year performance) in determining the success/failure of each career.

15

The SA and the SNA have allowed us to look into the events related to the university dispersion. For instance, these tools have allowed us to reveal how, almost all the changes, related to the medical courses take place in the early part of the career. In most cases, these changes are strategies applied by the students who fail the first time medical limited enrolment course test, and try again in subsequent years.

All this will lead to improper management of the resources by the University: while the medical courses are becoming increasingly competitive (students who attend medical test are not high school graduates but students already included in the university context), those of Pharmacy and MMPPNNSS courses constantly lose students.

The analysis is a pilot methodological study but it can easily be used for evaluation purposes. Universities, faculties or degree programs can adopt these tools in order to monitor the career dynamics (monitoring purpose) and to promptly intervene in order to contain the dispersion phenomena (action purpose).

References

Benvenuto, G., Decataldo, A., & Fasanella, A. (Eds.) (2012). *C'era una volta l'università? Analisi longitudinale delle carriere degli studenti prima e dopo la "grande riforma"*. Acireale–Roma: Bonanno.

Blanchard, P., Bühlmann, F., & Gauthier, J.A. (Eds.) (2014). *Advances in Sequence Analysis: Theory, Method, Applications. Series: Life Course Research and Social Policy (Vol. 2)*. New York, NY: Springer.

Blossfeld H.P., Hamerle A., & Mayer, K.U. (1989). *Event History analysis*. Hillsdale, NJ: Erlbaum.

D'Alessandro, G. (2015). *Proposte per lo studio della mobilità studentesca e l'analisi di dati longitudinali*. PhD thesis in Methodology of Social Science. Department of Communication and Social Research. Rome: Sapienza University of Rome.

Gabadinho, A., Ritschard, G., Müller, N.S., & Studer, M. (2009). Mining Sequence Data in R with the TraMineR package: A user's guide. Department of Econometrics and Laboratory of Demography, Geneva, University of Geneva. http://mephisto.unige.ch/pub/TraMineR/doc/1.4/TraMineR-1.4-Users-Guide.pdf. Accessed 19 March 2016.

Gabadinho, A., Ritschard, G., Müller, N.S., & Studer, M. (2011). Analyzing and Visualizing State Sequences in R with TraMineR. *Journal of Statistical Software*, 40(4), 1–37. doi:10.18637/jss.v040.i04.

McCormick, A.C., & Carroll, C.D. (1997). Transfer behaviour among beginning postsecondary students: 1989–1994. Department of Education, National Center for Education Statistics, ERIC Document Reproduction, Washington. http://eric.ed.gov/?id=ED408929. Accessed 11 December 2014.

MacIndoe, H., & Abbott, A. (2004). Sequence Analysis and Optimal Matching Techniques for Social Science Data. In M. Hardy, & A. Bryman (Eds.), *The Handbook od Data Analysis* (pp. 307–406). London–Thousand Oaks–New Delhi–Singapore: Sage.

Menard, S. (Ed.) (2007). *Handbook of longitudinal research: design, measurement, and analysis*. Amsterdam: Elsevier.

16

Mohr, J.W., & White, H.C. (2008). How to model an institution. *Theory and Society*, 37(5), 485–512. doi: 10.1007/s11186-008-9066-0.

Moscati, R., & Vaira, M. (Eds.) (2008). *L'università di fronte al cambiamento*. Bologna: il Mulino.

Ritschard, G., Gabadinho, A., Studer, M., & Müler, N.S. (2009). Converting between various sequence representations. In Z. Ras, & A. Darzinska (Eds.), *Advances in Data Management*, Series Studies in Computational Intelligence, vol. 223 (pp. 155–175). Berlin: Springer.

Scherer, S. (2013). *Analisi dei dati longitudinali. Un'introduzione pratica*. Bologna: il Mulino.

Studer, M. (2013). WeightedCluster Library Manual: A practical guide to creating typologies of trajectories in the social science with R. *LIVES Working Papers*, *24*, 1–32. doi:10.12682/lives.2296-1658.2013.24.

Studer, M., & Ritschard, G. (2014). A comparative review of sequence dissimilarity measures. *LIVES Working Papers*, *33*, 1–47. doi:10.12682/lives.2296-1658.2014.33.

The age of reproduction. The effect of university tuition fees on enrolment in Quebec and Ontario, 1946–2011

Benoît Laplante[1], Pierre Doray[2] and Nicolas Bastien[2]

[1] Centre Urbanisation Culture Société, Institut national de la recherche scientifique

[2] Centre interuniversitaire de recherche sur la science et la technologie, Université du Québec à Montréal

Abstract We are interested in the role tuition fees play in social reproduction. We use retrospective biographical data from a series of surveys on family events, long series on tuition fees and methods from event history analysis to study the effect of the level of tuition fees on university enrolment in two Canadian provinces, Quebec and Ontario. We focus on the variation of their effect on enrolment according to social origin, province, language and immigration status. Not considering age, the level of tuition fees increases enrolment for children of highly educated parents or immigrants, has no effect among the Quebec English-speaking, but decreases enrolment in all other groups. However, in most groups, the deterring effect increases with age. Among immigrants and their children, as well as among the Ontario English-speaking, the slope of the relation between the effect of tuition fees and age is markedly steep: In these groups, there seems to be a limited age window during which parents are willing to invest in their children's education.

1 Introduction

The recent raise of tuition fees in many Western countries – among them England, the United States, Germany and Canada – is renewing interest for the effect of tuition fees on the access to postsecondary education, and especially their role in social reproduction. The most obvious topic is whether increasing tuition fees reduces overall access, but a more complex question is whether the effect of tuition fees on enrolment is the same across social groups. Higher tuition fees could discour-

age the less wealthy to pursue higher education because of the cost and of a pessimistic estimation of future gains, but they could encourage the wealthier to enrol because cost is not a problem and future gains believed to be high thanks to reduced competition. If this were true, the combination of two opposite effects could lead to a null or only very negative effect of increasing tuition fees on overall enrolment. In other words, knowing more about the variation of the effect of tuition fees on enrolment across social groups is important for understanding the role of postsecondary education in social reproduction, but also for understanding the evolution of overall enrolment.

In this article, we examine the variation of the effect of tuitions fees on enrolment across social groups in two Canadian provinces over a 65-year period. We model the effect of tuitions fees on university access using individual biographical data from four cycles of the General Social Survey and aggregate data on Quebec and Ontario tuition fees spanning from 1946 to 2009. Our approach makes use of the non-monotonic variation of tuition fees over the period to disentangle the effect of tuition fees on enrolment from the general growth in enrolment that occurred over the period. It also allows us testing the equality of the effect of tuitions fees across social origin, language and immigration status. Our results show that the level of tuition fees has an impact on enrolling, and that this impact varies across social groups.

2 University education and social reproduction

Studies on the access to postsecondary education and, more generally, the sociology of education inequality as a whole, draw strongly on social reproduction theory. At the core of this view is the notion that despite the democratisation of postsecondary education, schooling is a key element of the intergenerational reproduction of the social structure. Since the seminal work on this topic (Bourdieu and Passeron, 1977, 1979; Bourdieu, 1984), numerous studies in numerous countries have used this perspective and benefited from its fruitfulness.

In its basic formulation, social reproduction theory focuses on social origin and family location in the class structure. Some extensions have proven important. Gender is now recognised as a structural difference of its own that induces differences in education pathways. While women have gained greater access to postsecondary education, their options still bear the mark of the gendered division of labour. (Duru-Bellat, 2001; Marry, 2004; Barone, 2011). In some societies, ethnicity, religion or language are additional components of the social structure. In the USA, race induces differences in education careers (MacLeod, 1987; Ogbu, 1994). In Canada, language – specifically, speaking French or English – is another source of social inequality, especially in education (Dandurand et al., 1980; Dandurand, 1986 et 1991; Laplante et al. 2014; Kamanzi and Doray, 2015). In Quebec, Dandurand was a precursor of a sort. As early as in the 1980s, he promoted that the

3

process of social reproduction was not driven only by social origin, but by gender and ethno-cultural differences as well. This insight is very close to what is now known at intersectionality (Adamuti-Trache and Andres, 2008; Collins and al., 2015; Beattie, 2002).

In contemporary Canada, immigration adds another dimension to the social structure. Canadian immigration policy overtly favours university graduates, with the consequence that the proportion of university graduates among immigrants is very high, especially among the older generations. Thus, a large fraction of children of immigrants is made of children of university graduates (Laplante et al. 2014).

3 Tuition fees as education policy

Tuition fees are an element of education policy. Assuming they reduce affordability and enrolment in university education, States may regulate them as a means of promoting accessibility (CMEC, 2007; Delaney and Kearney, 2015), This is, or has been, common in Europe. In North America, Quebec followed the European trend as part of a large-scale reform enacted in the 1960s. Tuition fees were 'frozen', which, in a context of moderate to high inflation, amounted to reduce them over time, this reduction being presented at the time as leading to their elimination in the near future. However, in Quebec as in many other jurisdictions, the will to curb public deficits and public spending more generally actually lead to reduce the public funding of universities and to increase their private funding by students and their families through increasing tuition fees. In Quebec as elsewhere, advocates of increasing tuition fees assured that higher tuition fees would not reduce accessibility because in jurisdictions where they had increased, for instance in neighbouring Ontario, enrolment had not decreased. Furthermore, as university graduates get higher wages because of their education, they should pay a large fraction of the cost of their education (e.g. Belzile, 2010; Kozhaya, 2004).

Not surprisingly, insights from scholarly research sometimes differ from those of advocacy. There is a large body of American literature on the effect of tuition fees on enrolment in postsecondary education. According to the meta-analysis by Leslie and Brinkman (1987), youth enrolment is more affected by fees than by financial help whatever its form. According to the studies they reviewed, increasing tuitions by 100 USD decreased enrolment by 0.7% among youth aged 18 to 24 and students from low-income families were more affected by changes in tuition fees. Heller (1997) updated Leslie and Brinkman's meta-analysis by including 15 more studies and drew the same conclusions. More recently, Hemelt and Marcotte (2008) found similar results.

The recent introduction or increase in tuition fees in some European countries has fostered new research. In the UK, Dearden, Fitzsimons and Wyness (2011) found that increasing tuition fees by 1,000 GBP was associated with a 3.9% de-

crease in enrolment among youth ages 18 to 29. Croxford and Raffe (2012) less students from Scotland and Northern Ireland in English universities, those who come from tend Scotland attend elite universities. Wales (2013) found similar results in England: increasing fees by 10% was associated with a 1.7% decrease in enrolment. Bolton (2014) reports that applications have dropped in England, but not in Scotland. Universities UK (2014) report lower enrolment especially in part-time studies. Sá (2014) found a decrease in applications especially for programs that lead to lower salaries and lower employment rates, but no evidence of a larger reduction for students from disadvantaged backgrounds.

In Germany, the decision to implement tuition fees belongs to the *länder*, which created a natural experiment. Taking advantage of this setting, Hübner (2012) found a negative effect of tuition fees on enrolment, larger than in previous studies for European countries, but similar to those reported for the USA. Dietrich and Gerner (2012) found that the introduction of tuition fees reduced the propensity of high school graduates to enrol at a university and favoured the vocational training option, whereas Bruckmeier and Wigger (2014) found no evidence that the introduction of tuition fees had a general negative effect on enrolment.

During the 1990s, as advocates of higher tuition fees gained influence on education policy, tuition fees increased simultaneously in most Canadian provinces. This stimulated Canadian research on the effect of tuition fees on enrolment. Most Canadian studies found that tuition fees have an impact on enrolment (Hui, 2004; Michael, 1999; Coelli, 2004; Johnson et Rahman, 2005; Neil, 2009; Hansen et Liu, 2013). Several had mixed findings (Corak et al, 2003; Drolet, 2005; Finnie et al., 2004; Frenette 2005). These studies focused on the 1990s, the period of widespread increase. They looked at the evolution of enrolment in a time where tuition fees are increasing, but they did not estimate the effect of tuition fees on enrolment. Still, Finnie et al. (2004) and Frenette (2005) found that increasing tuition fees had a greater impact on people from more vulnerable background. Moreover Frenette (2005) found increasing inequality in the access to university education in provinces where the deregulation had been more thorough: among people from low-income families, access to university education decreased more in provinces where tuition fees for programs leading to organised professions such as medicine, dentistry and accounting have been deregulated and thus increased the most. Two studies concluded that there is no relation between tuition fees and enrolment. Compared to the bulk of the studies, these two are outliers. One focused on a very short period (Rivard and Raymond, 2004), the other on a period during which tuition fees were almost stable (Christofides, Cirello and Hoy, 2001).

4 Tuition fees and enrolment in Quebec and Ontario

We are interested in the effect of tuition fees on enrolment and more specifically in the differences in their effect across social groups defined by social origin, lan-

5

guage and immigration. Comparing the effect of tuition fees in Quebec and in Ontario allows contrasting two societies that are reasonably similar, but whose governments have conducted different policies on tuition fees over the years (see below) and where language shapes the social structure in contrasted ways. The proportion of university graduates is the same within the French-speaking Quebec majority and within the English-speaking Ontario majority. However, in Quebec, the English-speaking minority is still concentrated in the upper classes and has a higher proportion of university graduates than the French-speaking majority, whereas the Ontario French-speaking minority is 'underrepresented' in the upper classes and has a lower proportion of university graduates than the English-speaking majority (Laplante et al. 2014).

Comparing tuition fees and enrolment rates in Quebec and Ontario to understand the relation between fees and enrolment may be tricky. First, the education systems are different. In Quebec, students graduate from secondary education after 11 years, go to college for two years and then may enter university where a typical program lasts three years. In Ontario, students go directly from high school to university, but a typical undergraduate program lasts four years. Second, Ontario secondary education underwent an important change in the 2000s: in 2003, the number of years students spent in primary and secondary education has been reduced from 13 to 12. Therefore, the number of students who enrolled into university in 2003 was about twice the number of the previous and the next year. Third, although most young Ontarians who enrol into university tend to do it in their early 20s, young Quebeckers tend to spread enrolment into university over all their 20s (Chenard and Doray, 2013). Because of this difference in timing, comparing Quebec and Ontario using enrolment rates computed for people aged between 18 and 24 excludes a significant portion of Quebec enrolment. Finally, there is no reason to limit the comparison to the 1990s and 2000s. Tuition fees have been an important part of education policies at least since the beginning of the expansion of university education in the 1960s and there are data that allow reconstructing the series on tuition fees and enrolment rates from this period.
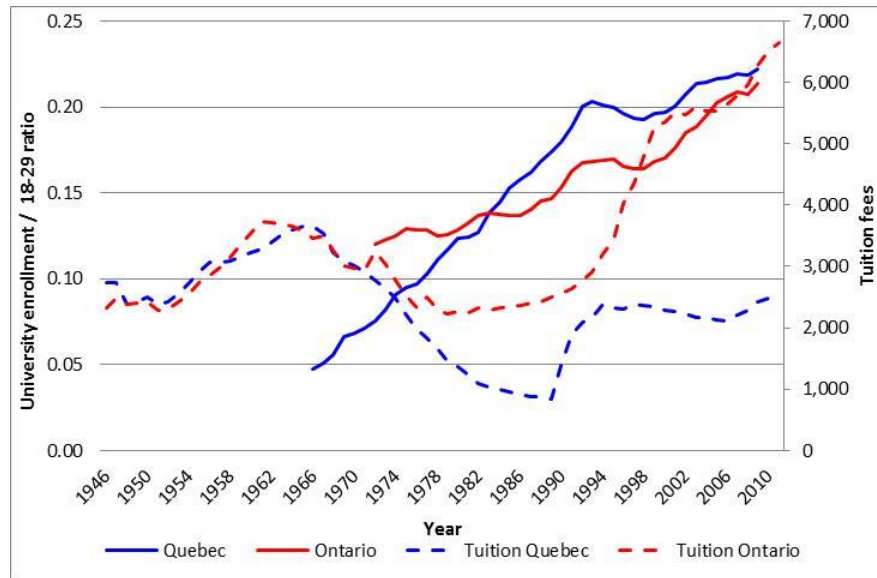
Fig. 1: Ratio of university enrolment to population aged 18–29, Quebec 1966–2009, Ontario 1972–2009, and average tuition fees, Quebec and Ontario 1966–2009 in 2011 constant dollars.[1]

Figure 1 graphs enrolment rates – defined as the proportion of the population aged 18–29 which attends university in a given year – and tuition fees – measured in 2011 constant dollars – in Quebec and Ontario from 1946 to 2011. From the beginning to the end of the series, enrolment rates grew in both provinces, from 4.7% to nearly 22.2% from 1966 to 2009 in Quebec, and from 12.0% to 21.4% from 1972 to 2009 in Ontario. The growth was not steady over the whole period. In both provinces, the enrolment rates levelled off or decreased slightly in the 1990s, and did not resume their previous trend until the end of the decade.

Over the same period, tuition fees in constant dollars did not follow a simple trend, but rather waxed and waned, conveying changes and continuities in education policies. In Quebec, they decreased from 1968 to 1989, actually remaining constant in current dollars, implementing a recommendation of the 'Commission

---

[1] Source. *Tuition fees*: Survey of higher education 1946–1948, 1950–1952, 1952–1954, Dominion Bureau of Statistics; University student expenditure and income in Canada, 1961–1962 , Dominion Bureau of Statistics; Tuition and living accommodation costs at Canadian degree-granting universities and colleges, 1966–1967, 1967–1968, 1968–1969, 1969–1970, 1970–1971, Dominion Bureau of Statistics; Tuition and Living Accommodation Costs for Full-time Students at Canadian Degree-granting Institutions , Statistics Canada, 1972--2011.
Population aged 18–29: Statistics Canada, Table 051-0001 & 051-0026 – Estimates of population, by age group and sex, Canada, provinces and territories, CANSIM.
Enrolment: Association of Universities and Colleges of Canada.
Note: Ontario enrolment figures have been smoothed to reduce the impact of the 2003 "double cohort".

7

royale d'enquête sur l'enseignement dans la province de Québec' (Royal Commission of Inquiry on Education in the Province of Quebec). Over that period, in 2011 constant dollars, they went down from nearly 3,700 CAD to about 800 CAD. The Quebec government increased tuition fees in steps from 1990 to 1994, raising them to nearly 2,400 CAD. They decreased anew in constant dollars between 1995 and 2006. From 2007 onwards, they increased steadily by 100 CAD each year, thus getting back in 2009 to what they were in 2004. The 2012 student protest followed that intent to increase further tuition fees so that, by 2016, they would have reached the level they were in 1968.

In 1966, the average tuition fees, in Ontario, were 3,500 CAD, slightly lower than in Quebec. The Ontario government kept tuition fees under strict control until the 1990s. They decreased in constant dollars from the late 1960s to the mid-1970s, and then levelled off until the beginning of the 1980s. They increased slowly through the 1980s, and sharply from the beginning of the 1990s onwards. By the end of the 2000s, they were closing to 6,300 CAD, about two and a half times those of Quebec.

Enrolment rates were increasing before the government of both provinces increased tuition fees in the 1990s. Soon after the increase in tuition fees, enrolment rates levelled off for a few years. Kozhaya (2004) dismisses that the Quebec increase in the tuition fees of the 1990s may have caused enrolment rates levelling off afterwards because the fees were not increasing when the enrolment rates levelled off. The argument is weak. The increase was sudden and substantial, and changed markedly the cost of university education. What happened in the following years is what could have been expected: breaking the increasing trend of university enrolment.

More interesting is that the effect of the increase in tuition fees was not to push down enrolment rates, but rather to stop their increase. Still more interesting is that they started increasing again a few years later. This should not come as a surprise. For an individual, enrolling into university is a decision that involves a variety of factors, among which tuition fees. At the aggregate level, however, the enrolment rate is a function of education policies not only on tuition fees, but also on the sustained increase of the supply of university education, and as well, of a strong trend in the economy towards increasing demand for university graduates. Enrolment rates levelling off soon after tuition fees had increased gives the naked eye a glimpse into the difference between the secular trend in increasing enrolment rates that depends on the transformation of the economy and the more mundane relation between tuition fees and enrolment rates.

Table 1. Descriptive statistics

	Quebec	Ontario
Proportion enrolling into university by cohort		
Before1936	0.044	0.054
1936–1950	0.199	0.170
1951–1974	0.560	0.524
1975–1990	0.197	0.252
Independent variables		
Cohort		
Before1936	0.126	0.122
1936–1950	0.204	0.184
1951–1974	0.465	0.471
1975–1990	0.205	0.222
Sex		
Male	0.492	0.489
Female	0.508	0.511
Sociolinguistic group		
Quebec French-speaking	0.924	
Quebec English-speaking	0.033	
Quebec immigrants	0.043	
Ontario English-speaking		0.687
Ontario French-speaking		0.054
Ontario immigrants		0.259
Social origin		
No PSE	0.691	0.629
Non-university PSE	0.157	0.166
University	0.152	0.206
Size		

Data from cycles 10, 15, 20 and 25 of the General Social Survey. Weighted estimation.

5 Hypotheses

Most studies on the effect of tuition on enrolment in higher education find some negative effect. Some studies found the negative effect to be more important among children of less favourable background, such as those who come from low-income families, have low-educated parents or live in lone-parent families. We are interested in the variation of this effect across social groups. We look at variation across social origin, as in some of the studies we reviewed, but also across groups defined by province, language and immigration status. As we explain in the previous section, there are reasons to believe that the effect of tuition fees may vary

across the traditional dimensions of social structure, at least in Quebec. The political turmoil associated with attempts at increasing tuition fees, larger in Quebec than in the rest of Canada, could be a consequence of different effects of tuition fees on enrolment in the groups that comprise the population in each province. Specifically, if tuition fees were a stronger barrier among French-speaking Quebeckers than among English-speaking ones, and if immigrants and their children – who account for a large fraction of Ontario population – were not sensitive to tuition fees in the same way as French-speaking Quebeckers are, the policy consequences of increasing tuition fees, as well as the political aspects of implementing such increase, could be very different in the two provinces. If it were so, the differences in the politics of tuition fees between Quebec and Ontario – and maybe between Quebec and the rest of Canada – could be accounted for by different effects of tuition fees across social groups, and the relative importance of these social groups in the population of each province.

6 Data and method

Verifying our hypotheses requires disentangling the effect of tuition fees from that of the general growth in enrolment that was part of the massification of higher education, but also modelling this effect in a way that allows it to vary across social groups defined by parents' education, language, province and immigration status. Furthermore, enrolling is an event that occurs or not over the life course, which itself requires proper modelling.

To disentangle the effect of tuition fees on enrolment from that of growth in enrolment, we use long series on tuition fees in which these vary non-monotonically, whereas we model the secular growth in the enrolment rate monotonically. We use event history analysis (i.e. survival models for the social sciences) to model enrolling as a life course event at the individual level. We model the effect of tuition fees conditional on parents' education and on membership in groups defined by province, language and immigration status. Finally, we allow the effect of tuition fees to vary across the life course – i.e. over age – within each of the groups we are considering.

6.1 Data

Our study requires data on persons who may have enrolled into university between 1946 and 2011, and on tuition fees in Quebec and in Ontario for the same period.

Ideally, we need individual data collected from a population or a probabilistic sample: age at enrolment in university studies, date of birth, province of birth, province of residence at the time of enrolment and a series of sociodemographic

characteristics including mother tongue. There are several sources of individual data on education in Canada, but none includes all the data we need. We use data from four 'cycles' – 1995, 2001, 2006 and 2011 – from a retrospective survey on family which Statistics Canada carries out every five years or so since 1995 within the General Social Survey (GSS) program. Each cycle uses a probabilistic sample of the Canadian population aged at least 15 and living in a province. These surveys are not designed to gather information on education, but they do collect most of the information we need. They have already been used for studying education dynamics in Canada by McIntosh (2009), and also by Sen and Clemente (2010) and Turcotte (2011) who did it combining data from several cycles.

Our data on tuition fees for the 1972–2011 period come from Statistics Canada's survey on 'Tuition and Living Accommodation Costs for Full-Time Students at Canadian Degree-Granting Institutions' (TLAC). These data are the average, weighted by enrolment figures and expressed in 2011 constant dollars, of the tuition fees paid by students enrolled full-time in undergraduate programs in each province. They are the best available estimates of the average tuition fees. For the 1946–1971 period, we gathered figures from publications of the Dominion Bureau of Statistics. These publications provide average tuition fees by year and by province, sometimes detailed by universities. We have not been able to gather complete annual series and have completed the series by interpolation within each province before converting all values in 2011 constant dollars. We provide references to these publications with Figure 1.

6.2 Method

We use an approach known, in the social sciences, as event history analysis. Our dependent variable is the age at which an individual enrols into a university program for the first time. We use the Cox's relative risk model (Kalbfleisch and Prentice, 2002: 42–43, 95–147). This model allows estimating the effect of independent variables on the dependent variable taking into account that some people never enrol into university.

The Cox's model may be written as

$$h(t) = h_0(t)e^{\mathbf{x}\boldsymbol{\beta}},\qquad(1)$$

where t represents age, $h(t)$ is the hazard, i.e. the probability of enrolling into university at age t if it has not occurred before, $h_0(t)$ is the "baseline" hazard – i.e. the relation between hazard and age for the "reference" individual, i.e. the person who belongs to the reference category of each independent variable–, $\mathbf{x}$ is the vector of independent variables and $\boldsymbol{\beta}$ is the vector of the coefficients associated with the independent variables.

Some of the equations we estimate include conditional relations that involve qualitative and quantitative variables. In these equations, the effects of trend and

tuition fees are not represented by a single coefficient, but rather by one coefficient for each category of social origin or each sociolinguistic group. These equations may be written as

$$h(t) = h_0(t)e^{\mathbf{x}\boldsymbol{\beta}}e^{A\mathbf{z}\gamma_A}e^{D\mathbf{z}\gamma_D}. \tag{2}$$

where A is the trend, D stands for tuition fees, $\mathbf{z}$ represents the categories of social origin or the sociolinguistic groups, γ_A represents the coefficients associated with trend for each category of social origin or each sociolinguistic groups, and γ_D represents the coefficients associated with tuition fees for each category of social origin or each sociolinguistic group.

Some of the equations we estimate include conditional relations involving time-varying effects. In these equations, the effect of tuition fees varies according to age in a different fashion for each category of social origin or each sociolinguistic group. These equations may be written as

$$h(t) = h_0(t)e^{\mathbf{x}\boldsymbol{\beta}}e^{A\mathbf{z}\gamma_A}e^{D\mathbf{z}\gamma_D}e^{DT\mathbf{z}\gamma_{DT}}. \tag{3}$$

where $\mathbf{z}$ represents again the three categories of social origin or sociolinguistic groups, vector γ_D represents the intercept of the relation between the effect of tuition fees and age for each category of social origin or each sociolinguistic groups, and vector γ_{DT} represents the slope of this relation for each category of social origin or each sociolinguistic group.

The GSS samples are probabilistic, but not simple random: their sampling design involves strata and sometimes clusters. Estimation must be done using sampling weights and standard errors must be estimated taking the sampling design into account. We use the sampling weights and we correct the standard errors using average design effects (Kish, 1995).

6.3 Variables

We define the birth cohorts so they represent, as much as can be, the evolution of the demographic, social, and economic context as well as the changes in education policies and in the education system. The oldest cohort, 'Before 1936', groups together people born before the baby boom; the second one, '1936–1950', the people born about the first half baby boom – this cohort came of age during the 30-year period of growth that followed the end of World War II; the third one, '1951–1974', the people born during the second half of the baby boom and a few years after it ended; the fourth one, '1975–1990', the people born later and old enough in 2011so that they could provide useful information for a study on university enrolment. Each cohort came of age in a different economic context. The two first cohorts reached the age at which people typically enrol in university for the first time before the expansion of the postsecondary education system in Quebec and Ontario, which occurred mostly in the late 1960s and the early 1970s. The two last

cohorts reached that age respectively at the time where the new system was taking shape or was already in place.

We define sociolinguistic groups by combining the respondent's place of birth, their parents' place of birth and mother tongue. The Quebec French-speaking and the Ontario English-speaking are people who were born in the province where they live. They speak the language of the majority, and have access to a well-developed network of universities in their own province. The effects of education policies are the most easily observed for these groups.

The Ontario French-speaking are the French-speaking people born in Ontario and living in Ontario at the time of survey. They are the linguistic minority and university training in their language in their province is in short supply: there is no French-language only university in Ontario and few academic institutions that offer university education in French. The Quebec English-speaking are the English-speaking people born in Quebec and living in Quebec at the time of survey. They are the linguistic minority, but, unlike the Ontario French-speaking, they have access to a complete supply of university training in their province and language: there are three English-language only universities in Quebec. In Ontario, there are no French-language only university and a limited number of programs given in French in bilingual or English- language universities.

We define and measure social origin through parents' education level. People are grouped in three categories: 1) neither father nor mother has any postsecondary education – abbreviated as "No PSE" in the tables –, 2) at least one of the two parents has a non-university postsecondary diploma – "Non-university PSE–, and 3) at least one of the two parents has a university diploma or degree – "University".

We include gender in all our equations.

Using the Cox model allows using time-varying independent variables. We use two such variables: tuition fees and 'trend'. As we explain above, tuition fees have a different value for each year and for each province. For a given individual, this variable takes a different value for each calendar year he or she was at risk of enrolling.

We are interested in the effects of tuition fees and of certain sociodemographic characteristics on individual enrolment. However, as we explain above, individual enrolment depends in part on the supply of university training and on demand for people having university education. There is no easy way to measure these. However, it is reasonable to assume that, unlike tuition fees that have fluctuated ion a non-monotonic fashion over the period we are studying, both the supply of university training and the demand for people having university education of them have increased in a monotonic way. In our equations, we account for this growth using a trend. This allows disentangling the effect of tuition fees from those of the expansion of university education and of changes in the labour market. The trend variable increments by one each year, and thus varies from 0, in 1946, to 65, in 2011.

13

Table 2. Enrolment into university according to selected sociodemographic characteristics, tuition fees and trend. Ontario and Quebec, 1946–2011. Cox model.

	1	2	3
Cohort [1975–1990]			
Before1936	0.403***		
1936–1950	0.837**		
1951–1974	0.881***		
Sex [Male]			
Female	1.009		
Sex and cohort [Male, 1975–1990]			
Female, Before 1936		0.331***	0.672*
Female, 1936–1950		0.886	1.48**
Female, 1951–1974		1.109	1.372***
Female, 1975–1990		1.573***	1.572***
Male, Before 1936		0.758*	1.532*
Male, 1936–1950		1.264**	2.11***
Male, 1951–1974		1.123	1.382***
Sociolinguistic group [Quebec French-speaking]			
Quebec English-speaking	1.077	1.068	1.070
Quebec immigrants	1.769***	1.763***	1.727***
Ontario English-speaking	1.039	1.041	1.076
Ontario French-speaking	0.826	0.821	0.848
Ontario immigrants	1.578***	1.592***	1.643***
Social origin [No PSE]			
Non-university PSE	1.825***	1.832***	1.787***
University	4.22***	4.239***	4.12***
Trend			1.015***
Tuition fees ($ thousands)			0.970

* $p<0.05$; ** $p<0.01$; *** $p<0.001$.Reference categories between brackets.
Coefficients expressed as relative risks ratios. Data from cycles 10, 15, 20 and 25 from the General Social Survey. Weighted estimation. Standard errors corrected using average design effect.

Age at enrolment was most delicate. Our data source records the age at which the highest diploma was awarded, not the age at enrolment in university. We do as is commonly done is such a case and determine the age at enrolment from the age at graduation. Dealing with people who never enrolled is easy. Dealing with people who got their diploma at the end of an uninterrupted trajectory is straightforward. Dealing with people who got their diploma later than the age at which such diploma would have been awarded to someone who attended school without interruption is more challenging. In such cases, we drew the age at enrolment from a normal distribution with mean at the most likely age at enrolment given the duration of the program and variance estimated from GSS data. Given that the draw is random, it does not induce a bias and can only increase the variance of the estimates of the coefficients associated with the independent variables.

7 Results

Table 2 shows results from the estimation of a series of equations relating the risk to enrol into university to sociodemographic characteristics, trend and tuition fees. The first equation provides the net effect of each characteristic. In the second equation, we combine sex and cohort. In the third, we add trend and tuition fees. Comparing equation 1 and 2 shows that the differences between men and women appear only when looking at their evolution across cohorts. Women's hazards increase from the oldest to the youngest cohort. Men's hazards peak within the 1936–1950 cohort, then decrease. In the youngest cohort, women's hazard is about one time and a half that of men. Immigrants, whether in Quebec or Ontario, are more prone to enrol in university than all other sociolinguistic groups. Enrolment is strongly related to parents' education level. Having at least one parent with a non-university postsecondary diploma increases the hazard of enrolling by about 80%. Having at least one parent with a university diploma increases it fourfold. According to equation 3, the hazard of enrolling increases as a function of age, i.e. over the life course, but tuition fees have no effect.

As the effect of sex that appears only when allowing it to vary across cohorts, the effect of tuition fees become apparent only when allowing it to vary according to dimensions of the social structure such as social origin and sociolinguistic groups. In equation 4 (Table 3), we estimate the effect of trend and tuition fees conditional on social origin. In equation 5, we estimate the effect of tuition fees conditional on social origin as a function of age. In equation 4 and 5, the effect of trend increases slightly with parents' education level. In equation 4, the effect of tuition fees varies according to parents' level of education. Tuition fees decrease the hazard of enrolment when parents do not have any university education, but increase it when at least one parent has a university diploma.

15

Table 3 Enrolment into university according to selected sociodemographic characteristics, tuition fees and trend. Ontario and Quebec, 1946–2011. Effect of tuition fees conditional on social origin. Cox model.

	4	5
Cohort and sex [Male, 1975–1990]		
Female Before 1936	0.675*	0.773
Female 1936–1950	1.623***	1.688***
Female 1951–1974	1.436***	1.629***
Female, 1975–1990	1.594***	1.591***
Male, Before 1936	1.549*	1.766**
Male 1936–1950	2.308***	2.388***
Male 1951–1974	1.446***	1.641***
Sociolinguistic group [Quebec French-speaking]		
Quebec English-speaking	1.135	1.136
Quebec immigrants	1.770***	1.830***
Ontario English-speaking	1.113*	1.086
Ontario French-speaking	0.875	0.854
Ontario immigrants	1.684***	1.644***
Trend by social origin		
No PSE	1.010***	1.010**
Non-university PSE	1.016***	1.017***
University	1.023***	1.025***
Tuition fees by social origin		
No PSE	0.811**	
Non-university PSE	0.932*	
University	1.100***	
Tuition fees according to age by social origin		
Intercept (γ_D)		
No PSE		0.963**
Non-university PSE		1.075***
University		1.264***
Slope (γ_{DT})		
No PSE		0.979***
Non-university PSE		0.980***
University		0.978***

* $p<0.05$; ** $p<0.01$; *** $p<0.001$. Reference categories are between brackets.

Coefficients expressed as relative risks ratios. Data from cycles 10, 15, 20 and 25 from the General Social Survey. Weighted estimation. Standard errors corrected using average design effect.
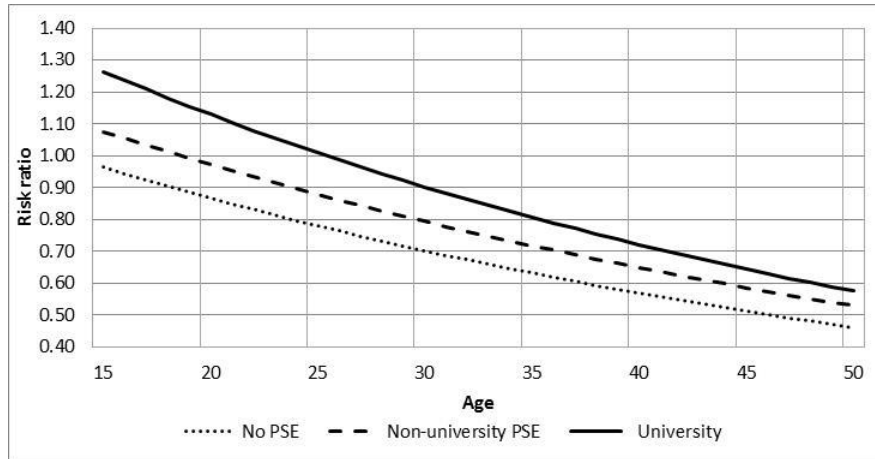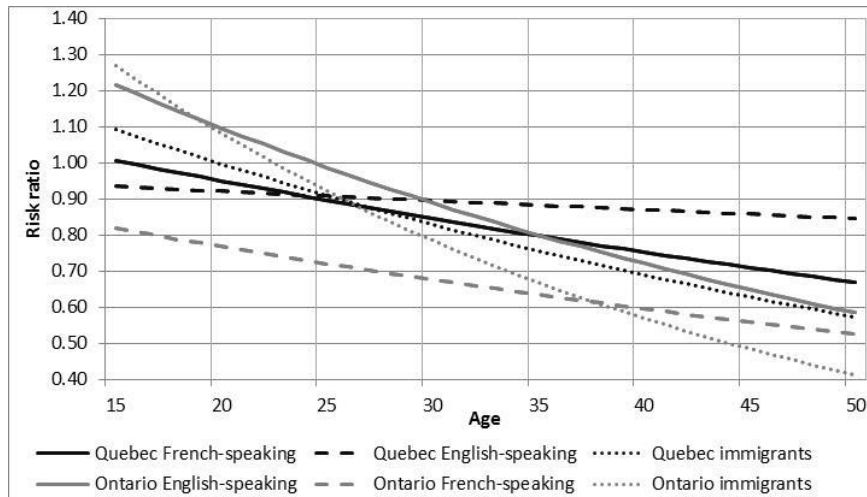
Fig. 2: The effect of tuition fees on enrolment according to age by social origin.

In equation 5, the effect of tuition fees on enrolment is modelled as a function of age for each category of social origin. The results are best understood when presented in graphic form. Figure 2 shows the functions defined by the intercepts and slopes of equation 5. The functions appear as curves rather than straight lines because we present the effects as risk ratios, which are easier to interpret than their logarithms. The relative risk is 1 for an individual aged 15 whose parents do not have any postsecondary diploma. The three curves come on top of each other and do not intersect. Over the whole age range, the hazard of enrolling is the highest when having at least one parent holding a university diploma and the lowest when having two parents who do not hold any postsecondary diploma. The gap between categories is the largest for young people and decreases with age. The negative effect of increasing tuition fees by 1,000 CAD is larger when having one parent with a non-university postsecondary diploma than when having at least one parent with a university diploma, and even larger when having two parents who do not hold any postsecondary diploma. The hazard ratio decreases with age within each category: in other words, the negative effect of a 1,000 CAD raise in tuition fees increases with age in all categories.

Table 4 is similar to table 3, but focuses on relations conditional on sociolinguistic groups rather than on social origin. Thus, in equation 6, we estimate the effect of trend and tuition fees conditional on sociolinguistic groups, and in equation 7, we estimate the effect of tuition fees conditional on sociolinguistic groups as a function of age. The effect of trend is positive for all sociolinguistic groups. It is relatively low for the Ontario English-speaking and relatively high for the Ontario French-speaking. This is not unexpected. The trend variable captures not only the growth in contextual factors such as the supply of the university education and the demand for university graduates, but also all residual increase in enrolment not explained by other independent variables. This is not a bug, it is a feature, as we

use this variable to capture all trends. Here it captures the catching up of French-speaking Ontarians as well as the fact that enrolment among the English-speaking Ontarians was relatively high already in the oldest cohorts (Laplante et al. 2014). Equation 6 shows that tuition fees reduce the hazard of enrolling among the Quebec French-speaking and even more among the Ontario French-speaking.

Figure 3 depicts the relation between the effect of tuition fees and age within sociolinguistic groups. The relative risk is 1 for a Quebec French-speaking individual aged 15. The three curves have very different intercepts and slopes, all negative, and they intersect. The slope of the Quebec English-speaking is practically zero. For them, increasing tuition fees has no effect on the hazard of enrolment. The slope of the Quebec French-speaking is negative, but its intercept is practically equal to that of the Quebec English-speaking: for them, the negative effect of increasing tuition fees increases with age. The intercept and slope of the Ontario English-speaking are greater than those of the Quebec French-speaking are: the negative effect of increasing tuition fees is greater among the Quebec French-speaking than among the Ontario English-speaking for the conventional age of university enrolment, but the gap closes down with age and could reverse after age 35. The slope of the Ontario French-speaking is about the same as that of the Quebec French-speaking, but their intercept is smaller: for them, the negative effect of increasing tuition fees is comparatively large at age 15 and increases still with age. The intercept of Quebec immigrants is greater than that of the Quebec French-speaking and even greater than that of the Quebec English-speaking, but their slope is also greater. At age 15, the negative effect of increasing tuition fees is not as strong for them as for the natives, but it increases faster with age. The same can be said of the Ontario immigrants. Their intercept is higher than that of the Ontario English-speaking, and thus the highest of all groups, but their slope is steep, the steepest of all groups. At age 15, the negative effect on increasing tuition fees is smaller among them than in any other group, but it increases faster with age.

Table 4. Enrolment into university according to selected sociodemographic characteristics, tuition fees and trend. Ontario and Quebec, 1946–2011. Effect of tuition fees conditional on sociolinguistic group. Cox model.

	6	7
Cohort and sex [Male, 1975–1990]		
Female Before 1936	0.692	0.804
Female 1936–1950	1.496**	1.576***
Female 1951–1974	1.399***	1.607***
Female, 1975–1990	1.577***	1.576***
Male, Before 1936	1.571*	1.813**
Male 1936–1950	2.137***	2.245***
Male 1951–1974	1.410***	1.619***
Social origin [No PSE]		
Non-university PSE	1.789***	1.806***
University	4.100***	4.116***
Trend by sociolinguistic group		
Quebec French-speaking	1.015***	1.016***
Quebec English-speaking	1.017**	1.017**
Quebec immigrants	1.023***	1.026***
Ontario English-speaking	1.007	1.007*
Ontario French-speaking	1.032***	1.031***
Ontario immigrants	1.019***	1.019***
Tuition fees by sociolinguistic group		
Quebec French-speaking	0.900**	
Quebec English-speaking	0.881	
Quebec immigrants	0.950	
Ontario English-speaking	1.038	
Ontario French-speaking	0.721**	
Ontario immigrants	1.025	
Tuition fees according to age by sociolinguistic group		
Intercepts (γ_D)		
Quebec French-speaking		1.008
Quebec English-speaking		0.935
Quebec immigrants		1.094
Ontario English-speaking		1.217***
Ontario French-speaking		0.818
Ontario immigrants		1.271***
Slopes (γ_{DT})		
Quebec French-speaking		0.988*
Quebec English-speaking		0.997
Quebec immigrants		0.982*
Ontario English-speaking		0.979***
Ontario French-speaking		0.987
Ontario immigrants		0.968***

* $p<0.05$; ** $p<0.01$; *** $p<0.001$. Reference categories are between brackets.
Coefficients expressed as relative risks ratios. Data from cycles 10, 15, 20 and 25 from the General Social Survey. Weighted estimation. Standard errors corrected using average design effect.

Fig. 3: The effect of tuition fees on enrolment according to age by sociolinguistic groups.

8 Discussion and Conclusions

In Quebec and Ontario, over the last two decades, tuition fees have mostly increased even in constant dollars. Enrolment has increased over that period. However, there are reasons to believe that increasing tuition fees decreases enrolment. There is also a set of sound reasons to believe that enrolment should have increased over this period even if the net effect of increasing tuition fees is to decrease enrolment. The proportion of foreign-born Canadians has been increasing steadily over that period. A large fraction of immigrants to Canada are selected on their education level, which increases the proportion of university degree-holders in the Canadian population and fosters enrolment of children from immigrants either because of simple intergenerational transmission, or because immigrants invest highly in their children's education as a means of social and professional integration (Zéroulou, 1988). The transformation of the economy towards knowledge-based production of value convinced parents and children of the virtues of higher education. The demand for higher education has been strong despite its rising cost. Disentangling the effect of tuition fees from those of other factors is an intricate problem.

As we saw earlier, some researchers acknowledged this difficulty. Some relied on comparing provinces with different levels of tuition fees and increases. Some used long series, as we do. Our strategy was using long series of tuition fees that include non-monotonic variation, and model the growth in enrolment net of the effect of tuition fees and of the other variables that we included in our equations.

This allowed us isolating the effect of tuition fees. However, this solved only the first part of the problem.

Enrolling into university is an event that occurs, or not, over the life course. Although a large fraction of students enrol in their early twenties, enrolling later has become more common, and, in Canada, more so in Quebec. Furthermore, previous researches lead to believe that the effect of tuition fees on enrolment could vary across social origins and across groups defined by their immigration status, language and province. Modelling the effect of tuition fees while taking all these aspects into account and especially its variation over the life course, had not been attempted before, although there were sound reasons to think it could a rather complex function of all these factors. Our results show that indeed, the effect of tuition fees varies across social groups and, within social groups, as a function of age.

Tuition fees have a negative effect on enrolment and this effect increases over the life course. When comparing groups define by their social origin, as measured by parents' level education, the effect of tuition fees varies in a relatively simple fashion. The effect of increasing tuition fees by 1,000 CAD is positive when having at least one parent with a university diploma, but negative when having one parent with a non-university postsecondary diploma, and more so when having two parents who do not hold any postsecondary diploma. However, these effects vary according to age with a negative slope within each group. Considering this variation provides a different picture. Tuition fees have a positive effect on enrolment for children from families where at least one parent has university education up to age 25, but have a negative effect on older children. They have a positive effect for children from families where at least one parent has non-university PSE up to age 20, but the effect turns negative afterwards. The effect is negative for children form families where parents do not have any postsecondary education and this effect increases with age.

When comparing sociolinguistic groups, the effect of tuition fees varies in a more complex way. For the Quebec English-speaking, increasing tuition fees has no effect on enrolment. For the Quebec French-speaking, the negative effect of increasing tuition fees increases with age. The negative effect of increasing tuition fees is greater among the Quebec French-speaking than among the Ontario English-speaking for the conventional age of university enrolment, but the gap closes down with age and could reverse after age 35. For the Ontario French-speaking, the negative effect of increasing tuition fees is comparatively large at age 15 and increases still with age. These differences do not come as surprises. The Quebec English-speaking are concentrated among the upper strata of society and money may not be the most important factor when considering enrolment. The cost of university education is a greater concern for the Quebec French-speaking, and a growing one as age increases, which is of some consequence in a group that is known for enrolling late. Tuition seem to be less of a problem among the Ontario English-speaking for the young, but it becomes one later on for those who did not enrol early in the life course.

21

Among children of immigrants, as among the Ontario English-speaking, the effect of tuition fees is strongly related with age. At younger ages, the effect of high tuitions fees, when comparing with native, seems to attract rather than deter, but their negative effect increases faster with age than among native groups. Apparently, parents, who are likely those who pay when the children are young, believe that money spent of university education is worth it for their children, but do not believe that it is worth spending money on older children or on themselves later. Immigrants, behave as if investing in education was worthwhile for the young, and presumably their children, rather than on adults.

The general finding is that the level of tuition fees has a negative effect on enrolment, except for children of highly educated parents and for children of immigrants. Even for these children, the effect becomes negative after the conventional enrolment age. Apparently, their parents are willing to pay some kind of education premium when they are young–maybe as long as they look 'promising'– but not afterwards. Our results suggest that in an era where university education contributes to social reproduction, this process is age-dependent and children from favoured background are not allowed missing the window of the age of reproduction.

The political implications are straightforward. Tuition fees are a real concern for children whose parents have little or no postsecondary education, especially if they do not enrol early. There may be several reasons for this, ranging from limited resources to limited understanding of the costs and benefits of higher education (Boudon, 1974; Usher, 2005). Nevertheless, whatever the mechanism that leads children of low-educated families to be less prone to enrol in higher education and whatever the motives for raising them, tuition fees seem to be a barrier to intergenerational mobility and increasing them may simply strengthen social reproduction.

The differences between sociolinguistic groups are revealing of the differences between Quebec and Ontario, and maybe most of Canada outside Quebec, on tuition fees. About 80% of the Quebec population belongs to a group for which tuition fees are a concern. For this very reason, increasing tuition fees in Quebec, whatever the motives for doing so, is not an easy political task. In their vast majority, Quebec English-speaking and immigrants – for whom tuition fees are not a concern – support the Liberal party, which supports increasing tuition fees. However, the French-speaking vote is spread across four parties. This configuration makes likely that tuition fees will remain a divisive political issue in Quebec for the near future. Things are different in Ontario, where a large fraction of the population belongs to groups for whom tuition fees are not a concern. This does not mean that high tuition fees do not have adverse effects on enrolment for children from low-educated families outside Quebec. However, it implies that promoting affordable university tuition fees as evidence-based policy, difficult in Quebec, is likely to remain even more difficult elsewhere in Canada. Advocacy for increasing tuition fees, despite not being based on solid evidence, matches very well the in-

terests of some social groups. Whether something similar could be found in England, the United States or Germany is an open question.

Acknowledgements

References

Abada, T., F. Hou, and B. Ram (2009) Ethnic differences in educational attainment among the children of Canadian immigrants, *Canadian Journal of Sociology* **34**(1), 1–28.

Abdurrahman, A., C. Wen-Hao, and M. Corak (2013) Intergenerational education mobility among the children of Canadian immigrants, *Canadian Public Policy* **39**(s1), 107–122.

Adamuti-Trache, M. and L. Andres (2008) Embarking on and persisting in scientific fields of study. Cultural capital, gender, and curriculum along the science pipeline, *International Journal of Science Education.* 30(12), 1557–1584.

Barone, C. (2011) Some things never change. Gender segregation in higher education across eight nations and three decades, *Sociology of Education* **84**(2), 157–176.

Beattie, I. R. (2002) Are all 'adolescent econometricians' created equal? Racial, class, and gender differences in college enrollment, *Sociology of Education* **75**(1), 19–43.

Belzile, G. (2010) *La hausse des droits de scolarité réduirait-elle l'accessibilité aux études universitaires?* Note économique. Montreal: Institut économique de Montréal.

23

Bonikowska, A. and F. Hou (2010) Reversal of fortunes or continued success? Cohort differences in education and earnings of childhood immigrants, *International Migration Review* **44**(2), 320–353.

Boudon, Raymond. 1974. *Education, Opportunity and Social Inequality.* New York: John Wiley and Sons.

Bolton, P. (2014) Tuition fee statistics. Standard note No. SN/SG/917. London: House of Commons Library.

Bourdieu, P. (1984) *Distinction: A social critique of the judgment of taste.* London: Routledge and Kegan Paul.

Bourdieu, P. and J.-C. Passeron (1977) *Reproduction in education, society and culture.* London; Beverly Hills: Sage Publications.

Bourdieu, P. and J.-C. Passeron (1979) *The inheritors. French students and their relation to culture.* Chicago: University of Chicago Press.

Bruckmeier, K., and B. U. Wigger (2014) The effects of tuition fees on transition from high school to university in Germany, *Economics of Education Review* **41**: 14–23.

Council of Minister of Education Canada (2007) *Literature review of postsecondary education affordability in Canada.* Toronto: CMEC.

Chenard, P. and P. Doray (2013) L'accessibilité aux études supérieures. Quelles démocratisations ?, in P. Chenard, P. Doray, L.-E. Dussault, and M. Ringuette (eds), *L'accessibilité aux études supérieures, un projet inachevé*, pages 49–68, Québec: Presses de l'Université du Québec.

Christofides, L. N., J. Cirello, and M. Hoy (2001) Family income and postsecondary education in Canada, *Canadian Journal of Higher Education* **31**(1): 177–208.

Christofides, L. N., M. Hoy, J. Milla, and T. Stengos (2012) *Grades, aspirations and post-secondary education outcomes.* IZA Discussion Paper No. 6867. Berlin: Forschungsinstitut zur Zukunft der Arbeit.

Coelli, M. B. (2009) Tuition fees and equality of university enrolment, *Canadian Journal of Economics* **42**(3): 1072–1099.

Collins, M., G. Collins, and G. Butt (2013) Social mobility or social reproduction? A case study of the attainment patterns of students according to their social background and ethnicity, *Educational Review* **67**(2): 196–217.

Corak, M., G. Lipps, and J. Zhao (2003) Family income and participation in post-secondary education, in C. Beach, R. Boadway, and M. McInnis (eds), *Higher Education in Canada*, pages 255–294, Montreal and Kingston: John Deutsch Institute and McGill-Queen's Press.

Crouse, J. T. (2015) Estimating the average tuition elasticity of enrollment for two-year public colleges, *American Journal of Economics* **5**(3): 303–314.

Croxford, L. and D. Raffe (2012) The impact of the 2012 tuition fee changes on student flows across the UK's internal borders. ERSC future of the UK and Scotland programme Working Paper No. 8. Edinburgh: The University of Edinburgh.

Dandurand, P. (1986) Rapports ethniques et champ universitaire, *Recherches so-ciographiques* **17**(1): 41–77.

Dandurand, P., M. Fournier, and L. Bernier (1980) Développement de l'enseignement supérieur, classes sociales et luttes nationales au Québec, *Sociologie et sociétés* **12**(1): 101–131.

De Broucker, P. (2005) *Getting there and staying there. Low-income students and post-secondary education*. Research Report No. W27. Ottawa: Canadian Policy Research Networks.

Dearden, L., E. Fitzsimons, and G. Wyness (2011) *The impact of tuition fees and support on university participation in the UK*. Discussion Paper No.126. London: Centre for the Economics of Education, London School of Economics.

Delaney, J. A., and T D. Kearney (2015) The impact of guaranteed tuition policies on postsecondary tuition levels. A difference-in-difference approach, *Economics of Education Review* **47**: 80–99.

Dietrich, H. and H.-D. Gerner (2012) The effects of tuition fees on the decision for higher education. Evidence from a German policy experiment, *Economics Bulletin* **32**: 2407–2413.

Drolet, M. (2005) *Participation in post-secondary education in Canada. Has the role of parental income and education changed over the 1990s?*. Catalogue No. 11F0019MIE–243. Ottawa: Statistics Canada.

Duru-Bellat, M. (2003) Actualité et nouveaux développements de la question de la reproduction des inégalités sociales par l'école, *L'orientation scolaire et professionnelle* **32**(4): 571–594.

Duru-Bellat, M., C. Marry, and A.k Kieffer (2001) La dynamique des scolarités des filles. Le double handicap questionné, *Revue française de sociologie* **42**(2): 251–280.

Finnie, R. (2014) Does culture affect post-secondary education choices?, *Higher Education Management and Policy* **24**(3): 57–85.

Finnie, R., C. Laporte, and E. Lascelles (2004) *Family background and access to post-secondary education. What happened over the 1990s?*. Catalogue. No. 11F0019MIE – 226. Ottawa: Statistics Canada.

Finnie, R., A. Sweetman, and E. Lascelles (2004) Who goes: The direct and indirect effects of family background on access to post-secondary education, in C. Beach, R. Boadway, and M. McInnis (eds), *Higher Education in Canada*, pages 295–338. Montreal and Kingston: John Deutsch Institute and McGill-Queen's University Press.

Finnie, R., R. E. Mueller, and A. Wismer (2012) *Access and barriers to post-secondary education: Evidence from the YITS*. Working Paper No. 2012–01. Ottawa: Education Policy Research Initiative, University of Ottawa.

Frenette, M. (2004) Access to college and university. Does distance to school matter?, *Canadian Public Policy* **30**(4): 427–443.

25

Frenette, M. (2005) *The impact of tuition fees on university access: Evidence from a large-scale price deregulation in professional programs.* Catalogue No. 11F0019MIE – 263. Ottawa: Statistics Canada.

Frenette, M. (2006) Too far to go on? Distance to school and university participation, *Education Economics* **14**(1): 31–58.

Hübner, M. (2012) Do tuition fees affect enrollment behavior? Evidence from a 'natural experiment' in Germany, *Economics of Education Review* **31**(6): 949–960.

Hansen, J. and X. Liu (2013) *A structural model of educational attainment in Canada.* IZA Discussion Paper No. 7237. Berlin: Forschungsinstitut zur Zukunft der Arbeit.

Heller, D. E. (1997) Student price response in higher education. An update to Leslie and Brinkman, *The Journal of Higher Education* **68**(6): 624–659.

Hemelt, S. W. and D. E. Marcotte (2011) The impact of tuition increases on enrollment at public colleges and universities, *Educational Evaluation and Policy Analysi*s **33**(4): 435–457.

Hui, T. S. (2004) *The US/Canada difference in postsecondary educational choice.* MPRA Paper No. 17995. Munich: Munich Personal RePEc Archive, Ludwig-Maximilians-Universität München.

Johnson, D. and F. Rahman (2005) The role of economic factors, including the level of tuition, in individual university participation decisions in Canada, *The Canadian Journal of Higher Education* **35**(3): 101–127.

Kalbfleisch, J. D. and R. L. Prentice (2002) *The statistical analysis of failure time data, second edition.* New York: John Wiley and Sons.

Kamanzi, C. and P. Doray (2015) La démocratisation de l'enseignement supérieur au Canada. La face cachée de la massification, *Revue canadienne de sociologie* **52**(1): 39–65.

Kamanzi, C., P. Doray, S. Bonin, A. Groleau, and J. Murdoch (2010) Les étudiants de première génération dans les universités. L'accès et la persévérance aux études au Canada, *The Canadian Journal of Higher Education* **40**(3): 1–24.

Kamanzi, C., P. Doray, J. Murdoch, S. Moulin, É. Comoé, A. Groleau, C. Leroy, and F. Dufresne (2009) *The influence of social and cultural determinants on post-secondary pathways and transitions.* Transitions Project Note 6. Montreal: The Canada Millennium Scholarship Foundation.

Kish, L. (1995) *Survey sampling.* New York NY: John Wiley and Sons.

Kozhaya, N. (2004) La hausse des droits de scolarité réduirait-elle l'accessibilité aux études universitaires?, Memorandum. Montreal: Institut économique de Montréal.

Kučera, M. (2008) *The educational attainment of second generation immigrants in Canada. Analysis based on the General Social Survey.* Catalogue No. SP-857-09-08E. Ottawa: Human Resources and Social Development Canada.

Laplante, B., P. Doray, N. Bastien, and P. Chenard (2014) A tale of two logics. Social reproduction and mobilisation in university access in Quebec, 1945–

2000, *British Journal of Sociology of Education*, online 8 August 2014. doi:10.1080/01425692.2014.933668.

Leslie, L. L. and P. T. Brinkman (1987) Student price response in higher education. The student demand studies, *The Journal of Higher Education* **58**(2): 181–204.

Marry, C. (2004) *Les femmes ingénieures. Une révolution respectueuse*. Paris: Belin.

McIntosh, J. (2009) Educational mobility in Canada: Results from the 2001 General Social Survey, *Empirical Economics* **38**(2): 457–470.

MacLeod, J. (1987). *Ai'tt no makin'it. Leveled aspiratons in a low-income neighborhood.* Boulder: Westview Press.

Michael, C. (1999) Rising tuition fees. Measuring the impact upon undergraduate enrolment in Ontario, 1977/78–1996/97, Paper presented at the meeting of the Canadian Institutional Research and Planning Association, Quebec City, 24–26 October 1999.

Moulin, L. (2010) *Coût individuel des études supérieures pour l'étudiant. Panorama international et enseignements.* Documents de travail du CEPN 2011–08. Paris: Centre d'économie de l'Université Paris-Nord, Université Paris-Nord.

Neill, C. (2009) Tuition fees and the demand for university places." *Economics of Education Review* **28**(5): 561–570.

Ogbu, J. U. (1994) Racial stratification and education in the United States. Why inequality persists, *Teachers College Record* **96**(2): 264–298.

Quirke, L. and S. Davies (2002) The new entrepreneurship in higher education. The impact of tuition increases at an Ontario university, *The Canadian Journal of Higher Education* **32**(3): 85–118.

Rahman, A., J. Situ, and V. Jimmo (2005) *Participation in post-secondary education: Evidence from the Survey of Labour and Income Dynamics*. Catalogue No. 81-595-MIE–036. Ottawa: Human Resources and Skills Development Canada.

Rivard, M. and M. Raymond (2004) *The effect of tuition fees on post-secondary education in Canada in the late 1990s*. Working Paper No. 2004–09, Ottawa: Department of Finance.

Sá, F. (2014) *The effect of tuition fees on university applications and attendance: Evidence from the UK*. IZA Discussion Paper No. 8364. Bonn: Forschungsinstitut zur Zukunft der Arbeit.

Sen, A. and A. Clemente (2010) Intergenerational correlations in educational attainment. Birth order and family size effects using Canadian Data, *Economics of Education Review* **29**(1): 147–155.

Terrail, J.-P. (1992a) Destins scolaires de sexe. Une perspective historique et quelques arguments, *Population* **47**(3): 645–676.

Terrail, J.-P. (1992b). Réussite scolaire. La mobilisation des filles, *Sociétés contemporaines* No 11–12: 3–89.

27

Turcotte, M. (2011) Intergenerational education mobility. University completion in relation to parents' education level, *Canadian Social Trends* No. 92: 37–43.

Universities UK (2014) *Patterns and trends in UK higher education*. London: Universities UK.

Usher, A. (2005) *A little knowledge is a dangerous thing. How perceptions of costs and benefits affect access to education*. Toronto: Educational Policy Institute.

Wales, P. (2013) *Access all areas? The impact of fees and background on student demand for postgraduate higher education in the UK*. SERC Discussion Paper No. 128. London: Spatial Economics Research Centre, London School of Economics and Political Science.

Zéroulou, Z. (1988) La réussite scolaire des enfants d'immigrés. L'apport d'une approche en termes de mobilisation, *Revue Française de Sociologie* **29**(3): 447–470.

Tuition Fees and Social Segregation

Lessons from a Natural Experiment a the University of Paris 9-Dauphine

Léonard Moulin[a], David Flacher[a] and Hugo Harari-Kermadec[b]

[a] CEPN – CNRS UMR 7234, Université Paris 13 Sorbonne Paris Cité
[b] IDHES – CNRS UMR 8533, École Normale Supérieure de Cachan

Extended abstract

Increases in tuition fees have been a major recent trend in higher education in many developed countries (OECD, 2014). In some countries, university tuition fees are long-established and have been the subject of research into the effect of tuition fees on the access to higher education, the pursuit of studies and the outcomes for students. In France, on the contrary, university tuition fees are set by the government, at a level that makes them almost free, contrary to the practice in most English-speaking countries.

In this article, we study the first experience of increased tuition fees in a French public university – the University Paris 9-Dauphine – and we assess the effect of these tuition fees on the academic pathways selected by this university and consequently on the characteristics of the student populations concerned. Dauphine was the first university to obtain the status of *grand établissement* (prestigious research and higher education institutions). This status, acquired in 2004, allows it to create what are called *diplômes de grand établissement*, Master's degree programmes for which the university is free to set the tuition fees. Some of the national Master's degree courses (for which the tuition fees are still set by the public authorities) were therefore transformed into *masters d'établissement*, for which the university sets the tuition fees. At Dauphine, the scale of tuition fees ranges from 0 to 4000 euros per year depending on the parents' income. The first courses concerned began in 2010/2011.

Although Dauphine can be seen as an experiment in a new system, the subject of this article – the link between university tuition fees, student pathways and outcomes – is particularly important in the new French university context. Having been given greater autonomy, French universities are facing a lack of financial resources; there is a strong temptation to collect additional funds by making students participate in their tuition costs. The subject is also important from a theoretical viewpoint: the results presented in the literature are often contradictory about the

effects of tuition fees. It is therefore possible that context plays a decisive role in the effects of tuition fees. From this perspective, a detailed study of France, with its institutions and conceptions of higher education that differ so greatly from those of the English-speaking countries, should be rewarding. In addition to this unusual national setting, our approach is original in that we consider the whole academic pathway, not just the access to education as a function of its cost. To our knowledge, this global approach is unique in the literature. And yet it seems difficult to exclude *a priori* the existence of path dependency, particularly at an advanced level of education like the Master's degree. This path dependency is expressed through the modes of selection used by the university and the applications and past choices made by students.

To study students' pathways, we use an optimal matching method (Abbott and Forrest, 1986). This is based on a calculation of distances between trajectories, on the basis of which we establish a typology of student pathways. We then seek to determine, using an unordered multinomial logit model, the extent to which the type of trajectory can be related to the socio-economic characteristics of the students and the extent to which the tuition fees increase at Dauphine may have changed the types of pathways favoured for admission to these Master's degrees at Dauphine. Lastly, we study the effect of this tuition fees increase on graduation outcomes. For some authors (see for example Gary-Bobo and Trannoy, 2008), a rise in success rates should be expected when tuition fees are increased: higher tuition fees lead to greater self-selection of students, who are more exacting as regards the quality of the services provided and more motivated (to avoid wasting the financial resources invested in their studies). From this perspective, we evaluate the effects of the tuition fees increase on outcomes, using the difference-in-differences method adapted for a non-linear model (Puhani, 2012).

The methodology used allowed us to identify four types of pathway and to bring to light the potential effects of segregation and inequality on student pathways generated by the introduction of tuition fees, and the absence of any positive effect of these elite programmes on graduation rates. More precisely, we shown that the introduction of tuition fees has had contrasting effects on the pathways of students selected by Dauphine. The pathways that procure the lowest probability of access to the fee-paying Master's are those with a relatively high proportion of students from disadvantaged social backgrounds and of scholarships on social criteria and with a low proportion of foreign students. Conversely, the pathways giving the highest probability of access to these fee-paying Master 2 programmes are those followed by the students from the most well-off social categories. Added to which, pathways with time spent outside the university system also favour enrolment in these Master 2 programmes, probably because the students who follow these pathways are more familiar with fee-paying studies.

The revelation of this segregating effect of the rise in tuition fees at Dauphine is all the more interesting because the rise was progressive. Scholarship students, although not directly concerned by the rise in tuition fees, are indirectly affected, since the academic pathways they follow are less likely to lead to one of these

3

prestigious Master's now they are fee-paying. This result emphasises the need to analyse the complex mechanisms that cause students from disadvantaged backgrounds to be penalised by the introduction of tuition fees, even when these tuition fees are determined according to the parents' income. More generally, it also raises questions about the specificity of Dauphine in the French university landscape, and the difficulties that ensue, before drawing conclusions as to whether the experiences of this establishment are more widely applicable. Its prestigious status as a quasi "grande école" and its particularly prosperous catchment area constitute one of the most favourable contexts for the apparently "successful" introduction of tuition fees (along with that of Sciences Po Paris). And yet, if a segregation effect can already be identified within the setting of this experiment, despite the fact that the tuition fees are determined according to the parents' income, it is highly probable that this effect will be much stronger if such a policy is extended to the national level. Finally, beyond the question of segregation effects, and as suggested by Flacher et al. (2013), it is essential to consider the issue of the introduction or increase in tuition fees within a broader context, taking into account the institutional frameworks specific to each country.

References

Abbott, A. and J. Forrest, (1986) Optimal Matching Methods for Historical Sequences, *Journal of Interdisciplinary History*, **16(3)**, 471-494.

Flacher, D., H. Harari-Kermadec, and L. Moulin (2013) Faut-il (vraiment) augmenter les frais d'inscription à l'université?, *Revue française d'économie*, **27(3)**, 145-183.

Gary-Bobo, R. and A. Trannoy, (2008) Efficient Tuition Fees and Examinations, *Journal of the European Economic Association*, **6(6)**, 1211-1243.

Puhani, P. (2012) The Treatment Effect, the Cross Difference, and the Interaction Term in Non-linear "Difference-in-Differences" Models, *Economics Letters*, **115(1)**, 85-87.

OECD (2014) *Education at a glance 2014: OECD indicators*. OECD Press: Paris.

A Typology of Delayed Graduation:

Using Sequence Analysis of Enrollment Data to Uncover Heterogeneous Paths to a Degree[1]

Andrew M Wallace

City University of New York, Office of Policy Research

Abstract   Taking more than the "traditional" amount of time to graduate is an increasingly common path for undergraduate students at both the baccalaureate and associate levels. The expectation that these degrees can be earned in four and two years respectively is less likely to hold for students, especially those at public universities. This study uses transcript data from a large, urban public university system to examine patterns of enrollment among delayed completers. For the purposes of the analysis, delayed graduation is defined as more than six years for a baccalaureate degree and four years for an associate degree.

Traditional Time to Degree is Increasingly Unrepresentative

Large proportions of undergraduates in American colleges and universities fail to complete a degree within the "ideal" time frame: two years for an associate degree and four years for a baccalaureate degree. Even when allowing a six-year time frame for graduation, nationwide only 21.6% of entrants to associate programs and 61.8% of entrants to baccalaureate programs have finished a degree (NCES 2011). Many policy makers and researchers view these low numbers as indicating serious flaws in our system of higher education, harming both the students involved and our nation's economic competitiveness (Goldin and Katz 2009). Additionally, graduation rates differ between ethnicities, with Black and Hispanic students graduating at lower rates compared to their White and Asian counterparts (Massey et al., 2011).

This paper will present an analysis of patterns of enrollment of students who earn a degree, but not necessarily "on-time." For the purposes of this analysis, on-time graduation is defined as earning a certificate or associate degree within eight se-

---

mesters or a baccalaureate degree within 12 semesters. Sequence Analysis is used for this paper instead of a more traditional, regression-based method in order to attempt to uncover possible heterogeneity in the paths students take to their graduation. The intent is to construct a policy-relevant typology of student pathways that could provide the basis for further analysis and eventual interventions to improve student outcomes. While other methods exist for uncovering heterogeneity, they would either measure other aspects of a students pathway or they lack the potential interpretability of a typology arrived at by Sequence Analysis. For example, a Growth Mixture Model of student credit earning would allow a researcher to look for subgroups within student populations that earn credits at different rates. But because it focuses on credits instead of graduation, it does not allow us to compare patterns of students who earn different types of degrees or multiple degrees in the same way that the state space described below does. A type of model that would allow us to directly measure time to degree would be Survival Analysis. However, any subgroups identified by using a mixture model version of a Survival Analysis would not be as directly interpretable as those uncovered by Sequence Analysis.

Using Sequence Analysis on Educational Data

The way that the university semester system discretizes time means that longitudinal educational data is especially well suited to applications of Sequence Analysis. Sequence Analysis requires that time be measured in ordered, discrete units as opposed to continuously. At each point of measurement in the data under analysis, each subject is in one and only one of the states described below. The data used to create the sequences comprises 20 semesters of enrollment and graduation variables. The total number of observations is 125,515. Each semester is coded as one of the following states:

- Enrolled Full-Time (12 or more credits) without earning a degree
- Enrolled Part-Time (fewer than 12 credits) without earning a degree
- Stopped Out Before a Graduation Outcome
- Transferred Out to a Non-System Institution
- Earned a Certificate Degree
- Earned an Associate Degree
- Earned a Baccalaureate Degree
- Not Enrolled, Post-Graduation

While there are ways of dealing with missing data in sequences, for the purposes of this analysis it is not a problem. This is because not being enrolled in a given semester is an influential factor in student outcomes. It is a source of information, not a source of missingness. Further, there are effectively two types of non-enrollment in this study: expected and unexpected. After a student graduates, we

would not expect them to still be enrolled in the university system. Many of the system's students come back for further education, but for the purposes of this study, those who have received at least one degree are a success. However, students who are not enrolled but have not yet received a degree are unexpectedly not enrolled.

A student may be enrolled in the semester in which they earn a degree, but for the purposes of this analysis, the type of degree they earn in a semester is far more important than the number of credits they were attempting in that semester. Because graduation is ultimately the outcome of interest, this aspect of a student's college career is the most salient feature of the state a student can be in during a given semester. For example, given three students observed in a particular semester:

- Student A: Attempted 15 credits, earned a baccalaureate degree
- Student B: Attempted 15 credits, but earned no degree
- Student C: Attempted 9 credits, earned a baccalaureate degree

I argue that student C is more similar at this point of observation to student A than is student B, regardless of the difference in credits attempted.

The fourth possible state, transfer to an outside institution, uses data collected by the university system from an outside source to measure whether or not a student who started within the system enrolled at a college or university outside of the system. The National Student Clearinghouse (NSC) provides enrollment records for students at participating colleges. For students who began college in the university system under study but did not earn a credential within the system, a query was made to the NSC to see if they turned up at another institution. This query was conducted by the institutional research office of the system and the results were provided in the data set used for analysis. The NSC also provides graduation data for students who earn a degree at participating colleges but the number of colleges that participates in the degree reporting is smaller than the number of colleges that report enrollment. Because their graduation coverage is not as comprehensive as their enrollment coverage, I am not including graduation outcomes at non-system colleges in the state space. This choice was made because the lack of equal coverage means that graduation at non-system colleges will necessarily be undercounted. For the purposes of this analysis, the fact that a student transferred to another college is sufficient information to differentiate them from students who remain within the system or who drop out of higher education entirely.

The effect of this on the coding schema that I use is that I don't have a way to differentiate between non-enrollment prior to earning a degree and non-enrollment after earning a degree for students that transfer. Because of this, I make the choice to code all semesters after which a student has an enrollment at a non-system college and does not subsequently return to the system as equivalent. Thus a semester

after the observed point of transfer in which a student is not enrolled is treated as an equivalent state to a semester for which the NSC has an enrollment record. This has the effect of eliding some differences in patterns for transfer students but I argue that this is more desirable than mistakenly lumping in students who graduate from a college that participates in enrollment report but not degree reporting to the NSC with those students who transfer to a non-system college and do not earn a degree.

This choice has the side effect of making transfer cumulative thus affecting its interpretation. A student who transfers out of the system after two years will have a sequence that ends with sixteen semesters of transfer, regardless of how many enrollment records were found in the NSC. Thus an analysis of the distribution of states over time may show higher proportions of transfers (as well as stop outs and non-enrollment post-graduation). It is important for the interpretation of this distribution to keep in mind that, for cumulative states, an increase in later semesters may indicate a higher incidence of earlier experience of that state, not that students are experiencing that state for the first time later in their career. So the proportion of students who are in a cumulative state such as transfer in nineteenth semester will include a mixture of those who experience that state for the first time in that semester and those who experienced it earlier.

The Sequence Analysis is completed using the TraMineR package in R (Gabadinho et al., 2011). The data set I use for this analysis provides historical, longitudinal data on all first-time freshman, undergraduate students at the system's campuses. Students who attended another college prior to entering the university system are not included in this data set. The data set includes students who entered between September of 1999 and September 2002. These cohorts were chosen because they are the ones for whom a 10-year window of opportunity to graduate exists in the data. Admittedly, a shorter time frame would allow for more cohorts to be included, but as noted above, students often graduate in a longer time frame than what is considered traditional. A time frame longer than 10 years is not possible at this time due to the constraints of the data set.

Degree Bands

Most research on higher education distinguishes between community college students and baccalaureate students. Of this, there is research that investigates how well community college students who transfer to the baccalaureate level fare compared to either their community college compatriots or to those students who started off as baccalaureate students. Not enough research includes the reverse phenomenon, downward transfer, in its analyses. In order to fully capture the

variation in student trajectories through higher education, the type of degree pursued must be measured both at entry and at exit.

To this end, I separate out these students into degree bands. Those students who ended up at the certificate level (as well as those who started there) are sufficiently few and outside of the analytic scope of this study that I am excluding them from analysis. Associate and baccalaureate degree attainment are the main focus of this analysis. The final analytic sample is separated into degree bands based on initial and final degrees sought. This allows me to include a more complete set of possible degree paths in the analysis. The degree bands used in this analysis are as follows:

- Baccalaureate at entry to and exit from the system (BA1)
- Baccalaureate at entry to and Associate at exit from the system (BA2)
- Associate at entry to and exit from the system (AA1)
- Associate at entry to and Baccalaureate at exit from the system (AA2)

As noted above, the total number of observations in the data set is 125,515. Of these, 69.01% entered the system initially seeking an associate degree. 30.99% initially sought a baccalaureate degree in the system. Of those who initially sought an associate degree, 63.66% were still pursuing an associate degree upon departure from the system, regardless of whether that departure was due to graduation, drop out, or transfer. 36.34% initially associate-seeking students transferred up to the baccalaureate level by the time of their departure from the system. Of those who initially sought a baccalaureate degree, 89.51% were still pursing a baccalaureate degree upon departure from the system and 10.49% transferred down to the associate level.

Describing the Patterns

Figure 1 shows that for initially baccalaureate students who stay at the baccalaureate level, the most common pattern is 7 semesters of full-time enrollment, followed by a semester in which they receive their baccalaureate degree. This pattern is followed by 10.4% of the BA1 population (3,618 students). 9 of the top 20 patterns for this degree band involve graduation. Further, all of those graduation patterns involve graduation within 6 years. Another 4 patterns in the top 20 involve transfer (the yellow blocks). Finally, there are 7 patterns in the top 20 that do not involve graduation or a transfer outcome. These are students who have dropped out of the higher education system during the window of analysis. The orange blocks represent semesters of non-enrollment before receiving some sort of degree. The reader will also note that there are two patterns that involve part-time enrollment (lavender). This type of enrollment is not very prevalent among the top 20 patterns for BA1 students.
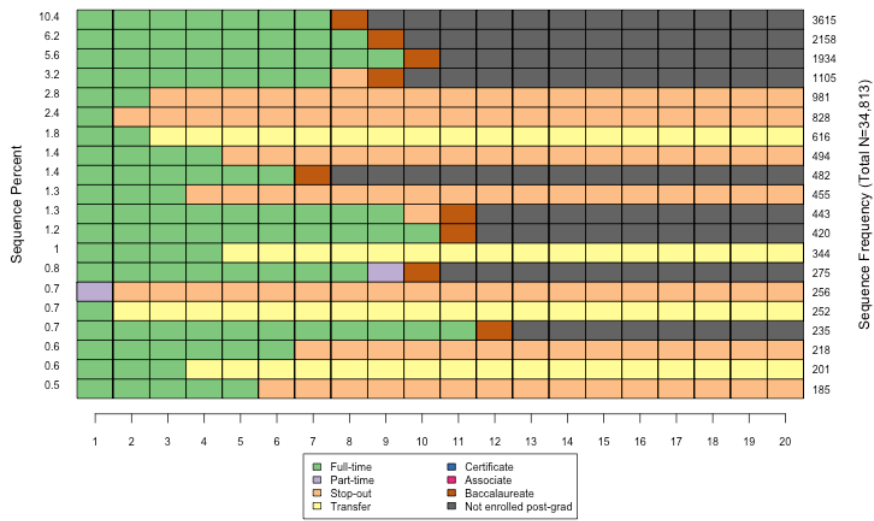
Fig. 1 Top Twenty Patterns, BA1
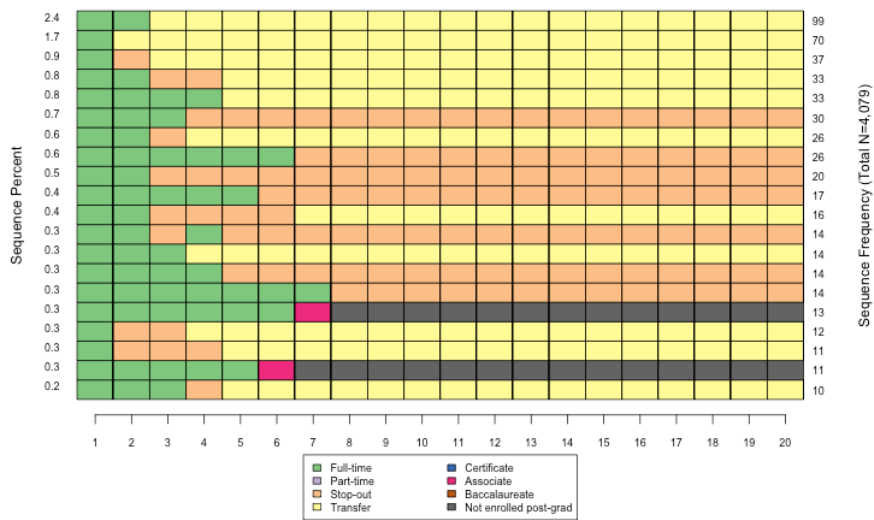
Fig. 2 Top Twenty Patterns, BA2
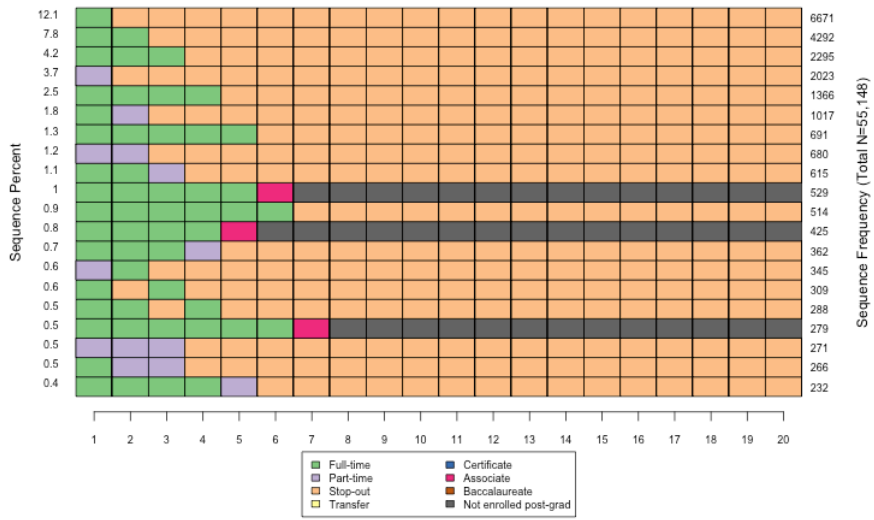
Fig. 3 Top Twenty Patterns AA1
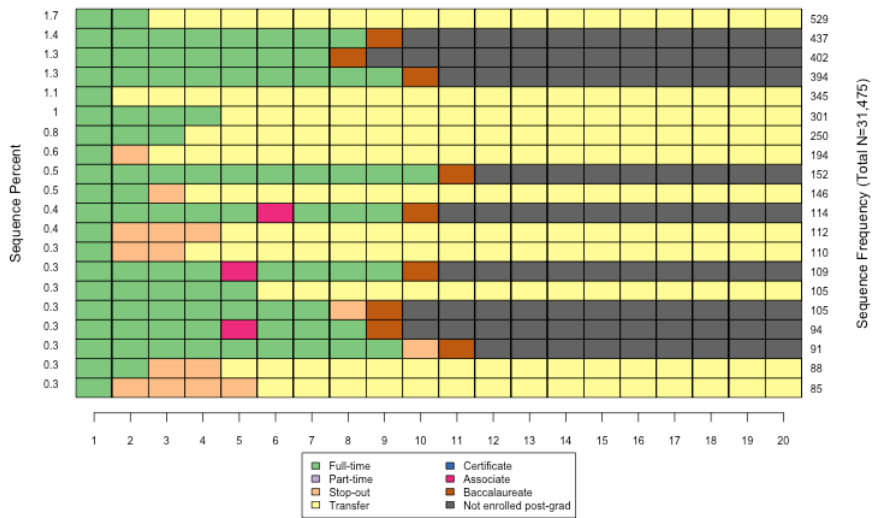
Fig. 4 Top Twenty Patterns, AA2

The top twenty patterns for the other degree bands told a rather different story. Figure 2 shows the top twenty patterns for BA2 students. Of the twenty patterns, only two involve graduation, seven involve dropping out, and eleven involve transfer. Only three of the top twenty patterns for AA1 students involve graduation and the rest involve dropping out (Figure 3). The patterns look better for AA2 students (Figure 4). Almost half of the patterns (nine) involve graduation and the remaining patterns involve transfer. None of these student's top twenty patterns involve dropping out.

As interesting as they are, the top 20 patterns for a degree band do not tell the whole story. The top 20 patterns for BA1, BA2, AA1, and AA2 only represent 44.52%, 12.75%, 42.56%, and 13.23% of their respective degree bands. These percentages indicate that students who stay at the degree level they started at are a lot more homogeneous than those who change level. This is evident from the fact that the top 20 patterns of those who stayed at the level they started at represent almost 45% and 43% of the students at the baccalaureate and associate levels respectively. On the other hand, of those who changed level, only around 13% of the students are represented by the top 20 patterns, regardless of starting level. The absolute number of students represented by these top 20 patterns is also worthy of note. For those who stayed at the same level that they started at, 15,516 and 23,520 students followed the top 20 baccalaureate and associate patterns respectively. For those who changed levels, 521 and 4,167 students are represented by the top 20 initially-baccalaureate and initially-associate patterns respectively.

Another way to examine the central tendencies of pattern data is to analyze the distribution of states over time. This distribution is not indicative of any sequence in the data much less the most common pattern. What the distribution can elucidate is general trends in state as the time window progresses. Figure 5 shows state distribution for the BA1 group of students (those who pursue a baccalaureate degree at entry to and exit from the system). In the figure, we can see that the proportion of students in the stop out state starts to grow in the second semester and peaks around the eighth semester. Around the same time as the peak of stop out, the baccalaureate graduation state (rust) starts to grow. Consequently, the proportion of non-enrollment post-graduation starts to grow, peaking at 57% in the twentieth semester.

The state distribution for the BA2 students shows smaller proportions of positive outcomes (Figure 6). The state of stopping out grows much more quickly and ends up being a larger proportion than it did for BA1 students. It peaks at 48% in the tenth semester. The proportion of transfer states is greater for this subgroup than the BA1 students, topping out at 41% in the twentieth semester. That said, there is some graduation and thereby non-enrollment post-graduation for these students starting after the seventh semester and growing slowly but steadily to 20% by the end of the analytic window.
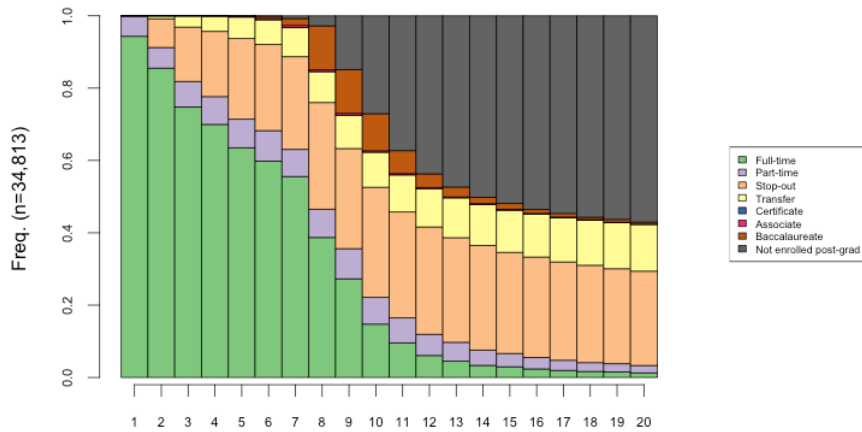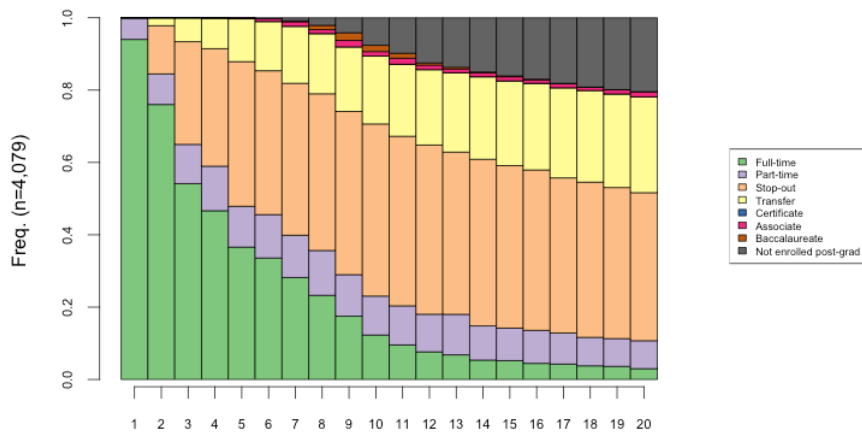
Fig. 5, State Distribution, BA1

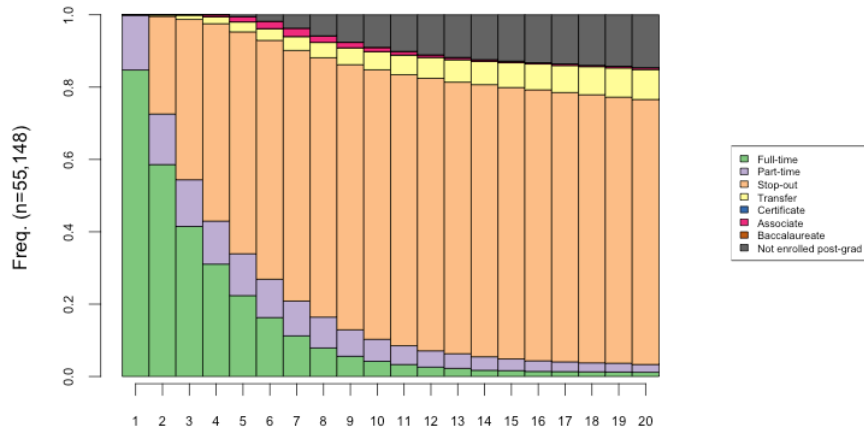**Fig. 6 State Distribution, BA2**

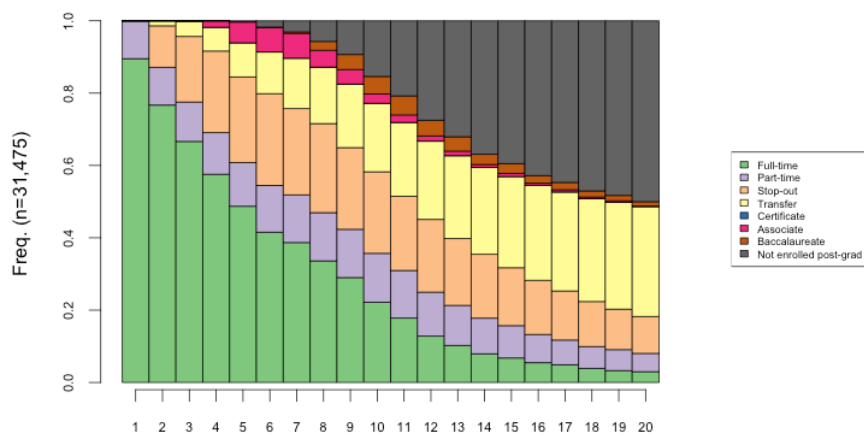**Fig. 7, State Distribution AA1**

Fig. 8, State Distribution, AA2

Figures 7 and 8 show the opposite trend in comparison for those who start at the associate level. Those who stay at the associate level have worse outcomes than those who transfer up to the baccalaureate level. This echoes findings in the literature that notes that students who manage to make it to the baccalaureate level from the associate level do as well or better than those who began at four-year colleges and stayed there. For AA1 students (those who were at the associate level at entry to and exit from the system), the proportion of students in the stop out state rises steeply beginning in the second semester then leveling out after the seventh semester. As a proportion of all states, stopping out peaks for these students at 75% in the tenth semester and remains around that level until the seventeenth semester where it declines slightly until the end of the analytic tracking window.

By contrast, AA2 students have increasing proportions of graduation, transfer, and non-enrollment post-graduation states starting in the fourth semester. Half of AA2 students are in a state of non-enrollment post-graduation by the end of the tracking window and a further 30% have transferred outside of the system.

Taken together, the top twenty patterns for a degree band, as well as the distribution of states over time gives a better indication of the central tendencies of the data than either measure alone. The top twenty patterns show commonalities in whole patterns while the distributions show changes in the composition of states in the subpopulation over time. The top twenty patterns point to the most common outcomes for students in a given group. The state distributions allow for comparisons at a given time point between groups. The differences between degree bands at a given time point as well as the distribution trends over time show how BA1 and AA2 students do systematically better than BA2 and AA1 students.

The average time spent in a given state is another useful measure of the central tendency of sequence data. Figures A.1 – A.4 in the appendix represent these measures calculated by degree band. On average, BA1 students spend slightly more than six semesters enrolled full-time and not quite six semesters not enrolled after graduating. BA2 students spend much less time enrolled full-time, more time stopped out, and much less time in post-graduation non-enrollment. AA1 students spend a distressingly long time in the state of being stopped out, which is likely due to dropping out early. In contrast, AA2 students average a similar amount of time enrolled full-time to BA1 students. They average less time not enrolled post-graduation but this is probably due to the effect of transferring programs (if not colleges) on delaying graduation.

While mean time in state tells us the proportion of a student's career that they spend in a given state, it does not tell us how long they spent in that state in a single spell. That means that a student who is enrolled full-time for ten consecutive semesters and then stops out for ten consecutive semesters will have the same average time in the full-time and stopped out states as a student who alternates be-

tween the two every other semester. Nonetheless, the average amount of time that a student spends in a state is still useful information.

Table 1 presents this information in another form: the proportion of a sequence that students average in a state. While Figure A.3 tells us that an AA1 student spends 12.88 semesters out of 20 stopped out, this table shows that accounts for 64.41% of a student's career. Compared to the 38.72% that BA2 students spend stopped out, the 23.97% that BA1 students spend stopped out, and the 17.77% that AA2 students spend stopped out, this paints a dismal picture for AA1 students. This table also shows that BA1 students spend the largest proportion of their time enrolled full-time and AA2 students have the largest proportion of part-time enrollment.

Table 1. Mean Time Spent in State by Degree Band

State	BA1	BA2	AA1	AA2
Full-time	30.97	23.80	15.06	29.01
Part-time	5.61	9.96	7.06	10.42
Stop Out	23.97	38.72	64.41	17.50
Transfer	8.70	17.36	4.85	17.77
Certificate	0.00	0.01	0.01	0.01
Associate	0.21	0.90	0.77	2.05
Baccalaureate	2.87	0.43	0.02	1.98
Non-Enrollment Post-Grad	27.66	8.83	7.81	21.26

Clustering the Patterns

It is important to note that, while there are generally accepted techniques for doing a cluster analysis, a certain amount of subjectivity is involved with a cluster analysis because of the necessity of choosing one algorithm over another. Further, the choice of a final number of clusters is ultimately based on theory and interpretability as much as it is upon any objective measure of cluster quality. Cluster analysis of sequence data requires that a matrix of distances be calculated to tell the researcher how close (by whatever measure) any two sequences are. The distance measure chosen for this analysis is the Optimal Matching distance as implemented by TraMineR. There is no consensus in the literature as to the single best way to weight the substitution, insertions, and deletions, but it is generally acknowledged that theory should be a guiding force in any weighting schema (Abbott and Tsay 2000; Gauthier et al., 2009; Lesnard, 2010). I chose to base the substitution cost on transition rates in order to have the difficulty of exchanging one state for another at any given time point be tied into how often this transition occurs from one time point to another. While this makes the weights less generalizable because the

clusters are more dependent on the transitions that occur in this particular data set, I argue that the benefit of not artificially imposing a substitution cost based on intuition outweighs the chance that the transitions occurring in this rather large data set are systematically different than those that might occur with a different data set. Tables A.5 – A.8 in the Appendix present the substitution cost matrices by degree band.

Once a distance matrix is calculated, a clustering algorithm needs to be chosen. In this analysis, I use Ward's method to cluster patterns because of its wide usage (Murtagh and Legendre, 2014). Table 2 presents a variety of measures of cluster quality. While the cluster quality measures produced by TraMineR suggested a two-, three-, or four-cluster solution, these clusters were not very informative and were certainly not policy-relevant. Essentially those cluster solutions tell us that students graduate, drop out, and transfer and not much more. In order to find more interesting patterns, I look at larger numbers of clusters to see what groups would emerge from the data. In these clusters, I find interesting patterns of degree completion, transfer, and dropping out. Looking at the various cluster solutions, I arrive at the set of cluster solutions presented in Table 3 as the best balance of interpretability and sample size. That is, I looked at representative sequences from each possible cluster solution (up to 20 clusters) and interpreted the story of the members of that cluster based on the representative sequences. Different numbers of cluster solutions were chosen for each degree band because the different degree bands had differing amounts of heterogeneity and with some of the degree bands (AA1 in particular), it took a higher number of clusters for interesting sequences to separate out from the larger clusters present in cluster solutions with fewer numbers of clusters.

Table 2 Quality Measures of Best Cluster Solutions by Degree Band[2]

	BA1		BA2		AA1		AA2	
	Clusters	Statistics	Clusters	Statistics	Clusters	Statistics	Clusters	Statistics
PBC	3	0.88	3	0.79	3	0.90	3	0.81
HG	3	0.98	3	0.91	3	0.99	4	0.94
HGSD	3	0.98	3	0.91	3	0.99	4	0.94
ASW	3	0.68	3	0.52	3	0.72	3	0.51
ASWw	3	0.68	3	0.52	3	0.72	3	0.51
CH	2	23492.5	3	1654.17	2	26219.23	2	13959.77
CHsq	3	66154.18	3	4511.72	3	85479.95	3	35738.92
HC	3	0.02	3	0.05	3	0.01	4	0.03

---

[2] The cluster quality measures are as follows: Point Biserial Correlation (PBC), Hubert's Gamma (HG), Hubert's Somer's D (HGSD), Average Silhouette Width (ASW), Average Silhouette Width – weighted (ASWw), Calinski-Harabasz index (CH), Calinski-Harabasz index squared (CHsq), and Hubert's C (HC). For details, see Studer (2013).

Table 3 Final Number of Clusters Chosen

Degree Band	Number of Clusters
BA1	10
BA2	8
AA1	12
AA2	8

Figures A.9 – A.19 in the Appendix present the representative sequences for all degree bands. For example, the subfigures in Figure A.9 show the ten representative sequences chosen by TraMineR to represent the ten clusters for the BA1 degree band in which graduation was the defining outcome. Unlike the top twenty sequences in Figures 1 – 4, the height of the representative sequences is proportional to how many students in the cluster followed a given sequence. The color scheme for this figure is the same as it was in the top twenty pattern figures.

Each subfigure within the larger figure represents a cluster that I have named based on the sequences presented. For example, in one cluster of BA1 students, the top ten sequences that TraMineR chose to represent this cluster all involved earning a baccalaureate degree by the sixth year from entry. As this was the characteristic that seemed to differentiate this cluster from the others, I labeled this cluster "On-time Graduates." Figure A.9(a) shows the top ten representative sequences for this cluster.

I have arranged the clusters in the overall figure in order of how long it took the students to reach the graduation outcome. The biggest difference between (a) and (b) is the how many semesters it took the students to get there (4-6 years instead of 6-7 years). They are both otherwise characterized by mostly full-time attendance (green) with some part-time attendance (lavender) and some stop out (orange). The clusters represented by subfigures (c) and (d) represent graduation outcomes that occur in the 6-8 year range. They are differentiated from each other by the manner in which the students got to this outcome in that time frame. Students in subfigure (c) took a leave that lasted 3-6 semesters and graduated immediately upon returning to the system. Students represented by subfigure (d) were enrolled for a significant amount of time attempting fewer than twelve credits (part-time) on their way to graduation. This shows a difference in the manner in which students achieved their outcome, not simply in the duration to that outcome as we saw between (a) and (b). Subfigure (e) shows patterns that represent students who stopped out relatively early (within 2.5 years) and took a long break (2 - 6.5 years) before coming back to finish their degree.

Table 4 Cluster Description and Frequencies, BA1.

Outcome	Cluster Name	N	Percent within Degree Band	Percent Coverage
	On-time Graduates	16,269	46.73	81.6
	6-7 year Graduates	2,176	6.25	52.4
	7-8 year Graduates with Gap	881	2.53	21.0
Graduates	6-8 year Graduates with lots of part-time	804	2.31	10.0
	7-10 year Grads with a long break	513	1.47	12.2
Transfers	Early Transfers	3,710	10.66	75.7
	Late Transfers	567	1.63	41.4
Drop Outs	Early Drop Outs	7,490	21.51	77.2
	Late Drop Outs	1,920	5.52	49.9
Other	Characterized by a lot of part-time	483	1.39	14.5

Table 5 Cluster Description and Frequencies, BA2

Outcome	Cluster Name	N	Percent within Degree Band	Percent Coverage
Graduates	Associate Degree in 3-5 years	398	9.76	39.4
	Associate Degree in 6-9 years	158	3.87	13.9
	Earn Baccalaureate then Seek Associate	317	7.77	32.2
Transfers	Early Transfers	727	17.82	69.7
	Late Transfers	304	7.45	52.3
Drop Outs	Early Drop Outs	1,049	25.72	41.3
	Late Drop Outs	563	13.80	26.6
	Porpoising Enrollment	563	13.8	8.0

Table 6 Cluster Description and Frequencies, AA1

Outcome	Cluster Name	N	Percent within Degree Band	Percent Coverage
	2-4 year Associate	5,776	10.47	64.9
Graduates	5-7 year Associate because of break	1,001	1.82	13.8
	5-7 year Associate because of part-time	794	1.44	22.1
	7-10 year Associate because of long break	565	1.02	12.9
	Early Transfers	2,168	3.93	71.6
Transfers	Middle Transfers	1,221	2.21	60.3
	Late Transfers	1,042	1.89	59.1
	Drop out within 2 years	27,541	49.94	100.0
	Drop out in 2-4 years	9,090	16,48	71.0
	Drop out in 3-5 years with part-time	3,327	6.03	49.3
Drop Outs	Drop out in 4-5 years mostly full-time	1,813	3.29	56.4
	Drop out in 6-7	810	1.47	14.7

Table 7 Cluster Description and Frequencies, AA2

Outcome	Cluster Name	N	Percent within Degree Band	Percent Coverage
	2–4.5 year Associate, 4–5.5 year Baccalaureate	5.585	17.74	60.2
	2.5–4 year Associate, 5–6.5 year Baccalaureate	8.476	26.93	28.0
Graduate	3–5 year Associate, 5.5–8 year Baccalaureate, mostly part-time	2,123	6.75	5.5
	3–5 year Associate, 8–9.5 year Baccalaureate	1,062	3.37	4.0
Transfers	Early Transfers	6,821	21.67	64.7
	Late Transfers	2.446	7.77	49.6
Drop Outs	Drop out by year 5, staggered	3,578	11.37	36.4
Other	Mostly characterized by a large break	1,384	4.40	4.3

Tables 4 – 7 present the cluster solutions in detail. They are arranged by degree band, general outcome (graduation, transfer, drop out, or other). The within-degree-band percentage of students who fall into the cluster is presented along with the percentage of coverage for the representative sequences in the Figure A.9 – A.19. Alexis Gabadinho and Gilbert Ritschard define the level of coverage for a set of sequences within a cluster as "the percentage of cases that are within the neighbourhood of at least one of the patterns in the set" (2013).

In these tables we can see that earlier onset of outcome tends to be more common than later onset in terms of the number of students in a cluster. As an example, the most common graduation outcome for AA1 students is receiving an Associate degree in two to four years, with 10.47% of students in the degree band falling within this cluster (Table 6). However, only a further 4.27% of AA1 students are categorized in the other three graduation clusters. Despite this, it is still important to take note of these students as they are largely ignored by metrics that only measure "traditional" time to degree. For this degree band, however, the biggest group of concern is the 27.541 students (49.94%) who are in the "Drop out within two years" cluster.

As the two to four cluster solutions suggested by the measures of cluster quality mentioned above indicate, there are three general trends in outcome across all four degree band: Graduation, Transfer, and Dropping Out. This is significant variation within these group and some differences across degree bands, but overall most clusters fell into one of these three categories.

However, within each outcome there is a definite distinction to be made between clusters based on the number of semesters it took students to reach that outcome. As an example in the BA2 cluster solution there are two clusters that are characterized by transfer outside of the university system under study. The difference between the clusters appears to be that one group of students transfers rather quickly after leaving the system while the other group takes a long break before transferring. The Early Transfers time to outcomes ranged from one to three semesters after their last enrollment in the system while the Late Transfers had a gap in enrollment lasting between five and fifteen semesters.

In some of the degree bands, there was more heterogeneity among those clusters that involve taking longer to reach the eventual outcome. In the AA1 cluster solution there were 4 graduation clusters. Figure A.14 presents these clusters ordered by time to outcome. Similar to the graduation clusters for the BA1 degree band, there is an on-time cluster, two delayed-graduation clusters, and a very delayed graduation cluster. As with the BA1, the difference between the delayed graduation clusters is between students who mostly attend part-time and those who take a break. Also similar to the BA1 cluster solution, the extremely delayed graduation cluster is due to a lengthy break in enrollment (as opposed to taking a number of short breaks).

Implications of the Cluster Solution

The clusters indicated by the analysis presented here have a number of potential implications both for our understanding of graduation and for what policies might be implemented to improve student outcomes.

That there are students who take longer than what is traditional to graduate is not surprising given the previous work done in this area. However, I was expecting that there would be a larger variety of differences in how students got to their delayed graduation than was found in the analysis. I expected that there might be students who came in and out of the higher education system, taking breaks in order to work or for some other reason. However, I find in general that students who took more than six years to graduate did so because they were part-time for a large portion of their career in the university system or they took a multi-semester break, after which they returned and finished the degree. The length of the break for delayed students surprised me.

This second group of delayed graduates is potentially interesting from a policy standpoint because the difference between them and a dropout is that they came back. This implies that further work can be done to explore the systematic differences between dropouts and those students who succeed despite an absence from higher education.

The cluster solution for the BA1 degree band along with those for the other degree bands support the typology of delayed graduation that I am proposing:

- Those who have no delay
- Those who have a short delay
- Those who have a medium delay with a break
- Those who have a medium delay with part-time enrollment
- Those who have a long delay with a long break

This typology is potentially useful as the basis for further research that would inform a variety of interventions aimed at helping student graduation faster. For example, following up with students who had a long or short break before graduation could help administrators understand the reasons that ultimately successful students were not able to complete a degree in a contiguous manner.

Further, the existence of these delayed graduates whose pathway to degree includes a break implies a potential intervention for those students who have

stopped out of the system and have not yet returned to complete their degree. Checking in on a student who nearly has enough credits to graduate but is not currently enrolled would reveal a wealth of actionable information. It could allow administrators to discover if a student has stopped out because of family obligations, financial considerations, or some other factor outside the control of the institution. If, however, the reason a student stops out just shy of the necessary credits for a degree is something within the scope of influence of the institution (class availability, tutoring, etc.), then the institution can attempt to provide the necessary services to help the student see their degree to completion.

On the other hand, students who drop out early in their career (a large proportion of AA1 students) would be helped by interventions earlier in the process. If a comprehensive survey of the reasons a student drops out early could be found by following up with those students who drop out early, the institution could tailor supports to students who face similar obstacles at or before entry to college.

The difference between paths to degree for delayed graduates and the similarity between a delayed graduate and a drop out who has nearly enough credits to graduate implies the need for interventions to bring these students back into the fold. It is important to note that a single type of intervention will be less likely to work than a variety of interventions targeted at students based on their pathway through college. A student who struggles early on and is in danger of dropping out is likely to need different supports than a student who is nearly done with their degree but is having trouble crossing the finish line. The typology proposed in this paper has the potential to act as a framework for understanding students' path to degree and where in their careers students might need additional supports to complete in a more timely manner. While it is important not to judge students who take longer to earn their degree if that is their choice or is the best possible solution given their life circumstance, the sooner a student completes their degree the sooner they can reap the economic benefits it will afford them.

References

Abbott, A. and Tsay, A. (2000). Sequence analysis and optimal matching methods in sociology review and prospects. *Sociological Methods and Research*, 29(1), 3–33.

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

Gabadinho, A., Ritschard, G. (2013). Searching for typical life trajectories applied to childbirth histories in in R. Levy and E. Widmer (eds) Gendered life courses – Between individualization and standardization. A European approach applied to Switzerland, 287 – 312.

Gauthier, J. A., Widmer, E. D., Bucher, P., and Notredame, C. (2009). How much does it cost? optimization of costs in sequence analysis of social science data. *Sociological Methods and Research*, 38(1), 197–231.

Goldin, C. D. and Katz, L. F. (2009). *The race between education and technology*. Cambridge, MA: Harvard University Press

Lesnard, L. (2010). Setting cost in optimal matching to uncover contemporaneous sociotemporal patterns. *Sociological Methods and Research*, 38(3), 389–419.

Massey, D. S., Charles, C. Z., Lundy, G., and Fischer, M. J. (2011). *The source of the river: The social origins of freshmen at America's selective colleges and universities*. Princeton, NJ.: Princeton University Press

Murtagh, F. and Legendre, P. (2014). Ward's hierarchical agglomerative clustering method: Which algorithms implement ward's criterion? *Journal of Classification*, 31(3), 274–295.

NCES (2011) Six-Year Attainment, Persistence, Transfer, Retention, and Withdrawal Rates of Students Who Began Postsecondary Education in 2003–04. U.S. Department of Education. Institute of Education Sciences, National Center for Education Statistics. https://www2.ed.gov/rschstat/research/pubs/prinindicat/prinindicat.pdf Accessed January 9, 2016

Studer, M. (2013). Weightedcluster library manual: A practical guide to creating typologies of trajectories in the social sciences with R. LIVES Working Papers, 24.

Appendix
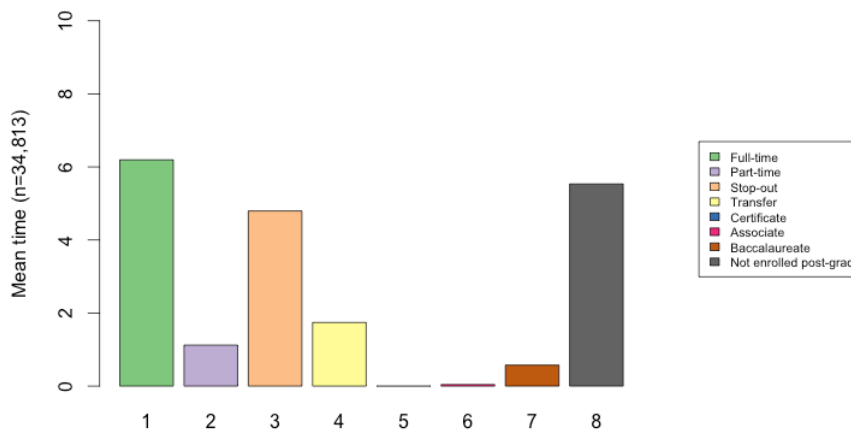
Fig. A.1 Mean Time in State, BA1
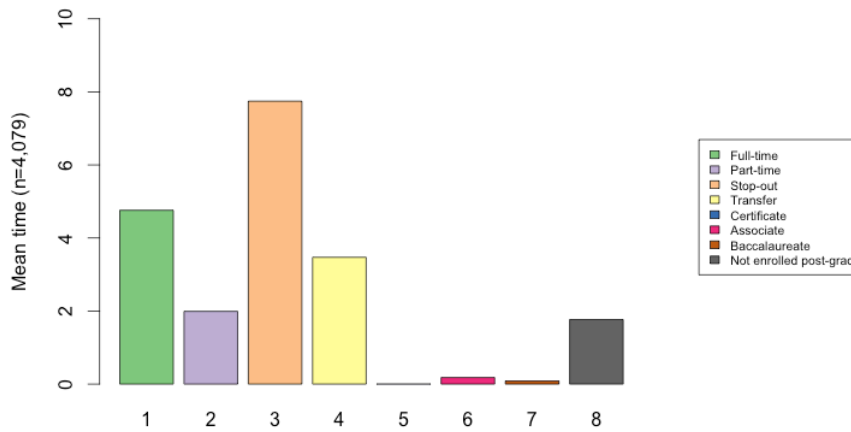
Fig. A.2 Mean Time in State, BA2.
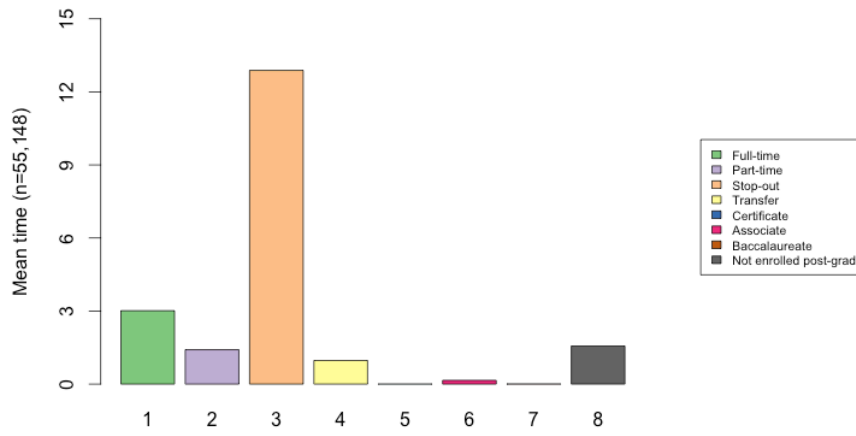
Fig. A.3 Mean Time in State, AA1.
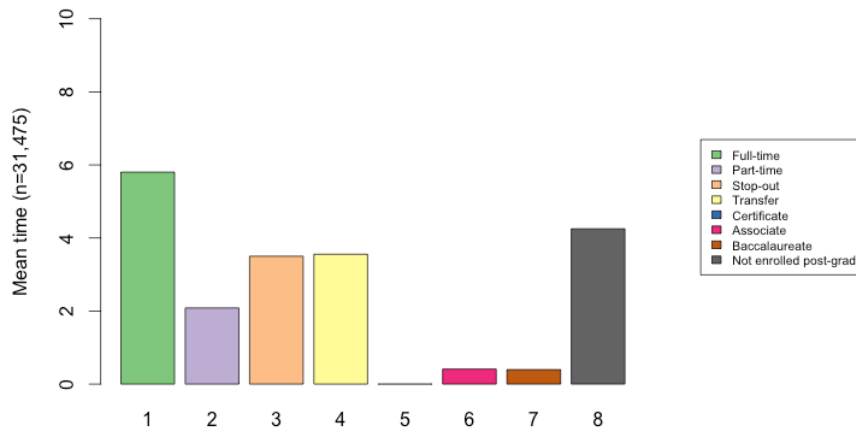
Fig. A.4 Mean Time in State, AA2.

Table A.5 Substitution Cost Matrix, BA1

	1	2	3	4	5	6	7	8
1	0.00	1.69	1.87	1.99	1.86	1.50	1.93	2.00
2	1.69	0.00	1.74	1.99	1.93	1.83	1.94	1.97
3	1.87	1.74	0.00	1.99	2.00	2.00	1.98	2.00
4	1.99	1.99	1.99	0.00	2.00	2.00	2.00	2.00
5	1.86	1.93	2.00	2.00	0.00	2.00	1.93	1.29
6	1.50	1.83	2.00	2.00	2.00	0.00	1.92	1.74
7	1.93	1.94	1.98	2.00	1.93	1.92	0.00	1.01
8	2.00	1.97	2.00	2.00	1.29	1.74	1.01	0.00

Table A.6 Substitution Cost Matrix, BA2.

	1	2	3	4	5	6	7	8
1	0.00	1.70	1.72	1.98	1.89	1.95	1.93	1.99
2	1.70	0.00	1.62	1.99	1.67	1.92	1.87	1.92
3	1.72	1.62	0.00	1.96	2.00	1.99	2.00	2.00
4	1.98	1.99	1.96	0.00	2.00	2.00	2.00	2.00
5	1.89	1.67	2.00	2.00	0.00	2.00	2.00	1.44
6	1.95	1.92	1.99	2.00	2.00	0.00	1.99	1.09
7	1.93	1.87	2.00	2.00	2.00	1.99	0.00	1.18
8	1.99	1.92	2.00	2.00	1.44	1.09	1.18	0.00

Table A.7 Substitution Cost Matrix, AA1

	1	2	3	4	5	6	7	8
1	0.00	1.71	1.70	2.00	1.82	1.94	1.96	2.00
2	1.71	0.00	1.60	2.00	1.83	1.93	1.82	1.98
3	1.70	1.60	0.00	1.99	2.00	2.00	2.00	2.00
4	2.00	2.00	1.99	0.00	2.00	2.00	2.00	2.00
5	1.82	1.83	2.00	2.00	0.00	1.96	2.00	1.39
6	1.94	1.93	2.00	2.00	1.96	0.00	2.00	1.08
7	1.96	1.82	2.00	2.00	2.00	2.00	0.00	1.22
8	2.00	1.98	2.00	2.00	1.39	1.08	1.22	0.00

Table A.8 Substitution Cost Matrix, AA2

	1	2	3	4	5	6	7	8
1	0.00	1.67	1.82	1.99	1.73	1.47	1.96	1.96
2	1.67	0.00	1.78	1.99	1.73	1.79	1.95	1.89
3	1.82	1.78	0.00	1.94	2.00	1.97	1.99	2.00
4	1.99	1.99	1.94	0.00	2.00	2.00	2.00	2.00
5	1.73	1.73	2.00	2.00	0.00	1.96	1.98	1.61
6	1.47	1.79	1.97	2.00	1.96	0.00	1.97	1.70
7	1.96	1.95	1.99	2.00	1.98	1.97	0.00	1.00
8	1.96	1.89	2.00	2.00	1.61	1.70	1.00	0.00

(a) On-time Graduates

(b) 6-7 year Graduates

(c) 7-8 year Graduates with Gap

(d) 6-8 year Graduates with lots of Part-Time

(e) 7-10 year Graduates with a long break
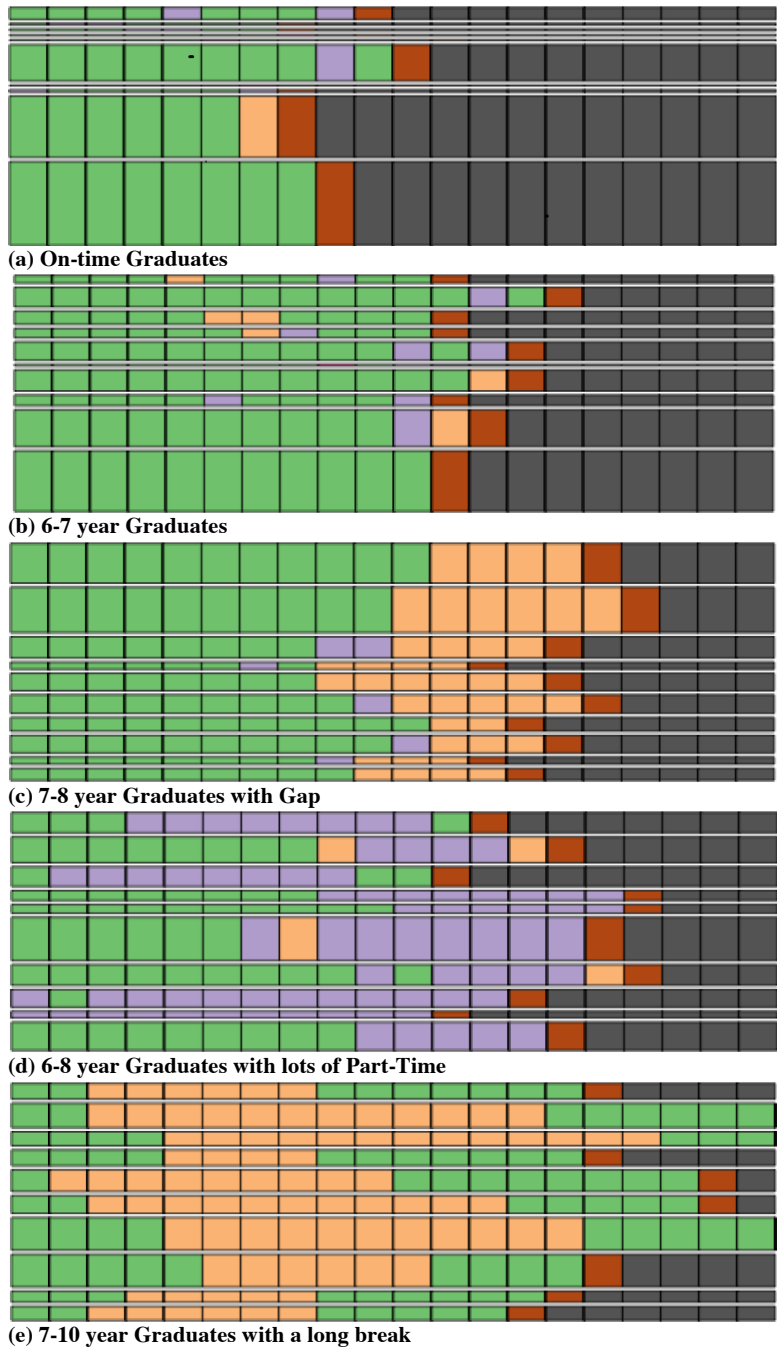
Fig. A.9 Representative Sequences, Graduation Group, BA1
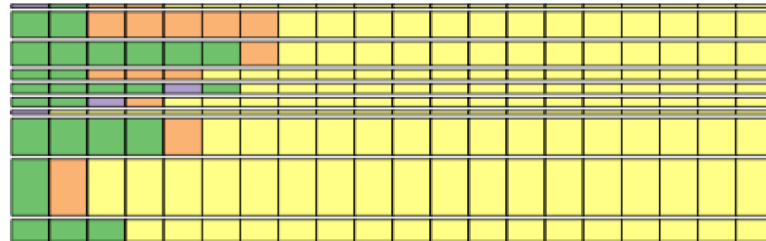
(a) Early Transfers
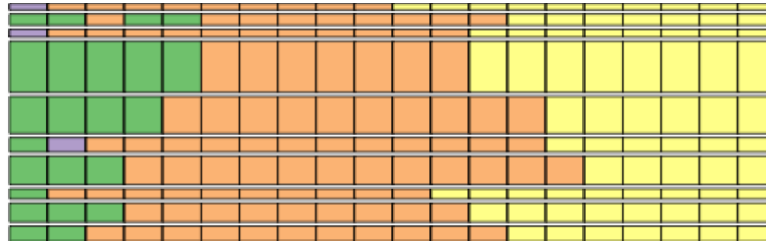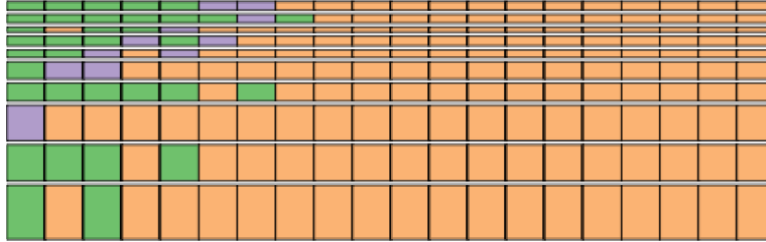
(b) Late Transfers

(c) Early Drop Outs
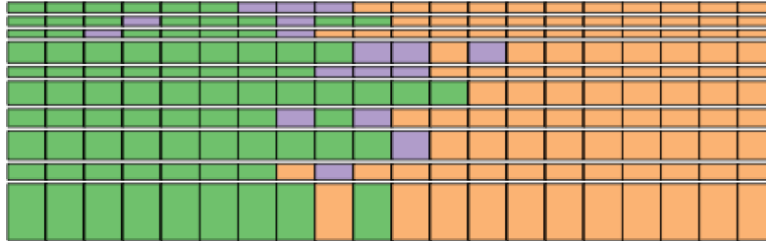
(d) Late Drop Outs
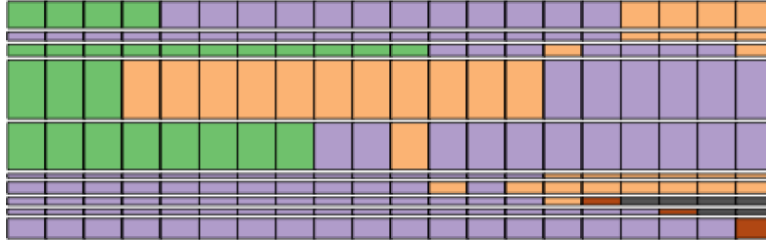
(e) Other, Characterized by a lot of part-time

Fig. A.10 Representative Sequences, Transfer, Drop Out, and Other Groups, BA1

(a) Associate Degree in 3-5 years

(b) Associate Degree in 6-9 years
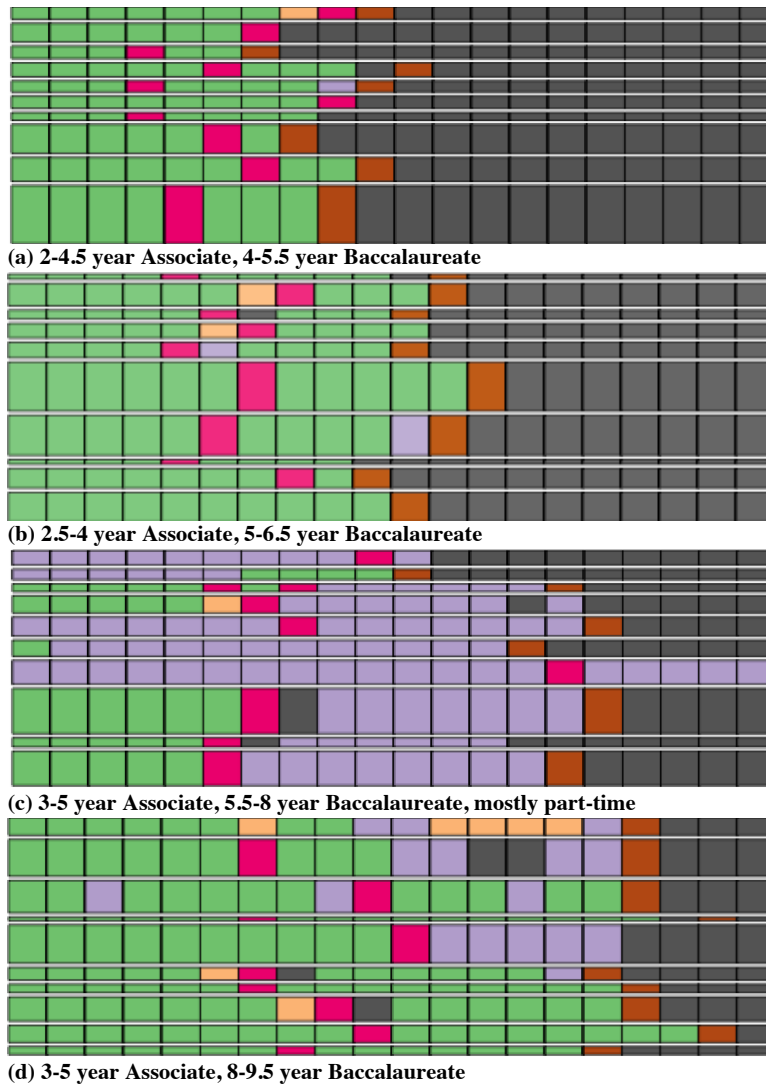
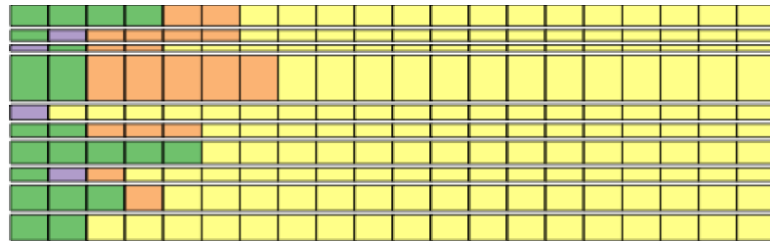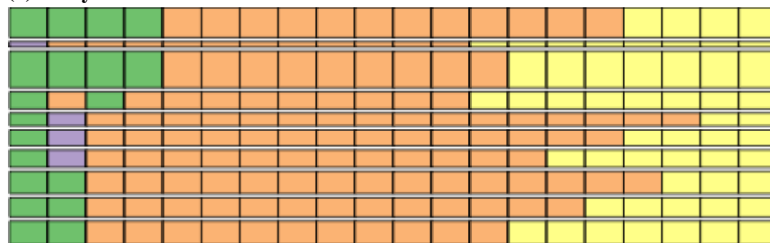**(c) Earn Baccalaureate then Seek Associate**
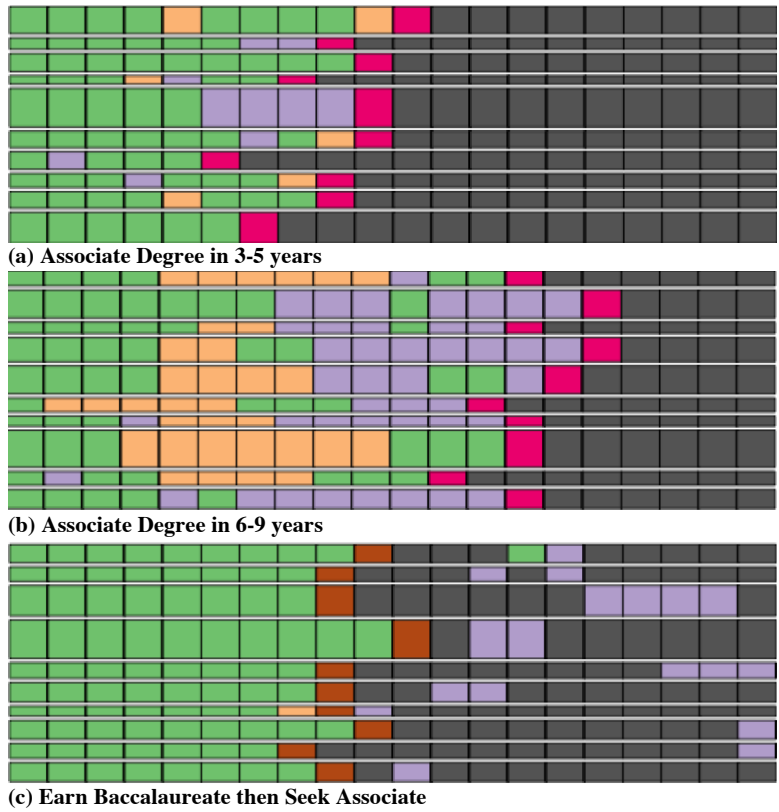
Fig. A.11 Representative Sequences, Graduation Group, BA2
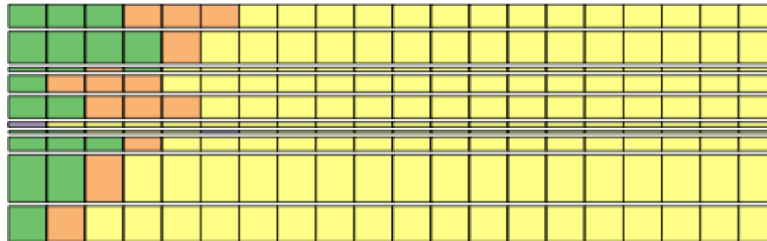
(a) Early Transfers
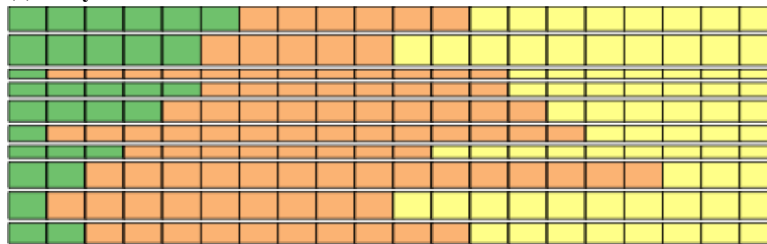
(b) Late Transfers

Fig. A.12 Representative Sequences, Transfer Group, BA2
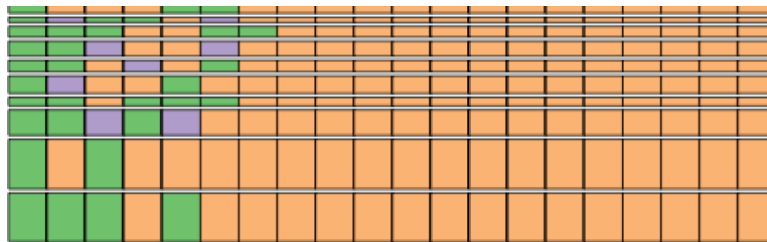
(a) Early Drop Outs
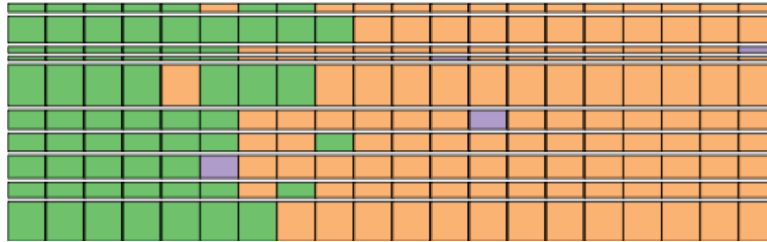
(b) Late Drop Outs
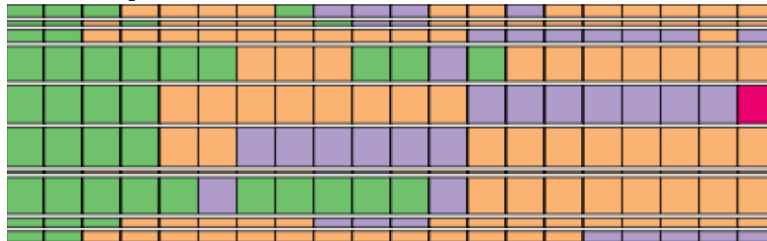
(c) Porpoising Enrollment

Fig. A.13 Representative Sequences, Drop Out Group and Other Group, BA2
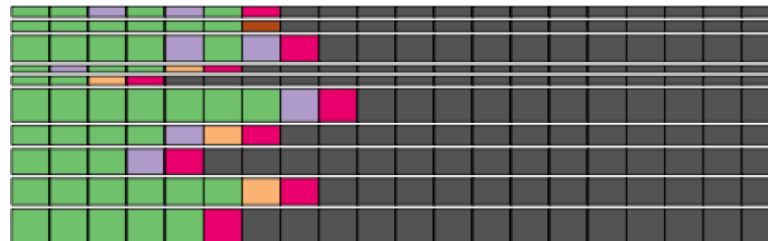
(a) 2-4 Year Associate
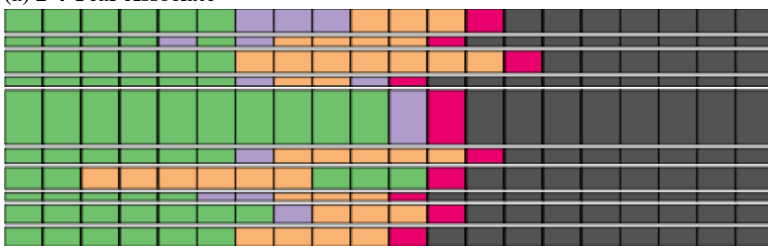
(b) 5-7 Year Associate because of break
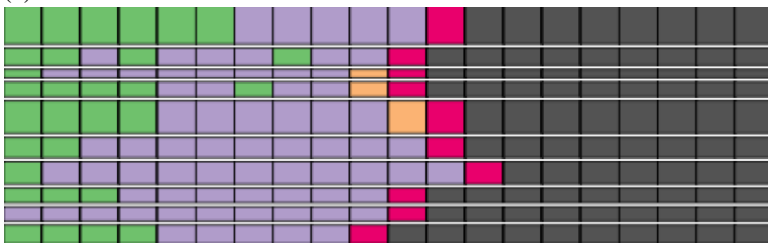
(c) 5-7 Year Associate because of part-time
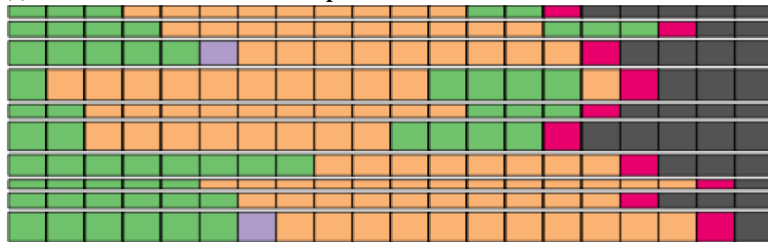
(d) 7-10 Year Associate because of long break

Fig. A.14 Representative Sequences, Graduation Group, AA1.

(a) Early Transfers

(b) Middle Transfers
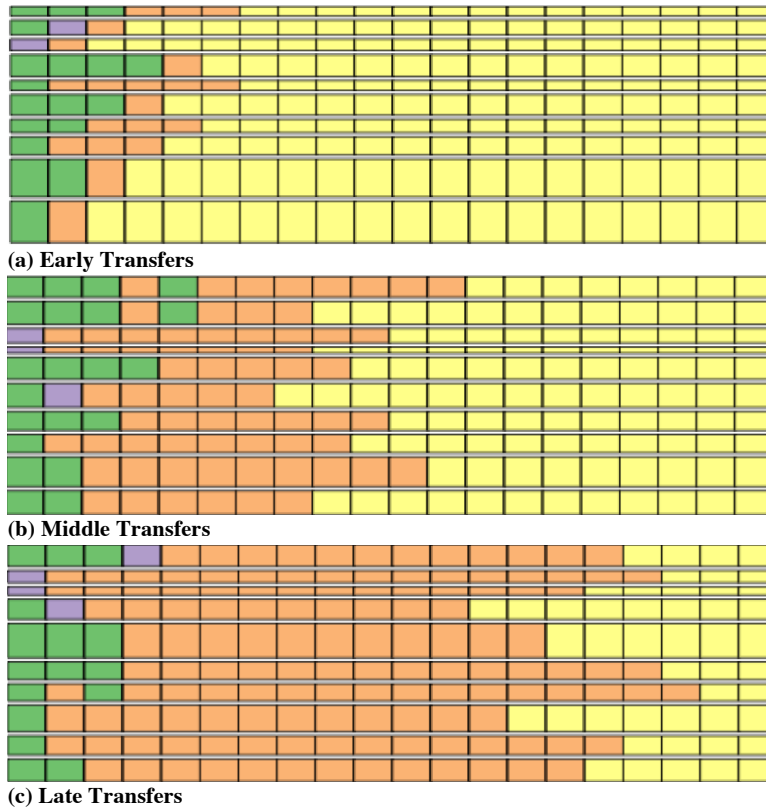
(c) Late Transfers

Fig. A.15 Representative Sequences, Transfer Group, AA1

(a) Drop out within 2 years

(b) Drop out in 2-4 years

(c) Drop out in 3-5 years with part-time enrollment

(d) Drop out in 4-5 years mostly full-time enrollment

(e) Drop out in 6-7 years
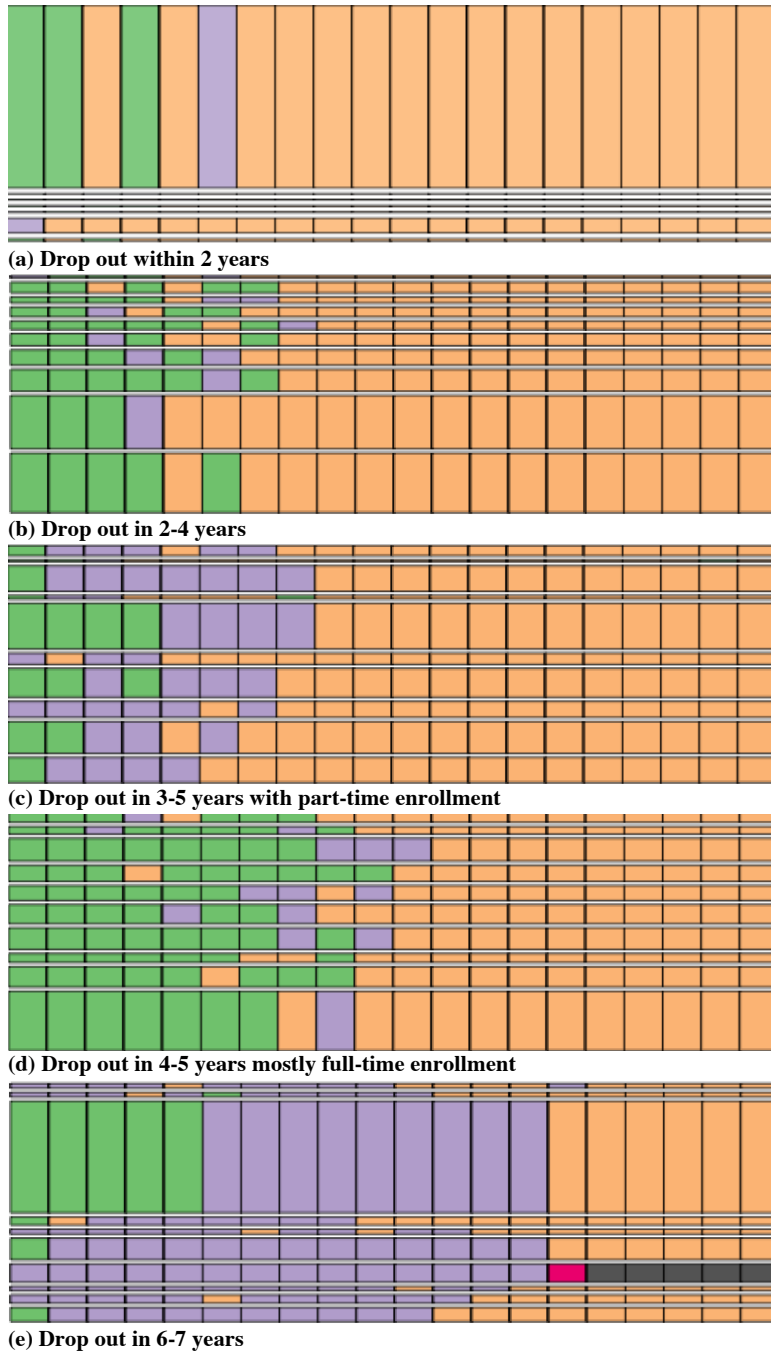
Fig. A.16 Representative Sequences, Drop Out Group, AA1

(a) 2-4.5 year Associate, 4-5.5 year Baccalaureate

(b) 2.5-4 year Associate, 5-6.5 year Baccalaureate

(c) 3-5 year Associate, 5.5-8 year Baccalaureate, mostly part-time

(d) 3-5 year Associate, 8-9.5 year Baccalaureate

Fig. A.17 Representative Sequences, Graduation Group, AA2.

(a) Early Transfers

(b) Late Transfers

Fig. A.18 Representative Sequences, Transfer Group, AA2
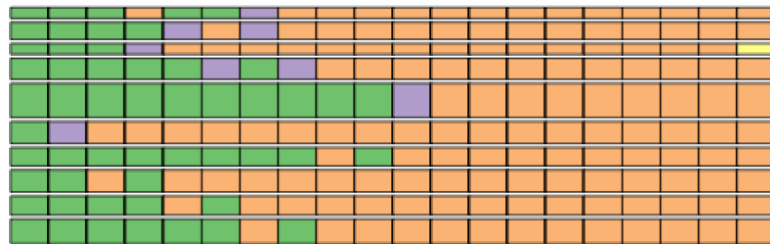
(a) Drop out by year 5, staggered
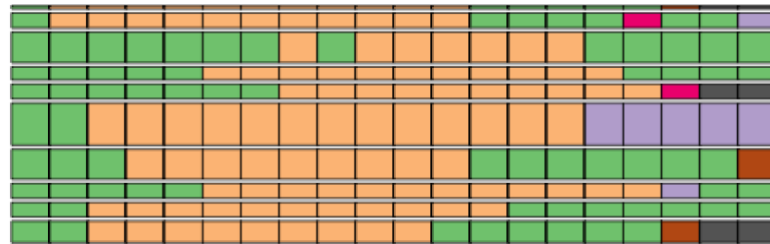
(b) Other, Mostly characterized by a large break

Fig. A.19 Representative Sequences, Drop Out and Other Group, AA2.

Session 14B: Sequence summary indexes

Measuring sequence complexity

A conceptual and empirical comparison of two composite complexity indices

Georgios Papastefanou

GESIS Leibniz Institute for the Social Sciences, Mannheim/Germany

For a causality oriented analysis of sequences, namely by estimating their covariation with exogenous social variables like social-economic status, gender, age or attributes of family of origin, structural pattern, which characterizes a sequence, has to be represented by a quantifying indicator. One way is to capture a sequence's complexity, by constructing an indicator with a specific quantitative range. Two prominent approaches, Gabadinho et al. (2011) and Elzinga (2010) propose each a different approach to incorporate crucial features of sequence patterning like variety (qualitative differentiation of states), variability (temporal differentiation of states as episodes) and regularity (repetition of subsequences) .

Gabadinho et al. (2011) propose this formula:

$$-1- \qquad C(s) = \sqrt{\left(\frac{q(s) * h(s)}{q_{max} * h_{max}}\right)}$$

As formula 1 indicates, two dynamic structural features are involved in the complexity index, namely the temporal variability on the one hand, as measured by the change of frequency and variety, as measured by the Shannon entropy. Moreover, this index takes into account the fact that individual sequences may vary in their lengths. Linking of the two normalized components change intensity and entropy is done by geometric mean, i.e. the square root of the product of normalized change intensity and normalized entropy.

The complexity index of T Elzinga (2010) is calculated using the following formula -2:

$$-2- \quad T(s) = log_2\left(\varphi \cdot \frac{s^2_{t,max}(s)+1}{s^2_t(xs)+1}\right)$$

where:

Log$_2$== logarithm to the base 2
φ == number of sub-sequences with distinct successive states
$s^2_{t,max}(s)$== maximum variance of episode durations in a sequence for a given number of episodes
$s^2_t(xs)$== variance of the durations of episodes within a given sequence

Für **$s^2_{t,max}(s)$**applies: $s^2_{t,max}(x) = \left(\ell_d(x)-1\right)\left(1-\bar{t}(x)\right)^2$

As both indicators want to be named as complexity indices, we will label the Gabadinho et al. approach as complexity C and Elzinga's definition as complexity T.

In a detailed conceptual analysis we discuss the foundation and restrictions of their components like transition rate, normalized entropy, number of distinct successive states and normalized episode duration variability. Further we examine interchangeability of C and T as is stated by Gabadinho et al (2011). We find – based on comparing C and T for nine systematically varied sequences – that there is nearly no co-variation between C and T.

Tabelle 1: Complexity indices C and T and their components for nine examplary sequences

sequence-Nr.	Full sequence pattern	Distinct states sequenc pattern	Transition rate	Normalized Entropy (sample-based)	Complexity C	Ranking by C	Phi (# DSS[1])	Variance factor	Complexity T	Ranking by nach T
1	AABBAABBAABB	ABABAB	0.45	0.39	0.42	7	33	6	7.63	1
7	AAAAAAAAAABBBAAAAAAAAAAB BBCCCC	ABABC	0.13	0.48	0.25	1	24	8.6	7.68	2
2	AABBCCBBAABB	ABCBAB	0.45	0.57	0.51	8	46	6	8.11	3
6	AAAAAAAAAABBBCCAAAAAAAAA ABBBCC	ABCABC	0.17	0.48	0.29	4	52	5.9	8.26	4
8	AAAAAAAAAABBBAAAAAAAAAAB BBCCCB	ABABCB	0.17	0.47	0.28	3	41	8.6	8.46	5
9	AABBAABBAABBAABB	ABABABAB	0.47	0.39	0.66	9	88	10	9.78	6
3	AAAAABBBBBAAAAABBBBBAAAAA BBBBB	ABABAB	0.17	0.39	0.26	2	33	81	11.38	7
5	AAAAABBBBBCCCCCAAAAABBBBB CCCCC	ABCABC	0.17	0.61	0.32	5	52	81	12.04	8
4	AAAAABBBBBCCCCCDDDDDEEEEEF FFFF	ABCDEF	0.17	1	0.41	6	64	81	12.34	9

note: 1) number of distinct successive sub-sequences

The calculated rank correlation coefficient of these sample sequences is 0.02. This very low correlaton means, that one gets quite a different complexity ranking of these sequences, if one uses T or C.

For a more extensive test of the substitutability of T and C we did an empirical analysis of 2000 sequences of leisure activities on Sunday, based on the German Time Use Survey of 2001/2002. As a starting point we take the issue of complexity of the personal leisure time on weekends (Papastefanou, Gruhler 2014). For this purpose, we make use of the data collected in the time use survey of the Federal Statistical Office from 2001-2002. As alphabet of leisure activities on Sunday following activities are defined: reading, listening to music, watch television, computers, pursue hobbies, sports, and the residual category "other activities". As indicators of the socio-structural situation, the following variables are taken in account: gender, age, marital status, household income (interval categories), household size, general secondary education, vocational education and occupational status. For ease of interpretation, we restrict the target group to persons aged over 17 years who are employed full-time.

First, we find a high co-variation of both complexity indices C and T: for the group of full-time employees over 17 years the Pearson correlation coefficient is r = .94. The two composites complexity indices so appear to be interchangeable.

But a multivariate modeling of T and C of leisure time sequences on Sunday as outcome of social determinants like gender, age, family status, net household income, household size, education status and occupational status reveal significant differences between the models (see table 2).

Tabelle 2: Socio-demographic covariates of complexity scores and of their components (separate OLS-Regressions, b, standard error)

VARIABLES	(1) Komplexität C	(2) Komplexität T	(3) Wechselrate	(4) rel. Entropie	(5) Log₂(phi)	(6) Log₂(Varianzfaktor)
gender	-0.0283***	-1.359***	-0.0133***	-0.0654***	-0.766***	-0.622***
	(0.00472)	(0.245)	(0.00384)	(0.00870)	(0.141)	(0.181)
age	0.000281	0.0226*	0.000106	0.000995**	0.0216***	0.0152
	(0.000259)	(0.0136)	(0.000211)	(0.000477)	(0.00778)	(0.00991)
single	0.0105*	-0.0727	0.00463	0.0118	0.00494	-0.208
	(0.00635)	(0.331)	(0.00517)	(0.0117)	(0.190)	(0.243)
divorced	-0.00510	0.00606	-0.00782	-0.00822	-0.0378	-0.178
	(0.00746)	(0.384)	(0.00605)	(0.0137)	(0.220)	(0.284)
widowed	-0.0141	0.100	-0.00992	-0.00983	-0.274	0.658
	(0.0210)	(1.072)	(0.0174)	(0.0393)	(0.615)	(0.816)
separated	-0.0174	-0.879	-0.00292	-0.0356	-0.412	-0.0629
	(0.0183)	(0.954)	(0.0151)	(0.0343)	(0.547)	(0.712)
Size of household	-0.000589	-0.0905	-0.000336	-0.00119	-0.0300	-0.0180
	(0.00185)	(0.0960)	(0.00151)	(0.00343)	(0.0551)	(0.0712)
Middle educational status	0.00167	0.356	-0.000253	0.00224	0.351**	0.0659
	(0.00532)	(0.278)	(0.00431)	(0.00977)	(0.160)	(0.203)
High educational status	0.0110*	0.939***	0.00698	0.0192	0.722***	0.420*
	(0.00649)	(0.339)	(0.00527)	(0.0119)	(0.195)	(0.248)
Fachschule/Meister	-0.00107	0.204	-0.00334	0.00116	0.162	0.0590
	(0.00669)	(0.351)	(0.00541)	(0.0123)	(0.201)	(0.255)
University degree	-0.00629	-0.00485	-0.00310	-0.0135	0.0646	-0.00153
	(0.00782)	(0.407)	(0.00630)	(0.0143)	(0.234)	(0.297)
Civil servant	0.0220***	0.769**	0.0116**	0.0410***	0.520**	0.363
	(0.00708)	(0.368)	(0.00570)	(0.0129)	(0.211)	(0.268)
White collara	0.0233***	1.083***	0.0162***	0.0527***	0.696***	0.914***
	(0.00614)	(0.319)	(0.00497)	(0.0113)	(0.183)	(0.234)
Blue collar	0.0274***	1.388***	0.0190***	0.0538***	0.892***	1.001***
	(0.00694)	(0.363)	(0.00561)	(0.0127)	(0.208)	(0.264)
apprenticeship	0.0167	0.916	0.00877	0.0433**	0.311	1.007**
	(0.0112)	(0.584)	(0.00902)	(0.0204)	(0.335)	(0.424)
Military/ civil service	0.0167	1.064	0.00264	0.0338	0.716	-0.304
	(0.0206)	(1.074)	(0.0160)	(0.0363)	(0.616)	(0.754)
Constant	0.177***	1.472	0.0868***	0.352***	4.719***	-4.772***
	(0.0201)	(1.050)	(0.0163)	(0.0370)	(0.602)	(0.767)
Observations	2,088	1,998	2,170	2,170	1,998	2,170
R-squared	0.033	0.032	0.016	0.045	0.043	0.021

In brackets: Standard error. *** $p<0.001$, ** $p<0.01$, * $p<0.1$

We find, that in the model of the complexity of T a linear effect of age is estimated to be significant, which is not estimated significantly in the model of complexity of C. For example, the estimation of the model of the complexity C suggests that unmarried persons show significant higher complexity of leisure activities on Sunday as married ones. Also, it should be noted that the Model T is to determine a clearer differentiation in the differences between the professional status groups officials, employees and workers. According to this model, workers had the highest complexity, while the three status groups are not essentially different from each other for the model with C.

 In sum it seems, that T and C might represent (at least partially) substantially different processes of sequence patterning. This assumption finds support in modeling the social effects on the components of C and T separately. In sum, we conclude that it might be more adequate to analyse the components of C and T separately instead of their joint incorporation into C and T, because the components especially like normalized entropy and number distinct successive subsequences seem to represent different processes of sequence differentiation like variety and regularity.

References

Elzinga Cees H., (2010), Complexity Of Categorical Time Series, *Sociological Methods Research February 2010 vol. 38 no. 3 463-481*

Gabadinho, A., G. Ritschard, N. S. Müller and M. Studer ,2011, Analyzing and Visualizing State Sequences in R with TraMineR. *Journal of Statistical Software* **40**(4), 1-37.

Measuring early employment insecurity and its effects

Margherita Bussi and Jacqueline O'Reilly

University of Brighton

Abstract In this paper we investigate the early employment insecurity of young people's career by applying the "index of complexity". This index measures the entropy and the number of transitions within a sequence (Gabadinho et al 2010). The entropy indicates the distribution of different positions within a sequence, hence its higher of lower degree of predictability; the number of transitions reveals the instability of the trajectory. This index is used to described the quality of young people's trajectories in the UK using the UKHLS, also known as Understanding Society. The index is calculated for the trajectories covering the first four waves (2009-2013). The index is then used to unveil the relationship between a young person's past trajectory and their labour market position in the fifth wave (2013-2014). We expect that young people with unstable trajectories are more likely to be out of the labour market or in temporary employment. This will confirm the scarring effect of early employment precariousness.

Keywords: trajectories, early employment insecurity, index of complexity, scarring, UKHLS

1   Introduction

A large body of research recognised that first transitions in the labour market have become more hectic, less straightforward and more individualised (Vickerstaff 2006). In particular, transitions from temporary contracts into permanent contracts and their effects on the younger generations have found an increasing interest and stimulated prolific research (Scherer 2004; Gebel 2010; Ortiz 2010). Moreover, the crisis of the last few years has brought the challenge of youth employment again high on the agenda as the crisis has exacerbated the risk of downwards transitions or getting trapped into unemployment or inactivity for the younger generation (European Commission 2014; O'Reilly et al 2015).

Several studies have investigated the impact of the initial job position (Gebel 2010; Nickell et al 2002) as well as the impact of spending long spells in unemployment at the beginning of the work career (Hammarström and Janlert 2002; Strandh et al 2014). Findings confirm that first negative labour market experiences can have far-reaching negative impact on labour market opportunities but also several aspects of life (Cable et al 2008; Schmelzer 2011; Weish and Lewis 1998). This effect is known as "scarring effect" (Bell and Blanchflower 2011; Gregg and Tominey 2005; Tumino 2009).

2   Early job v. early employment insecurity

When condensing the increasing attention towards precariousness and churning (Standing 1999) particularly regarding young people, we argue that it is more interesting to switch the attention towards the concept of early-employment insecurity rather than of early-job insecurity. While the latter is limited to the single, though important, first work experience; the first is believed to better grasp the precariousness and the individual negotiated choices embedded in multiple changes overtime (Fuller 2009). Moreover, when considering the negative

consequences of precariousness - such as the deterioration of human capital, stigma, negative signalling and the status dependency (Ayllón 2013) -there is an important temporal dimension of accumulation that should not be overlooked and can be accounted for in longitudinal analysis.

3   Describing trajectories: some indicators

Looking at trajectories implies adopting a longitudinal perspective and adapted quantitative methods. Investigating trajectories has become increasingly common in the social sciences and in particular in life-course analysis looking at careers and social status (Abbott and Tsay 2000, Halpin and Cban 1998; Widmer and Ritschard 2009). A second generation of sequence analysis is being developed to answer the first criticisms and go beyond some initial limits (Aisenberg and Fasang 2010; Scherer 2001).

The quality of the trajectories investigated as been mainly assessed in terms of membership to clusters where one or more patterns could be easily identified. However, cluster analysis, on top of its arbitrary capacity of finding patterns even in unstructured data (Scherer 2001), has also the limit that the cluster membership is similar for all those individual observations belonging to that cluster. This implies that it does not provide a degree of quality for each single individual.

To our best knowledge, three contributions developed a measure to establish the quality of trajectories: the "turbulence" indicator developed by Elzinga and Liefbroer (2007), the "index of complexity" by Gabadinho et al (2010) and the volatility and integrative capability by Brzinsky-Fay (2007).

4   The quality of trajectories and scarring effect

Several aspects can be taken into account when looking at the trajectories: the type of statuses experience, their timing, the spell length of the statuses, the sequencing and the total time spent in a specific status (Studer and Ritschard 2014). Among these aspects, we expect that the number of statuses experienced and their distribution (i.e. time spent in specific status) are important in determining the instability and predictability of the trajectories. For this reason, we chose the "index of complexity" to measure the precariousness of a trajectory because it it brings together two important aspects: *entropy* – i.e. the distribution of statuses within a sequence - and the *number of transitions* (Gabadinho et al 2010); besides, it implies manageable calculations and quite straightforward interpretation.

Taking into account the relevance of trajectories for future labour market outcomes and the current state of the art of the tools used for measuring the quality of trajectories, we argue that:

H: The more complex trajectories are, the more likely a young person will be to be unemployed, inactive or in temporary employment in the fifth wave.

5   Data analysis

The data used are a subsample drawn from the first wave of the UK Household Longitudinal Survey (Understanding Society) of young people, excluding students, in the early work career. Young people's trajectories are investigated over four years and their labour market positions are grouped into six categories: « being employment on a permanent contract », « being employed on a temporary contract », « being in education/training/apprenticeship scheme », « being unemployed », « being inactive » (excluding maternity leave), and « other ».

The « index of complexity » (Gabadinho et al 2010) will be calculated for each individual trajectory and descriptive statistics will be used to show differences of the index according to the main socio-demographic variables.

The second step will be to adapt the index of complexity in order to make it suitable for being used as an explanatory variable in a series of logistic regressions. Because it is based on entropy and the number of transitions, a labour market trajectory only composed of a single employment spell will have an index of complexity equals to 0: no transitions and, as a results, no different proportions of different labour market statuses. A similar result would be obtained for a trajectory of a young person who spent the whole observation period in unemployment. The index of complexity would be, again, equal to 0. A solution based on a weighted entropy will be investigated where a subjective/utility of the spell is accounted for (Kannappan 1980) in order to account for the subjective utility of the spells.

Finally, to investigate the effect of the complexity of trajectories, the modified index of complexity will be then included, as an independent variable, in a set of logistic regressions investigating the probability of being in employment or outside the labour market in the fifth wave of the survey. The regressions will establish whether, all other things being equal, our hypothesis is validated or not.

6   Expected results

In line with the literature, we expect that young people with lower education, with a migrant background and with a fragile household position (e.g. namely single parents) will be more likely to have complex/precarious trajectories. Some other relationships between the quality of the trajectory and individual's living environment are also likely to be explored.
We also expect to find that a higher degree of complexity as a negative effect on the probability of being employed in the last year of observation (fifth wave).

7   References

Abbott, A. & Tsay, A. (2000). Sequence analysis and optimal matching methods in sociology review and prospect. *Sociological methods & research*, 29(1), 3-33.

Aisenbrey, S. & Fasang, A. E. (2010). New Life for Old Ideas. The 'Second Wave' of Sequence Analysis - Bringing the Course Back into the Life Course. *Sociological Methods & Research*, 38(3), 420-462.

Ayllón, S. (2013). Unemployment persistence: not only stigma but discouragement too. *Applied Economics Letters*, 20(1), 67-71.

Bell, D.N. & Blanchflower, D.G. (2011). Young people and the Great Recession. *Oxford Review of Economic Policy*, 27(2), 241-267.

Brzinsky-Fay, C. (2007). Lost in transition? Labour market entry sequences of school leavers in Europe. *European Sociological Review*, 23(4), 409-422.

Cable, N., Sacker, A., Bartley, M. (2008). The effect of employment on psychological health in mid-adulthood: findings from the 1970 British Cohort Study. *Journal of Epidemiology and Community Health*, 62(5).

Duckworth, K, & Schoon, I. (2012). Beating the Odds: Exploring the Impact of Social Risk on Young People's School-to-Work Transitions during Recession in the UK. *National Institute Economic Review*, 222(1), R38-R51.

Elzinga, C.H. & Liefbroer, A.C. (2007). De-standardization of family-life trajectories of young adults: A cross-national comparison using sequence analysis. *European Journal of Population/Revue européenne de Démographie*, 23(3-4), 225-250.

European Commission. (2014). Labour market transitions before and during a severe economic downturn: some evidence from micro-economic data. *Social Europe - EU Employment and Social Situation*. Supplement March 2014.

Ferrie, J., Martin, E., Shipley, J., Marmot, M.G., Stansfeld, S., Davey Smith, G. (1998). The health effects of major organisational change and job insecurity. *Social Science & Medicine* 46(2), 243-54.

Fuller, S. (2009). Investigating longitudinal dimensions of precarious employment. In L. F. Vosko, M. MacDonald, I. Campbell (Ed.), *Gender and the Contours of Precarious Employment:* Routledge.

Gabadinho, A., Ritschard, G., Studer, M., Müller, N.S. (2010). Indice de complexité pour le tri et la comparaison de séquences catégorielles, *Revue des nouvelles technologies de l'information RNTI*. 19, 61-66.

Gangl, M. (2002). Changing labour markets and early career outcomes: labour market entry in Europe over the past decade. *Work, Employment & Society*, 16(1), 67-90.

Gautié, Jérôme. 2009. *Le chômage*. Paris: La Découverte.

Gebel, M. (2010). Early career consequences of temporary employment in Germany and the UK. *Work, Employment & Society*, 24(4), 641-660.

Gregg, P. and Tominey, E., 2005. The wage scar from male youth unemployment. *Labour Economics*, *12*(4), pp.487-509.

Halpin, B. and Cban, T.W., 1998. Class careers as sequences: An optimal matching analysis of work-life histories. *European sociological review*, *14*(2), pp.111-130.

Hammarström, A. and Janlert, U., 2002. Early unemployment can contribute to adult health problems: results from a longitudinal study of school leavers. *Journal of Epidemiology and Community Health*, *56*(8), pp.624-630.

Kannappan, P., 1980. On some functional equations from additive and nonaddistive measures (III). Stochastica, 4(1), pp.15-22.

Nickell, S., Jones, P. and Quintini, G., 2002. A picture of job insecurity facing British men. *The Economic Journal*, *112*(476), pp.1-27.

O'Reilly, J., Eichhorst, W., Gábos, A., Hadjivassiliou, K., Lain, D., Leschke, J., McGuinness, S. Kureková, L.M. et al. (2015). Five Characteristics of Youth Unemployment in Europe. *Flexibility, Education, Migration, Family Legacies, and EU Policy*, 5(1).

Ortiz, L. (2010). Not the right job, but a secure one over-education and temporary employment in France, Italy and Spain. *Work, Employment & Society*, 24(1), 47-64.

Scherer, S. (2001). Early career patterns: A comparison of Great Britain and West Germany. *European sociological review*, 17(2), 119-144.

Scherer, S. (2004). Stepping-stones or traps? The consequences of labour market entry positions on future careers in West Germany, Great Britain and Italy. *Work, Employment & Society*, 18(2), 369-394.

Schmelzer, P. (2011). Unemployment in early career in the UK A trap or a stepping stone?. *Acta Sociologica*, 54(3), 251-265.

Standing, G. (1999). *Global Labour Flexibility: Seeking Distributive Justice*. Basingstoke: Macmillan.

Strandh, M., Winefield, A., Nilsson, K., Hammarström, A. (2014). Unemployment and mental health scarring during the life course. *The European Journal of Public Health*, 24(3), 440-445.

Studer, M., & Ritschard, G.. (2014). A comparative review of sequence dissimilarity measures. *LIVES Working Papers*, doi:10.12682/lives.2296-1658.2014.33

Tumino, A. (2015). The scarring effect of unemployment from the early '90s to the Great Recession. *Institute for Social and Economic Research*, 05.

University of Essex. Institute for Social and Economic Research and National Centre for Social Research, Understanding Society: Waves 1-5, 2009-2014 [computer file]. Colchester, Essex: UK Data Archive [distributor], December 2015. SN: 6614.

Vickerstaff, S. (2006). Life Course, Youth, and Old Age. In P. Taylor-Gooby, J.O. Zinn (Ed.) *Risk in social science*. Oxford: Oxford University Press.

Weich, S. & Lewis, G. (1998). Poverty, unemployment, and common mental disorders: population based cohort study. *BMJ* 317(7151), 115-19.

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

Wu, L. L. (2000). Some Comments on Sequence Analysis and Optimal Matching Methods in Sociology: Review and Prospect, *Sociological Methods and Research* 29(1), 41-64.

Binary Sequence Dynamics applied to Career Quality

Anna Manzoni and Irma Mooi-Reci

Abstract In this paper we propose a measure to quantify successfulness in binary sequences that can be meaningfully interpreted as sequences of Successes and Failures. In order to operationalize the concept of successful and less successful sequences, we formulate some general properties that a measure of successfulness must adhere to, construct a measure that does fulfill these requirements and, finally, show that the measure can be modeled in a theoretically meaningful way. Furthermore, we show an application of our measure to model the successfulness of binary sequences on the labor market.

Anna Manzoni
North Carolina State University 27695-8107 Raleigh, NC, USA
e-mail: anna.manzoni@gmail.com

Irma Mooi-Reci
University of Melbourne, Australia

Turnover of individuals with similar career sequences as predictor of employer change

Katja Dlouhy, Torsten Biemann

University of Mannheim, Germany

Abstract Occupational career patterns are conceptualized as sequences that consist of individuals' occupational states and employer changes. These sequences are often similar, as many careers are path dependent and follow general patterns. Our hypothesis is that employee turnover can be predicted by employer changes of individuals with similar career trajectories. We derived 1,651 career sequences that incorporate 20 years of individuals' occupational positions from a large national panel. The similarity of career sequences was assessed with the optimal matching method. We then used the resulting similarity measures as weights for a novel predictor of individuals' employer changes. In support of our hypothesis, employer changes in similar career sequences predicted turnover. The method introduced in this study could help in reinforcing the use of prospective, longitudinal designs in career literature.

1. Introduction

Research on employee turnover has a long tradition in the organizational sciences. Numerous studies analyzed how individual, organizational, and contextual factors influence individuals' turnover intentions and actual turnover (Cotton & Tuttle, 1986; Crossley, Bennett, Jex, & Burnfield, 2007; Griffeth, Hom, & Gaertner, 2000). Accordingly, meta-analytic evidence offers a comprehensive picture of relevant antecedents, for example job satisfaction, pay, job content, or alternative job opportunities (Heavey, Holwerda, & Hausknecht, 2013; Tett & Meyer, 1993; Zimmerman, 2008). Following the rationale of behavioral consistency, the number of previous employer changes has been shown to be a valid predictor of voluntary turnover, too (Judge & Watanabe, 1995). However, reducing the working life to the total number of previous employer changes might be an oversimplification, as an individual's occupational career contains much richer information.

Next pages blanked out on authors' request.

2. Theory

2.1. Career patterns as manifestation of job-related personal dispositions

2.2. Path dependence

2.2.1. Path dependence in organizations

2.2.2. Path dependence in careers

3. Method

3.1. Sample

3.2. Measures

3.3. Data analytic strategy

3.3.1. Optimal matching analysis

3.3.2. Join count analysis

3.3.3. Logistic regression with similarity weights

4. Results

4.1. Join count analysis results

4.2. Logistic regression analysis results

13

4.3. Simulation of sequences without career patterns

5. Discussion

5.1. Prediction of employee turnover

5.2. Limitations and future research

5.3. Conclusion

References

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

Aisenbrey, S., & Fasang, A. E. (2010). New life for old ideas: The "second wave" of sequence analysis bringing the "course" back into the life course. *Sociological Methods & Research,* 38(3), 420–462.

Allen, D. G., Shore, L. M., & Griffeth, R. W. (2003). The role of perceived organizational support and supportive human resource practices in the turnover process. *Journal of Management*, 29(1), 99–118.

Anselin, L. (2013). *Spatial econometrics: methods and models*. Dodrecht: Springer.

Bell, N., Schuurman, N., & Hameed, S. M. (2008). Are injuries spatially related? Join–count spatial autocorrelation for small-area injury analysis. *Injury Prevention*, *14*(6), 346–353.

Biemann, T., & Datta, D. (2014). Analyzing Sequence Data: Optimal Matching in Management Research. *Organizational Research Methods*, 17(1), 51–76.

Biemann, T., Zacher, H., & Feldman, D.C. (2012). Career Patterns: A Twenty-Year Panel Study. *Journal of Vocational Behavior*, 81(2), 159–170.

Bivand, R. S., Pebesma, E., & Gómez-Rubio, V. (2013). *Applied spatial data analysis with R*. New York: Springer.

Bivand, R. S., Hauke, J., & Kossowski, T. (2013). Computing the Jacobian in Gaussian spatial autoregressive models: An illustrated comparison of available methods. *Geographical Analysis,* 45(2), 150–179.

Bivand, R.S., & Piras, G. (2015). Comparing Implementations of Estimation Methods for Spatial Econometrics. *Journal of Statistical Software,* 63(18), 1–36.

Cliff, A. D., & Ord, J. K. (1981). *Spatial processes*. London: Pion.

Cotton, J. L., & Tuttle, J. M. (1986). Employee turnover: A meta-analysis and review with implications for research. *Academy of Management Review*, 11(1), 55–70.

Crossley, C. D., Bennett, R. J., Jex, S. M., & Burnfield, J. L. (2007). Development of a global measure of job embeddedness and integration into a traditional model of voluntary turnover. *Journal of Applied Psychology*, 92(4), 1031–1042.

Dlouhy, K. & Biemann, T. (2015). Optimal matching analysis in career research: A review and some best-practice recommendations. *Journal of Vocational Behavior,* 90, 163–173.

Felps, W., Mitchell, T. R., Hekman, D. R., Lee, T. W., Holtom, B. C., & Harman, W. S. (2009). Turnover contagion: How coworkers' job embeddedness and job search behaviors influence quitting. *Academy of Management Journal*, 52(3), 545–561.

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

Garud, R., Kumaraswamy, A., & Karnøe, P. (2010). Path dependence or path creation?. *Journal of Management Studies,* 47(4), 760–774.

Gerhart, B. (1990). Voluntary turnover and alternative job opportunities. *Journal of Applied Psychology*, 75(5), 467–476.

Ghiselli, E. E. (1974). Some perspectives for industrial psychology. *American Psychologist,* 2, 80–87.

Griffeth, R. W., Hom, P. W., & Gaertner, S. (2000). A meta-analysis of antecedents and correlates of employee turnover: Update, moderator tests, and research implications for the next millennium. *Journal of Management*, 26(3), 463–488.

Heavey, A. L., Holwerda, J. A., & Hausknecht, J. P. (2013). Causes and consequences of collective turnover: a meta-analytic review. *Journal of Applied Psychology*, *98*(3), 412–453.

Huang, Q., & Sverke, M. (2007). Women's occupational career patterns over 27 years: Relations to family of origin, life careers, and wellness. *Journal of Vocational Behavior*, 70(2), 369–397.

Judge, T. A. (1993). Does affective disposition moderate the relationship between job satisfaction and voluntary turnover?. *Journal of Applied Psychology*, 78(3), 395–401.

Judge, T. A., & Watanabe, S. (1995). Is the past prologue?: A test of Ghiselli's hobo syndrome. *Journal of Management,* 21(2), 211–229.

Kabos, S., & Csillag, F. (2002). The analysis of spatial association on a regular lattice by join-count statistics without the assumption of first-order homogeneity. *Computers & Geosciences,* 28(8), 901–910.

Koch, J. (2011). Inscribed Strategies: Exploring the Organizational Nature of Strategic Lock-In. *Organization Studies,* 32(3), 337–363.

Koch, J., Eisend, M., & Petermann, A. (2009). Path dependence in decision-making processes: Exploring the impact of complexity under increasing returns. *Business Research,* 2(1), 67–84.

Maertz, C. P., Griffeth, R. W., Campbell, N. S., & Allen, D. G. (2007). The effects of perceived organizational support and perceived supervisor support on employee turnover. *Journal of Organizational Behavior*, 28(8), 1059–1075.

McVicar, D., & Anyadike-Danes, M. (2002). Predicting successful and unsuccessful transitions from school to work by using sequence methods. *Journal of the Royal Statistical Society: Series A*, 165(2), 317–334.

Meyer, J. P., Stanley, D. J., Herscovitch, L., & Topolnytsky, L. (2002). Affective, continuance, and normative commitment to the organization: A meta-analysis of antecedents, correlates, and consequences. *Journal of Vocational Behavior*, 61(1), 20–52.

Morrowa-Howell, N. (1994). The M word: Multicollinearity in multiple regression. *Social Work Research,* 18(4), 247–251.

Ng, T. W., Eby, L. T., Sorensen, K. L., & Feldman, D. C. (2005). Predictors of objective and subjective career success: a meta-analysis. *Personnel psychology*, 58(2), 367–408.

Ng, T. W., Sorensen, K. L., Eby, L. T., & Feldman, D. C. (2007). Determinants of job mobility: A theoretical integration and extension. *Journal of Occupational and Organizational Psychology*, 80(3), 363–386.

Park, T. Y., & Shaw, J. D. (2013). Turnover rates and organizational performance: A meta-analysis. *Journal of Applied Psychology*, 98(2), 268–309.

Salgado, J. F. (2002). The Big Five personality dimensions and counterproductive behaviors. *International Journal of Selection and Assessment,* 10, 117–125.

Savickas, M. L. (2002). Reinvigorating the study of careers. *Journal of Vocational Behavior,* 61(3), 381–385.

Shartle, C. L. (1959). *Occupational information*. Englewood Cliffs: Prentice-Hall.

Stark, D., & Vedres, B. (2006). Social Times of Network Spaces: Network Sequences and Foreign Investment in Hungary1. *American Journal of Sociology*, 111(5), 1367–1411.

Sydow, J., Schreyögg, G., & Koch, J. (2009). Organizational path dependence: Opening the black box. *Academy of Management Review*, 34(4), 689–709.

Sydow, J., Windeler, A., Müller-Seitz, G., & Lange, K. (2012). Path constitution analysis: A methodology for understanding path dependence and path creation. *Business Research,* 5(2), 155–176.

Tett, R. P., & Meyer, J. P. (1993). Job satisfaction, organizational commitment, turnover intention, and turnover: path analyses based on meta-analytic findings. *Personnel Psychology*, 46(2), 259–293.

Timmerman, T. A. (2006). Predicting turnover with broad and narrow personality traits. *International Journal of Selection and Assessment,* 14(4), 392–399.

Tobler, W. R. (1970). A computer movie simulating urban growth in the Detroit region. *Economic Geography*, 46, 234–240.

R Core Team. (2015). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.

Rosenbaum, J. E. (1979). Tournament mobility: Career patterns in a corporation. *Administrative Science Quarterly,* 24(2), 220–241.

Vergne, J. P., & Durand, R. (2010). The missing link between the theory and empirics of path dependence: conceptual clarification, testability issue, and methodological implications. *Journal of Management Studies*, 47(4), 736–759.

Vinkenburg, C. J., & Weber, T. (2012). Managerial career patterns: A review of the empirical evidence. *Journal of Vocational Behavior*, 80(3), 592–607.

Wagner, G. G., Frick, J. R., & Schupp, J. (2007). The German Socio-Economic Panel Study (SOEP): Scope, evolution and enhancements. *Schmollers Jahrbuch,* 127(1), 139–169.

Woo, S. E. (2011). A study of Ghiselli's hobo syndrome. *Journal of Vocational Behavior,* 79(2), 461–469.

Session 16A: Employment

Professoral Career Patterns between Academia and the Corporate World

Applying sequence analysis to the study of academic autonomy[1]

Pierre Benz, Felix Bühlmann & André Mach

Institute of Social Sciences (ISS) and Institute of International, Political and Historical Studies (IEPHI), University of Lausanne

1. Introduction

The rise of collaborations between scientific research and private industry holds a central place in contemporary debates on the autonomy of science and higher education. These collaborations can take various forms and evolve in time. This paper focuses on Swiss professors' professional careers as a dimension of the ties that participate to define the changing boundaries between academic and private sector.

Little is known about historical trends of relations between universities and the economic sphere in Switzerland. If the academic field does not escape the movement of "economization" of higher education at the end of the 20th century (Jost, 2015: 130), it is likely that collaborations took other forms before, especially for engineering sciences, chemistry or, later, biotechnologies and life sciences. In a study about Swiss pharmaceutical research in the 20th century, Bürgi (2011) shows the importance of the ties between academic research and private industries in the development of chemistry and biotechnologies during the whole century. Focusing on the first part on the century, Tamm (1997) and Simon (1997) note that academic research and private interests have converged since the early development of chemical industry, notably through the recruitment of professors in private research institutes.

2

1.1 The Changing Career Patterns of Professors

The scientific field has never been entirely autonomous from exogenous dynamics as the production of knowledge relies on extra-academic resources in various proportions. Autonomy therefore becomes a central concept for sociology of science to broach classification struggles to impose a definition of science and disciplines in their historical dimension (Gingras, 2012: 294).

The recent research on collaborations between academic and private sphere mostly focuses on the network dimension (i.e. co-publications, co-signatures of patents or research mandates[2]). Our purpose is to explore another dimension of connections through the study of professors' careers. By using sequence analysis, we open an innovative perspective on the analysis of collaborations between scientific research and the private sector, both biographically and historically. Our research revolves around three main points of inquiry:

First, how do careers evolve biographically in terms of phases, turning points, rhythm? Second, how do careers change depending on historical period, i.e. how do they reflect institutional transformations? Third, can typical careers be identified for specific disciplines, or groups of disciplines?

Furthermore, concentrating on professor's careers allows us also to shed new light on the changing autonomy of the academic field. Following Weber's definition of openness and closure of social relations (Weber, 1995), we propose to consider professional careers as a dimension of exchange at the interface between scientific and economic spheres, and thereby as an indicator of autonomy. As defined by Lamont and Molnar (2002), *boundary-work* is either a way to protect professional autonomy against outside powers (political, economic), as well as a form of social control. Hence, we can analyse careers as a specific dimension of boundary-work through collaborations across fields.

This conceptual framework allows us to focus also on disciplinary boundaries, as we expect careers to be strongly differentiated between disciplines. Gingras (2012) states that the homogeneity of careers is an indicator of the degree of autonomy of disciplines. The more a scientific field is institutionalised, the more careers are buoyed. Again, careers function as a strategy to maintain scientific autonomy and at the same time as a way to strongly limit subversion strategies (Gingras, 2012: 290). A focus on typical career trends constitutes a way to comprehend the legitimacy of different types of capital among peers.

1.2 Why study the case of the EPFL

To study the careers of professors we concentrate on the case of the Ecole polytechnique fédérale de Lausanne (EPFL). The EPFL is one of two federal technical universities in Switzerland (Jost, 2015). It was recently rated in the top 15 of world's

---

[2] For some examples, see Grossetti & Milard, 2003; Callon & Gamberini, 2000; Barrier, 2014.

3

universities (top 5 of European universities)[3] and it is widely renowned for cultivating close relations to the private sector. From 2000 to 2014, 192 start-ups were established, 129 inventions disclosed and 99 priority patents were filed. Moreover, the "EPFL Innovation Park" hosts currently several big companies such as Axa Technology services, Credit Suisse, Logitech, Siemens, Merck-Serono or Nestlé.

We chose to divide the history of the EPFL into three periods: The first period lasts from 1969 to 1980. Engineering sciences are developed together with scientific disciplines (chemistry, physics and mathematics), although all diplomas are delivered as engineer's degree. During this first period, the EPFL remained the most disconnected from the University of Lausanne and the Swiss academic field in general.

The second period, from 1980 to 2000, is characterized by the progressive concentration of activities on engineering sciences, and the development of microengineering and computer sciences. Collaborations between EPFL and the University of Lausanne revive. Several projects of coordination are also developed with other Swiss universities. Biology is not a priority, in contrast to engineering physics and chemistry (Leresche et al., 2012: 194). Although some structures of collaboration with industries had been established since the 1970's, they strongly increase from 1990. In year 2000, the EPFL hosts 47 private companies, including 37 start-ups, which have been founded at the EPFL (Pont, 2010: 106).

The last period starts in 2000 with the reorientation of the EPFL toward the development of biology and biotechnologies. Partnerships with industrial firms expand strongly, with a concomitant growth of new start-ups. The number of professors increases from around 150 in 2000 to more than 200 in 2010, significantly in the Faculties of basic sciences (chemistry and physics) and life sciences[4].

2. Data and strategy

We built an original dataset that includes all full, extraordinary and associate professors of the EPFL active at least at one of three benchmarks: 1980, 2000 and 2010. These benchmarks refer to the three periods presented in the precedent chapter.

Based on a positional approach (Mills (2010 [1956]) the inclusion criterion of our sample is the position held, namely the position of full, extra-ordinary or associate. We attempted to gather data on the entire career of each professor, including academic and extra-academic functions.

---

[3] http://information.epfl.ch/chiffres (02.01.2016)

[4] http://vpri.epfl.ch/files/content/sites/vpplnew/files/shared/Stat%202014/Personnel%202014.pdf
(12.01.2016)

4

Table 1. Total number of professors of the EPFL

1980	2000	2010
N=112	N=142	N=211

Table 1. Indicates the total number of professors holding a position of full and extraordinary (associate) professor for year 1980, 2000 and 2010. Some individuals may be present on two different dates if their mandate runs, for example, from 1978 to 2004.

We examined careers from the age of 20 to the age of 65 for all professors. The alphabet includes all the functions held within and outside the academic field. For each professor, we collected data for the year of nomination and departure of the function, the nature of the function, place and country of workplace. The alphabet is composed of 8 states.

Academic positions are divided into three states, related to the hierarchy of positions within academic institutions. The first state includes all full professors. The second state includes extra-ordinary and associate professors. The third includes all other academic positions namely post-doctoral fellowship, lecturer, senior lecturer, assistant professor, research positions, head of research.

Extra-academic positions are pooled into one single state: positions as engineer, researcher, head of research, director of research within private laboratories or companies and executive directors. Non-executive functions such as member of the board of directors are not taken into account; they are not considered to be full-time activities.

The third set of states is related to multipositionality, namely holding one of the three academic positions at the same time with an extra-academic position: ordinary professor + extra-academic position, extra-ordinary or associate professor + extra-ordinary position, other academic position + extra-ordinary position.

The last state is related to educational period running from the age of 20 to the date of the graduation. Namely doctoral degree or high degree diploma such as Licence, Master, diploma or equivalent in the case individual does not hold a doctoral degree.

Finally, missing data are added as a state when at the beginning or when resulting from a lack of data during the career. Missing data at the end of the career are not taken into account when they due to still going on career The data on careers are completed by a set of biographical indicators such as date and place of birth and death, sex, nationality and educational degree (graduate diploma, PhD, discipline, place and year of acquisition).

5

3. First results and perspectives

Our analytical strategy is based on sequence analysis, mainly optimal matching and clustering. Our results on data for 1980[5] (N=112) relies on a TramineR clustering using a Ward method and a standard matrix, all substitution costs set at 1 (optimal matching using a constant method, including missing data). The sequences are defined from the age of 20 to the age of 65.

The results distinguish three relatively distinct types of careers: "academic careers" (Type 2, N=54), "extra-academic careers" (Type 1, N=43) and "part-time careers" (Type 3, N=15). Careers type 1 is characterized by a relative short period of education and a relative long career outside the academic field prior to the nomination of professor. What is more, out of 43 trajectories, 20 indicate a nomination as full professor directly following the extra-academic career and 18 indicate a nomination as extra-ordinary professor directly following the extra-academic career. Type 2 is composed of careers that follow a "full academic career". Only eight trajectories include extra-academic functions with a meantime of two years. In comparison, the time of extra-academic functions is of 12 years in average for type 1. Type 3 contains trajectories of part-time careers until the age of 65, mostly extra-ordinary professors and some full professors keeping an extra-academic activity until the end of their career.
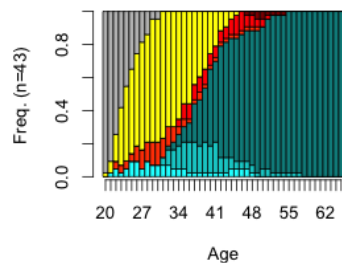
Fig. 1: Extra-academic careers (Type 1)　　　　**Fig. 2:** Academic careers (Type 2)
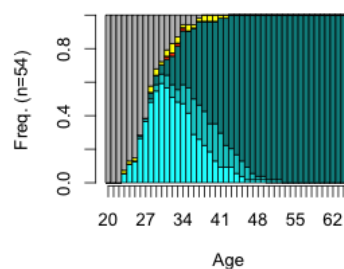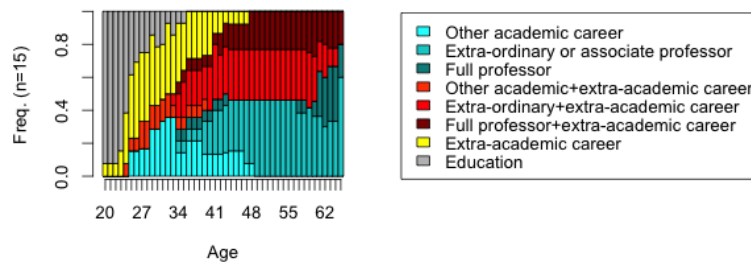

Fig. 3: "Part-time professorship" (Type 3)

[5] In the final paper we will have comparative data on all three years: 1980, 2000 and 2010.

6

To complete this first analysis for year 1980, we crossed three variables with each clusters, as shown in table 2.

Table 2. Characteristics of the three career types

Variable	Type 1	Type 2	Type 3	Ø
Doctoral degree YES	0.30	0.89	0.47	0.60
Natural sciences YES	0.14	0.50	0.20	0.32
Foreign nationality YES	0.14	0.24	0.13	0.19

Table 2. All variables are dummies. Doctoral degree YES includes all professors holding a doctoral degree (60%). Natural sciences YES include chemistry, physics and mathematics department (32%) *versus* departments of engineering and architecture. Foreign nationality YES (19%) gathers all non-Swiss nationalities (18 Europe, two USA, one India). Total variable ratio is calculated through total of the variable on total population.

Comparing the characteristics of the three types, the percentage of professors with doctoral degree is much higher for Type 2 than for the average. Also the natural sciences and foreign professors are clearly overrepresented in Type 2. In comparison, Type 3 and particularly Type 1 are mostly composed of Swiss professors, engineering sciences and a relative low percentage of professors holding a doctoral degree.

This first descriptive analysis shows three interesting results. First, the average of professors holding a doctoral degree is of remarkably low 60%. Thus we may argue that at the EPFL, an extra-academic career is a substitute for a doctoral degree, and we may ask to what extend this is true for 1980. Second, the EPFL appears not to be as internationalized in 1980 as it is today, with 19% of foreign professors. This percentage is very likely to increase for the recent period together with a diversification of countries. Third, regarding closer to disciplines, the results show two different patterns. On one hand, natural sciences, highly represented in Type 2, are mostly linked with doctoral degrees and foreign origin. On the other hand, a low percentage of doctoral degrees together with a high proportion of Swiss professors and engineering sciences characterize the Type 1. We expect this opposition between natural sciences and engineering sciences to decrease in the recent period.

Finally, these first results bring us to formulate broader questions about career patterns and academic autonomy: What is the frequency and the biographical pattern of professor's shifts between the corporate sector and the academic domain? Do professor change between business and academic mainly in the beginning of their career or only once they enjoy a stable academic position? Can we observe parallel careers between the private and the academic field?

7

5. References

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

Abbott, A. (1995b). Things of Boundaries. *Social Research*, 62(4), 857-882.

Barrier, J. (2014). Partenaires particuliers : financements sur projet et travail relationnel dans les réseaux de collaboration science-industrie. *Genèses*, 94, 55-80.

Blanchard, P., Bühlmann, F., Gauthier, J.-A. (2014). *Advances in Sequence Analysis: Theory, Methods, Applications*. Springer.

Bourdieu, P. (1984). *Homo academicus*. Paris: Minuit.

Bourdieu, P. (1994). *Raisons pratiques. Sur la théorie de l'action*. Paris: Seuil.

Bourdieu, P. (2001). *Science de la science et réflexivité*. Paris: Raisons d'agir.

Bühlmann F. (2008). The Corrosion of Career? - Occupational Trajectories of Business Economists and Engineers in Switzerland. *European Sociological Review* 24(5), 601-616.

Bürgi M., Strasser B. J. (2010), Pharma in Transition: New Approaches to Drug Development at F. Hoffmann-La Roche & Co, 1960–1980. In V.M. Quirke and J. Slinn (Eds), *Perspectives on 20th-Century Pharmaceuticals* (pp. 391-432). Oxford: Peter Lang.

Bürgi M. (2011). *Pharmaforschung im 20. Jahrhudert. Arbeit an der Grenze zwischen Hochschule und Industrie.* Zurich : Chronos Verlag.

Cosandey, M. (dir). (1999). *Histoire de l'Ecole Polytechnique Lausanne : 1953-1978.* Lausanne: Presses Polytechniques et Universitaires Romandes.

Dlouhy, K., Biemann, T. (2015). Optimal matching analysis in career research : A review and some best-practice recommendations. *Journal of Vocational Behaviour*, 90 : 163-173.

Gingras, Y. (2002). Les formes spécifiques de l'internationalité du champ scientifique. *Actes de la recherche en sciences sociales*, 141-142: .31-45.

Gingras, Y. (2003b). Mathématisation et exclusion : socio-analyse de la formation des cités savantes. In J.-J. Wunenburger (Ed.), *Gaston Bachelard et l'épistémologie française* (pp.115-152). Paris : PUF.

Gingras, Y. (2012). Le champ scientifique. In F. Lebaron & G. Mauger (Eds), *Lectures de Bourdieu*. (pp. 279-294) Paris: Ellipses.

Lamont, M., Molnar, V. (2002). « The Study of Boundaries in the Social Sciences » in *Annual Revue of Sociology*, 28, 167-195.

Lamy, E., Shinn, T. (2006). L'autonomie scientifique face à la mercantilisation. Formes d'engagement entrepreunarial des chercheurs en France. *Actes de la recherche en sciences sociales*, 164, 23-50.

Leresche, J.-P. *et al.* (2012). *Gouverner les universités. L'exemple de la coordination Genève-Lausanne (1990-2010).* Lausanne : Presses polytechniques et universitaires romandes.

Mills, C. W. (2012 [1956]). *L'élite au pouvoir*. Marseille: Angone.

Owen-Smith, J. & Powell, W. W. (2004). Carrières et contradictions en sciences de la vie : réponses du corps académique aux transformations de la connaissance et de ses utilisation. *Sociologie du travail*, 46, 347-377.

Pont, M. (2010). *Chronique de l'EPFL 1978-2000. L'âge d'or de l'ingénierie.* Presses polytechniques et universitaires romandes.

Simon, C. (1997). Chemie in der Schweiz. Dimensionen eines Themas. In Busset. T., Rosenbusch A., Simon C. (Eds.), *Chemie in der Schweiz. Geschichte der Forschung und der Industrie* (pp.25-59). Basel : Christoph Merian Verlag.

Tanner, J. (1997). Universität und Industrie. Die Entwicklung der Chemie in Basel. In Busset. T., Rosenbusch A., Simon C. (Eds.), *Chemie in der Schweiz. Geschichte der Forschung und der Industrie* (pp.59-77). Basel : Christoph Merian Verlag.

Weber, M. (1995 [1922]). *Economie et société*. Paris: Pocket.

Pathways to the Power Elite - Career trajectories in the core of the Danish Networks of Power

Christoph Houman Ellersgaard, Anton Grau Larsen, Lasse Folke Henriksen and Jacob Aagaard Lunding
All of the Department of Business and Politics, Copenhagen Business School

This paper explores the career trajectories of the 423 individuals identified as the power elite in the core of the Danish Elite Network (Ellersgaard 2015; Larsen 2015) through sequence analysis. We identify patterns in four distinct sequences - sectorial, occupational, organisational and geographical - enabling us to explore the relationship between career path and current position within the different sectors of the power elite. The four sequences are used to explore how elite cohesion and oppositions are created through four different mechanism: 1) inter-sectoral mobility 2) career slope 3) character formation through organisational adaption and 4) location in relation to the national and international power centers.

Methodologically, we identify career segments based on optimal matching (Abbott & Hrycak 1990; Blair-Loy 1999) of the career trajectories in the four different types of sequences mentioned above of the 423 individuals in the core of the elite network. We investigate the alignment between these multiple sequences. Furthermore, we identify other social characteristics of the segments such as their gender, social background and educational profile. Finally, we explore the relationship between cohesion based on social ties and cohesion based on similarities in career trajectory through distances in sequence patterns and sociometric geodestics derived from the network database (Ellersgaard & Larsen 2015). These analysis serves to explore cohesion and fragmentation across the elite based on career trajectories within the type of sequences.

The inter-sectoral mobility is analysed to discuss to what extent career path reflect current sector affiliation. Michael Hartmann (2010: 292) defines inter-sectoral mobility as the degree to which elite individuals have careers that span several of the key sectors in a society, which in turn creates larger homogeneity among the individuals at the very top. These patterns of revolving doors described by Mills (1956: 287) as 'the heavy traffic between the … structures, often in very intricate patterns' or 'the interchangeability of positions' add to the cohesion of the power elite. The exchange of a key position within one field to a key position in another has also been described though the term pantouflage by Pierre Bourdieu (1996). Furthermore, we access the flow - and the direction of the flow - between sectors (Denord, Lagneau-Ymonet & Thine 2011:34). The Danish case is particu-

2

larly interesting as it is expected that there is comparatively little inter-sectoral mobility (Hartmann 2010).

The career slopes for each individual towards the position at the apex of the elite network is used to assess the velocity of career ascent. By identifying which occupational position and the hierarchy of promotions needed, the career progression of elite individuals can be compared across sectors. Again following Hartmann (2000) and Mills (1956) this may be tied to socialization within the same classes as dominate the sector or field, the elite individual is rising though. Furthermore, career slopes reveal the average time investment need to achieve a dominant position in the sector or field (cf. Bourdieu 1986).

The organisational adaptation follows the number and type of organisations within the career trajectories of elite individuals. We explore whether individuals are shaped by the same type of organisations, i.e. banks or the ministry of justice, thus creating a similar life experience, outlook and character within this group (Mills 1956). In particular, we investigate the importance of career positions in academies (Cappelli & Hamori 2005: 25), that is, certain firms or institutions endowing their former employees with an aura of excellence such as consulting giant McKinsey & Company or the grand corps in France (cf. Bourdieu 1996). Furthermore, we investigate how indicators of involvement in the political or organisational field may later lead to entry in these sectors.

The movement towards the power centre follows the location of employment of the elite individuals. Two different aspects are of interest here. First, the dynamics between center and periphery within Denmark. Who have to move around in the provinces before returning to the power centre of the capital and who manage to enter the elite network without having geographical ties to the capital. Second, to what extent have individuals gathered cosmopolitan capital (Weenink 2008, Bühlman, David & Mach 2013). By looking at the specific area of employment or education abroad we differentiate between the status of areas based on their position in world system theory (Chase-Dunn, Kawano & Brewer 2000) and in the World City System (Alderson & Beckfield 2004).

In conclusion, we compare segments derived from each of the four types of sequences and discuss how the current sector affiliation of the elite individual is reflected in his or her career trajectory. Do careers and experiences intertwine, creating potential for shared understandings, personal ties and thus even stronger cohesion within the core group of the Danish elite network or do careers follow distinct paths in different sectors.

Our preliminary analysis suggests that career patterns overlap substantially with the current sector affiliation of the members of the elite networks. This even applies to careers within the very same organisations in these sectors. In particular,

3

the senior civil servants, the chief executives of the largest corporations, the scientists and the union leaders follow distincts paths. However, two groups appear to deviate from the sector specific career patterns. First, managers of organisations that themselves cross between sectors such as state-owned enterprises, national research centers or university based tech corporations. Second, a small group of multipositionals (cf. Boltanski 1973) bridge between the distinct career patterns found in different sectors. This group is further characterized by also holding the most central network positions, having attended the most typical university programmes for the elite and a having long careers. Thus we find a core within the core (cf. Denord, Hjellbrekke et al 2011) who holds the most senior positions within the field of power and both through current and former ties are well-connected to all key sectors within the Danish welfare state.

References:

Abbott, Andrew, og Alexandra Hrycak. 1990 «Measuring Resemblance in Sequence Data: An Optimal Matching Analysis of Musicians' Careers». *American Journal of Sociology* 96 (1): 144–85.

Alderson, Arthur S., og Jason Beckfield. 2014. «Power and Position in the World City System». *American Journal of Sociology* 109 (4): 811–51.

Blair-Loy, Mary. 1999. «Career Patterns of Executive Women in Finance: An Optimal Matching Analysis». *American Journal of Sociology* 104 (5): 1346–97.

Boltanski, Luc. 1973. «L'espace positionnel. Multiplicité des positions institutionnelles et habitus de classe». *Revue française de sociologie* 14 (1): 3–26.

Bourdieu, Pierre. 1986. «The forms of capital». in *Handbook for Theory and Research for the Sociology of Education*, edited by John G. Richardson, 241–58. New York: Greenwood.

Bourdieu, Pierre. 1996. *State Nobility - Elite Schools in the Field of Power.* Cambridge: Polity Press.

Bühlmann, Felix, Thomas David, og André Mach. 2013. «Cosmopolitan Capital and the Internationalization of the Field of Business Elites: Evidence from the Swiss Case». *Cultural Sociology* 7 (2): 211–29.

Cappelli, Peter, og Monika Hamori. 2005. «The new road to the top». *Harvard Business Review*: 25-32.

Chase-Dunn, Christopher, Yukio Kawano, og Benjamin D. Brewer. 2000. «Trade Globalization since 1795: Waves of Integration in the World-System». *American Sociological Review* 65 (1): 77–95.

Denord, François, Paul Lagneau-Ymonet, og Sylvain Thine. 2011. «Le champ du pouvoir en France». *Actes de la recherche en sciences sociales* 190: 24–57.

Denord, François, Johs. Hjellbrekke, Olav Korsnes, Frédéric Lebaron, og Brigitte Le Roux. 2011. «Social capital in the field of power: the case of Norway». *The Sociological Review* 59 (1): 86–108.

4

Ellersgaard, Christoph Houman. 2015. «Elites in Denmark: Power Elites and Ruling Classes in a Welfare State». Ph.D. Dissertation, Department of Sociology, University of Copenhagen, 2015

Ellersgaard, Christoph Houman, og Anton Grau Larsen. 2015. «Data Exchange Network: The Danish Elite Network». *Connections* 35 (1).

Hartmann, Michael. 2000. «Class–specific habitus and the social reproduction of the business elite in Germany and France». *The Sociological Review* 48 (2): 241–62.

Hartmann, Michael. 2010 «Elites and Power Structure». in *Handbook of European Societies*, edited by Stefan Immerfall og Göran Therborn, 291–323. New York: Springer.

Larsen, Anton Grau. 2015. «Elites in Denmark: Identifying the elite». Ph.D. Dissertation, Department of Sociology, University of Copenhagen.

Mills, Charles Wright. 1956. *The Power Elite*. Oxford: Oxford University Press.

Weenink, D. 2008. «Cosmopolitanism as a Form of Capital: Parents Preparing their Children for a Globalizing World». *Sociology* 42 (6): 1089–1106.

Methodological approaches to profiling and modelling disadvantaged employment pathways. An application to employment trajectories in Australia

Danilo Bolano and Michele Haynes

Abstract This paper investigates the employment pathways of men and women in Australia and profiles the characteristics of individuals who are at risk of disadvantage defined by unstable employment histories and frequent transitions into unemployment. The paper focuses on the transitions of respondents, aged 15-64 years, between different employment states over a span of 13 years (2001-2013) using panel data from the Household, Income and Labour Dynamics in Australia survey. To describe employment trajectories and to analyse the likelihood of transition from one employment state to another, we will rely on sequence analysis and two probabilistic models that might account for state dependence: dynamic multinomial logit random effects models and Markov models. Sequence analysis was used to identify typology of employment pathways and the associated sociodemographic characteristics and intergenerational links. The preliminary results confirm the presence of gender differences in employment. Women are more likely to be employed part time or not in the labour force while men are more likely to experience stable employment trajectories of full time work. Women also experience a slightly higher proportion of employment transitions. Among the sociodemographic and background factors considered, mature age workers, those with health problems and less educated parents, in particular father-son and mother-daughter links, are mainly at risk of experiencing unstable employment pathways.

1 Introduction

Trends in unemployment are of interest due to the financial and social costs borne by individuals. Some of the cost are absorbed by welfare payments but many of which

---

Danilo Bolano
Institute for Social Science Research, The University of Queensland, e-mail: d.bolano@uq.edu.au

Michele Haynes
Institute for Social Science Research, The University of Queensland e-mail: m.haynes@uq.edu.au

2   Danilo Bolano and Michele Haynes

are faced directly by families and communities who may not have the capacity to absorb them. Unemployment not only means a loss of economic security but also feeds poverty and social dislocation as the ties of civil society are severed. In general, unemployment removes an important set of social relationships without which many people have little support or security in dealing with often profound economic hardship. The longer the duration in transition from unemployment to employment, the more these hardships are exacerbated. Therefore unstable employment histories with long periods of unemployment can be considered as a measure of social disadvantage.

Unstable employment history can lead to labor market disadvantages too. The longer the period of time spent in unemployment, the lower an individual's likelihood of gaining employment. On the supply-side of the labor market, employers may be reluctant to employ the long-term unemployed, in which case economic policies aimed at reducing unemployment may have limited success (ABS, 2001). And on the supply side, absences from employment are associated with loss of human capital, under use and underdevelopment of skills (Vickery, 1999), limited occupational advancement (Burbidge, 2005). These can lead to the deterioration of labour market prospects, or scarring, in the form of relatively poor prospects of finding work (ABS, 2001; Saunders and Brown, 2004), loss of income whilst unemployed and lower wages on return to the workforce (Arulampalam, 2001; Green and Leeves, 2009; Stevens, 1997).

A range of personal characteristics are known to be associated with transition in and out the labor force and the duration of unemployment spells. In particular, we focus on the difference in employment trends for men and women. Women have often indicated a preference for part-time work as a way of attempting to combine child bearing and child rearing, home and non-child related caring responsibilities, for which women still bear the overwhelming responsibility (ABS, 2009; Hakim, 2000).

The paper has three objectives. First, to describe the different employment pathways of men and women in Australia. Second, to profile individuals at greater risk of disadvantage by demographic characteristics and family background, where disadvantage is defined in terms of unstable labor market history and transitions out of the labor force. Third, to review and compare statistical approaches to analyze employment trajectories. These approaches include sequence analysis, generalized linear mixed model in the form of a dynamic multinomial logit random effects model with lagged dependent variable to account for state dependence and Markov-based models.

2 Data and Methods

We use data from the first thirteen waves of the Household, Income and Labour Dynamics in Australia (HILDA) panel survey. It is a household-based panel study which began in 2001. Information are collected annually from each person in the

household aged 15 and more. The first wave was carried out from a sample of 13,969 individuals from 7,682 households (Watson and Wooden 2002).

This paper investigates employment trajectories and potential sociodemographic and family determinants of disadvantaged employment transition pathways. Therefore, we select only individuals of working age (15 to 65 years old) who participated in at least to two waves of the HILDA survey since Wave 1 (mean number of observations per individual = 9.92). The final sample consists of 10,941 respondents (52.14% of women, n=5,705). The mean age at the first observation was 38.5 years (median = 39).

The response variable is employment status defined by five categories: 'Being Full Time Student' (EduFT - defined as those who are enrolled in full time education and not employed), 'Employed Full Time' (FT), 'Employed Part Time' (PT), 'Looking for a job (LookingJob)', 'Not in the Labor Force' (NotLabourForce). Non responses have been included as an additional category ('Non Respond'). In the final version of the paper, we will analyse in more details the missingness patterns.

32.8% of respondents (3,584 individuals) are in the same employment state over time and 67.5% of them (2,419 individuals) declare to be employed full time at each wave (it is worth recalling now that the number of observations for each individual can vary from 2 up to 13). Pooling together the 13 waves of HILDA survey (Table1), we observe the predominance of situations of full time employment (51.2%) and part time jobs (21.7%). However, the share of times spent outside the labor force is also reasonably high (20.27%).

Table 1 Distribution of states, pooled waves 1-13

	Frequency	Percentage
Full time student (EduFT)	1,776	1.64
Full time employed (FT)	55,557	51.19
Part time employed (PT)	23,503	21.66
Looking for a job (LookingJob)	2,953	2.72
Not in labour force (NotLabourForce)	21,994	20.27
Non respond (NonRespond)	2,740	72.53
Total data points	108.523	

To investigate potential sources of disadvantage in employment transitions, a set of demographic and human capital characteristics of the respondents, health condition, socio-economic indicators (Socio-Economic Index for Areas and AUSEI), relevant life transitions (transition to parenthood) and socio-economic backgrounds (parental educational level, socio-economic status, parents occupation) will be included in the final analysis. The paper aims not only to identify the key factors associated to employment transitions between working statuses but also to review and compare different approaches for analyzing individual patterns. We will consider both descriptive tools (sequence analysis) as in this abstract and probabilistic models as Markov models and generalized linear mixed models with random effects (see e.g., Haynes et al. 2008).

4   Danilo Bolano and Michele Haynes

3 Clustering using Sequence Analysis and Preliminary Results

With a holistic approach (Billari, 2001), the unit of interest is the entire individual trajectory and holistic studies mainly rely on sequence analysis (SA, Abbott, 1995). Using a sequence analysis approach, working trajectories are described as an ordered sequence of labour market states. Looking at the entire set of individual transitions among states occurred over time (potentially over an entire lifetime career) as a unique sequence of observations, sequence analysis can discover, describe and explain different individual life course patterns being more informative then focusing on one single working transition.

SA has been applied to study different types of social processes. Studies on work trajectories and pathways from school to work include McVicar and Anyadike-Danes, 2002; Martin et al., 2008, Fasang 2010. As far as we know, the only extensive study on employment trajectories in Australia using sequence analysis is Fry and Boulton (2013). This paper extends their work in several ways. The paper uses all the waves of HILDA survey available at the time of writing this abstract (13 waves, 2001-2013), compares different competitive statistical models and takes a life course approach. The papers controls for demographic characteristics, life events and socio-demographic background to identify the factors that are more likely associated to employment mobility and situation of working disadvantages.

The length of employment history considered for each individual can vary from 2 up to 13 years. For 5,246 individuals—47.9% of our sample—we have complete sequences. The most frequent pattern is being full time employed. Such stable working condition concerns 1,253 individuals. Only 220 respondents (2% of the entire sample) are outside the labour force and not in education during the entire period from 2001 and 2013.

Table 2 Most common working subsequences. Ordered by frequency for the whole sample.

Subsequence	Entire sample (n=10,941)		Female (n=5,705)		Male (n=5,236)	
	Freq	Percent	Freq	Percent	Freq	Percent
Full Time Employed over the 13 waves	1253	11.45	260	4.56	993	18.96
FT- FT	226	2.07	73	1.28	153	2.92
Not in the labour force over the 13 waves	220	2.01	169	2.96	51	0.97
NotLabourForce NotLabourForce	187	1.71	122	2.14	65	1.24
FT FT - FT	153	1.4	50	0.88	103	1.97
NotLabourForce- NotLabourForce- NotLabourForce	128	1.17	79	1.38	49	0.94
Part Time Employed over the 13 waves	96	0.88	96	1.68	-	-

The stability in working trajectories in Australia is confirmed looking at the transitions between labour market states over two consecutive waves (Figure 1) and the most common (sub)sequences reported in Table 2. On average, people experiences less than 2 working transitions (mean of 1.848) over the life span considered and

in 3,952 cases (45.19% of the total transitions) the respondent remains in the same labour market states (i.e., we observe no transitions) over two consecutive periods.
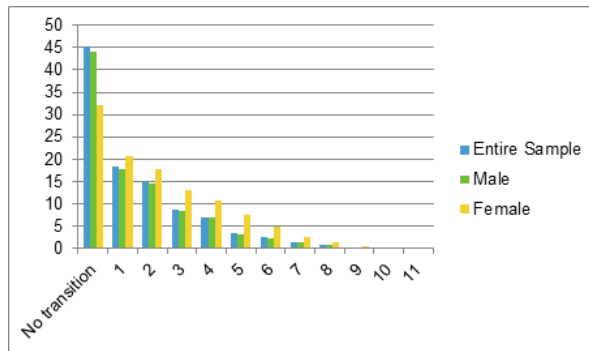

Fig. 1 Distribution of working transitions over two consecutive periods. Entire sample and divided by gender.

It is not surprising that women seem to follow more complex working trajectories experiencing a higher percentage of working transitions (see Figure 1) and spending more time outside the labour market or in a part time job (Figure 2). As reported in Table 2, only 4.56% of women declare to be full time employed from 2001 and 2013 against 19% of men. And, more than 6% of women spent at least two consecutive periods outside the labour force. In particular, almost 3% of women between 15 and 65 years old included in our sample never entered in the labour market in any of the waves considered.

Figure 3 shows graphically the working trajectories of 10 individuals (top right panel for the entire sample and bottom panels by gender) and the cross sectional distribution of states for the whole sample (top left panel). Figure 4 shows instead the differences between genders in the most common working trajectories. The R package TraMineR (Gabadinho et al., 2011) has been used for the analysis. As discussed before, respondents tend to be in the same state over time but men are more frequently in a stable full time position (violet in the graph) while women are commonly either outside the labour force (blue state) or in a part time job over time (red state). Moreover, the pictures seem suggesting a greater between-individual variations among employment trajectories for women.

Sequence analysis can be used not only to describe and represent trajectories but also to compare patterns and cluster individuals in homogeneous groups according to their pathways. Using the Optimal Matching distance as a measure of similarity between working trajectories (Abbott and Forrest 1986), we have identified four typologies for each gender. For women, the first cluster represents those are most time in a part time job (1,406 individuals). The second cluster, that is the most common one (n=1,913), groups women who are frequently outside the labor force during the period of observations. Women employed in full time o jobs are represented in

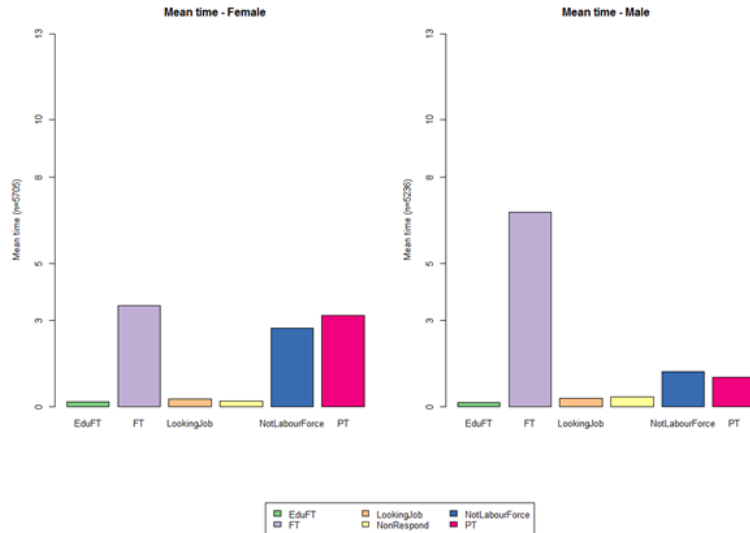6   Danilo Bolano and Michele Haynes

Fig. 2 Mean time (number of waves) spent in each working status. Divided by gender.

group 3 (n=1,664). The remaining 722 female participants characterized by higher variability in working trajectories are grouped in cluster 4.

For men, the clusters are less differentiate with cluster two (n=1,292) and three (n=1,729) representing situations of long term full time employment. Cluster 4 includes 578 individuals who are outside labor force for several years. Finally, cluster 1 groups shorter sequences and individual who has several periods of part time job. Further analysis will be performed accounting for the patterns of missing responses.

More sophisticated techniques in sequence analysis allow to analyze the factors that better discriminate between different pathways. Using discrepancy analysis, we might test for instance the share of discrepancy between employment trajectories explained by human capital (see Studer et al. 2011 for a detailed discussions on discrepancy analysis). According to our first results (not shown in this abstract), the education level of respondents explains 1.3% of discrepancy in working trajectories in Australia among male respondents and 1.6% among women. Similarly, the socio-demographic background explain less than 2% of discrepancy between pathways. On the other hand, the age of respondents explains 10% of the differences among trajectories observed for men.

To identify profiles at risk of having unstable and uncommon working trajectories a regression tree model can be used. The regression trees (Figure 6 and Figure 7) allow to test the association between the observed working trajectories and a set of individual characteristics. The socio-demographic background plays a central role in explaining the differences observed in the trajectory (i.e., between individual variations) determining the first split in the tree. Individuals with a better social background (i.e., a parent with a minimum level of education) experiences longer

Employment pathways in Australia　　　　　　　　　　　　　　7

Fig. 3 Graphical representations of working trajectories in Australia. Cross-sectional distribution (top panel left-hand side). 10 random working trajectories (top right). Individual sequences by gender on the bottom.
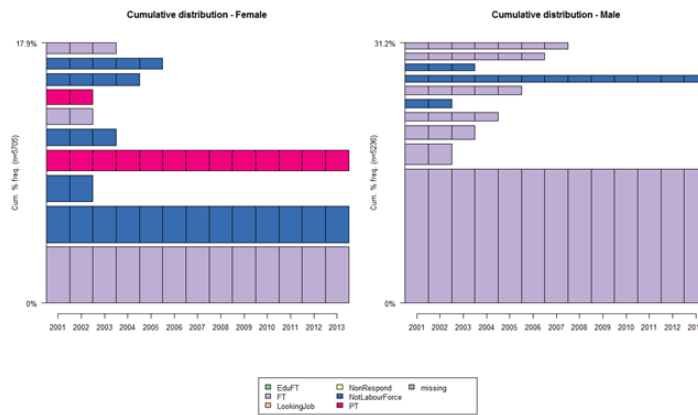

Fig. 4 Most common working trajectories. 2001-2013. Divided by gender.

periods of full time or part time job (Figure 6). The age of the respondents is, as expected, another key factor. Older cohorts are more likely retired or outside labor force (the blue state in the graphs) with respect to respondents between 14 and 54

8   Danilo Bolano and Michele Haynes
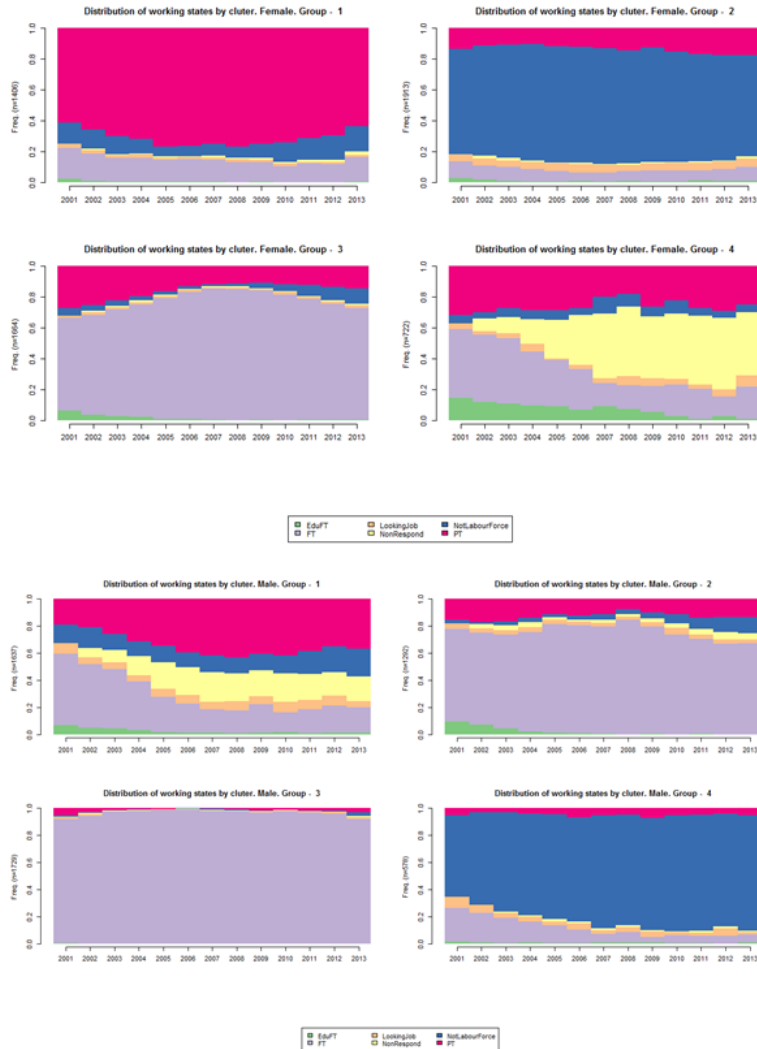
Fig. 5 Cross sectional distribution of working states by cluster. By gender.

years old. Finally, the health condition has a strong influence on labor force partici-
pation. For instance, looking at the bottom right corner of Figure 6, a male respon-
dent aged 14 - 54 with a father with a minimum level of education, is more likely
outside the labor force or in part time job if reports a fair or poor health condition.

Employment pathways in Australia
9
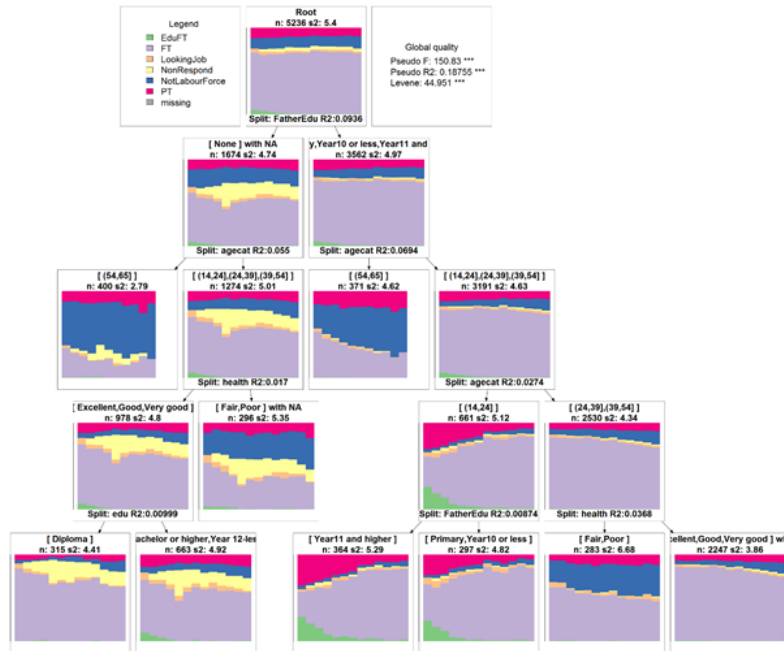
Fig. 6 Regression tree for working trajectories. Male respondents (n=5,236)

Thus these preliminary results confirm that relevant gender differences in labor force participation and stability in working trajectories still exist nowadays in Australia. It is interesting to notice another gender-related factor with respect to the family background of the respondents. The results show that a significant intergenerational relationship through same gender parent-child. For a male respondent, it is the level of education of the father that is the more associated with a stable working trajectory. For a female respondent, the participation to the labor force seems associated with the level of education of the mother. Such father-son and mother-daughter relationship should be investigated further also taking into account the missing information on parent level of education. On the other hand, educational level seems to be less relevant such as area-level indicators of disadvantages (SEIFA). However more sophisticated analysis will be performed in the final version of the paper to explore better on the effect of sociodemographic and background characteristic on working trajectories in Australia.

As briefly shown, Sequence Analysis allows to represent, describe and identify the factors associated to unusual and unstable employment trajectories. However, as most data mining techniques, it does not make any assumption about the 'social process' that have generated the trajectories. We will then explore alternative (stochastic) methods such as Markov models (see Billard 2001 for an introduction) and generalized linear mixed effect models (GLMM. See for instance Wooldridge 2002, Haynes et al. 2008). Using SA and stochastic models we might be able to
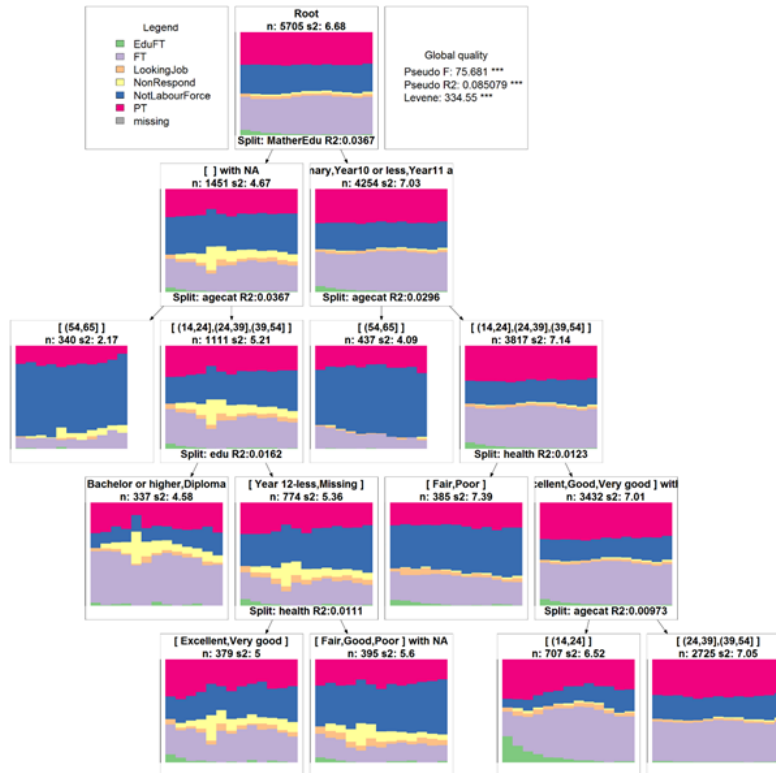
10   Danilo Bolano and Michele Haynes

Fig. 7 Regression tree for working trajectories. Female respondents (n=5,705)

identify both between-individual differences in employment trajectories and within individual variations.

In the Markovian perspective for example, life trajectories are considered as the result of a stochastic process in which the probability of occurrence of a particular state or event depends on the sequence of states observed so far and, eventually, on a set of covariates. In other words, considering life trajectories as sequences of mutually exclusive states—e.g., sequences of employment statuses—the Markovian perspective focuses on the successive transitions and attempts to depict the life history of an individual by means of the probabilities to switch to the different states of interest given the state history lived so far. The focus is more on the transitions than on the entire trajectories. By the means of probabilistic models, we might then investigate the effect of the socio-economic background of the respondents in transitions out of unemployment (i.e., the probability of observing the transition 'Out of labor force - Employed') and the accumulation of human capital as potential protective factor against 'falling out' the labor force (i.e., of observing the transition 'Employed - Not Employed'). Then, we might profile individuals at greater risks of disadvantages.

Employment pathways in Australia   11

The generalized linear mixed model (GLMM) often used to analyse nominal variables with repeated observations is the dynamic multinomial logit random effects model. For such model, the response distribution is defined conditionally on the random effects that are assumed to arise from a multivariate distribution. For a conventional multinomial logit model, let Y_{it} denote the tth observation for individual i with J possible states (6 working states in our case), $Pr(Y_{it} = j|X_{it})$ is the probability of being in state j at time t given a set of exploratory variables. Among them, we will also include the employment status in the previous wave (e.g., lagged dependent variable) in order to include explicitly the time dependence between observations.

The multinomial model can be expressed as follows

$$\pi_{itj} = Pr(Y_{it} = j|X_{it}) = \frac{e^{X_{it}\beta j}}{\sum_{k=1}^{J} e^{X_{it}\beta k}}$$

The logit model pairs each response category with an arbitrary baseline category. For instance if the first response (state j_1) is set as reference, the multinomial logit model will have the form

$$\log\left(\frac{\pi_{itj}}{\pi_{it1}}\right) = X_{it}\beta j \quad j = 2, 3, ..., J$$

We might interpret the logit as sort of utility function in a decision making process. We define the utility of choosing a particular response, a certain employment state, by the random variables U_{itj} $(j = 1, , J)$, with the function $U_{itj} = X_{it}\beta_j + e_{itj}$. An individual will choose the response j, for example he/she will decide to not participate to the labor market, if the utility in staying in this state is the greatest. In other words if $U_{itj} = \max_{1 \leq k \leq J} U_{itk}$.

We will also consider another model specification introducing individual-specific random effects to model sprurious dependence. The random effects $\alpha_i = \{\alpha_{ij}, ..., \alpha_{iJ}\}$ capture non-observable individual effects that are specified to arise from a multivariate normal distribution. The multinomial logit random effect model is then defined as

$$\log\left(\frac{\pi_{itj}}{\pi_{it1}}\right) = X_{it}\beta j + Z_{ij}\alpha_{ij} \quad j = 2, 3, ..., J$$

Where Z_{ij} denote a vector of coefficients for the random effects.

The paper intends not only to illustrate different methodological approaches for analyzing longitudinal data but also to show how such approaches, investigating the socio-demographic and background factors leading to situation of disadvantages, may provide guides to policy makers who are interested in addressing labor market disadvantages.

References

1. Abbott, A. (1995). Sequence Analysis: New Methods for Old Ideas. *Annual Review of Sociology* 21, 93-113.
2. Abbott, A. and Forrest (1986). Optimal Matching for Historical Sequences. *Journal of Interdisciplinary History*. 16(3). 471-794.
3. ABS. (2001). Duration of Unemployment: Recent Definitional Changes. *Canberra: Australian Bureau of Statistics. Cat. No. 6203.0.*
4. ABS. (2009). Trends in Household Work. , in Australian Social Trends, March 2009. *Canberra: Australian Bureau of Statistics. Cat. No. 4102.0.*
5. Arulampalam, W. (2001). Is Unemployment Really Scarring? Effects of Unemployment Experiences on Wages. *The Economic Journal, 111(475), F585-F606.*
6. Billard, L. (2001). Markov Models and Social Analysis. In N. J. Smelser, P.B. Baltes (eds.), International Encyclopedia of the Social & Behavioral Sciences, pp. 9242 9250. Pergamon, Oxford.
7. Billari, F. C. (2001). Sequence Analysis in Demographic Research. *Canadian Studies in Population, Special Issue on Longitudinal Methodology* 28(2), 439- 458.
8. Burbidge, A. (2005). Australian Labour Market Transitions: Empirical Evidence. *University of Melbourne - No. CEDA paper: 82*.
9. Fasang, A. (2010), Retirement: Institutional Pathways and Individual Trajectories in Britain and Germany, *Sociological Research Online*, 15(2)
10. Fry J., and Boulton C. (2013), Prevalence of Transition Pathways in Australia. *Productivity Commission Staff Working Paper*.
11. Gabadinho, A., G. Ritschard, N. S. Mller and M. Studer (2011). Analyzing and Visualizing State Sequences in R with TraMineR. *Journal of Statistical Software* 40(4), 1-37.
12. Green, C.P. and Leeves, G.D. (2009), Scarring, Habituation and Social Exclusion: Working Histories in Secure and Insecure Employment. *School of Economics Discussion Paper No. 385*. The University of Queensland
13. Hakim, C. (2000). Work-Lifestyle Choices in the 21st Century: Preference Theory. Oxford: Oxford University Press.
14. Haynes, M. Western, M., Yu, L. and Spallek, M. (2008). Analysing Nominal Data from a Panel Survey: Employment Transitions of Australian Women. Paper prepared for the 103rd Annual Meeting of the American Sociological Association, 1-4 August, 2008, Boston, USA.
15. Martin, R. and Morrison., P.S. (Eds.) (2001). *Geographies of labour market inequality*. TSO Publishing London.
16. McVicar D., and Anyadike, M. (2002). Predicting successful and unsuccessful transitions from school to work by using sequence methods. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 165(2), 317-334.
17. Saunders, P., and Brown, J. (2004). Explaining welfare to work transitions among the unemployed. *Australian Journal of Labour Economics*, 7(3), 395-410.
18. Stevens, A. H. (1997). Persistent Effects of Job Displacement: The Importance of Multiple Job Losses, *Journal of Labor Economics*, 15(1), 165-188
19. Studer, M., Ritschard, G., Gabadinho, A. & Mller, N.S. (2011), Discrepancy analysis of state sequences. *Sociological Methods and Research*, 40(3), 471-510.
20. Vickery, J. (1999). Unemployment and Skills in Australia *Reserve Bank of Australia (No. 1999-12)*.
21. Watson, N., and Wooden, M., (2002) The Household, Income and Labour Dynamics In Australia (HILDA) Survey. *HILDA Project Technical Paper Series. No, 1/02, May 2002*.
22. Wooldridge, J.M. (2002). Econometric Analysis of Cross Section and Panel Data. Cambridge, MA; The MIT Press.

Session 16B: Methods III

Latent-transition approach to evolution of household debt possession patterns in Poland

Piotr Białowolski

University of Turin

Abstract

Based on latent transition approach we investigate evolution of debt possession patterns in Poland. We obtain interteporally comparable segments of debt holders and show highly significant role of autoregressive and life-cycle factors in shaping transitions between the segments. With data from three waves (2011 – 2015) and over 36,000 responses from the biennial panel study Polish households – Social Diagnosis - we show that: (1) transition probabilities from any of the indebted states to any other indebted state are higher than transition probabilities to non-active state; (2) households located in a given segment are most likely to remain in the segment, if not for other factors; (3) transitions between segments are shaped by socio-economic covariates - age of household head, income and number of household members; (4) their role differs very considerably between the segments showing that different socio-economic traits shape the evolution of segments; (5) probabilities of mortgage debt, debt for durables or renovation are positively related to income and in line with life cycle predictions, which indicates the role of long term factors in their acquisition; (6) consumption debt is less age related and in some groups also inversely related to income, which shows its short term dependence. The results of the study indicate also that (1) relationship between debt and income depends on relative prevalence of segments on the market and (2) that influence of life-cycle factors can be missing at the aggregate level if mortgage debt does not dominate.
Keywords: household debt, segmentation, latent transition

2

1. Introduction

We could picture a situation, where household financial markets operate without frictions and market imperfections remain absent. In such a case there would be no need to distinguish between different kinds of household assets or liabilities. In extreme, as (Bertola & Hochguertel, 2007) claim "there would not even be a need to distinguish owning from renting, …, savings from borrowing, as each of them could be costlessly and instantly converted from one form to another and from there to consumption." Yet, it is commonly observed that households use credit to finance their purchases but also to bridge temporary drops in income in a very diversified way (Białowolski, 2014). In addition, they rely on a very complex process of acquisition of financial products, which is especially visible in the case of debt (B. Kamleitner & Kirchler, 2007; Bernadette Kamleitner, Hoelzl, & Kirchler, 2012; Kirchler, Hoelzl, & Kamleitner, 2008). In consequence, they end up with differentiated credit instruments addressing their specific needs (Bertola & Hochguertel, 2007).

The scope for consumption motives, which are eventual cause for credit behaviour, was primarily addressed by Keynes (1936). He pointed to Enjoyment, Short sightedness, Generosity, Miscalculation, Ostentation and Extravagance as primary motives for consumption. However, as the life-cycle theory (Friedman, 1957; Modigliani & Brumberg, 1954) became the workhorse of economics, recollection of the behavioural nature of consumption (and credit) proposed by Keynes started to fade. However, the life-cycle theory encountered many problems in explaining household financial behaviour both with respect to saving and debt. Remedies to the household misbehaviour in the basic model comprised various amends and covered among others introduction of liquidity constraints (Hall & Mishkin, 1982; Jappelli & Pagano, 1989), durable goods in utility (Browning, 1989; Mankiw, 1985), uncertainty mostly associated with future income path (Parker & Preston, 2005; Zeldes, 1989) or habit formation (Campbell & Cochrane, 1999; Carroll, Overland, & Weil, 2000). Even though these theory developments seemed to introduce

3

new and more sophisticated framework to analyse life-cycle behaviour, they still failed to provide sufficient evidence that life-cycle approach is the proper one to analyse household behaviour. Liquidity constraints might be in force but still there might be a group of households simply unwilling to participate in any saving or credit activities, which is corroborated by various survey data (Białowolski & Dudek, 2014). Durable goods play an important role but in the current world the durability of various goods – probably apart from house/apartment – has shortened significantly (Davidson & Bates, 2002) rendering a diminished role for the durables in shaping the life-cycle behaviour with respect to savings and debt. Finally, there is little support for habit formation (Dynan, 2000) or an important role of uncertainty.

Some answers to the puzzling financial market behaviour of households were addressed by introduction of behavioural concepts into the analysis. Developments in the area of self-control pointing to impatience of consumers (Thaler & Shefrin, 1981) led to better understanding of consumer behaviour. Yet, the most recent trend is towards reconciliation of the two approaches, where two types of behaviour (life-cycle and behaviourally driven) meet. One of the approaches was proposed by (Bertaut, Haliassos, & Reiter, 2009), who with the accountant-shopper model try to solve the puzzle of existence of a groups of individuals holding both short-term debt and liquid assets. They suggest that each consumer makes consumption choices in two stages (Bertaut et al., 2009). In the first stage, when she adopts a role of accountant, plans expenditures and makes basic payments. In the second stage, the shopper comes into the stage. He is also rational but less patient. As the accountant is aware of shoppers attitudes, he does not want to leave too much space to the shopper and thus holds positive short-term debt. Based on this approach it is possible to show that households would maintain positive short-term debt even in the light of increasing incomes and financial assets. In the same strain remains the approach of Kahneman (2011) comprising two systems, where duality of human decision making approach is captured with two selves - rational undertaking life-cycle choices and impulsive prone to various behavioural biases.

In this paper we show that neither purely life-cycle nor purely behavioural approach are adequate for an analysis of household financial behaviour

4

with respect to debt taking. Our goal is to show that there is a finite number of segments, which define in probabilistic terms household debt behaviour. Till present there has been very little evidence presented that such patterns can be discerned. Works of (Białowolski, 2014; Gunnarsson & Wahlund, 1997; Viaud & Roland-Lévy, 2000) comprise exceptions to this rule. By showing that segments are intertemporally comparable we shape way for analysis of transitions between segments. Distinction of comparable groups allows us to verify, whether postulated by theory relationships like curvilinear shape of debt vs. age (Bernadette Kamleitner et al., 2012; Ngwenya & Paas, 2012; Paas, Bijmolt, & Vermunt, 2007) or lower use of debt among single person households (Chien & Devaney, 2001) are observed universally or present only in certain groups of debt takers. With segmentation approach it is also possible to disentangle confusing relations between income and credit, as households within groups are allowed to behave differently. It is especially important in light of ambiguity in the behaviour of aggregates. Crook (2006) indicates to positive relationship between income and debt, while Gunnarsson & Wahlund (1997) indicate lack of such relationship.

In order to meet the goals, adopted approach should enable to distinguish the number of distinct patterns of behaviour (groups), investigate invariability of these patterns, and intertemporal transitions between them. Additionally, it is essential to provide valid arguments for the source of market evolution and distinguish between household migration between groups associated with inertia and intrinsic modes of credit use, but also investigate a role of household socio-economic characteristics. Hence, we adopt latent transition approach, which enables to capture not only multitude of paths with respect to credit, which can be either life-cycle or behaviourally driven, but also allows to capture intrinsic fuzziness of approach to credit of a given household. With such an approach it is possible to show existence of a mixture of behaviours even at the most basic, household level. To evaluate the credit paths we use data from the largest household panel in Poland – Social Diagnosis Survey and limit our attention to its last three waves 2011, 2013, and 2015.

This paper features at least two innovative points. First, to the best of our knowledge, it is the first approach based on latent transition models,

5

which tries to evaluate stability of changes between patterns of various debt possession, while at the same time investigating plausibility of the assumption of intertemporal pattern comparability. Second, to the best of our knowledge, it is the first attempt to show that households might exhibit both behaviourally and life-cycle driven paths of their credit market behaviour.

To meet these objectives, the paper is organised as follows: in Section 2, we present background for the analysis describing the situation in the area of household debt in Poland – the market subject to subsequent analysis in the paper. A detailed description of methods, including the dataset from the Social Diagnosis Survey, selection of measures and methods used for the analysis follows in Section 3. In Section 4, we describe results obtained with multi-group latent class models and latent transition models. We investigate the determinants of households' participation in the credit market and latent class (segment) membership, while taking into account strength of both transitions and socio-economic variables. Section 5 presents the conclusions of the study.

2. Background – data and measures

Household indebtedness in Poland rose quickly from the beginning of 2000 till the onset of the financial crisis (Figure 1). The indebtedness of households went from merely 12% of the Polish GDP in 2003 to 35% in 2015. During the period, the ratio of households' debt to GDP in the EU27 countries averaged at the level of 60% (Pyykkö, 2011) and, from this perspective, the indebtedness of Polish households remained low. Between 2003 and 2009 there was a rapid growth in the penetration rates of credit in all areas – consumer, housing and other. However, just after the outburst of the financial crisis the consumer credit market started to contract in relative terms. Nevertheless, household debt still grew till 2011 due to mortgages. It should be stressed that in the period 2009 – 2011 Polish carry traders were considerably affected by depreciation of the Polish zloty and significant share of indebtedness growth during that period can be attributed to this phenomenon.

6

Figure 1 Debt-to-GDP ratio (mortgage, consumer and other debt), share of indebted households in Poland.
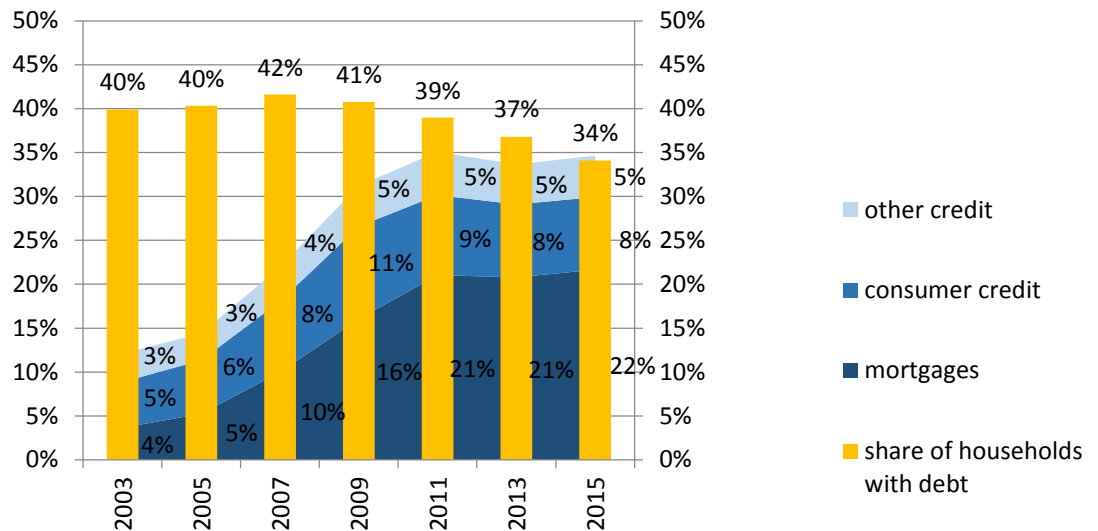
Source: National Bank of Poland, Polish Central Statistical Office.

The level of household indebtedness in Poland did not exceed debt levels observed in any of the developed economies. Additionally, the growth trend of debt-to-GDP stopped, which requires further scrutiny and poses additional questions about the reasons for that. However, from the perspective of further analysis the household debt-to-GDP ratio seems to have achieved relative stability during the period 2011 – 2015, which allows to pick that period as reference frame for analysis of debt possession patterns.

Results of the largest panel study in Poland – the Social Diagnosis (Czapiński & Panek, 2015) – confirm that since 2009 a constant decrease in the use of credit market is observed among Polish households (Figure 2). Trends indicate that, although there is a constant growth of total value of household debt, the share of participating households is shrinking. It has decreased from 41.6% at its peak in 2007 to 34.1% in 2015. For the purpose of analysis we limit our attention to three waves of the survey. They cover the state of households' credit portfolios for the period from 2011 – 2015 and are gathered in a panel-type study.

7

During the period of interest the number of households participating in the Social Diagnosis Survey ranged between 11,703 in 2015 and 12,354 in 2013. Ca. 60% of households remained in the panel between consecutive waves.

For the purpose of our study we utilize description of household debt possession patterns in three dimensions: debt source, objectives for taking debt and the value of the debt. The evolution of the share of households with respect to all of the dimensions is presented in Table 1.

Data in Table 1 confirms the decline in the share of debt holders among Polish households between 2011 and 2015. With respect to the debt value, there is a constant trend of increase, which is manifested by growing number of shares of households with debt exceeding their yearly incomes. However, with respect to the source of credit little variability has been noted apart from a considerable increase in the share of households using other financial institutions between 2011 and 2013. As noted by (Bertola & Hochguertel, 2007) household credit portfolios are not very diversified with respect to the product and source, which is confirmed by overwhelming dominance of banks among debt sources. However, household debt behaviour seems to significantly differ with respect to the purposes the debt serves (Białowolski, 2014). With respect to credit target, there is still visible an upward trend in possession of mortgage – indebtedness for house/flat purchase - but also more indebted households tend to finance their holidays. On the other hand consumption related credit starts to decline. It is hard to explain this behaviour solely by the life-cycle patterns as expected high path of income growth strongly favours increase in the share of debt takers and the use of debt for various purposes – also consumption related.

8

Table 1   The percentage of households with respect to the value, source and target of a loan/credit (among borrowers) between 2011 and 2015

		Wave		
		2011	2013	2015
Share of debt holders		39.0%	36.8%	34.1%
debt value (in the value of monthly income, among debt holders)	< 1	22.1%	19.4%	16.7%
	1 – 3	22.6%	20.2%	19.9%
	3 – 6	16.7%	17.0%	15.7%
	6 – 12	14.8%	13.1%	11.5%
	above 12	23.8%	30.3%	36.2%
credit source (among debt holders)	Banks	90.6%	87.8%	90.6%
	other financial institutions	11.8%	15.8%	12.3%
	family/friends	5.1%	5.4%	4.4%
credit target (among debt holders)	current consumption expenditures (e.g., food, clothing)	17.5%	15.5%	13.5%
	fixed charges (e.g., house maintenance)	8.0%	8.2%	6.2%
	purchase of durable goods	36.7%	34.0%	32.3%
	purchase of a house/flat	18.0%	22.8%	26.1%
	renovation of a house/flat	31.1%	30.2%	30.2%
	medical treatment	6.4%	7.1%	6.2%
	purchase/rent of working equipment	2.6%	3.0%	2.9%
	Vacation	2.6%	3.0%	3.8%
	repayment of previous debts	8.0%	8.2%	6.7%
	development of own business	5.6%	4.6%	5.0%
	education/training	3.3%	4.3%	3.8%
	other purposes	10.5%	11.7%	12.0%

Source: Own calculations in Mplus based on data from The Social Diagnosis.

9

3. Methods

3.1. Statistical approach

The segments of households with respect to debt possession patterns are revealed with latent-class modelling, its multi-group extension and latent transition modelling. These techniques allow for accounting for an unobserved heterogeneity in the multidimensional data set. Dimension reduction is primary performed with introduction of a single latent variable describing segments.[1] Multi-group approach (McCutcheon, 2002) is an extension of latent class modelling, which enables additional testing of segment homogeneity and allows for direct comparison of segments between groups. In this paper, different groups correspond to different time points of analysis. Latent transition modelling enables to track changes in segment membership and relate them to socio-economic characteristic.

A multi-group latent class model can be defined with N manifest variables $A_1 A_2 .... A_N$ (answers to N questions), each having M_i ($m_1=1..M_1$; $m_2=1..M_2$; ...;$m_N=1..M_N$) answer categories, one latent variable X with $k=1,...,K$ classes and one grouping variable T with $t=1,...,L$ groups. In this setting, it is possible to define L cross-tables each with N dimensions that represent interrelations between manifest variables in each group (in our case at each time point). Including latent variable X leads to the following form of the model:

$$\pi_{m_1 m_2 ... m_N kt}^{A_1 A_2 ... A_N X|T} = \pi_{kt}^{X|T} \pi_{m_1 kt}^{A_1|XT} \pi_{m_2 kt}^{A_2|XT} ... \pi_{m_N kt}^{A_N|XT} \quad (1),$$

---

[1] More detailed description of the latent class modelling can be found in (Muthén, 2004) or (Białowolski, 2014). Its advantages over other segmentation techniques are well depicted in Vermunt & Magidson (2002). Estimation of the latent class models is performed with a maximum likelihood estimator following the EM algorithm, in which the information on latent class membership is considered missing and thus is derived from the data (Muthén, Shedden, & Spisic, 1999).

10

where $\pi_{m_1 m_2 \ldots m_N kt}^{A_1 A_2 \ldots A_N X | T}$ defines the conditional probability that a respondent with the set of answers ($m_1, m_2, \ldots, m_N$) given in period t belongs to latent class k, while $\pi_{kt}^{X|T}$ defines the conditional probability of belonging to class k given period t, and $\pi_{m_i kt}^{A_i | XT}$ defines the probability of providing answer m_i to item A_i given class membership (k) and given the period of analysis (t). Latent class models in such a specification are based with an assumption of local independence, which implies that the answers to manifest questions ($A_1, A_2, \ldots, A_N$) are independent of each other, given the latent class k.

Multi-group latent class allows to establish comparability of groups between periods and can be described by the following formula:

$$\pi_{m_1 m_2 \ldots m_N kt}^{A_1 A_2 \ldots A_N X | T} = \pi_{kt}^{X|T} \pi_{m_1 k}^{A_1 | X} \pi_{m_2 k}^{A_2 | X} \ldots \pi_{m_N k}^{A_N | X} \tag{2}$$

In this specification, the indicator variables – answers to questions – are not directly dependent on the grouping variable (time). The understanding of latent classes (segments), as expressed by its indicators (questions), is invariant of the grouping variable. At this level of measurement invariance, a change in the probability of answering a given question depends only on the latent class membership (not on the time). However, latent class membership probability can change between time points.

Finally, the most elaborate form of analysis reported here is latent transition model, where class membership in a given point of time depends on both previously reported class membership and other socio-economic covariates. As it is only the class membership that is subject to autoregressive processes and influenced by socio-economic characteristics, probability of class membership is only subject to influence by those factors:

11

$$\pi_{kt}^{X|T} = \frac{e^{thresh_{k,t}+\sum_{j=1}^{J}\alpha_{j,k,t-1}\cdot x_{j,t-1}+\sum_{p=1}^{K-1}\beta_{p,k,t-1}\cdot\upsilon_{p,t-1}}}{1+\sum_{i=1}^{K-1}e^{thresh_{i,t}+\sum_{j=1}^{J}\alpha_{j,k,t-1}\cdot x_{j,t-1}+\sum_{p=1}^{K-1}\beta_{p,i,t-1}\cdot\upsilon_{p,t-1}}}$$

(3),

where $\{x_{1,t-1},...,x_{J,t-1}\}$ is a set of explanatory variables in period t-1, while $\alpha_{j,k,t-1}$ represents the estimated parameters, which are set to zero for a selected, reference class. Additionally, $\{\upsilon_{1,t-1},...,\upsilon_{K-1,t-1}\}$ represents a set of binary variables representing class membership in period t-1 with $\beta_{p,k,t-1}$ being parameters associated with transition from class p to k between t-1 and t.

3.2. Modelling strategy

Our approach to analyse debt possession patterns is based on the following assumptions: (1) there is a possibility to detect distinct patterns of credit use and those patterns are comparable intertemporally, (2) distinct patterns are driven, with varying strength, by both autoregressive processes and socio-economic characteristics of households. To verify these assumptions we use the following step-wise approach. At first, we determine the number of classes and subsequently verify the hypothesis of equal intertemporal meaning of latent patterns in the area of household debt possession patterns. Second,[2] we check the influence of segment membership in t-1 on segment membership in t and subsequently search for covariates of latent patterns.

---

[2] Because the transitions are multiple we modify the approach traditionally used for estimation of the latent transition model. We conduct the analysis stepwise allowing the description of classes to be defined in the first step and not influenced by transition ($\beta_{p,k,t-1}$) and socioeconomic covariates ($\alpha_{j,k}$). In order to maintain the latent character of class membership, we multiply impute the class membership following the latent class membership probabilities obtained in the first step. For this purpose we use set of 10 multiple imputations.

12

At each step of the analysis, we adopt the approach based on Bayesian Information Criterion (BIC) (Schwarz, 1978) following the procedure: (1) the optimal number of groups is established in the model with measurement invariance (i.e. satisfying eq. 2) and without measurement invariance (eq. 1); (2) For the final solution, a quality check with the entropy measure is performed;[3] (3) the selected segmentation model is first checked for importance of previous period latent class membership, it is done with the use of information on previous period class membership $\{v_{1,t-1},...,v_{K-1,t-1}\}$ and $\beta_{p,k,t-1}$ coefficients are being estimated; (4) the class movement coefficients are checked for equality in the intertemporal setting, i.e. constraint $\forall_{p\in 1,...,K-1;k\in 1,...,K-1}\beta_{p,k,1} = \beta_{p,k,2}$ is imposed on the parameters; (4) socio-economic covariates of class membership are introduced to the model in the stepwise manner starting from age of household head, and following by income, number of household members, education of household head, and place of residence, (5) each covariate is first introduced unconstrained, i.e. diverse impact between periods is possible, then constrained solution in intertemporal setting is checked $\forall_{j\in 1,...,J;k\in 1,...,K}\alpha_{j,k,1} = \alpha_{j,k,2}$, finally insignificant estimates are constrained to zero and those for adjacent categories, if observed to be insignificantly different between each other, are constrained equal.

4.   Results

4.1. Number of classes

To detect the number of homogeneous segments at all time points, latent class models are initially estimated separately for 2011, 2013 and 2015. During the estimation process, it was established that the best

---

[3] Each household is classified into its most likely class, and then a table is constructed with rows corresponding to households classified into a given class and column entries give the conditional probabilities of belonging to a given class (Muthén 2004). The entropy measure is defined to vary between zero and one, and entropy values close to one indicate clear classifications based on the model (Muthén, 2004). The detailed formula for the entropy measure can be found in Muthén (2004).

fitting model for 2011 are those with 10 classes, and for 2013 and 2015 the best-fitting model have 9 classes (see Appendix 1).[4] In order to ascertain comparability of latent classes in the intertemporal setting two types of models are estimated - models with unconstrained conditional response probabilities and models with constrained ones. In the latter specification, the probabilities of class membership in different groups (time points) can be compared because the meaning of the latent classes is preserved for all of the periods of the analysis. The values of the BIC for the two specifications of the model are presented in Table 2.

Table 2   BIC for latent class models with unconstrained and constrained response probabilities

BIC	Heterogeneous – latent classes have different meaning	partially homogeneous – latent classes are comparable between periods
different no. of classes	278608.3	---
11 classes	279168.5	275013.8
10 classes	278709.9	**274969.8**
9 classes	278957.9	275798.3

Source: Own calculations in MPlus.

Based on the results, the best fitting model is the 10-class specification with constrained response probabilities, i.e. intertemporally comparable latent classes. The fit of the model also proves to be very good, which is confirmed by the value of the entropy measure (0.976). Item response probabilities, calculated in line with Formula (2) for each latent class in the final model, are presented in Table 3. The final model comprises an implicit description of the latent classes of households in Poland in the credit market. There are ten distinct segments of households active in the credit market (classes 1 – 9) and one segment not active on the market. In order to understand transitions between different segments, we need to provide a more detailed description of those:

---

[4] Due to very large number of observations it is highly unlikely that the effect observed by Lukočienė, Varriale, & Vermunt (2010) was present. They state that the smaller the sample size, the less likely that one finds the correct number of classes.

Table 3   Response probabilities in latent classes (2011 – 2015)

Response probabilities		c.1	c.2	c.3	c.4	c.5	c.6	c.7	c.8	c.9	c.10
credit value (in the value of monthly income)	Zero	0	0	0	0	0	0	0	0	0	1
	< 1	.026	.349	.081	.063	.357	.417	.222	.29	.156	0
	1 – 3	.035	.294	.14	.118	.297	.331	.244	.323	.235	0
	3 – 6	.053	.174	.22	.185	.182	.143	.205	.204	.207	0
	6 – 12	.084	.099	.225	.193	.092	.064	.161	.109	.191	0
	above 12	.801	.085	.333	.441	.072	.045	.169	.074	.212	0
credit source	Banks	.994	1	.995	.99	0	.267	1	1	1	0
	other financial institutions	.027	0	.017	.318	.712	1	.058	.042	.041	0
	family/friends	.021	.003	.016	.191	.318	0	.013	.037	.013	0
credit target	current consumption expenditures (e.g., food, clothing)	.011	.041	.011	.503	.367	.052	.061	.623	.038	0
	fixed charges (e.g., house maintenance)	.006	.004	0	.379	.203	0	.007	.271	.004	0
	purchase of durable goods	.116	1	.246	.494	.115	1	.129	.033	.236	0
	purchase of a house/flat	1	.003	.052	.091	.04	.004	0	.014	.014	0
	renovation of a house/flat	.111	0	.132	.468	.324	.129	0	.104	1	0
	medical treatment	.006	.017	.024	.232	.117	.03	.024	.23	.034	0
	purchase/rent of working equipment	0	.002	.319	.031	.009	.011	.008	.005	.002	0
	Vacation	.01	.012	.024	.173	.041	.031	.012	.038	.017	0
	repayment of previous debts	.009	.013	.104	.524	.1	.005	.026	.121	.033	0
	development of own business	.008	.002	.622	.051	.014	.005	0	.005	.001	0
	education/training	.009	.015	.035	.177	.043	.021	.031	.083	.024	0
	other purposes	.015	0	.074	.291	.155	.014	1	.019	.053	0

Source: Own calculations in MPlus based on data from The Social Diagnosis.

Class 1 (MORTGAGE DEBTORS) – Households that are extremely highly indebted (80% possess debt exceeding their annual incomes). These households' sources of credit are mostly banks (99.4%) and rarely other institutions. These households comprise the first group of those indebted to finance the purchase of a house or flat (100%). They rarely also use debt to finance the renovation of a house or flat (11.1%) or to purchase durables (11.6%).

Class 2 (BANK FINANCED DURABLE GOODS CONSUMERS) – Households that have a relatively low value of debt (64.3% with debt below their quarterly incomes). These households' acquire their debt from banks (100%) and almost never from elsewhere. The debt is devoted almost solely to purchases of durables (100%).

Class 3 (HOUSEHOLD-BUSINESS DEBTORS) – Households with debts above average that were acquired from a bank (extremely rarely supported by a loan from other source). The debt is designated for development of own business (62.2%), purchase of working equipment (31.9%) and sometimes for the purchase of durables (24.6%).

Class 4 (OVERINDEBTED CONSUMERS) – Households that have high probability of debt in many categories with respect to the purpose and very often with a high value of debt (63.4% with debt exceeding their semi-annual incomes). These households' sources of credit are mainly banks (99%), but they also often search for credit from other financial institutions (31.8%) and from their friends and family (19.1%). In this group, there is a very high probability of credit for current consumption (50.3%) and for repayment of previous debts (52.4%) but also for fixed charges (37.9%), the purchase of durables and renovation of a flat (each of the last two approaching 50%). Due to a very high value of debt and the goals associated with current consumption (or repayment of debts), this group of households can be classified as over-indebted.

Class 5 (NON-BANKING SECTOR CONSUMERS) – Households that have a low value of debt (65.4% below their quarterly incomes) but who acquire it from outside the banking sector. 71.2% declare loans from other financial institutions and 31.8% from private persons. These households devote their loans mostly to consumption (36.7%) and fixed charges (20.3%) but also to renovation of a flat (32.4%), purchase of durables (11.5%) and medical treatment (11.7%).

16

Class 6 (NON-BANKING SECTOR DURABLE GOODS CONSUMERS) – Households with lowest value of debt (74.8% with debt below their quarterly incomes). These households' acquire their debt from other financial institutions (100%) and sometimes from banks (26.7%). The debt is devoted almost solely to purchases of durables (100%), rarely to renovation of flat (12.9%) or current consumption (5.2%).

Class 7 (OTHER PURPOSE DEBTORS) – Households that have slightly below-average value of debt. These households' acquire their debt from banks (100%), and they very rarely support the debt with a loan from other sources. They use the debt to finance other purposes (100%) and rarely to purchase durables (12.9%).

Class 8 (LOW-VALUE CONSUMPTION DEBTORS) – Group of households with below average value of debt (61.3% report debt below their quarterly incomes). These households' sources of credit are almost entirely banks (100%). Although the debt is low, there is a very high probability of debt for current consumption (62.3%) or fixed charges (27.1%). Repayment of previous debts is present in the group of 12.1% members of the segment. Additionally, financing of health care expenditures is present among 22.3% of group members.

Class 9 (RENOVATION DEBTORS) – Households that have a slightly lower than average value of debt. These households' acquire their debt from banks (100%) and rarely from elsewhere. The debt is always devoted to renovation of apartment (100%). Sometimes supported by the purchase of durables (23.6%).

Class 10 (NON-ACTIVE) – Households with no debt.

There is a distinct diversity of credit market participation patterns with very different values of debt, sources and motives for its acquisition. However, each segment is characterised by a distinguishable primary motive for debt possession or clear-cut combination of motives. It is natural that for larger sums, households tend to switch towards the banking sector. It is especially visible in acquisition of debt for housing purposes, which is usually associated with considerable debt values. However, with respect to financing consumption or small durable purchases one can observe competing strategies and banks face competition from the non-financial institutions.

17

4.2. Evolution of segments in time

Comparable pattern of credit market behaviour enables to trace the evolution of the composition of the market between 2011 and 2015. In Figure 2, we present the results of the latent class segmentation providing the composition of the group of active debt holders.

Figure 2 Debt possession patterns among Polish households between 2011 and 2015
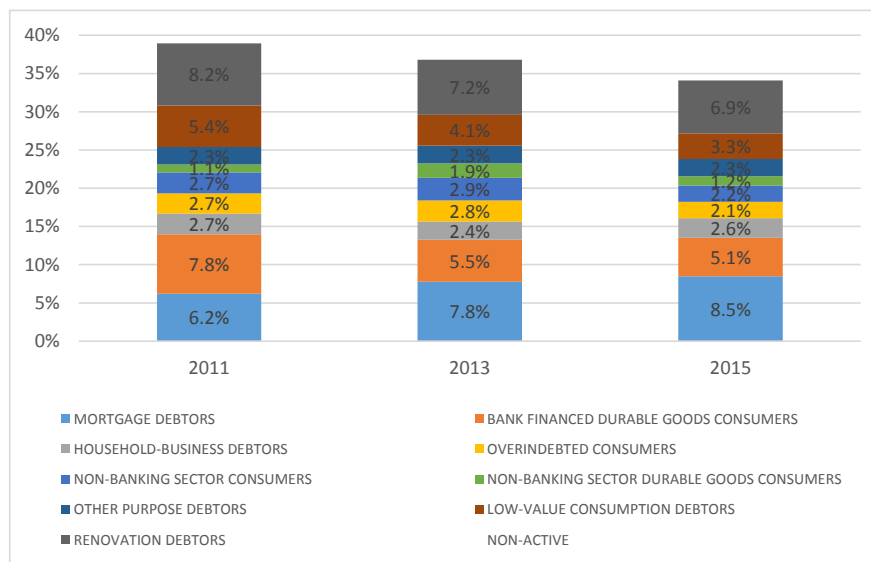
During the period of analysis, the share of indebted households was decreasing. However, also the relative role of segments has been changing. MORTGAGE DEBTORS became the dominant group of debt holders, while the BANK FINANCED DURABLE GOODS CONSUMERS have been constantly losing their market share. The groups of households indebted for pure consumption or even over-indebted have been also declining. The declining role of non-banking forms of borrowing is confirmed by the gradual disappearance of the group of households that borrows from the non-banking sector.

In the latent transition approach, the total change in the market structure is attributed to the time evolution of the market and socio-economic characteristics of households that influence participation in the credit market. Following the modelling strategy presented in 3.2 in the model we first

18

included transitions between latent states. As the model with intertemporally equal transition coefficients proved to be superior in terms of BIC, table 4 depicts odds ratios for transitions between classes in consecutive periods.

Table 4   Odds ratios for transitions between latent classes in consecutive periods

		Class membership in t								
		c.1	c.2	c.3	c.4	c.5	c.6	c.7	c.8	c.9	c.10
Class membership in t-1	c.1	94.25***	5.27***	9.47***	15.39***	2.46*	3.06**	7.92***	4.02***	12.40***	ref.
	c.2	2.51***	9.23***	4.98***	4.62***	4.03***	6.23***	6.06***	4.53***	5.73***	ref.
	c.3	3.43***	5.14***	52.46***	9.54***	2.57**	1.97	8.83***	6.63***	5.02***	ref.
	c.4	17.89***	8.00***	13.44***	108.09***	15.67***	8.13***	18.71***	19.99***	15.93***	ref.
	c.5	1.54	3.35***	1.07	9.68***	16.64***	6.29***	6.39***	4.51***	5.10***	ref.
	c.6	4.23***	5.77***	1.72	11.00***	7.36***	10.39***	4.62***	3.48***	5.47***	ref.
	c.7	7.40***	7.21***	7.58***	15.86***	5.97***	2.49*	23.69***	7.32***	7.17***	ref.
	c.8	6.62***	5.85***	5.16***	24.29***	7.36***	3.99***	8.97***	21.85***	8.70***	ref.
	c.9	7.27***	5.64***	5.20***	14.94***	6.46***	5.32***	6.94***	7.29***	21.28***	ref.
	c.10	ref.	ref.	ref.	ref.	ref.	ref.	ref.	ref.	ref.	ref.

Source: Own calculations in MPlus based on data from The Social Diagnosis.
Note: *** significant at 0.01 level; ** significant at 0.05 level; * significant at 0.1 level.

Estimates of the odds ratios – mostly significant and all larger than one – indicate that being a debt holder increases transition to the same or other debt holding segment with respect to the NON-ACTIVE segment. It shows a behavioural trait in household behaviour indicating that being a debt holder works in the direction of higher debt holding probability in the following period. It shows that once a household enters in a debt holding pattern, it is more likely to carry on with debt in the following periods. One can also observe very high probabilities of transition to the same debt holding state. Especially high odds ratios are recorded for households remaining in the MORTGAGE DEBTORS, HOUSEHOLD-BUSINESS DEBTORS and OVERINDEBTED CONSUMERS states. In all cases the odds ratios for remaining in the same state vs. transitioning to the NON-ACTIVE state exceed 50.[5] Although in the case of

---

[5] It should be underlined that the state of overindebtedness does not necessarily have to lead to default on debts. It is more likely income shocks (unemployment) or negative household events (e.g. divorce) that stimulate difficulties with repaying debts (Jentzsch &

19

MORTGAGE DEBTORS it seems quite natural to remain in a given state as mortgages are products with extremely long maturities, in the case of OVERINDEBTED CONSUMERS very high transition probability indicates strong persistence of debt.

4.3. Influence of socio-economic factors on segment evolution

Additionally to the autoregressive nature of transitions between debt possession states, there is a natural tendency of the state membership to be influenced by the socio-economic variables. Following the procedure depicted in 3.2 we started the evaluation of potential influence of age of the household head, income and number of household members. Among those only the first three were found significant factors influencing the class transition. The odds ratios for transition to a latent in the final model with a set of covariates are presented in Table 5.

In the case of seven out of nine[6] classes the probabilities of transition to a debt holding segment from NON-ACTIVE was at least partially influenced by age. MORTGAGE DEBTORS tend to be extremely strongly influenced by age and it is much more likely to transition to the class for households with head aged 45 or less than for households with older heads. The odds for over 65 year olds are more than six times smaller than for the reference group (25 – 34 years). Very strong life cycle patterns are also observed in the case of HOUSEHOLD-BUSINESS DEBTORS and OVERINDEBTED CONSUMERS. However, in the case of the latter group the odds for being in the class drop significantly only after the age of 55. Moderate life cycle patterns are observed in the groups of BANK FINANCED DURABLE GOODS CONSUMERS and RENOVATION DEBTORS. Both these groups exhibit much lower transition probabilities after the head reaches age of 45 years. There are no groups in which a reversed debt-age pattern would have been obse**rved.**

Riestra, 2006). Consequently, very high transition probabilities pointing to remaining in this state.

[6] The last class serves as reference class.

Table 5 Odds ratios for transitions to a given latent class for socio-economic variables

		c.1	c.2	c.3	c.4	c.5	c.6	c.7	c.8	c.9	c.10
no. of people in household (4 – ref.)	1	0.61***0.62***		0.52**	---	---	---	---	---	0.75***	ref.
	2	0.74**	0.78**	1.00^f	---	---	---	---	---	0.75***	ref.
	3	1.00^f	1.00^f	1.00^f	---	---	---	---	---	1.00^f	ref.
	5 or more	1.00^f	1.00^f	1.71***	---	---	---	---	---	1.00^f	ref.
real income per equivalent unit in PLN (above 1500 PLN up to 2000 PLN – ref.)	below 500	0.40***0.67**		---	---	3.49***	1.00^f	---	1.34**0.54***		ref.
	500 – 999	0.59***1.00^f		---	---	2.23***	1.00^f	---	1.34**1.00^f		ref.
	1000 – 1499	1.00^f	1.00^f	---	---	1.62***	1.00^f	---	1.34**1.00^f		ref.
	2000 – 2999	1.71***1.00^f		---	---	1.00^f	0.48***	---	0.64**1.00^f		ref.
	3000+	3.14***1.00^f		---	---	1.00^f	0.48***	---	0.64**1.00^f		ref.
age of the head of house-hold in years (35 – 44 years – ref.)	up to 24	1.00^f	1.00^f	1.00^f	1.00^f	---	1.00^f	---	1.00^f	1.00^f	ref.
	25 – 34	1.00^f	1.00^f	1.00^f	1.00^f	---	1.00^f	---	1.00^f	1.00^f	ref.
	45 – 54	0.50**0.77**		1.00f	1.00f	---	1.00f	---	1.00f	0.79**	ref.
	55 – 64	0.38***0.75***		0.57***	0.70**	---	1.00f	---	1.00f	0.80**	ref.
	65 and over	0.15***0.51***		0.29***	0.42***	---	0.61***	---	0.79**0.47***		ref.

Source: Own calculations in Mplus based on data from The Social Diagnosis; f = fixed parameters.

However, in the case of NON-BANKING SECTOR CONSUMERS and OTHER PURPOSE DEBTORS there is no distinct relationship with age. It strongly suggests that behaviour in these groups is not driven by life cycle relationship but rather depends on something else – like behavioural traits of household.

With respect to income, also significant differences between groups are observed. As the debt value is measured relatively to the household income, the larger the influence of income, the more it is expected that exclusion and uncertainty to play a role. Households with relatively low incomes are banned from some types of credit but also experience anxiety related to higher probability of job loss and repayment problems. These two phenomena manifest the most in the group of MORTGAGE DEBTORS. The odds for transitioning to the segment are over three times higher for households with incomes above 3000 PLN than for the reference group (1500-2000 PLN). The group of MORTGAGE DEBTORS is the most affected by the incomes in the direction supported by extensions to the life cycle theory associated with liquidity constraints and uncertainty. In other groups the impact of incomes on participation membership is significantly lower, with examples of some exclusion related to very low incomes in the groups of BANK-FINANCED DURABLE GOODS CONSUMERS and RENOVATION DEBTORS and other examples where the influence of incomes is reversed - NON-BANKING SECTOR CONSUMERS, NON-BANKING SECTOR DURABLE GOODS CONSUMERS and LOW-VALUE CONSUMPTION DEBTORS. In the case of BANK-FINANCED DURABLE GOODS CONSUMERS and RENOVATION DEBTORS there is a notable exclusion of very low income earners, which is consistent with the fact that households from these groups have acquired their debts always from banks, which are very cautious in supplying debt to very low-income households. A different pattern noted from non-banking sector consumers NON-BANKING SECTOR CONSUMERS suggests that these consumers tend to satisfy their consumption needs but fail to obtain financing from banks so they are strongly motivated to search for alternatives – the lower incomes they have the more likely is their transition to the segment. In the group of NON-BANKING SECTOR DURABLE GOODS CONSUMERS the role

22

of low-incomes is not discriminatory but the high incomes tends to limit activity in the sector. Finally, LOW-VALUE CONSUMPTION DEBTORS, who are often indebted in banks, have a limited ability to obtain credit if exhibiting low incomes, so the odds for participation do not increase with decreasing incomes so much as in the NON-BANKING SECTOR CONSUMERS but participation in the group is strongly limited among higher income households.

The last relation that turned out to be significant is related to the number of household members. However, it proved to be significant only for the groups of MORTGAGE DEBTORS, BANK-FINANCED DURABLE GOODS CONSUMERS, HOUSEHOLD-BUSINESS DEBTORS and RENOVATION DEBTORS. Households with more people have to face larger needs – especially related to housing (mortgages and renovation), transportation (cars comprising large durables) or fostering their small business. Thus, their transitions to the groups of debt holders mentioned above are self-sustaining with the assumptions of life-cycle theory. In other groups, the motives for taking debt are less life-cycle related and thus there is no observed relationship between the number of household members and the odds for being in the group.

5.   Conclusions

In this article, we presented an analysis of households' behaviour with respect to the possession of debt in the longitudinal perspective. We tracked household indebtedness patterns and traced its sources to life-cycle patterns but also to self-persistent behaviour related to transition drivers related to previously exhibited indebtedness pattern. In the scope of the analysis, we identified nine distinct patterns of debt possession in Poland and one group of households without debt. With a latent transitions framework, we demonstrated that models in which the rules of segmentation of the market remain constant in all the periods of analysis are superior to the models in which the rules change over time, which enabled to conduct a valid tracking of time evolution of debt behaviour.

One of the most interesting developments from the study is that there is a very strong persistence in credit behaviour. It is not only related to

23

staying within one group of debt behaviour but also shows that transitions between any group of debt holders is more likely than transitions to the NON-ACTIVE state. Yet, the role of life-cycle factors cannot be ruled out. There is a very strong presence of those factors in the group of MORTGAGE DEBTORS. In this group all factors that proved to be significant (age, income and number of household members) point to a very strong life-cycle behaviour and additionally indicate that extensions of the LCPIH associated with liquidity constraints play a crucial role. It is also visible, but to a lesser extent, in the purchases of durables and renovation expenditures with the use of bank credit. However, groups of households for which the main driver of credit behaviour is consumption are much more inclined to exhibit an inverse probability of debt to income relation, which although to some extent is supported by the findings of the life-cycle theory with respect to the consumption smoothing but also, in the light of very high probabilities of transitioning to the same classes, shows that such behaviour might be intrinsically driven. In some groups transitions cannot be resolved by the life cycle model. It is especially the case of OTHER-PURPOSE DEBTORS but also to some extent manifests itself within the group of OVERINDEBTED CONSUMERS. Their debt behaviour pattern is neither income motivated nor are they liquidity constrained, it does not depend on the number of household members. Transition to this group seems however to be constant during the working life time and only diminish close to retirement age.

However, this paper faces also some limitations. It is still unclear whether the discovered debt behaviour patterns are universal or only observed in Poland, which was subject to analysis. Second, Polish household credit market is still affected by limited availability of credit in the first years of the transition to the market economy. Thus many households did not have a chance to obtain credit during the period, which due to strong persistence of credit might have effects even now.

24

References

Bertaut, C. C., Haliassos, M., & Reiter, M. (2009). Credit Card Debt Puzzles and Debt Revolvers for Self Control. *Review of Finance*, *13*(January), 657–692. http://doi.org/10.1093/rof/rfn033

Bertola, G., & Hochguertel, S. (2007). Household Debt and Credit : Economic Issues and Data Problems. *Economic Notes*, *36*(2), 115–146. http://doi.org/10.1111/j.1468-0300.2007.00181.x

Białowolski, P. (2014). Patterns of debt possession among households in Poland – a multi-group latent class approach. *Bank & Credit*, *45*(2), 79–104. Retrieved from http://www.bankikredyt.nbp.pl/content/2014/02/bik_02_2014_01_art.pdf

Białowolski, P., & Dudek, S. (2014). *Sytuacja na rynku consumer finance - I kwartał 2014*. Gdańsk.

Browning, M. (1989). The Intertemporal Allocation of Expenditure o Non-durables, Services, and Durables. *Canadian Journal of Economics*, *22*(1), 22–36.

Campbell, J. Y., & Cochrane, J. H. (1999). By Force of Habit: A Consumption-Based Explanation of Aggregate Stock Market Behavior. *Journal of Political Economy*, *107*(2), 205. http://doi.org/10.1086/250059

Carroll, C. D., Overland, J., & Weil, D. N. (2000). Saving and Growth with Habit Formation. *The American Economic Review*, *57*(3), 415–426. http://doi.org/10.1257/jep.6.3.79

Chien, Y.-W., & Devaney, S. a. (2001). The Effects of Credit Attitude and Socioeconomic Factors on Credit Card and Installment Debt. *Journal of Consumer Affairs*, *35*(1), 162–179. http://doi.org/10.1111/j.1745-6606.2001.tb00107.x

Crook, J. (2006). The Demand and Supply of Household Debt: A Cross Country Comparison. In G. Bertola, R. Disney, & C. Grant (Eds.), *The Economics of Consumer Credit*. Cambridge, Massachusetts: The MIT Press.

Czapiński, J., & Panek, T. (2011). Social Diagnosis.

Davidson, W. R., & Bates, A. D. (2002). The retail life cycle. In A. M. Findley & L. Sparks (Eds.), *Retailing: The evolution and development of retailing* (pp. 164–277). Routledge, Taylor & Francis Group.

Dynan, K. E. (2000). Habit Formation in Consumer Preferences: Evidence

25

from Panel Data. *The American Economic Review*, *90*(3), 391–406. Retrieved from http://www.jstor.org/stable/117335

Friedman, M. (1957). *A Theory of the Consumption Function*. Princeton, NJ.

Gunnarsson, J., & Wahlund, R. (1997). Household financial strategies in Sweden: An exploratory study. *Journal of Economic Psychology*, *18*(2-3), 201–233. http://doi.org/10.1016/S0167-4870(97)00005-6

Hall, R. E., & Mishkin, F. S. (1982). The Sensitivity of Consumption to Transitory Income : Estimates from Panel Data on Households. *Econometrica*, *50*(2), 461–481.

Jappelli, T., & Pagano, M. (1989). Consumption and Capital Market Imperfections : An International Comparison. *The American Economic Review*, *79*(5), 1088–1105.

Kahneman, D. (2011). *Thinking, fast and slow*.

Kamleitner, B., Hoelzl, E., & Kirchler, E. (2012). Credit use: psychological perspectives on a multifaceted phenomenon. *International Journal of Psychology : Journal International de Psychologie*, *47*(1), 1–27. http://doi.org/10.1080/00207594.2011.628674

Kamleitner, B., & Kirchler, E. (2007). Consumer credit use: a process model and literature review. *Revue Européenne de Psychologie Appliquée/European Review of Applied Psychology*, *57*(4), 267–283. http://doi.org/10.1016/j.erap.2006.09.003

Keynes, J. M. (1936). *The General Theory of Employment, Interest, and Money*. Palgrave Macmillan. Retrieved from http://cas.umkc.edu/economics/people/facultyPages/kregel/courses/econ645/Winter2011/GeneralTheory.pdf

Kirchler, E., Hoelzl, E., & Kamleitner, B. (2008). Spending and credit use in the private household. *The Journal of Socio-Economics*, *37*(2), 519–532. http://doi.org/10.1016/j.socec.2006.12.038

Lukočienė, O., Varriale, R., & Vermunt, J. K. (2010). The Simultaneous Decision(s) About the Number of Lower- and Higher-Level Classes in Multilevel Latent Class Analysis. *Sociological Methodology*, *40*(1), 247–283. http://doi.org/10.1111/j.1467-9531.2010.01231.x

Mankiw, N. G. (1985). Consumer Durables and the Real Interest Rate. *The Review of Economics and Statistics*.

McCutcheon, A. L. (2002). Basic Concepts and Procedures in Single and Multiple-Group Latent Class Analysis. In J. A. Hagenaars & A. L. McCutcheon (Eds.), *Applied Latent Class Analysis* (pp. 56–85). New York, NY: Cambridge University Press.

26

Modigliani, F., & Brumberg, R. (1954). Utility Analysis and the Consumption Function: An Interpretation of the Cross-Section Data. In K. Kurihara (Ed.), *Post-Keynesion Economics*. Rutgers University Press.

Muthén, B. O. (2004). *Mplus Technical Appendices*. Los Angeles: Muthen&Muthen.

Muthén, B. O., Shedden, K., & Spisic, D. (1999). *General latent variable mixture modelling.Technical report*.

Ngwenya, M. a., & Paas, L. J. (2012). Lifecycle effects on consumer financial product portfolios in South Africa: An exploratory analysis of four ethnic groups. *Journal of Economic Psychology*, *33*(1), 8–18. http://doi.org/10.1016/j.joep.2011.09.008

Paas, L. J., Bijmolt, T. H. a., & Vermunt, J. K. (2007). Acquisition patterns of financial products: A longitudinal investigation. *Journal of Economic Psychology*, *28*(2), 229–241. http://doi.org/10.1016/j.joep.2006.06.006

Parker, J. A., & Preston, B. (2005). Precautionary Saving and Consumption Fluctuations. *The American Economic Review*, *95*(4), 1119–1143. Retrieved from http://www.jstor.org/stable/4132708 Accessed:

Pyykkö, E. (2011). Trends in European Household Credit. Solid or shaky ground for regulatory changes ? *ECRI Commentary No. 7*.

Schwarz, G. (1978). Estimating the Dimensions of a Model. *The Annals of Statistics*, *6*(2), 461–464.

Thaler, R. H., & Shefrin, H. M. (1981). An Economic Theory of Self-Control. *Journal of Political Economy*, *89*(2), 392–406. http://doi.org/10.1086/260971

Vermunt, J. K., & Magidson, J. (2002). Latent Class Cluster Analysis. In J. A. Hagenaars & A. L. McCutcheon (Eds.), *Applied Latent Class Analysis* (pp. 89–106). Cambridge University Press.

Viaud, J., & Roland-Lévy, C. (2000). A positional and representational analysis of consumption. Households when facing debt and credit. *Journal of Economic Psychology*, *21*(4), 411–432. http://doi.org/10.1016/S0167-4870(00)00011-8

Zeldes, S. P. (1989). Optimal Consumption with Stochastic Income : Deviations from Certainty Equivalence. *The Quarterly Journal of Economics*, *104*(2), 275–298. Retrieved from http://www.jstor.org/stable/2937848

27

Appendix 1.   Latent class model BICs for the periods of analysis.

No. of latent classes	BIC for unrestricted model		
	2011	2013	2015
2 classes	76549.666	75154.144	64020.901
3 classes	74588.002	72179.615	61457.201
4 classes	73199.089	70580.761	60192.197
5 classes	72343.302	69800.102	59746.254
6 classes	71686.767	69372.04	59605.198
7 classes	71205.349	69206.536	59275.087
8 classes	70948.634	68832.209	59052.865
9 classes	70692.260	**68620.199**	**58961.783**
10 classes	**70471.353**	68746.679	58975.377
11 classes	70474.126	---	---

Source: Calculations in Mplus 6.1.

Think Fast, Feel Fine, Live Long: A 29-Year Study of Cognition, Health, and Survival in Middle-Aged and Older Adults

Stephen Aichele, Patrick Rabbitt, and Paolo Ghisletta

Abstract In a 29-year study of 6203 individuals with ages ranging 41–96 years at initial assessment, we evaluated the relative and combined influence of 65 mortality risk factors—which included socio-demographic variables, lifestyle attributes, medical indices, and multiple cognitive abilities. Reductions in mortality risk were most associated with higher self-rated health, being female, fewer years smoking, and smaller life span decrements in processing speed. Thus, two psychological variables—subjective health status and processing speed—were among the top survival predictors. We suggest that these psychological attributes, unlike more narrowly defined risk factors, are indicative of (and influenced by) a broad range of health-related behaviors and characteristics. Information about these attributes can be obtained with relatively little effort or cost and—given the tractability of these measures in different cultural contexts—may prove expedient for prevention, diagnosis, and treatment of conditions related to increased mortality risk in diverse human populations.

Stephen Aichele · Patrick Rabbitt · Paolo Ghisletta
Faculty of Psychology and Education Sciences, University of Geneva, Switzerland
e-mail: stephen.aichele@unige.ch

A phase-type model of cohabiting union duration

Jean-Marie Le Goff

Centre for Life Courses and Inequality Studies (Lines), University of Lausanne, and Swiss National Center of Competencies in Research Overcoming Vulnerability: Life Course Perspectives (Lives), Switzerland

Abstract   We propose a phase-type model to analyze the duration of non-conjugal cohabiting unions. This model is a compartment model with two competing events: the marriage with the partner and the separation of the couple. We suppose that a non-marital union can be characterized by two hidden phases. The first begins at the start of the union and at each moment, people have the possibility to move from this first phase to the second phase during which hazard rates of marriage and of separation differ from the first phase. Investigations on data from the British 1958 National Child Study and the 1970 British National Study show that the proposed model fits well with data about the first cohabiting union of interviewed people. Results show that processes of marriage and separation differed between the two cohorts.

1 Introduction

Our aim in this paper is to investigate cohabiting union duration with a phase-type model (Aalen, 1995; Aalen & Gjessing, 2001; Lindqvist, 2013; Lindqvist & Amundrustad, 1998). Phase-type models are duration models with two peculiarities: first, they are compartment models in which one or several states of a studied system are hidden, i.e. non-observed; second, transition hazard rates from one state to another, hidden or not, are considered to be constant (Cox, 1962; Aalen, 1995). Phase-type models are much developed in reliability theory as well as in health sciences, for example to analyze the process of degradation of a machine or

the progression of an illness (Aalen et al, 2008; Lindqvist & Amundrustad, 1998). They are potentially interesting in life course research in which notions of phases and transitions between phases are important (Levy & the Pavie team, 2005).

The model we propose to analyze cohabiting union duration is a competing risk model with two kinds of cohabitation termination: the marriage of the couple or its separation. In the next section (section 2), we sketch the interest to develop an approach of cohabiting unions with a phase-type model. In the following section (section 3), the proposed model is presented. This section is followed by a brief presentation of data we used in this paper; first, the 1958 National Child Study, and second, the 1970 British National Study (section 4). In section 5, we present and discuss our results and finally we offer a conclusion.

2 Marriage and cohabiting union formation as composed of different hidden phases

Compartment models with hidden phases are scarce in the literature of family demography. One exception is the model of first marriage that was proposed by Coale and McNeil in the early 1970s (Coale & McNeil, 1972). In this model, it is considered that people progress in a succession of different social states before the marriage: access to the marriage market at the end of adolescence, period of partner search, dating, and engagement (figure 1). All these intermediary states, in the model of Coale and Mac-Neil, present the peculiarity of being "hidden" because of their difficulty to be delineated in data collection (Coale & McNeil, 1972; Coale, 1977). This model can be considered as a kind of Erlang process (Cox, 1962) in the sense that marriage can occur only if persons cross all of these hidden different stages before being married.
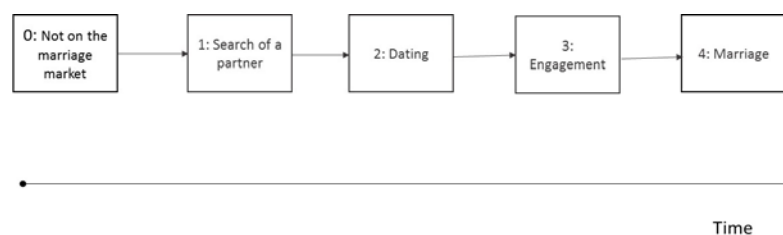
Fig. 1: The Coale and McNeil model of marriage as a compartment model

According to Coale and McNeil (1972), their model is very well suited to marriage distribution in developed countries from the Second World War to the beginning of the seventies. However, in a lot of these countries, the process of marriage and union formation transformed strongly starting in the seventies with the apparition and diffusion of extra-marital unions. Several authors proposed a sche-

ma-type of evolution of links between marriage and cohabitations in which each of these two kinds of union changed of their meaning (Villeneuve-Gokalp, 1990; Toulemon, 1997; Manting, 1996). In a first period, some precursors adopted cohabitation as an alternative to the marriage. These precursors originated from the contestation milieu in youth during the end of sixties and the beginning of seventies. They envisaged extra-marital cohabitation as an alternative to bourgeois marriage (Manting, 1996). Often, these precursors were students, and lot of them, starting from the end of seventies, became managers or exercised intellectual professions (Villeneuve-Gokalp, 1990). It is from this milieu that cohabitating unions diffused to middle and lower classes of societies. The meaning of cohabitation in the life course changed during this period of diffusion. It became a prelude to the marriage; couples experimented with living together before eventually deciding to marry. In this perspective, marriage remains an engagement to form a family. Starting from the seventies and eighties, several countries experimented with an increase in extra-marital births, which indicated that several couples no longer wished to marry when they wanted to have children. Cohabitation replaced marriage.

This approach, in which cohabitation is considered to have had a different succession of meanings according to its diffusion in a population, is often proposed by authors that are next to the theory of the second demographic transition (Lesthaeghe, 1995). According to this theory, demographic changes observed in developed countries since the end of the baby boom are related to the passage of an industrial to a post-industrial society. However, an alternative to this theoretical approach of cohabitation diffusing through society is present in the literature (Reiss & Lee, 1988; Perelli-Harris & Gerber, 2011). In this approach, the choice of cohabitation is constrained by economic reasons. Because marriage is expensive and because young people from lower classes often experiment with precarious jobs in the labor market, couples prefer to cohabit before marriage.

Whatever the characteristics of extra-marital unions in a country or in a social group (chosen or constrained, preceding a marriage or alternative to it), its emergence and its diffusion means a change in the process of marriage as a succession of hidden stages, as it was initially proposed in the Coale and McNeil model. In a country or in a social group in which the extra-marital union is a norm accepted by a large segment of the population, one can consider that the phase of entry on the union market is a phase that occurs before the cohabitation, as well as the phase of dating, while engagement seems to form a stage that occurs after the entry into cohabitation. This phase could have split into several phases. In this paper, we hypothesize that there are two hidden phases: a first phase corresponding to a period of trial that precedes the second phase, which can be considered as a "true" period of engagement. During this second phase of engagement, couples decide to marry, if the cohabiting union is a prelude to marriage, or remain living in cohabitation, if cohabiting union has the meaning of an alternative to the marriage. This conception of cohabiting unions into two hidden phases is the base of the model we propose to develop.

4

3 A compartment model for analyzing cohabitation duration

Manting (1996) and Toulemon (1997) envisaged that the hazard rate of marriage of a cohabiting couple was related to the meaning of cohabitation. In this paper, we will enlarge this point of view by: 1) taking into account that a cohabiting union can lead not only to a marriage but to a separation; and 2) supposing that the shape of the distribution of marriage and separation across the time is related to the social meaning of a cohabiting union.

We propose to develop a compartment model in which a cohabiting union is separated into two hidden phases. From each of these two phases, there is the possibility to end this cohabiting union with a marriage or a separation (figure 2). There are three possible events from the first phase of the cohabiting union: marriage, separation, and transition to the second phase of cohabiting union. From the second phase, there is the possibility to experiment a marriage.

....
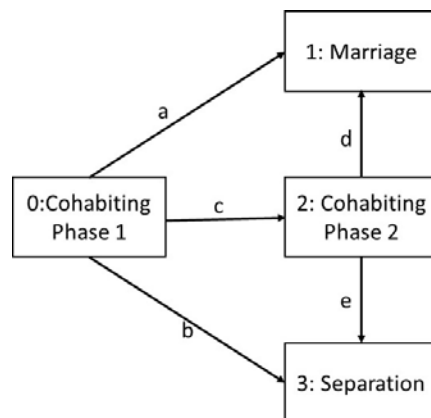
Fig. 2: phase-type model of cohabitation duration

The difficulty with such a model is that durations in the first phase and in the second phase are, by definition, non-measured. We do not have any information that allows us to know when a person is in the first phase or in the second phase of a cohabiting union. The only information we can generally collect in demographic surveys is the duration of each cohabiting union of respondents and how this unions ended (if it ended), with a marriage or a separation. By hypothesis, we will consider that each transition rate is time constant across the time, which means that the process is Markovian. Transition hazard rates of marriage from the first and the second phases will be denoted *a* and *d*, those of separation, *b* and *e*, while the transition rate from the first to the second phase will be denoted *c* (figure 2). As there is the possibility to experiment with both events from each stage of the cohabiting union, the proposed model belongs to the family of the Coxian models (Aalen, 1995).

A matrix approach allows for estimating the cumulative distribution functions (CDF) of transition to marriage and separation (Aalen, 1995; Lindqvist & Amundrustad, 1999; Lindqvist, 2013). Suppose a transition Markovian time infinitesimal matrix, in which each line i and each column j represent the different states possible. In our case, this transition matrix M can be represented by:

$$M = \begin{bmatrix} -a-b-c & c & a & b \\ 0 & -d-e & d & e \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

0 means that there is no transition possible from the state i to the state j, or on the diagonal of the matrix that any individual can leave the state i. It has been shown that the exponential of a phase-type model matrix times t gives the set of cumulative distribution function from one phase to another phase (Neuts, 1981). In the present case:

$$\exp(tM) = \begin{bmatrix} P_{00}(t) & P_{01}(t) & P_{02}(t) & P_{03}(t) \\ 0 & P_{11}(t) & P_{12}(t) & P_{13}(t) \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

where $P_{ij}(t)$, $i{\neq}j$, represents the cumulative probability to have transited from phase i to phase j at time t, while $P_{ii}(t)$ represents the survival probability in phase i at time t. In the present case, the cumulative distribution functions $P_{02}(t)$) and $P_{03}(t)$ are of most interest for us, as they represent the cumulative probabilities at time t from the beginning of the union at time t_0 for, respectively, the marriage and separation. We used possibilities of symbolic computations implemented in the software Mathematica to develop formulas of each cumulative distribution function starting from the matrix M (Aalen, 1995; Wolfram Alpha, 2016). The basic hypothesis of a constant hazard rate of transition from state i to state j can be considered very simplistic, as it means that a transition is without memory of the past, at least of the duration of the first phase of the cohabiting union. Even if the starting matrix is simple, formulas of cumulative distribution of marriage $P_{02}(t)$) and separation $P_{03}(t)$ appear to be quite tedious[1]:

---

[1] As an alternative, cumulative distributions of marriage and separation can be written with a matrix equation (Lindqvist, 2013). This alternative form presents the advantage to be flexible in a sense that a large family of phase-type models can be written by this way. However, the process of estimation of such general models in which there is the necessity to compute numerically the matrix expo-

6

$$P_{02}(t) = -\exp(-(d+e)t) * c * \frac{d}{((a+b+c-d-e)*(d+e))} + \frac{a*d+c*d+a*h}{((a+b+c)*(d+e))}$$

$$+ \exp(-(a+b+c)t) * \left(\frac{-(a+c)*d-a*e}{((a+b+c)*(d+e))} + c * \frac{d}{((a+b+c-d-e)*(d+e))} \right)$$

$$P_{03}(t) = -\exp(-(d+e)t) * c * \frac{e}{((a+b+c-d-e)*(d+e))} + \frac{b*d+c*e+b*h}{((a+b+c)*(d+e))}$$

$$+ \exp(-(a+b+c)t) * \left(\frac{-(d+e)*b-c*e}{((a+b+c)*(d+e))} + c * \frac{e}{((a+b+c-d-e)*(d+e))} \right)$$

The sojourn function in cohabiting union $S(t)$ (in phase 1 or in phase 2) can then be computed by:

$$S(t) = 1 - P_{02}(t) - P_{03}(t)$$

Despite the complexity of formulas, such an approach with two hidden phases during the cohabitation allows envisaging different meanings of cohabiting unions, according to the level of initial hazard rates. We propose here four simulated and non-exhaustive scenarios, each of them differing by their shape of cumulative distribution of marriage and separation (table 1 and figure 3).

In the first scenario, hazard rates of marriage and separations are high compared to the transition to the second phase (hazard rate of marriage being higher in this example than the one of separation), while they become low when couples enter the second phase of cohabiting unions (figure 3, type 1). Such a set of hypotheses about the different hazard rates would correspond to a type of cohabitating union in which remaining in this union after a trial period corresponding to the first phase of cohabiting unions is scarce. In this case, the model is very similar to an exponential model with two competing events (Aalen et al., 2008). Cumulative distributions of marriage and cohabitation increase very quickly at the beginning of the process. This increase is much slower in the scenario of a higher hazard rate of hidden transition from the first phase of cohabiting union to the second one and the same level of hazard rates of marriage and separation from the first and second phases of cohabiting unions (figure 3, type 2). Such a scenario would correspond to a model of cohabiting unions as an alternative to the marriage in which a lot of couples after a trial period decide to remain cohabitant (transition to the second phase), marriages and separations becoming rare. In this case, processes of marriage and separation are slowed down by the fact that many couples enter the second phase of cohabiting unions. Processes are not terminated at the end of the period of observation—because of marriages and separations that occur when couples are in the second phase of cohabiting union—even if hazard rates are low

nential is very long at the opposite of the estimations of the model as it is expressed by this equation.

7

during this second phase. A third scenario is opposed to the first one (figure 3, type 3). It corresponds to a model in which there is a trial period during the first phase of cohabiting union in which marriages as well as separations are scarce. After this trial period, there is a transition to the second phase of cohabiting union, in which hazard rates of marriage and separations become high. In this case, during the first period in which a majority of couples are yet in the first phase of cohabiting union, the cumulative distribution for each event increases slowly and then increase more quickly as soon as couples enter the second phase of cohabiting unions. The fourth type corresponds to a scenario in which the first phase of cohabiting unions correspond to a trial period, during which a lot of couples separate (figure 3, type 4). This trial period ends when couples enter the second phase of cohabiting union from which a lot of couples marry. In such a scenario, the cumulative distribution of separation is higher than the one of the marriage while a majority of couples are in the first phase of cohabitation. But when the remaining majority of cohabiting couples have transited to the second phase of the cohabiting union, the cumulated proportion of marriage can become higher than the one of separation.
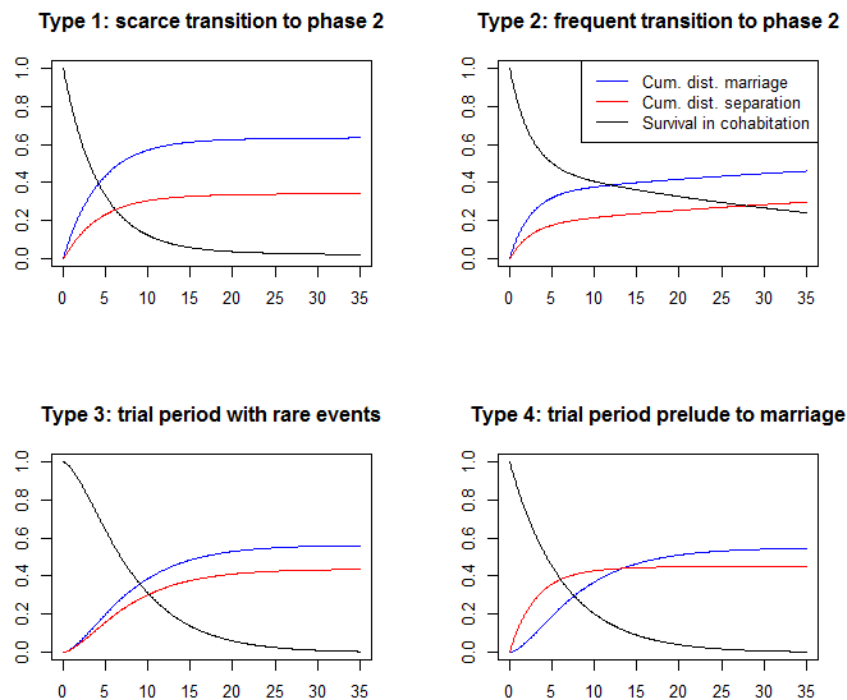
Fig. 3: Different scenarios of the role of cohabiting unions

Table 1: Different scenarios of hazard rates of the model of cohabitation

	Type 1	Type 2	Type 3	Type 4
a	0.15	0.15	0.01	0.01
b	0.08	0.08	0.01	0.15
c	0.01	0.20	0.20	0.20
d	0.01	0.01	0.20	0.15
e	0.01	0.01	0.15	0.01

Other scenarios can be envisaged, such as the opposite of the fourth type in which the hazard rates of marriage and separations are respectively high and low during the first phase of the cohabiting union, and the inverse during the second phase. However, all scenarios show that the rate of hidden transitions from the first phase to the second one plays a major role in the shape of cumulative distributions of marriage and separation.

We then draw several ideal types of cohabiting unions with two hidden phases. It becomes interesting to test the fit of the model on real data. Which scenario of cohabiting union corresponds to reality?

4 Data and model estimation

We estimated our model on first cohabiting union duration in Great Britain. Data we used come from two very similar survey databases: the National Child Development Study (NCDS) and the 1970 British Cohort Study (BCS70). The NCDS is a panel survey that interviews more or less regularly men and women (N=circa 17 000) that were born on a specific week of March 1958 (CLS, 2015). Several topics have been developed in questionnaires of the different waves, often in relation to the phase of the life course of respondents: early development, childhood, health transition to adulthood, professional career, and unions and marriage. The BCS70 follows persons born during a specific week of 1970 with aims very similar to the ones of the NCDS. These two surveys are managed by the Center for Longitudinal Studies (CLS) in London. Data of these two surveys are downloadable from the website of UK data service, the data archive center in the UK (UK data service, 2015).

Among all data files available, we specifically use for each survey a file on the history of unions that was built by the CLS team (Hancock, 2011a, 2011b). In these data files, union spells of each respondent were recorded with information about the type of union at its beginning (marriage or cohabiting union) and, in the case of a cohabiting union, the outcome of this union (marriage, separation, or remaining cohabitant at the last interview). We then selected all first unions that began with a cohabiting union. We computed durations from the beginning of the union to marriage or to separation for respondents that end their union with one of these events. For those who did not experiment with any of these events during

their observation, censored duration taken into account was the time between the beginning of the union and the last interview. We also censored the rare cases of death of the partner during the cohabiting union at the moment of the death. Durations were computed originally in months, but we converted them to years. Distributions of events and censoring in each cohort for men and women are reported in table 2.

Table 2: Events and censoring in each survey (in %)

	Marriage	Separation	Censoring	N
Cohort 58 men	67.06	7.02	25.92	2878
Cohort 58 women	68.72	5.58	25.70	2436
Cohort 70 men	50.37	32.96	16.67	4648
Cohort 70 women	54.59	32.10	13.31	4825

For the sake of simplicity, we will identify these two databases as the 1958 cohort and the 1970 cohort. To be born in 1958 means that those who experiment with a cohabitation union in the seventies were akin to early adopters. Forty-five percent of first unions in this cohort were cohabiting unions at their beginning, without significant differences between men and women. Other unions were direct marriages and are not taken into account in our investigations. This proportion reaches 85% in the 1970 cohort, which means that the cohabiting union became common for people of this cohort. We can then expect that the meaning of a cohabiting union will differ between the two cohorts. In the case of cohort 58, we suppose that the cohabiting union could have signified a prelude to marriage or an alternative to the marriage. In the case of the cohort 1970, it could have signified a trial period.

For each survey, we estimated the five parameters by the likelihood maximization method. In the case of a single event, each individual contributes to the likelihood by the density $d(t)$ (which is the derivative of the cumulated distribution function of the event) if she experimented with the event and by the survival function $S(t)$ if she did not. In the present case of competing events, the likelihood equation takes into account the density of the distribution of each event $d_{02}(t)$ and $d_{03}(t)$. These two densities are the derivatives of respectively $P_{02}(t)$ and $P_{03}(t)$. Despite the complexity of formulas for cumulative distributions, their derivatives are not complicated to compute. The likelihood equation can then be written as:

$$L_{(a,b,c,d,e)} = \prod_{i:\delta_i=2}^{n} d_{02}(T_i) \prod_{i:\delta_i=3}^{n} d_{03}(T_i) \prod_{i:\delta_i=0}^{n} S(T_i)$$

where n represent the total of individuals; $\delta_i = 2$ if the individual is married, $= 3$, if he is separated, and $= 0$ if the duration is truncated. T_i represents the duration associated with an individual. As usual, this is in fact the logarithm of this equation that is maximized. In order to avoid estimating negative hazard rates, we in fact will estimate logarithms of each hazard rate.

10

5 Results

Parameters were estimated in each cohort for men and women. We first estimated in each sub-population Aalen-Johansen estimators of the cumulative distribution of marriage and separation and survival function in cohabiting unions for men and women of each cohort (figure 3). Aalen-Johansen estimation is a non-parametric method that generalizes the Kaplan-Meier estimator in cases of competing event and multi-state models (Aalen & Johansen, 1978). It computes the cumulative distribution for each transition and the survival function in each "visible" state. In the present case, the cumulative distribution functions of marriage and separation as well as the "survival" function in a cohabiting union are estimated. The package *etm* in R was used to estimate these non-parametric estimations (Alignol et al., 2011). Results show that separations are rare in the 1958 cohort for men as well for women, while the marriage is more important for both sexes. A large part of cohabitations remain in this kind of union. In the case of the 1970 cohort, cumulative distributions of these two event were more balanced, even if marriages remained more important than separations. Remaining in cohabiting unions is less important in proportion at the end of the window of observation than in the case of the 1958 cohort.

Afterwards, we estimate our phase-type model of cohabiting unions on each sample. We used the command *mle* in the library *stats4* in R to make likelihood estimations of the five parameters of the model. Results show very good fits of the model on each sample, which gives good reasons to validate the hypothesis of a cohabiting union divided into two phases (figure 4).

A view of estimated coefficients allows for understanding processes of marriage and separations (table 3). In the case of the 1958 cohort, the hazard rate of transition from the first phase of cohabiting union to marriage and to the second phase are high, while the transition to separation is rare. When persons are in the second phase, marriages become rare while separations remain scarce. Such results correspond well to the second scenario that we proposed (figure 3 and table 2), except that during the first phase, separations are rather rare. Cohabiting unions seem to have two roles: first a prelude to the marriage and, second, an alternative to the marriage. Processes are very similar between men and women.

In the case of the 1970 cohort, transitions to marriage and even to separation are high, while the transition to the second phase became scarcer, in comparison to the 1958 cohort (table 3). When people are in this second phase, marriages and separations are rare. It corresponds more to the first ideal type of cohabiting union that we proposed (figure 3 and table 2). As in the first cohort, results show two types of cohabitations. The first one is a trial period before the marriage, in which couple can marry or break. The second cohabiting union, corresponding to the second phase, is an alternative to the marriage. As in the 1958 cohort, there are no differences in behaviors between men and women.
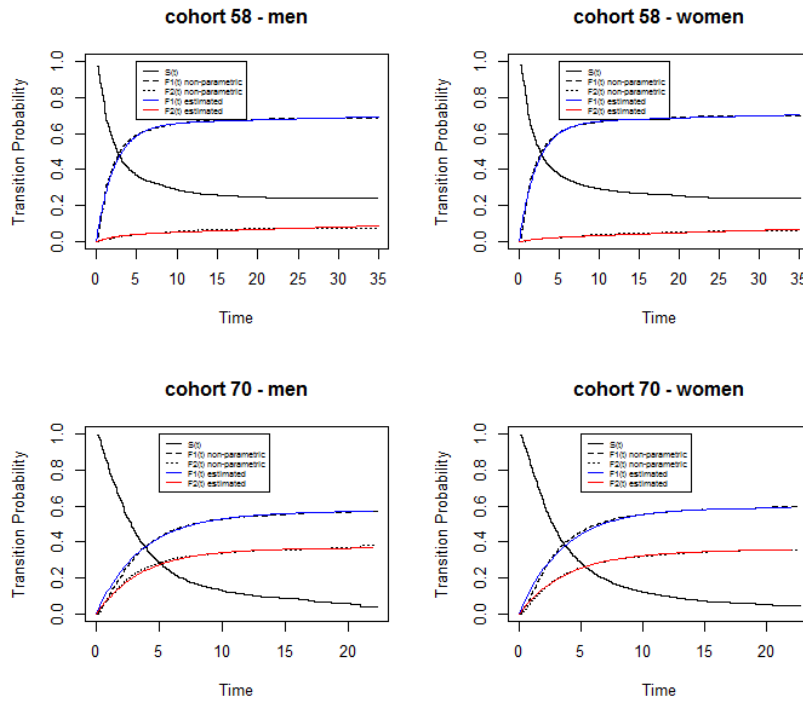
11

Fig. 4: Fit of estimated coefficient

Table 3: Estimated coefficients and hazard rate of each transition

			Cohort 58			
	Men				Women	
log(hazard)	Estimations	Std.error	hazard rate	log(hazard)	Estimations	Std.error	hazard
a	-1,23	0,03	0,293	a	-1,18	0,03	0,308
b	-4,01	0,11	0,018	b	-4,58	0,17	0,010
c	-1,98	0,05	0,138	c	-1,92	0,05	0,146
d	-5,35	0,20	0,005	d	-5,37	0,18	0,005
e	-5,19	0,13	0,006	e	-5,20	0,12	0,006
			Cohort 70			
log(hazard)	Estimations	Std.error	hazard rate	log(hazard)	Estimations	Std.error	hazard
a	-1,85	0,02	0,157	a	-1,80	0,02	0,166
b	-2,27	0,03	0,103	b	-2,37	0,03	0,094
c	-3,62	0,18	0,027	c	-3,77	0,18	0,023
d	-4,10	0,55	0,017	d	-5,27	1,26	0,005
e	-4,86	0,65	0,008	e	-3,80	0,32	0,022

6 Conclusion

A phase-type approach of cohabiting union duration allowed us to propose some scenarios of links between distribution of union duration and the social meaning of these unions. The estimation of the model on real data permitted to understand the evolution of this kind of union in the British context. At the opposite of a more traditional approach in event history or survival analysis in social sciences, especially in life course research, we emphasized our purpose on the shape of the hazard rate while we neglected the effect of individual characteristics on the hazard rate, as it is often proposed in the literature *via* the hypothesis of the proportionality of the hazard rate, such as in the Cox model. Expectations about effects of individual characteristics on each hazard rate, however, also can be developed in a phase-type model by hypothesizing, for example, that each hazard rate depends on these characteristics, as well as an estimating coefficient associated with each of these characteristic. The classical proportionality assumption can be called upon for estimation of the effect of characteristics on the different hazard rates. Note that, in the present case, we just further tested our model, not only in distinguishing men and women, but also by differentiating them by their age at the beginning of the union or by the social class (observed by the profession of the father or the mother of the respondent). Results obtained did not reveal significant differences between categories.

In the first section of this paper, we indicated in our review of the Coale and McNeil model that phases crossed by persons during the process leading to marriage were difficult to delineate in a quantitative survey. It is also the same in the case of the two phases we envisaged in our example of cohabitating union. The notion of hidden or unobserved phases, however, requires some verification in order to avoid mistakes. At best, the development of a phase-type model and its fit to duration quantitative data should be accompanied by other investigations, such as qualitative interviews, in order to verify the existence and the meaning of the hidden phases.

13

References

Aalen O.A. (1995). Phase type distribution in survival analysis. *Scandinavian Journal of statistics*. 2284):447-463.

Aalen A. O., Borgan H. & Gjessing H. (2008). *Survival and event history analysis*. A process point of view. New-York Berlin: Springer.

Aalen A. O. & Gjessing H. (2001). Understanding the shape of the hazard rate. A process point of view. *Statistical Science*, 16(1): 1-14.

Aalen O.O. & Johansen S. (1978). An empirical transition matrix for nonhomogeneous Markov chains based on censored observations. *Scandinavian Journal of Statistics*. 5: 141–15.

Alignol A., Schumacher M. & Beyersmann J. (2011). Empirical Transition Matrix of Multi-State Models: The etm Package. *Journal of Statistical Software*. 38(4): 1-15.

CLS (2015). Center for Longitudinal Studies. Home of the 1958, 1970, Millennium and Next Steps Cohort Studies. http://www.cls.ioe.ac.uk/Default.aspx. Retrieved 22/12/2015.

Coale A.J. (1977). The development of new models of nuptiality and fertility. *Population*, 32(1): 131-154.

Coale, A. J. and McNeil. D. R. (1972). The Distribution by Age of the Frequency of First. Marriage in Female Cohort. *Journal of American Statistical Association*. 67:743-749

Cox D.R (1962). *Renewal theory*. London: Methuen; New York: J. Wiley.

Hancock M. (2011a). *National Child Development Study. Partnership Histories 1970-2008, A Guide to the datasets*. London: Centre for longitudinal Studies.

Hancock M. (2011b). *British National Cohort (1970). Partnership Histories 1986-2008, A Guide to the datasets*. London: Centre for longitudinal Studies.

Lesthaeghe, R. (1995). The second demographic transition in Western countries: An interpretation. In: Mason, K.O. and Jensen, A.-M. (Eds.). *Gender and family change in industrialized countries*. Oxford: Clarendon Press: 17-62.

Levy R. and the Pavie team (2005). Why Look at Life Courses in an Interdisciplinary Perspective. In Levy R., Ghisletta P., Le Goff J.-M., Spini D. and Widmer E. (Eds). *Towards an Interdisciplinary Perspective on the Life Course* (10:3-32). Amsterdam-Boston: Elsevier. Coll Advances in Life Course Research.

Lindqvist, B.H. (2013), '*Phase-Type Distributions for Competing Risks*'. Paper. Special Topic Session on Statistical analysis of competing risks data. 59thISI World Statistics Congress. Hong Kong, China, 25-30 August 2013.

Lindqvist B. H. & Amundrustad H. (1998). Markov models for periodically tested components. *Proceedings of European and reliability conference*. Trondheim.

Manting D. (1996). The Changing *Meaning* of *Cohabitation* and Marriage. *European Sociological Review*, 12 (1): 53-65.

Neuts M. H. (1981). *Matrix geometric solutions in stochastic models. An algorithmic Approach*. Baltimore: John Hopkins University Press

Perelli-Harris B. and Gerber T. P. (2011) Nonmarital childbearing in Russia: second demographic transition or pattern of disadvantage? *Demography*, 48, (1), 317-342.

Reiss I. L., & Lee G. R. (1988). *Family systems* in America (4th Ed.). New York: Holt, Rinehart and Winston

Toulemon L. (1997). Cohabitation is here to stay. *Population: An English Selection*. 9: 11-46.

UK data services (2015). Home page. https://www.ukdataservice.ac.uk/. Retrieved 22/12/2015.

14

Villeneuve-Gokalp C. (1990). Du mariage aux unions sans papier. Histoire récente des transformations conjugales. *Population*. 45(2): 265-297.

Wolfram Alpha (2016). *Wolfram Alpha. Computational knowledge engine*. http://www.wolframalpha.com/. Retrieved 12/01/2016.

Author index