Comparing Qualitative Harmonic Analysis and Optimal Matching. An Exploratory Study of Occupational Trajectories

Nicolas Robette[*] and Nicolas Thibault[**]

Modelling the life course in terms of interactions between different areas of individual activity generates complex datasets which are difficult to analyse in quantitative terms. One way of addressing biographical complexity is to focus on certain landmark transitions. Another option is to describe the entire trajectory from start to finish. The shift from analysis of transitions to the typological description of entire trajectories is a very useful change of perspective, which leads from the particular to the general. This is the approach proposed here by *Nicolas ROBETTE* and *Nicolas THIBAULT*. Comparing qualitative harmonic analysis, born out of French tradition, and optimal matching developed in genetics, the authors make a timely evaluation of these two rapidly developing approaches.

Following the example of the *Triple biographie* survey conducted by INED in 1981, event history surveys are designed to collect retrospective data on complete individual trajectories, often on a yearly basis. These data provide a basis for describing the life course in statistical terms. The *Biographies et entourage* (event histories and contact circle) survey was carried out by INED in 2001 on a sample of 2,830 individuals born between 1930 and 1950, representative of the corresponding Paris region population at the time of the survey (Lelièvre and Vivier, 2001). Its main aim was to improve knowledge of respondents' residential and family mobility in relation to that of their contact circle. It also produced interesting data on respondents' occupations that warranted more in-depth analysis. The various occupations held by respondents over their life course were recorded on a retrospective annual timesheet. Each stage was then coded using the INSEE socio-occupational categories (SOC). Although occupational trajectories can, in theory, cover the entire range of possibles, they nonetheless exhibit strong trends that reflect concomitant changes in society (Marchand and Thélot, 1997).

The aim of this article is to determine the best way to process these trajectories, qualified as complex in the sense that (i) all the states which characterize them may recur over time and (ii) there

[*] Institut national d'études démographiques, Paris.
[**] At the time of writing, Nicolas Thibault worked at the Institut national d'études démographiques, Paris.
Translated by Catriona Dutreuilh

are numerous possible transitions[1] between these states. One option, which fits the paradigm of event history analysis (Courgeau and Lelièvre, 1996), is to use exploratory statistical methods to identify life-course typologies that capture individual trajectories in their entirety, and not simply in terms of the events they comprise (Billari, 2001). Various typological methods exist (Grelet, 2002): chi-square distance, Euclidian distance (Espinasse, 1993), composite indicators, etc. Two of them emerge in the literature as particularly suited to this type of data, i.e. capable of systematically describing the sequence of events and their duration. The first is qualitative harmonic analysis (QHA), a factor analysis method developed by French statisticians in the 1980s and which takes account of time. The second is optimal matching[2] (OM), a set of algorithmic techniques imported from the life sciences by American sociologists in the late 1980s. In this article, we will discuss the comparative advantages of the two techniques for analysing the occupational trajectories of male respondents in the *Biographies et entourage* survey (INED, 2001).

1. Exploratory analysis of occupational trajectories. To what purpose?

Processing complex trajectories

A range of standard statistical tools can be used to analyse time spent (survival) in a given state: non-parametric estimations are used to measure survival (Kaplan and Meier, 1958 ; Nelson, 1972 ; Aalen, 1978), and semi-parametric (Cox, 1972) or parametric[3] models are used to measure the impact of individual characteristics (which are also social properties) on survival in a particular state. These models are meaningful for durations clearly defined by unambiguous start and end dates. However, they are not designed to describe individual trajectories characterized by a complex sequence of changes of state, i.e. when the analysis concerns a succession of repeatable events with numerous possible transitions between states (GRAB, 2006), as is the case for occupational trajectories, for example. How, then, can we explore this type of data?

The problem of repeatable events

Modern longitudinal methods, which can be applied at the individual level, have already proved useful for studying labour market trajectories. Using historical data from the French

[1] For example, in our study there are 9 states and hence 81 possible transitions.

[2]

[3] Parametric duration models assume that survival in a given state obeys a determined law, which it itself time-dependent.

employment agency (Agence nationale pour l'emploi, ANPE), we can, for example, study the labour market reintegration of unemployed persons on the basis of individual characteristics (Degenne and Lebeaux, 1999), and in relation to the benefits and support received by each individual (Crépon, Gurgand and Dejemeppe, 2005). Duration models must be used because we are estimating the time between registering as unemployed in the ANPE database and returning to employment. This definition of the duration of unemployment serves to measure the effectiveness of employment policies. However, it disregards the fact that return to employment may be short-lived, and followed rapidly by a further period of unemployment. In other words, duration models cannot directly capture the repeatability of events.[4]

Likewise, when career interruptions are analysed, their duration is generally taken to be the difference between the start of the *n*th period of inactivity and the date of subsequent resumption of activity. However, when analysing the complexity of intermittent career trajectories, the analysis of a single event is not sufficient, since individuals may interrupt their career several times over a lifetime. A variety of methods have been suggested for dealing with repeated episodes. First, labour market exits can be classified by exit order. Birth order is traditionally used to study fertility, but unlike the order of births in a family, the order of labour market exits has no particular significance. In terms of the working career, what counts is their duration, or their aggregate duration (Desplanques and Saboulin, 1986; Lelièvre, 1987; Cambois and Lelièvre, 1988). All in all, this method is not entirely satisfactory as the duration of inactivity periods is specified in the same way as in the first definition (time elapsed between start of *n*th period of inactivity and subsequent resumption of activity). Second, the set of periods can be taken as a whole by considering people who have experienced one or more inactivity periods as a level of aggregation in a multi-level model (Courgeau, 2000). This avenue of research has yet to be explored and rendered operational.

The problem of multiple states and transitions

The problem of defining duration becomes even more acute when the state in question is not binary, but can be characterized by a variety of situations: membership of various socio-occupational categories (SOC), full-time or part-time work, unemployment, economic inactivity.

[4] The solution adopted by Gurgand, Crépon et Dejemeppe (2005) is to consider that the return to employment is not effective until a threshold period in employment has elapsed.

Demographers, for example, are interested in measuring the duration of activity and relating it to other events such as birth of children. They ask the classic question of how labour force participation, of women especially,and fertility interact. Biographical methods appear better suited for this purpose than cohort analysis, which is based on the assumption that events are independent. However, according to M. Kempeneers and É. Lelièvre (1991), they face three interlinked problems: defining the state whose duration are estimated and the exogenous variables, defining the population exposed to risk and defining the study interval. The choices of definition affect the robustness of the reasoning applied and make the conclusions dependent upon the initial, often implicit, assumptions. The authors therefore recommend a more descriptive analysis taking account of all periods of activity from age 15. This gives rise to a new type of study which not only records labour market exit dates, but also follows life event histories over the entire life course.

2. Building a typology of complex occupational trajectories

The trajectories studied

The statistical unit of classification is the individual trajectory: clusters are formed on the basis of ressemblance between trajectories. Certain choices must be made regarding (i) the study population; (2) the study interval and (iii) construction of the analysis variables.

Only men's working careers are analysed here (n = 1,341). Classification of the entire population is possible, but has two drawbacks. First, the list of states is different for men and women: military service affects men only. Second, when both sexes are handled together, the necessary simplication involved in classification[5] may sometimes mask the career differences between men and women

As the individuals leave the observation at the survey date, the data are right-censored[6]. With duration models it is possible to control the effect of truncation, but not the descriptive statistics. There is no technical reason to prevent qualitative harmonic analysis (Barbary and Pinzon Sarmiento, 1998) or optimal matching methods (Macindoe and Abbott, 2004) from being applied to trajectories of

[5] A description of trajectories based on the entire population nonetheless shows the similarities between certain male and female careers. But this does not affect our preference for treating men and women separately. If the careers have similarities, we will identify similar types, as is the case for higher-level occupations. If the trajectories are different, we avoid smoothing sex-related disparities.

[6] By construction, the surveys being retrospective, the survey date is the only source of right-censoring: individuals who are dead or who have migrated are not observed. The data collection was sufficiently exhaustive to prevent left-censoring, i.e. the presence of individuals whose life-event history is only known after a particular date.

varying lengths. However, by doing so, we move away from the description of a single population over time. The description of individual trajectories implies that all individuals are observed over an identical period, delimited by the same boundaries.

Our study, for example, concerns occupational mobility between the age of 14 – the legal school-leaving age for the cohorts studied – and age 50,[7] the age of the youngest respondents.[8] The analysis could have been extended beyond age 50 by working on a survey sub-population. Its usefulness would have been limited, however, as the number of respondents drops rapidly with age: only 67.4% of the population are still in the sample at age 55, 42.5% at age 60, 21.3% at age 65 and 4.6% at age 70.

We describe the various states that constitute the trajectories using the eight socio-occupational categories defined by INSEE. A more detailed breakdown of states would not be especially useful, since changing from a clerical worker to a sales worker does not have the same significance as moving up to a higher-level occupation. By construction, the sample does not include any retirees, as the description ends at age 50. We should thus have reasoned on the basis of seven SOCs, i.e. the six active categories and the category of "other economically inactive persons". However, we thought it would be useful to divide this latter category between students and persons who are inactive for family or health reasons.[9] A further category was needed to describe the trajectories of men who had experienced the Algerian war. We therefore added the "military conscript" category, based on the 1954 list of SOCs but which has since been removed.

2.2 Creating a typology by qualitative harmonic analysis

Harmonic analysis is a branch of mathematics widely applied in physics and biology. Its use in the social sciences is more recent, and dates back to the 1970s (Deville, 1974, 1977). The aim at that time was to use the notion of duration to explain social phenomena via data on individual trajectories. As pointed out by Deville (1977, p. 18): "Faced with such abundant data, the statistician

[7] Certain respondents reported being in employment from age 8, because they looked after animals on a farm for example, and only occasionally went to school. We will ignore these episodes in the analysis as respondents were asked to begin at age 14. This choice is consistent with the INSEE definitions of the SOCs, which only apply to persons aged over 15.

[8] We reason in terms of aged reached (difference in years) and not exact age. We know the respondents' exact age at the time of the survey but biographical events are recorded by age reached.

[9] It also includes unemployed persons who have never worked (but who are "active" under the INSEE definition), though such cases can be ignored as they concern only 3 stages out of the 19,930 stages recorded in the survey. This number is low because of the annual nature of data.

feels somewhat perplexed. Increasingly complex tables become uninterpretable without the help of "automatic" analysis methods. So he tries to define an analysis method to help him get the most from his data. Here, "the most" has a specific, quantifiable, meaning, linked to the method he applies". This technique was then adapted for the exploratory statistical analysis of complex trajectories (Deville et Saporta, 1980; Deville, 1982), and became known as qualitative harmonic analysis (QHA).

QHA involves defining an observation period, dividing it into a finite number of intervals then measuring for each individual the proportion of time spent in each of the states in each interval. A correspondence analysis on the matrix[10] thus formed provides a means to summarize the information by selecting the factors with the highest inertia (Deville, 1982). This eliminates statistical "noise" without eliminating the individual. Progress in classification techniques has made it possible, among other things, to use ascending hierarchical classification to determine trajectory typologies (Barbary, 1996; Degenne, Lebeaux and Mounier, 1996; Barbary and Pinzon Sarmiento, 1998). The typology thus constructed takes account not only of the sequence of states, but also the periods spent in the different states. It also keeps all individuals in the analysis.

QHA has rarely been used until now, due to a lack of suitable data. J.-C. Deville (1982), for example, created an ad-hoc sample on the basis of rejected questionnaires from the INSEE family surveys of 1962 to 1975.[11] The sample is therefore by no means representative and its sole purpose was to provide a field of application for a method which until then had none. Over a twenty-year period, the method was used only once on French data[12] (Degenne, Lebeaux and Mounier, 1996). This research focuses on a part of the occupational trajectory with the aim of studying labour market integration. In fact, it was not until the emergence of renewed interest in event history data collection that practical applications were developed, first on Latin American data (Dureau et al., 1994; Barbary, 1997; Barbary and Pinzon Sarmiento, 1998). The *Biographies et entourage* survey enables us to track

[10] With individuals in the rows and time spent in each state during each interval in the columns.

[11] The sample comprises women married three times or more and whose conjugal history was deemed too complex for analysis.

[12] Note that A. Degenne, M.-O. Lebeaux and L. Mounier (1996) also introduced a variable representing transitions between states into the matrix: the nine groups we identify would give rise to 81 transitions. In addition to the time spent in each state at each period, the number of occupational transitions of each type are recorded for each individual. Here, as our aim is to compare the two methods, we will limit discussion to qualitative harmonic analysis in the strict sense of the term.

migration trajectories using this method (Bonvalet, Bringé and Robette, 2008), but also entire working careers.[13]

Constructing the harmonic matrix

Once the study period has been defined, it must be divided into intervals for analysis. As the data are annual, a breakdown into one-year periods may appear to be the most natural choice, best suited to processing the event history information. This is not in fact the most efficient procedure as it produces a table with a majority of empty cells, adversely affecting the quality of analysis.[14] Conversely, if the number of intervals is too small, a valuable part of the available information will be lost. A compromise must thus be achieved when deciding the number of intervals.

Another necessary choice concerns the amplitude of intervals. These amplitudes do not have to be of equal length. On the contrary, certain periods in life, most often during youth, are punctuated by a large number of occupational changes, while other periods are much more stable. Choosing narrow intervals is problematic for periods of frequent change, since two sequences which are identical but whose timing is slightly different may be seen as very different. All in all, interval size must be matched to the overall pace of changes of state (Florette, 1988). Following a series of tests, we decided on a breakdown of ten intervals corresponding to the deciles of the distribution of occupational changes by age. [15]

For each individual, we calculate the proportion of the duration of each interval spent in each of the possible states.[16] We then perform a correspondence analysis on the matrix obtained, followed

[13] In 1989, the INSEE career and mobility survey included incomplete records of life event histories: respondents were asked about their current occupation on five different dates between 1960 and 1989. It thus provided a partial record, but did not follow individuals continuously throughout their working career. It was nonetheless used to produce a typology of occupational trajectories using the k-means method, on the basis of information on mobility between two successive dates (Goux, 1991). But as there is no information on the types of employment occupied (or inactivity) at fixed dates, the problem of capturing duration is not resolved.

[14] We observe 1,341 individuals who can experience 9 states over *p* periods. The study matrix will have 1,341 lines. If we choose an annual breakdown of the period between ages 14 and 50, the matrix will have 9 x 37 = 333 columns. By construction, the individual cannot experience more than one state per year. Each row thus contains 37 "ones" and 296 "zeros". Out of the 1,341 x 333 = 446,553 cells of the table, only one-ninth will be filled.

[15] We can, of course, choose the annual values closest to the theoretical deciles. The chosen deciles were thus 18, 20, 22, 24, 26, 29, 33, 38 and 43 years.

[16] We have ten intervals and nine states. We thus created 90 state variables. The final matrix comprises 1,341 rows and 90 columns.

by an ascending hierarchical classification based on the first 25 factors which carry 70% of the inertia. This reduces data heterogeneity with minimum information loss.[17]

2.3 Constructing a typology by optimal matching

The optimal matching method is based on a set of dynamic algorithms used mainly in molecular biology to analyse similarities between DNA sequences. It was introduced into the social sciences by Andrew Abbott in the 1980s (Abbott and Forrest, 1986; Abbott and Hrycak, 1990). It serves to measure the degree of similarity or dissimilarity between sequence pairs. This is done by evaluating the "cost" of transforming one sequence into the other. This transformation involves different operations: insertion (an element is inserted into the sequence), deletion (an element is removed) and replacement (an element is replaced by another). In practice, optimal matching is based on just two elementary operations. In the first operation, known as *indel* (a contraction of insertion and deletion), the same cost is assigned to both insertion and deletion, so that changing from sequence 1 to sequence 2 by inserting an element is equivalent to changing from sequence 2 to sequence 1 by deleting one.[18] The second operation, replacement, is a combination of an insertion and a deletion,[19] though the cost of a replacement does not have to be double that of an indel: the replacement cost is chosen by giving priority to reducing the distance either between sequences that are identical but which occur at different times, or to sequences that occur simultaneously but have one or more differences.

The cost of a series of operations is equivalent to the sum of costs of the corresponding elementary operations. The distance between two sequences is thus defined as the minimum cost of transforming one sequence into another. Specific dynamic algorithms ensure that the minimum cost is rapidly obtained (Sankoff and Kruskal, 1983). By matching all sequence pairs, a distance matrix is created which can then be used to group the most similar sequences, using classification methods for example, and obtain a typology. Choosing the costs of elementary operations is a key stage in optimal matching analysis. It is this freedom to determine costs which gives the method its flexibility and adaptability (Lesnard et Saint-Pol, 2004).

[17] The tests performed show that the typology classes obtained are relatively stable when 50-95% of inertia is conserved. Few individuals change class when the share of information is modified.
[18] For example, the difference between sequences ABA and AA can be eliminated either by removing B from ABA or adding B to AA.
[19] For example, transforming ABA into AAA involves deleting a B then inserting an A.

Constructing the cost matrix

We work on sequences of occupations held by respondents over annual periods. Replacement costs can be varied according to the elements replaced. Although certain researchers prefer to use fixed replacement costs, for lack of theoretical grounding on the question (Dijkstra and Taris, 1995), many studies adopted differentiated replacement costs based on hypotheses specific to the study topic: the more similar the elements, the lower the replacement cost. For example, in some studies of working careers, the replacement costs are based on the relative hierarchical positions of different occupations (Stovel et al., 1996; Halpin and Chan, 1998; Blair-Loy, 1999; Scherer, 2001; Solis and Billari, 2002). An alternative solution is to derive replacement costs from probabilities of transition between elements: the lower the transition probability, the higher the replacement cost (Rohwer and Pötter, 2005). In the absence of any preset hierarchy between socio-occupational categories, we chose the second solution.

We then focus on the relation between replacement cost and indel cost. According to the literature, several options are available to us. First, as a replacement is equivalent to a combined insertion and deletion, the indel cost can be set at half the replacement cost. We can then set the indel cost at a value slightly above half the maximum replacement cost, thereby avoiding the use of indels. This approach is justified when priority is given to the position of the elements in the sequence (i.e. the timing of events), rather than their respective order. If priority is given to the sequencing of elements, it is preferable to set the indel cost at 1/10 of the maximum replacement cost (Macindoe and Abbott, 2004). In a trajectory characterized by intra-generational mobility, the order of the various states is crucial. We therefore chose the second option (Appendix 1).

The coding of trajectories with each method is illustrated in Box 1.

Box 1

Representing trajectories with OM and QHA

Let us take 7 years of a career as an example (from age 14 to 20): at ages 14 and 15, the respondent is a student; at ages 16-18 he is a manual worker, and at ages 19-20 he holds a higher-level occupation. To simplify the presentation, we will limit the analysis to these three states (S, M and H). This trajectory is represented differently by the two methods.

- With QHA, if the period is divided into two intervals, one of 4 years (ages 14-17) and one of 3 years (ages 18-20), we get the following matrix:

Age 14-17			Age 18-20		
S	M	H	S	M	H
0,5	0,5	0	0	0,33	0,67

- With OM, we get the sequence SSMMMHH in which each letter represents the respondent's state in a particular year. The sequence could be changed to SSSMMHH by inserting an M and deleting an S or by replacing an M with an S. We select the sequence of operations with the lowest cost.

3. Comparing the two techniques for occupational trajectories

Similar resulting cluster distributions

Working on all the male occupational trajectories, we chose two partitions[20] (6 and 10 clusters of trajectories) which offered a good compromise between the imperative of synthesis and the need to present the heterogeneity of individual trajectories. The 6-cluster solutions obtained with QHA and OM are relatively similar. Five of the clusters have very comparable profiles, though their sizes are slightly different (Table 1).

[20] The clusters were calculated using Ward's criterion. We did not consolidate the partition, by means of k-means technique for example, so as to obtain different partition levels of a single cluster and thereby facilitate comparison: the 6-cluster partition is a regrouping of the 10-cluster one.

Table 1. Six-cluster career distribution of men from the Paris region, 1930-1950 cohorts[21]

Nbr	Cluster	QHA	OM
1	Intermediate occupations	29%	27%
2	Manual workers	26%	26%
3	Higher-level occupations	26%	26%
4	Clerical and sales workers	13%	9%
5	To self-employed	5%	5%
6	From intermediate to higher-level occupations	–	6%
7	From farming to manual occupations	1%	–
	Total	100%	100%

Population: 1,341 working careers of men in the 1930-1950 cohorts living in the Paris region at the time of the survey
Source: *Biographies et entourage* survey (INED, 2001).

Cluster 1: Intermediate occupations (29% with QHA, 27% with OM)

The men in this category are characterized by long periods spent in intermediate occupations. Various different trajectories exist however. A first sub-cluster includes men who have spent time in a large number of SOCs, though its numerically small. A second comprises men who began their career as clerical workers, before moving quite quickly into an intermediate occupation and, in some cases, later holding a higher-level occupation. A similar career profile is observed among men who started as manual workers. These two sub-clusters contrast with the trajectories of men who started their career in an intermediate occupation. Among these men, those who entered a higher-level occupation before age 50 can be distinguished from those who did not. The interval between completion of education and entry into an intermediate occupation is quite short. In many cases, these men entered this SOC through their qualifications rather than through promotion.

Cluster 2: Manual workers (26% with both QHA and OM)

Men who have spent time as manual workers are found in several clusters: they may have remained in this SOC for varying lengths of time, and at different periods of their life. Cluster 2

[21] The category names indicate their main characteristic, making their description more readily understandable. But a typology is no more than a summary of the wide variety of trajectories and each category is internally heterogeneous. This internal heterogeneity can be illustrated using graphs such as chronograms, presenting the distribution of individuals by state at each age (the example of the six-category OM typology is given in appendix 2). Another type of graph, commonly known as a "carpet" can be used to represent individual trajectories, but is not very legible unless printed in colour. An example showing the OM "higher-level occupations" category is nonethless given in appendix 3. It reveals the wide range of different trajectories within a single category: the transition from student to higher-level occupation occurs at a variable age, often after one or more episodes in other states.

includes men who were manual workers over a large proportion of the observation period, even if they were in other SOCs at some time in their working career. Some manual workers move directly into intermediate occupations by becoming technicians or supervisors, sometimes through promotion. Yet this cluster is strongly characterized by its limited upward mobility.

Cluster 3: Higher-level occupations (26% with both QHA and OM)

This cluster includes respondents who entered higher-level occupations at a young age, generally from the start of their working career. Trajectories differ according to age of entry into this group.

Cluster 4: Clerical workers (13% with QHA and 9% with OM)

Type 4 includes a quite heterogeneous set of trajectories, though this should not be interpreted as a weakness of our method. On the contrary, type 4 includes only men who have all spent a relatively long period as clerical workers (14.1% of the population). The sub-clusters which appear at different partition levels of the hierarchical clustering show the different forms that these periods may take. We observe, in particular, a group of men who became clerical workers after working as manual workers in their youth. A second group comprises men who started out as clerical workers and who then moved up into intermediate occupations and, in some cases, to higher-level occupations. A third comprises men who remained in clerical occupations for most of the period.

Cluster 5: Towards self-employed (5% with both QHA and OM)

This cluster mainly comprises individuals who became self-employed. Most often, they previously worked in manual, clerical, intermediate or higher-level occupations for a varying length of time. They thus reflect a certain form of occupational mobility.

Cluster 6: From intermediate to higher-level occupations (6% with OM)

Individuals in this cluster experience upward intra-generational mobility, from intermediate to higher-level occupations. This relatively homogeneous cluster is only identified by OM.

Cluster 7: From farmers to manual workers (1% with QHA)

This last cluster, specific to QHA, comprises individuals who were farm workers in their youth. At the start of their trajectory, most of them work as family helpers, then very quickly, between ages 20-30, they move on to become manual workers in most cases.

The six-cluster QHA and OM solutions seem to be based on the state occupied for the longest period. For this reason, mobility is only very marginally identified (clusters 5, 6 and 7). A ten-cluster distribution provides a more detailed picture (Table 2).

Table 2. Ten-cluster career distribution of men from the Paris region, 1930-1950 cohorts

Nbr	Cluster	QHA	OM
1	Higher-level occupations	25%	26%
2	Intermediate occupations	22%	16%
3	Manual workers	19%	19%
4	Clerical and sales workers	12%	6%
5	From manual to intermediate occupations	10%	6%
6	To self-employed	5%	5%
7	From farming to manual occupations, before age 25	2%	-
8	From farming to manual occupations, after age 25	1%	-
9	Economically inactive	1%	-
10	Self-employed	2%	-
11	From intermediate to higher level occupations	-	6%
12	From manual worker to intermediate occupation or self-employed	-	7%
13	From clerical to intermediate occupations	-	5%
14	From manual to clerical occupations	-	2%
	Total	100%	100%

Population: 1,341 working careers of men in the 1930-1950 cohorts living in the Paris region at the time of the survey.
Source: *Biographies et entourage* survey (INED, 2001).

The ten-cluster distributions are fairly similar. In particular, although the sizes sometimes differ, the first six clusters (which represent 93% of respondents with QHA and 78% with OM) have very comparable profiles. Clusters 1-4 include the stable trajectories of higher, intermediate, manual and clerical occupations. Clusters 5 and 6, on the other hand, correspond to occupational mobility: transition from manual to intermediate occupation, and to the status of self-employed.

The other QHA clusters are small and reflect marginal trajectories. Clusters 7 and 8, for example, both concern individuals who started in farming occupations and then become manual workers, age at transition being the distinction between the two. Cluster 9 is of particular interest as it is the only one to place emphasis on economic inactivity: it comprises men who have experienced one or more periods of inactivity whose total duration is significantly longer than is the case for men in the other clusters.

Last, the clusters identified by OM only are larger and concern mobile trajectories only: from intermediate to higher-level occupations; from manual to intermediate or self-employed (transition after age 35); from clerical to intermediate; from manual to clerical. OM thus seems to be more sensitive to transitions and is capable of capturing intra-generational mobility processes in relatively homogeneous clusters.

The emergence in the ten-cluster distributions of mobility trajectories that are invisible in the six-cluster distributions illustrates the importance of exploring a trajectory cluster at different partition levels, to see, for example, whether the population grouped in a particular trajectory type could be usefully subdivided into several separate types. A distribution with a large number of clusters reveals a broad range of trajectories, but may be difficult to characterize. The aim of a cluster approach is not only to explore complex data, but also to summarize the broad variety of individual behaviours. The choice of the number of clusters thus represents a compromise between concision and exhaustivity, depending on the research hypotheses and questions under study. These results represent basic analysis of occupational trajectories that can be refined by matching different individual characteristics (birth cohort, place of birth, father's occupation, etc.) with the respondent's career type using descriptive statistics. Regression models including the trajectory cluster as a dependent or independent variable can also be used, although the diachronic nature of the relations of causality calls for a careful approach.

Measuring differences

The degree of similarity between the cluster distributions obtained using qualitative harmonic analysis and optimal matching can be illustrated using a correspondence matrix which cross-tabulates the population distributions obtained with two different partitions. It shows whether the individuals of a cluster from the first partition are concentrated in a single cluster of the second partition or scattered between several different ones. Here, the six-cluster OM distribution spans the rows and the QHA distribution spans the columns (Table 3). The correspondence matrix of the ten-cluster distributions is given in Appendix 4.

Table 3. Correspondence matrix between OM and QHA six-cluster distributions

		QHA						
		1	2	3	4	5	6	Total
OM	1	291	16	4	45	5	1	362
	2	7	320	0	5	13	7	352
	3	17	5	311	18	0	0	351
	4	2	11	1	98	0	5	117
	5	13	0	9	2	49	0	73
	6	59	1	23	2	1	0	86
	Total	389	353	348	170	68	13	1 341

Population: 1,341 working careers of men in the 1930-1950 cohorts living in the Paris region at the time of the survey.
Source: Biographies et entourage survey (INED, 2001).

The diagonal of the table illustrates the strong similarity between the clustering obtained with the two methods, with the exception of cluster 6, which is small.[22] A correspondence rate summarizing the degree of similarity between the two partitions can be calculated by summing the mode of each line and of each column and dividing it by the double of the total sample size. We thus calculate the mean of the correspondence of the OM clustering to the QHA clustering and of the QHA clustering to the OM clustering:

$$\sigma = \frac{\sum_i \max_j (n_{ij}) + \sum_j \max_i (n_{ij})}{2N} \qquad [1]$$

where n_{ij} represents the number of trajectories belonging simultaneously to the la i^{th} cluster of the first partition and the j^{th} cluster of the second partition, and where *N* is the total population.

The correspondence between the distributions obtained with the two methods is 82% for a six-cluster partition and 75% for a ten-cluster partition. The two methods thus give strongly converging results, a good indication of their robustness.

There are several nuances, however. On the basis of our observations, we can formulate hypotheses as to the comparative advantages of each method. Comparison of the six- and ten-cluster distributions reveals several differences, which can be explained in two ways.

[22] 86 for OM and 13 for QHA.

First, the OM clusters are more sensitive to transitions, notably at the start of the trajectory when changes of occupational state are more frequent. They specifically distinguish between stable trajectories and mobile ones, and in this respect are more homogeneous than the QHA clusters. This is because OM analyses the sequence year by year, while QHA measures the proportion of each period spent in each state and does not include the sequence of transitions from one state to another in the analysis.

In addition, QHA produces several smaller clusters grouping very specific types of trajectory, i.e. differing substantially from the others by the nature and sequence of states over the trajectory as a whole. We observe, for example, a cluster comprising individuals who have experienced short and scattered periods of inactivity, along with trajectories including one or more periods in a farming occupation. By contrast, the OM partition tends to incorporate these marginal trajectories into larger clusters.

Comparative advantage of each method

QHA provides a means to select for clustering only the main factors resulting from factor analysis. It is thus possible to eliminate some of the "noise" contained in trajectory data, i.e. to tease out the most structuring characteristics of the information, thus making it easier to interpret. But analysts cannot know what part of the statistical information is ignored as a consequence. They can only identify ex-post the factors disregarded by the procedure. In addition, by dividing the observation period into intervals of varying length, attention can be focused on the stages where the events of interest to the analyst are most frequent. This may be useful for occupational trajectories in which most events occur before age 30. For example, QHA provides a means to concentrate on the periods when specific events occurred and on the duration of states. In particular, the occupations held for the longest durations have a stronger influence on cluster distribution than in OM, thus placing emphasis on socio-occupational stability, which still remains high today (Goux, 1991).

Conversely, the OM analysis conserves the full detail of the sequence, rather than simplifying it (Lesnard et Saint-Pol, 2004). It is thus the analyst's task to fix the costs on the basis of theoretical hypotheses, thereby favouring one pattern of trajectory clustering over another. Depending on choice of indel cost, for example, interest can be focused on the type of transition or rather on the timing of an event. Compared to QHA, optimal matching thus places more emphasis on the sequence of events

and the types of transition occurring within it. Consequently, it gives more weight to transitions from one SOC to another, and is able to detect intragenerational occupational mobility

Conclusion

Event history surveys are designed to provide data for statistical analysis of individual trajectories. Yet the promoters of these surveys themselves acknowledge that the results are often disappointing. For example, in a book on quantitative event history surveys published by the Groupe de réflexion sur l'approche biographique (GRAB) the authors admit that "[...] the potential of these data is not being fully exploited. It is true that the considerable energy expended in producing these data and acquiring the skills required to process them effectively could and should have been more productive, more widely applied and disseminated" (GRAB, 1999, p. 46, quoting M. Bottai). Indeed, the questionnaires are long, and hence costly to administer, while the samples are necessarily limited to a few thousand individuals. The information collected covers a wide variety of events and a small number of individuals, and is therefore difficult to analyse.

Clustering methods overcome this difficulty by offering a processing tool that captures the logic of individual trajectories in quantitative terms. They provide an effective means to exploit these promising data sources for demographers and social scientists in general. Yet we need to find out more about the methods available to us, some of which are still relatively unknown, and whose comparative advantages have rarely been tested. Our work shows that qualitative harmonic analysis performs better when attention focuses on the duration of certain stages and when we need to know more about the states in which an individual remains for the longest period of time. Optimal matching, on the other hand, is the better choice for analysing trajectories by type of transition and when occupational mobility is the focus of attention. But individual event history data are highly structured and the main types of trajectory emerge, whatever the method applied.

Acknowledgements: We would like to thank Valérie Golaz, Maryse Marpsat and Thibaut de Saint-Pol for their valuable comments on the first version of this article. The present version owes much to their suggestions. Our thanks also to Vincent Cardon for his careful rereading of the present version.

Appendices

Appendix 1. Matrix of OM replacement and indel costs

	farm	*self-emp*	*h-occ*	*int-occ*	*cler*	*manu*	*mil serv*	*inact*	*stu*
farm	0,000	1,992	2,000	2,000	1,990	1,895	1,999	2,000	1,954
self-emp	1,992	0,000	1,990	1,986	1,987	1,978	1,997	1,987	1,991
h-occ	2,000	1,990	0,000	1,971	1,990	1,997	1,972	1,964	1,912
int-occ	2,000	1,986	1,971	0,000	1,961	1,976	1,966	1,948	1,853
cler	1,990	1,987	1,990	1,961	0,000	1,970	1,972	1,962	1,896
manu	1,895	1,978	1,997	1,976	1,970	0,000	1,947	1,905	1,782
mil serv	1,999	1,997	1,972	1,966	1,972	1,947	0,000	1,986	1,947
inact	2,000	1,987	1,964	1,948	1,962	1,905	1,986	0,000	1,980
stu	1,954	1,991	1,912	1,853	1,896	1,782	1,947	1,980	0,000

Abbreviations: farm = farmer; self-emp = self-employed; h-occ = higher-level occupation; int-occ = intermediate occupation; cler = clerical and sales worker; manu = manual worker; mil serv = military service; inact = inactive; stu = student

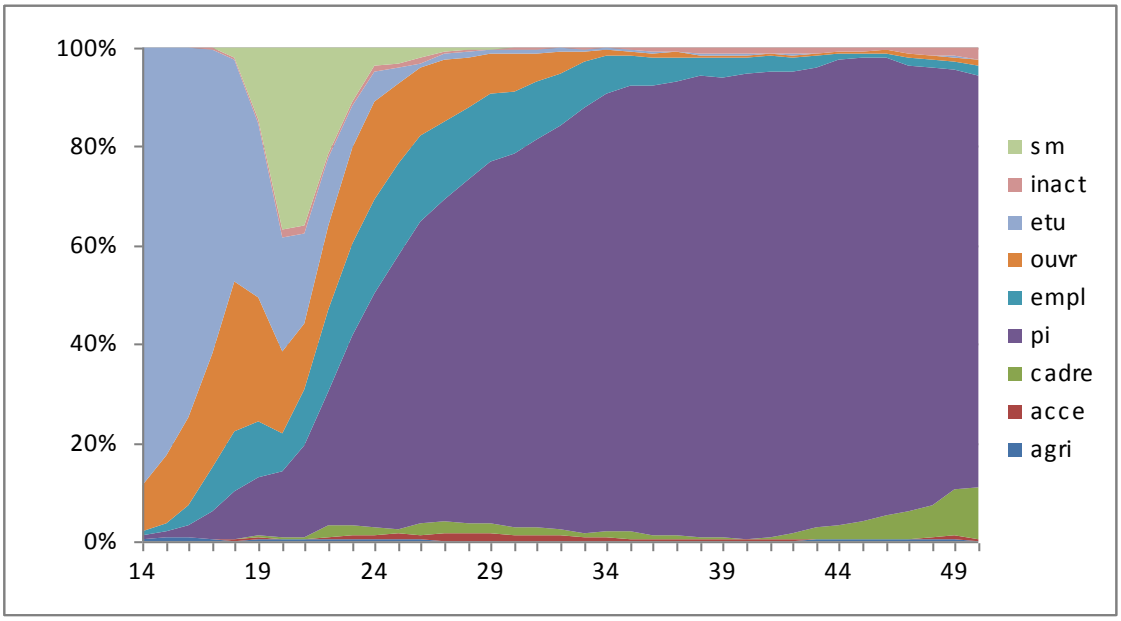
Indel = 1,01

Population: 1,341 working careers of men in the 1930-1950 cohorts living in the Paris region at the time of the survey.

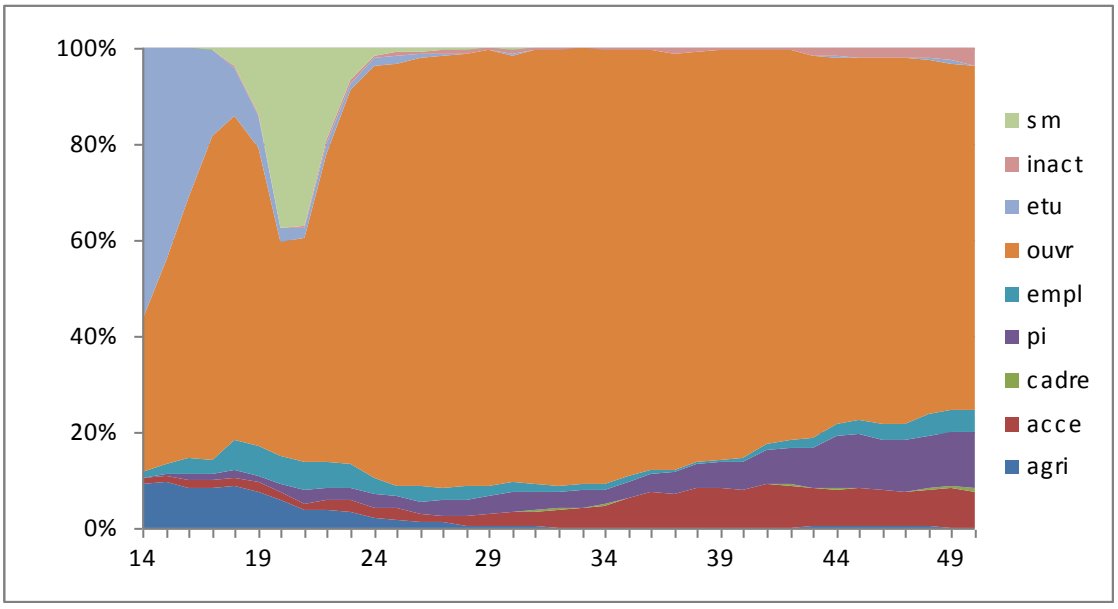
Source: Biographies et entourage survey (INED, 2001).

Appendix 2. Chronograms of the OM six-cluster distributions (proportion of individuals of a given cluster in each category by age)

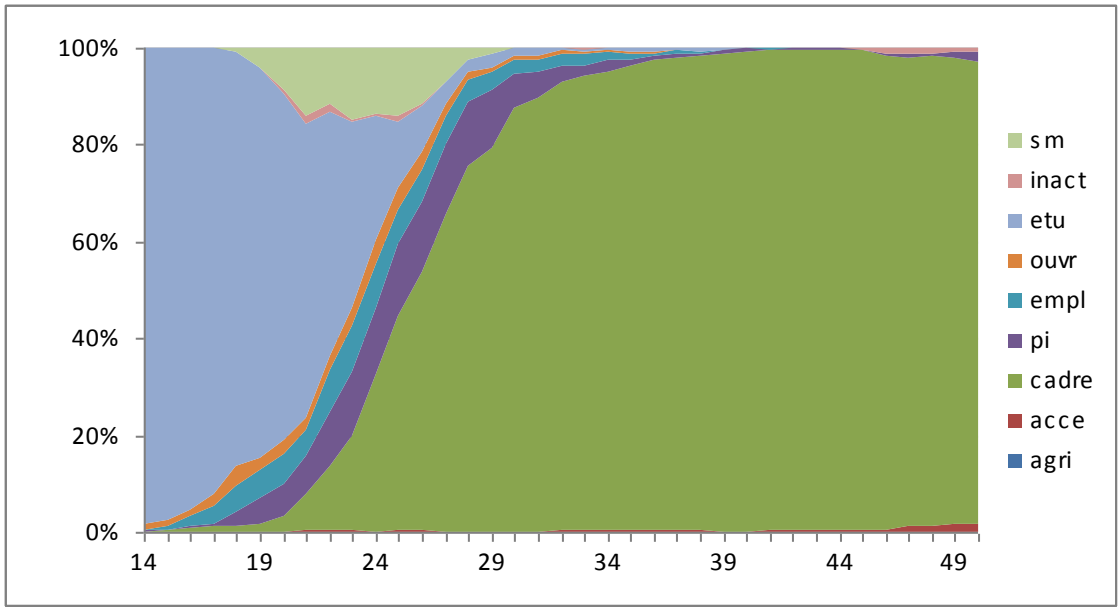
Cluster 1



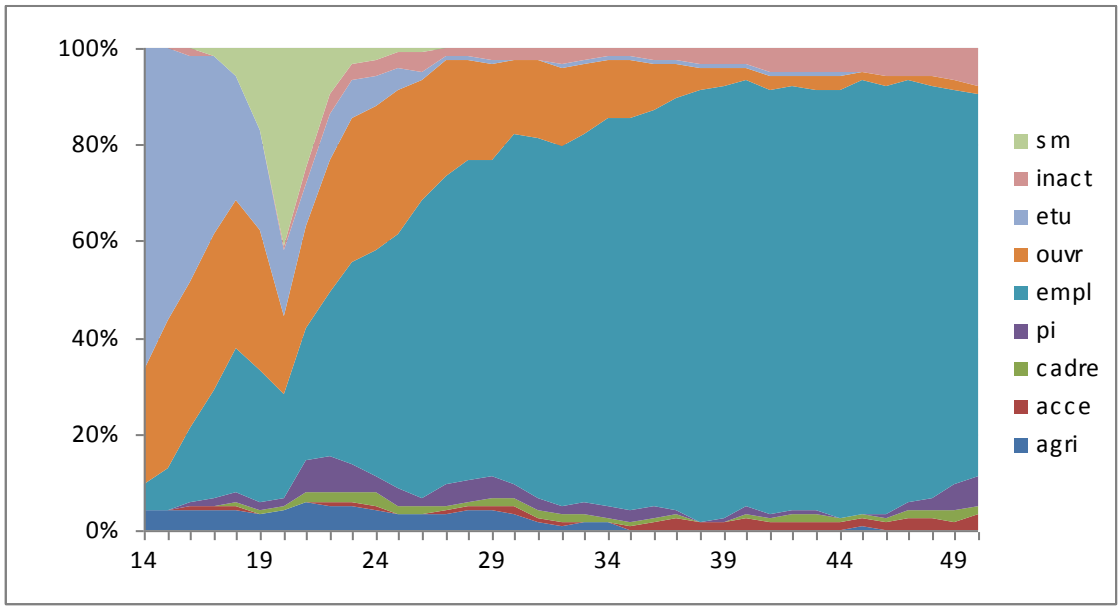
Cluster 2



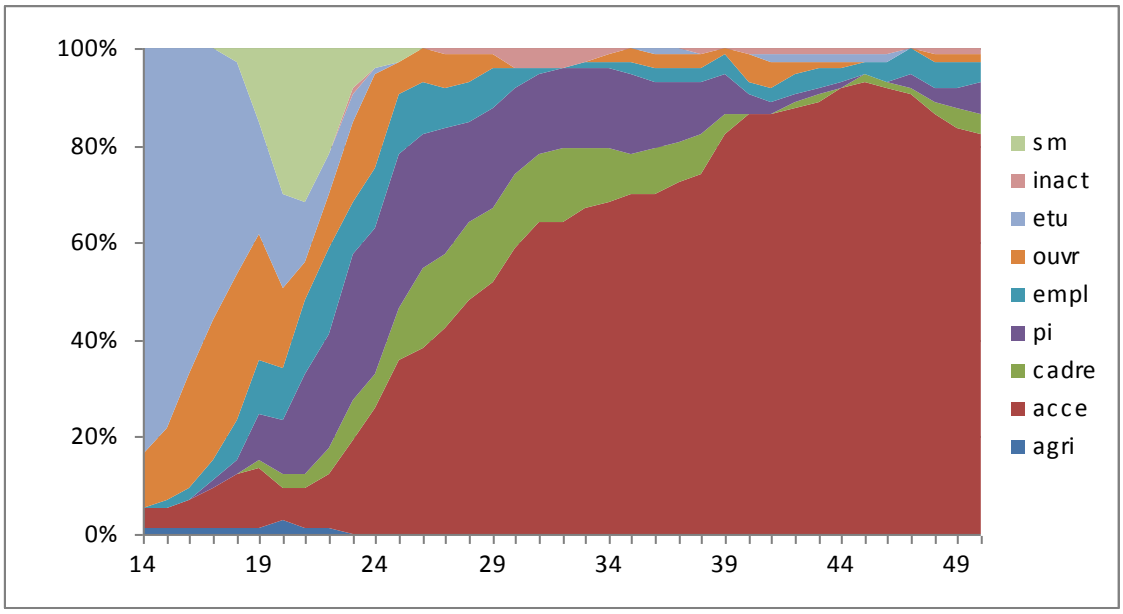
Cluster 3



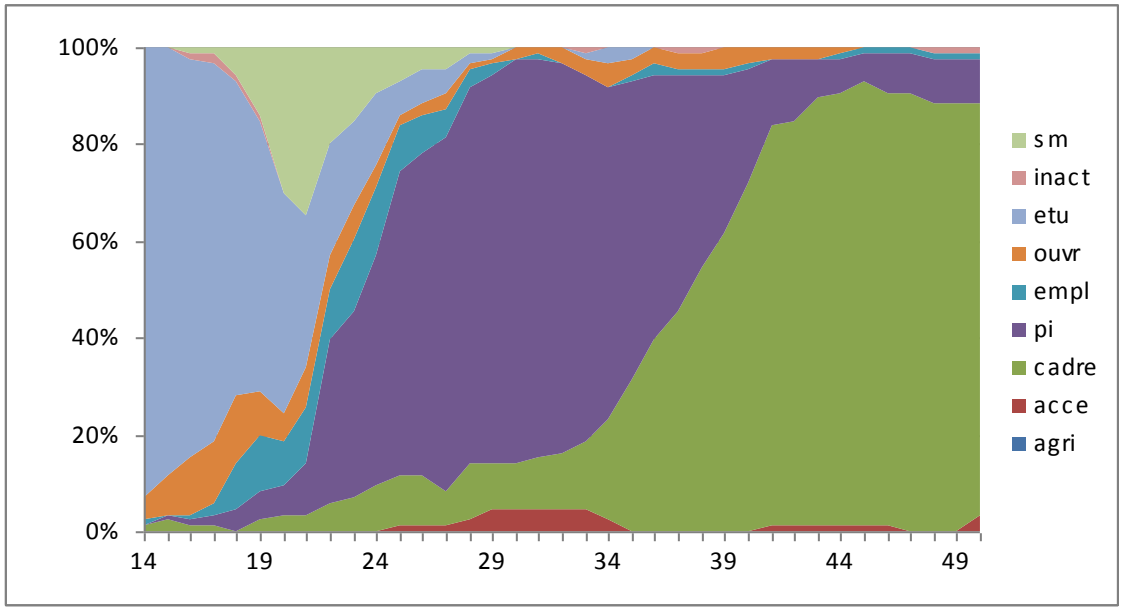
Cluster 4



Cluster 5



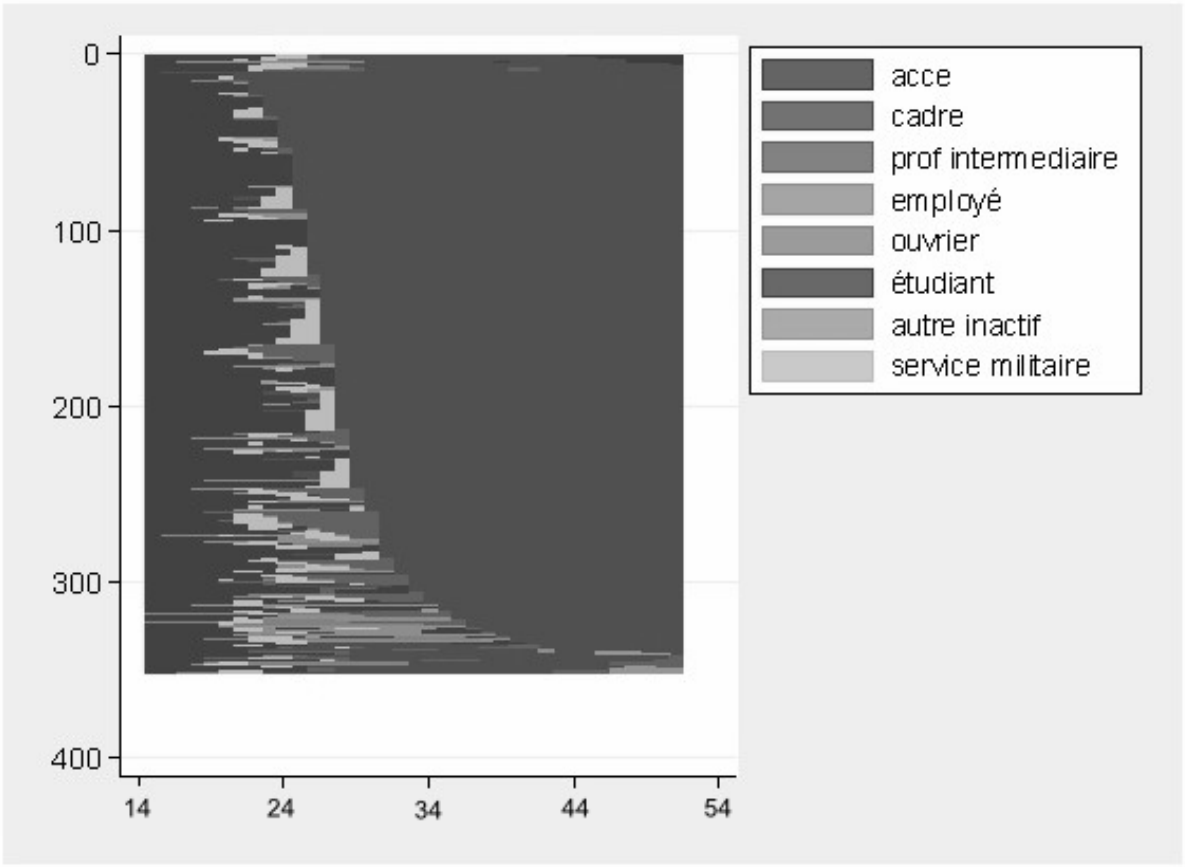
Cluster 6



Abbreviations: farm = farmer; self-emp = self-employed; h-occ = higher-level occupation; int-occ = intermediate occupation; cler = clerical and sales worker; manu = manual worker; mil serv = military service; inact = inactive; stu = student

Population: 1,341 men in the 1930-1950 cohorts living in the Paris region at the time of the survey.
Source: Biographies et entourage survey (INED, 2001).

Appendix 3. Carpet of OM cluster 3 (higher-level occupations)



Abbrevations: ***

Each line represents an individual trajectory, and each episode in the trajectory is represented by a segment whose length corresponds to duration and whose colour corresponds to state.

Population: men in the 1930-1950 cohorts living in the Paris region at the time of the survey.
Source: Biographies et entourage survey *(INED, 2001).*

Appendix 4. Correspondence matrix between OM and QHA ten-cluster distributions

		QHA										
		1	2	3	4	5	6	7	8	9	10	Total
OM	1	304	17	0	18	9	0	0	0	3	0	351
	2	4	187	0	0	15	0	0	0	2	0	208
	3	0	0	231	1	0	0	20	5	1	0	258
	4	1	2	0	83	0	0	0	0	1	0	87
	5	0	8	0	0	72	1	1	1	0	1	84
	6	1	10	0	0	0	41	0	0	0	21	73
	11	22	51	0	2	9	2	0	0	0	0	86
	12	0	6	27	4	24	24	3	0	0	6	94
	13	0	18	0	45	6	0	0	0	0	1	70
	14	0	0	5	13	3	0	2	4	3	0	30
	Total	332	299	263	166	138	68	26	10	10	29	1 341

Population: 1,341 working careers of men in the 1930-1950 cohorts living in the Paris region at the time of the survey.
Source: Biographies et entourage survey *(INED, 2001).*

References

Aalen O., 1978, “Nonparametric inference for a family of counting processes”, *The Annals of Statistics*, 6(4), pp. 701-726

Abbott A., Forrest J., 1986, “Optimal matching methods for historical sequences”, *Journal of Interdisciplinary History,* 16(3), pp. 471-494.

Abbott A., Hrycak A., 1990, “Measuring ressemblance in sequence data: An optimal matching analysis of musicians’ careers”, *American Journal of Sociology*, (96), pp. 144-185.

Barbary O., 1996, *Analisis tipologico de datos biograficos en Bogota*, Bogota, Universidad Nacional de Colombia, 254 p.

Barbary O., 1997, “Analisis estadistico de datos biograficos : metodos, ejemplos y perspectivas en el estudio de itinerarios migratorios” in J. A. Bustamante, D. Delaunay, J. Santibanez, *Medicion de la migracion internacional*, Tijuana, Documento de trabajo del Colegio de la Frontera Norte.

Barbary O., Pinzon Sarmiento L.M., 1998, “L’analyse harmonique qualitative et son application à la typologie des trajectoires individuelles”, *Mathématiques informatiques et sciences humaines*, 144, pp. 29-54

Billari F., 2001, “Sequence analysis in demographic research”, *Canadian Studies in Population*, Special Issue on Longitudinal Methodology, 28(2), pp. 439-458.

Blair-Loy M., 1999, “Career patterns of executive women in finance: An optimal matching analysis”, *American Journal of Sociology*, 104(5), pp. 1346-1397.

Bonvalet C., Bringé A., Robette N., 2009, “Les trajectoires géographiques des Franciliens : un exemple de complémentarité qualitatif-quantitatif”, in Actes du colloque *Approches quantitatives et qualitatives des mobilités*, Namur, AISLF, forthcoming.

Cambois M.A., Lelièvre É., 1988, “Durée d’activité et interruption de carrière des femmes âgées de 45 ans à 64 ans en 1981”, *Population*, 43 (3), pp. 669-675.

Courgeau D., 2000, “Vers une analyse biographique multiniveau”, in Actes des *Journées de méthodologie statistique*, Insee, 4-5 December, http://jms.insee.fr/site/files/documents/2008/658_1-JMS2000_S4-5_COURGEAU.PDF

Courgeau D., Lelièvre É., 1996, “Changement de paradigme en démographie”, *Population,* 51 (3), pp. 645-654.

Cox D. R., 1972, “Regression models and life tables (with discussion)”, *Journal of Royal Statistical Society*, B34, pp. 187-220.

Crépon B., Gurgand M., Dejemeppe M., 2005, “Counseling the unemployed: Does it lower unemployment duration and recurrence ?”, *Document de travail*, Centre d’études de l’emploi, 40.

Degenne A., Lebeaux M.-O., 1999, *Étude sur les sorties du chômage, comparaison jeunes et adultes*, Rapport pour le Commissariat général du plan, Caen, Lasmas.

Degenne A., Lebeaux M.-O., Mounier L., 1996, “Typologies d’itinéraires comme instrument d’analyse du marché du travail”, in A. Degenne, M. Mansuy, G. Podevin, P. Werquin (eds.), *Typologie des marchés du travail, suivi et parcours*, Document du CEREQ, 115, pp. 27-42.

Desplanques G., Saboulin M. (de), 1986, “Activité féminine : carrières continues et discontinues”, *Économie et Statistique*, 192-193, pp. 51-62.

Deville J.-C., 1974, “Méthodes statistiques et numériques de l’analyse harmonique”, *Annales de l’Insee*, 15, pp. 3-101.

Deville J.-C., 1977, “Analyse harmonique du calendrier de constitution des familles en France”, *Population*, 32 (1), pp. 17-63.

Deville J.-C., 1982, “Analyse de données chronologiques qualitatives : comment analyser des calendriers ?”, *Annales de l’Insee*, 45, pp. 45-104.

Deville J.-C., Saporta G., 1980, “Analyse harmonique qualitative”, in E. Diday (ed.), *Data Analysis and Informatics*, Amsterdam, North Holland Publishing, pp. 375-389.

Dijkstra W., Taris T., 1995, “Measuring the agreement between sequences”, *Sociological Methods & Research,* (24), pp. 214-231.

Dureau F., Barbary O., Elisa Florez C., Hoyos M. C*.,* 1994, *La observacion de las diferentes formas de movilidad : propuestas metodologicas experimentadas en la encuesta de movilidad espacial en el area metropolitana de Bogota,* Paris, Orstom, CEDE workshop (Montevideo) 27-29 October 1993 : “Nuevas modalidades y tendencias de la migracion entre paises fronterizos y los processos de integracion”, 31 p.

Espinasse J.-M., 1993, “Enquêtes de cheminement, chronogrammes et classification automatique”, Note du Lhire, 19 (159).

Florette A., 1988, *Approximation et choix du découpage dans le cadre de l’analyse harmonique qualitative*, Post-graduate dissertation, Ensae, Paris.

Goux D., 1991, “Coup de frein sur les carrières”, *Économie et Statistique*, 249, pp. 75-87.

GRAB, 1999, *Biographies d’enquête. Bilan de quatorze collectes biographiques*, Paris, INED (Méthodes et Savoirs, 3), 340 p.

GRAB, 2009, *Fuzzy States and Complex Trajectories. Observation, Modelization and Interpretation of Life Histories*, Paris, INED (Méthodes et Savoirs 6), 176 p.

Grelet Y., 2002, “Des typologies de parcours. Méthodes et usages”, *Document Génération 92*, (20), 47 p.

Halpin B., Chan T. W., 1998, “Class careers as sequences: An optimal matching analysis of work-life histories”, *European Sociological Review*, 14 (2), pp. 111-130.

Kaplan E., Meier P., 1958, “Nonparametric estimation from incomplete observations”, *Journal of American Statistical Association,* vol. 53, pp. 457-481.

Kempeneers M., Lelièvre É., 1991, “Analyse biographique du travail féminin”, *Revue européenne de démographie*, 7, pp. 377-400.

Lelièvre É., 1987, “Activité professionnelle et fécondité : les choix et les déterminations chez les femmes françaises, de 1930 à 1960”, *Cahiers québécois de démographie*, 16, pp. 209-236.

Lelièvre É., Vivier G., 2001, “Évaluation d’une collecte à la croisée du quantitatif et du qualitatif. L’enquête Biographies et entourage”, *Population*, 56 (6), pp. 1043-1074.

Lesnard L., Saint-Pol T. (de), 2004, “Introduction aux méthodes d’appariement optimal (Optimal Matching Analysis)”, *Document de travail du Crest,* (15), 30 p.

Macindoe H., Abbott A., 2004, “Sequence analysis and optimal matching techniques for social science data”, in Hardy Melissa, Bryman Alan*, Handbook of Data Analysis*, London, Sage, pp. 387-406.

Marchand O., Thélot C., 1997, *Le travail en France. 1800-2000*, Paris, Nathan, 269 p.

Nelson W., 1972, “Theory and application of hazard plotting for censored failure data”, *Technometrics*, vol. 14, pp. 945-965.

Rohwer G., Pötter U., 2005, “TDA's user manual”, http://www.stat.ruhr-uni-bochum.de/tman.html

Sankoff D., Kruskal J. (eds.), 1983, *Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison*, Reading, Addison-Wesley, 408 p.

Scherer S., 2001, “Early career patterns: A comparison of Great Britain and West Germany”, *European Sociological Review*, 17 (2), pp. 119-144.

Solis P., Billari F., 2002, “Work lives amid social change and continuity: occupational trajectories in Monterrey, Mexico”, *Max Planck IDR Working paper*, 2002-009, 52 p.

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), pp. 358-399.

Event history surveys provide a means to analyse large numbers of complete individual occupational trajectories. A variety of statistical methods have been developed to measure the time spent in a give state as a function of individual characteristics. Until the 1990s, exploratory data analysis to describe the full complexity of trajectories was rarely mentioned in the literature. Qualitative harmonic analysis and optimal matching are two exploratory methods that can be used to build typologies of complex individual trajectories that take account of both the sequence and the duration of events. They are used here to classify the working careers of male respondents of the *Biographies et entourage* survey (INED, 2001), with the aim of comparing the respective advantages of each technique.

Nicolas Robette, Institut national d'études démographiques, 133 boulevard Davout, 75980 Paris Cedex 20, tél : 33 (0)1 56 06 22 56, courriel : nicolas.robette@ined.fr