Sequence Analysis for Political Science

Philippe Blanchard and Olivier Fillieule
Lausanne University, IEPI-CRAPUL

Draft version. Do not cite.

Abstract
Several areas of political research deal with sequences, that is, successions of standard categorical states or events: political sociology, evolution of regimes, analysis of speeches, geopolitics, comparative studies, or elections. At least three kinds of longitudinal methods, popular in political science, may attempt at treating political longitudinal objects: regression models, event history analysis and time series analysis. Yet, none can unfold the three dimensions of categorical time series, that is, the nature of the states/events composing the sequences, their order and length. Sequence analysis, with the optimal matching algorithm as a core tool, was specifically designed to this task. It is now commonly used in sociology and demography, and more and more in geography and history. This pragmatic, state-by-state comparison of sequences does not make any assumption about an underlying process that would generate sequences. The paper first defines sequences and their empirical applications. Then it details the principles of sequence analysis and its canonical steps. It shows how sequence analysis connects to and/or competes with other multivariate methods, before giving an overview of advanced issues and available software. To illustrate how fruitful this approach can be for political science, we apply it to a retrospective survey conducted among members of the main French activist organization mobilizing against AIDS.

Key-words
Sequence analysis; method; social time; optimal matching.

Corresponding address
philippe.blanchard@unil.ch
Lausanne University, IEPI, Vidy, CH-1015 Lausanne
www.uni.ch/iepi

Paper delivered at 6th ECPR General Conference, Reykjavik, August 2011
Panel 581: "Non a Priori Modelling Methods for Political Science"

Sequence Analysis for Political Science

Sequence analysis (SA) is a method to process sequence data, sequences being defined as series of states or events in the trajectories of statistical individuals. SA includes tools to code and format sequences, to compare them by pairs, to cluster them, to represent them in alphanumeric and graphical forms, to calculate specific statistics for sequences and groups of sequences, to mine sequences and to extract prototypical sequences. SA has been developing since the mid-1980s in the social sciences, after being imported from genetics and computer science. It has spread quickly to sociology and demography, and is now developing in history, geography, anthropology and political science. Some strong and original results have been achieved by means of SA, regarding for example the transition from education to work life and from work life to retirement. In this paper, we wish to show that SA presents a large potential for political science, through many, largely unexplored empirical applications. SA could soon enter the standard statistical toolbox, as a pragmatic, mainly descriptive approach to time-related phenomena. This paper intends to show the diversity of potential applications of SA, how to implement it and how it complements other methods.

1. Definitions

a. Sequences, alphabet, states/events

A sequence is a *succession* of elements chosen inside an *alphabet*. Sequences are time series with categorical data. The *succession* usually follows the order of time. In most studies, the sequence describes an individual trajectory in any domain of life (family, work, religion, leisure activities…), inside a particular organization or group of organizations (a company, an association, an artistic network…), or inside a social group (defined by ethnicity, geographical location, age…). In the most dominant usage, sequences are called "biographies", "careers", "personal histories" or "processes of (individual) development". But sequences may also describe the trajectory of collective entities, such as families that go through typical steps in their growth and in their internal organization (mere couple, one child, more children, separation of parents or divorce, children leaving the parental home…) or populations that undergo typical stages of development (e.g. before/during/after the demographic transition). Non human statistical individuals may also be considered, such as the media coverage of a controversy revealed through varying successive angles or stages (rumour, official revelation, controversy, and decline).

Sequences are usually considered in the long-term and measured in years, following the demographers' perspective, but other measurements of time can be used. Sequences of activities or trips may occur over the course of a day or a week, using time-budget data (Lesnard 2004). Sequences of economic activities may span over years or decades. Succession may also concern less obviously time-related elements, such as the rhetorical steps in a speech, symbols in a text, steps in a dance, verses in a prayer or a song, gestures in a ceremony, and so on. In these cases, time is implicit: a speech unrolls inside time, a text is both written and read in time, dances and songs develop in time, and ceremonies unfold in time. In the latter cases, the time axis is an underlying source of order.

Subsequences are sections of sequences. The more common subsequences two sequences share, and the longer these common subsequences are, the more similar the two sequences are. A less obvious way to consider subsequences is when they are punctuated with gaps: *nonsuccessive common subsequences* are made of the same states of events, in the same order, but with one or several gaps that may not be located at the same place in the subsequence. For example, the experience of professional downgrading can be fast, from an initial position of full-time executive (FTE) to a final situation of unemployment (U), or slow, the two extreme states being separated by some years spent in a part-time executive position (PTE) or in a full-time subaltern position (FTS). The two sequences <FTE→FTE→FTE→U> and <FTE→FTE→FTE→PTE/FTS→PTE/FTS→PTE/FTS→U> share the *common nonsuccessive subsequence* <FTE→FTE→FTE→U>. A *stable subsequence,* also called an *episode*, is a subsequence made of two or more identical states/events. To put it differently, it is a state that lasts several units of time or a repeated event. This is the case with <FTE→FTE→FTE> in the previous example.

b. Sequence analysis

Sequence analysis is the systematic study of a population of sequences. Before it developed as such, some work on sequences has been carried out in a less systematic fashion, that is, without formalizing sequences as above and without expliciting formal algorithms of treatment. A whole literature in history and sociology has been dealing with sequences, such as macro-historical theories of political, economic and social development (Abbott 1990a, 1995b).

In a more restricted fashion, since the 1990s, SA has come to designate a methodological package with five objectives. The first is to describe and to represent sequences. Descriptive statistics are a major part of modern statistics, although often neglected in aid of causal modelling. The visualization may be exhaustive or partial, depending on the size of the sample. It may use alphanumeric codes, if one wants to decipher a limited number of chosen subsequences, or graphs, if

one needs to grasp a general shape, trend or pattern inside the sample at hand or a subsample of individuals. In the case of sequences, the description has to be global in order to understand the coherence of the trajectory. This coherence is both objective, with former institutional and material situations influencing the following situations, and subjective, each individual behaving in accordance with her former behaviours, as well as with her anticipations and plans for future activities. It is driven by individual legacies, by ideals, plans and strategies, but also by social norms, statuses, networks and organizations.

Secondly, SA aims at comparing and classifying sequences. Along an inductive approach, one may ask: what groups emerge from our population? what social or ideological cleavages give it a structure? A deductive approach is also possible: how different are two category-groups? which of a set of groups is the closest to a defined pattern? Beyond the strict description of two sequences in order to check for common subsequences and common proportions of the same states, SA proposes a way to measure the amount of commonality and difference, by means of optimal matching analysis (OMA).

A third goal is sequence mining, that is, searching inside the population of sequences. One may want to look for sequence patterns defined *a priori*. For example, in a sample of individual voting sequences gathered over several ballots from polling stations records, which party affiliation is more likely to be associated with non-regular voting? One may also look for the dominant pattern inside one given group. What most common career paths do the leading executives of a company go through, before/during/after their stay in the company? Another question regards which pattern(s) of sequence discriminate between two or more given groups. Is there a significant difference between months in daily travelling sequences in a city, in terms of means of transportation? What moment of the standard family cycle differentiates the most generations born after the big changes of the 1960s-1970s from previous generations?

Finally, sequence analysis can explain trajectories. The analyst may search for the causal relationship between external variables and clusters of similar trajectories that come out of SA: do women and men have similar work trajectories or political involvements? How much does your religious background impact the timing of coupling, marrying, bearing children in your own family? Reversely, clusters of sequences may explain some individual features. For example, the past trajectory of patients in terms of diseases and medical treatments constitutes a decisive clue for an epidemiologist to predict what other diseases they might catch in the future. A sequence approach is usually used with benefit when non-sequenced variables fail to account for the phenomenon under study. For example, Hollister (2009) assesses the quality of typologies of early occupational sequences through their ability to predict work status five years after the end of the sequences. In a fully

"sequenced" perspective, SA may also provide both sides of the causal link, such as explaining the type of voting trajectory by the previous educational trajectory.

Following these five goals implies using diverse tools that offer complementary views on sequences. It opposes to the strictly numerical perspective of a reduction of SA to an automated calculation of distance between sequences (OMA), as some authors do. Yet, some options have been chosen in the 1990s that now dominate the field, in particular regarding the way OMA is used and combined with other tools. We will show that alternative options inside Abbott's perspective remain open.

2. Sequence analysis and the social sciences

Sequence analysis was originally used in computer science to detect dissimilarities between long strings of codes. Then it was exported by biocomputing specialists, who automatically compare DNA strings and assess their degree of dissimilarity. Social sequences obviously show structures similar to DNA strings, but they differ in several manners. The main difference is empirical: social sequences are much shorter (by million times in the case of human DNA), and their alphabet is often larger than the four DNA nucleobases (by one to ten times in general). It is also more complex: states/events result from debatable conventions and interpretations, while nucleobases are objectively distinguished from each other. Secondly, the genetic code encrypted in DNA strings is the product of the species' evolution. Hence, at least at some time in the history of bioglogy, decoding DNA has meant uncovering a historical process behind the genetic material. On the contrary, social scientists usually do not make hypotheses about a process underlying sequences. Social or political trajectories are made of continuities and changes that depend on a complex mix of decisions and constraints, although some structural factors influence them. A third difference is that social sequence analysts are not as interested as geneticians in quantifying global differences between sequences. Although dissimilarity indices help them distinguish large divides in a sequence population, their ultimate goal is rather the specific role certain states/events and specific subsequences play in the social process at hand.

Social scientists put the accent on the stages of coding and on the adaptation of the method to the data at hand. Each analysis relies on specific empirical and theoretical constraints, each alphabet mixes some objective observations and some conventional decisions. Social scientists also insist more than biologists on the interpretation of the results of SA.

Social sequences are made of three basic dimensions: the *nature* of the successive states, chosen among the alphabet; the *order* in which they occur; their *duration*, that is, the duration of constant subsequences. SA takes these three dimensions into account, with the possibility of weighting them, according to the research design. For example, duration may be ignored when processing sequences from a survey that does not provide precise dates of transition, or if duration is seen as a less relevant factor than order.

Beyond the nature of states, their order and duration, SA helps uncover more complex aspects of time that can be of crucial theoretical importance. The first is the moment at which states/events are experienced, in other words, their location along the time (or pseudo-time) axis. For example, studying the sequences of lynching of black people in the Deep South between 1882 and 1930, Katherine Stovel (2001: 863) keeps in the analysis both the transhistorical sequential patterns along which protest and violence develop, and the historical moments at which they happen. Alternatively, the analyst may neglect the historical factor during the calculation and reintroduce it afterward as a covariate or in a durational model (Lemercier 2005: 10).

The second complex aspect of time is recurring patterns (MacIndoe and Abbott 2004). SA uncovers them, either as successive or nonsuccessive common subsequences. The optimal matching algorithm agregates sequences that look the same, in the first place ones that partly or completely *match*. Some populations of sequences are simple enough to provide numerous common patterns, while others, with complex alphabets and/or complex variations in sequential structures, do not provide any. In the latter case, OMA clusters sequences on the basis of shorter, less obvious similarities. The interpretation stage consists in making sense of unevenly obvious similarities between sequences.

The third complex time effect is generations, that is, the fact that some individuals experience the same event(s) or have been socialized in comparable environments, which makes them behave similarly on the long term. From a sequential perspective, generations combine the individuals' biological or organisational age, that is, the length of sequences, and common sequence patterns. Our case study (section 7) gives an example of generational divides inside sequences of activist involvement.

3. Applications to political science

SA for social sciences first developed in the sociology of job careers, with applications to German 18th century musicians (Abbott and Hrycak 1990), to the historical development of American

psychiatry (Abbott 1991), banking and finance careers (Stovel, Savage and Bearman 1996; Blair-Loy 1999) and to intragenerational class mobility among Irish and British workers (Halpin and Chan 1998). At the same time, more exotic applications were tried, such as the development of welfare legislative activity among countries (Abbott and DeViney 1992), the evolution of the design of sociology articles and the emergence of modern scientific writing (Abbott and Barman 1997) and the order of steps in British folk dances (Abbott and Forrest 1986). But since the early 2000s the main stream clearly focused on life course analysis, that is, the study of work, family and residential trajectories. Sociologists and demographers in this field test hypotheses about the (de)standardization and the individualization of life trajectories in modern societies, often using large survey data sets (see for example: Gauthier 2007; Pollock 2007; Aisenbrey and Fasang 2010).

Usual work and family careers are mainly composed of ordinal alphabets, with little reversibility to lower positions. Their structure had already been well investigated by specialists of social stratification. Demographers and life course experts insist on rather simple statuses regarding coupling, parenting and residential changes. Introducing alternative time scales, other sequence units (groups, organizations, institutions), other time scales, below and above yearly measures, and taking historical time into consideration implies specific methodological options. By contrast with life course studies, political trajectories often shift the focus from strictly biographical dynamics to organisational and historical dynamics. Political scientists are more interested in the context in which individual trajectories unfold. The concern for power and domination also adds complexity to the samples under study: individuals do not evolve independently a in global social context, they often bear relationships of cooperation or conflict with each other, adding network structures in the analysis.

As a consequence, extending SA to political science topics requires adjustments. Several areas of political research deal, explicitly or not, with sequences. Political sociology is interested in studying the career of all kinds of political actors. At the individual level (elected people, higher civil servants and activists), the alphabet may be composed of mandates, party positions, social statuses, or of composite states, such as degrees of political involvement or of political success. At the organisational level (parties, trade unions, social movement organizations, think tanks), the alphabet may rely on *ad hoc* scales of development, or on measures of the diffusion of the organization's agenda in public opinion and government spheres. Sequences in political sociology may depend on a life-long perspective, but also on a shorter view on transitional periods, for example when actors enter or leave the political field, or when they convert their resources from one sector to the other. On a macrosociological perspective, sequence analysis may help understand and classify the long-term evolution of regimes or their transition to democracy. Similarly, geopolitical sequences may be defined as the stages national, infra-national or supra-national levels entities go through during crises.

Another domain of political science with interest for sequences is comparative studies. For example, standard stages in the diffusion of reforms may be observed and compared between countries or regions. In election studies, stages of opinion formation during campaigns can be compared across time, across space and between geographical levels, through measures of the participation probability and of the vote itself. Finally, any written or spoken discourse being a string of characters, words, sentences and paragraphs, SA provides new insights into the strategic succession of rhetorical steps, or in specific combinations of themes.

In the list of empirical cases above, most have not been identified as sequences, or have been treated with little systematicity. Moreover, the list is not closed. To date, five applications of SA that can be deemed close to political science. Abbott and Deviney (1992) studied the historical timing in which the old industrial countries passed five main social laws - worker's compensation; sickness and maternity benefits; old-age, invalidity and death supports; family allowances; and unemployment insurance. SA makes it possible to assess three main hypotheses proposed in the previous literature, respectively of individual-level effects, of the diffusion of welfare programs and of a world-historical process. Stovel (2001)'s investigation of lynching sequences goes beyond single historical narratives, showing that SA can extract sequences of violence with similar timing inside counties of the South. Lemercier (2005) proposed a classification of the careers of French economic elites in the 19th century, according to their successive and/or simultaneous positions in the main national regulatory authorities. Fillieule and Blanchard (Blanchard 2005, 2010; Fillieule and Blanchard forthcoming) deal with the careers of social movements activists. Using a retrospective individual survey among the present and former members of French organizations mobilizing against AIDS, they examine militant trajectories in the context of parallel private and public trajectories. The last SA study of interest for political science is Buton, Lermercier and Mariot (under review)'s. They used signature lists from a French polling station to track the voting behaviour of 1,800 persons over 44 ballots. Participation rate appear to follow homogeneous sequential patterns at household level.

These published studies are far from covering the whole range of potential applications. Other fieldworks are in progress: a study of 42 regime crises, measured through a yearly democratic index composed of the degree of political participation, the openness and competitiveness of executive recruitment, and constraints on the chief executive (Casper and Wilson); a research project about the trajectories of French Jews deported by Nazis during the Second World War (Zalc); a study on the careers of Spanish ministers (Real-Dato) and a comparative study of the careers of former students of the London School of Economics and of the Paris Institut d'Etudes Politiques (Scot). Applications should grow more diverse over the coming years, the formal criteria a population of N sequences should satisfy to be relevant for SA being little restrictive:

1. N should be reasonably low, relatively to calculation possibilities (for example, N = 10,000 in Robette 2010). For heavy SA procedures, a population of excessive size may be sampled, sequence analyzed, then reunited.

2. N should be high enough to justify the cost of an automated treatment (N = 16 in Abbott and Deviney 1992).

3. The alphabet should be reasonably small compared to N, so as to avoid excessively idiosyncratic cases.

4. Sequences may contain a reasonable proportion of missing and uncertain elements.

5. Sequences may be simultaneous, if they are not, they should have some kind of synchronization, either historically or biographically or along any another relevant time axis.

6. Sequences may be the same length or not. If they are not, either length variation can be neutralized, or used as a discriminating factor between sequences.

Alongside these empirical criteria, the analyst should naturally also consider theoretical relevance: the research question has to deal with timing, duration, order, in sum, with sequencing.

4. Competing methods for longitudinal data and their limits

One may try to treat sequence data by means of usual multivariate statistical methods: regression model on successive (synchronic) time-points, correspondence/factor analysis, and cluster analysis. Indeed, time is included in these approaches, with multiple time-points, in the form of historical or biographical moments. The main weakness of all these approaches is that each time-point is independent. Hence, duration and order are ignored. None of these methods keeps the individual sequential logic, that is, sequences made of specific successive steps, with specific durations and in a specific order. Regression models on several time-points $t_1 \ldots t_T$ including time as an ordinal independent variable do account for time flow, but just as a linear factor of the global causal model with conclusions on $[t_1 \ldots t_T]$.

Specific longitudinal methods are better candidates to process sequences. In event history analysis (EHA), time is treated as the length before the transition occurs. For example, one might predict the time needed before a member of a party or a social movement organization leaves, given this member's profile: age, sex, socioeconomic status, activity inside the organization, etc. Duration is taken into account with high precision, including censored individuals, but one main drawback: the

focus being on the transition from one category to another, the complexity of successive steps in a trajectory, in terms of nature, order and duration, is neglected.

Time series analysis (TSA) takes time seriously, inasmuch as what happens at time t_T to variable v_1 depends on what happened at $t_{1...T}$ to $v_{1...p}$. Autoregressive, integrated, moving average models, in particular, involve autoregressive mechanisms, time lags and moving averages. Yet, this method deals with the evolution and correlation of strictly continuous variables on very frequent time points, while many social and political sequences are categorical and cannot be observed more than a few tens of times. Moreover, TSA aims at finding a unique, "simple stochastic generator that effectively fits an entire sequence" (Abbott 1995: 104). This would mean that sequences are the consequence of an underlying process, an hypothesis most analysts in the social and political sciences might not want to make.

As a whole, the alternative methods discussed above depend on a restrictive approach to longitudinal data. The individuals/variables structure of the general linear model fundamentally limits the understanding of narratives, that is, phenomenons that develop along time. Mathematically, the space of all possible sequences is largely empty because "most things that could happen don't" (Abbott 2001: 16, 166-169). This emptiness justifies a procedure that searches for local orders inside the space of sequences, instead of disassembling the states/events into "combinations of (supposedly) independent variable properties", as the linear model does (*ibid.:* 169). The real sequences contained in a given population can be compared to fish swimming in a lake. A systematic three-dimension search, like a randomly inspecting sonar would do, is not fruitful because fish follow a limited number of routine paths around weeds and rocks at some preferred depth. It is more efficient to track the fish, study their preferred paths and see how these paths distinguish from each other. As a consequence, the data structure is not made of individuals and their attributes (the variables) but of events/states that occur to individuals, with a strong accent put on the context of these occurrences. Moreover, instead of analyzing a sequence point after point, SA treats it as a whole, thus respecting the global dynamic of the trajectory.

5. Principles of the optimal matching algorithm

Let us consider a population of *N* sequences with an alphabet of *p* states. The OM algorithm can be summed up as follows:

a. The algorithm compares sequences by pairs and associates a *dissimilarity index*, or distance index, to each pair. Dissimilarity is the sum of all differences between the two sequences. For example, the

five-state sequences S_1=<abbbc> and S_2=<abccc> differ regarding steps 3 and 4: <bb> *vs* <cc>. Their dissimilarity would be 2, out of a maximum of 5.

b. The *cost* of the comparison (replacement) of a state b in S_1 by a state c in S_2 may vary, according to the degree of dissimilarity attributed to b and c. For example, if S_3=<abacc>, cost(b,a)=2 and cost(b,c)=1, then $d(S_1,S_3)=2+1=3$. This elementary operation is called a *substitution*.

c. Substitution costs *SCs* are gathered in a *substitution cost matrix SCM* of dimensions p x p. For practical reasons, SCM is symmetric along its main diagonal: cost(a,b)=cost(b,a). The main diagonal is only made of 0s (cost(a,a)=0), the cost of replacing a state by itself, called a *match*. Several methods help set up SCM. One may use exogenous knowledge about the cost of transitions, that is, of moving from a state to another. This solution is the most satisfying theoretically, it give more substantial value to the calculation of distances. One may also compute SCs from the transition rates in the data or in a reference population: the more frequent a transition (a move from state a to state b), the lowest its cost. This method is the most logically consistent source of SCs, especially if the knowledge about objective transitions is insufficient. A last method keeps all SCs constant, for the sake of methodological and theoretical economy. It leads to satisfying results under certain restrictive empirical conditions.

d. One may introduce time lags in the comparison. For example, if S_4=<aabac>, S_3 and S_4 share the same subsequence <abbb>, which makes them very obviously similar, except S_4 is one state lagged. Instead of three substitutions in steps 2, 3 and 4, we may add an *a* as first state and delete fifth state *c*, with a cost of $d(S_1,S_4)=1+1=2$. Insertions and deletions of states are abbreviated as *indels* and indel costs as *ICs*. Using indels, which is called *aligning* the sequences, gives a global character to the comparison, by catching common patterns, whatever their location in the sequences. Two sequences appearing very dissimilar from a year-by-year viewpoint but sharing a large proportion of common subsequences might generate a low distance by means of a few ICs.

e. Substitutions account for the nature of the states experienced, while indels account for the varying moment at which common patterns occur. Balancing SCs and ICs means weighting these two dimensions of time. SCs usually vary between states (see c) but ICs are usually constant. Both substitutions and indels account for duration: the longer an episode, the more elementary operations are used, either SCs or ICs, and the more the distance grows.

f. If distinct series of operations compete to compare two sequences, the algorithm selects the cheapest series. This is why the algorithm is called the *optimal matching algorithm*. Cheap comparisons often come out of sequences that share common subsequences, either successive or not. The Needlemann-Wunsch algorithm, invented for the comparison of proteins' structure, provides a rigorous and quick

calculation of the optimal comparison. It is based on reverse selection of cheapest way inside the diagram of all successive cheapest operations (Needleman and Wunsch 1970). The optimal distance results of a trade-off between substitutions and indels, weighted by the balance among SCs.

g. Alternative algorithms for the comparison of sequences exist. The following changes have been proposed: using other elementary operations, such as swaps, that transform <ab> into <ba> (Dijkstra and Taris 1995), or using only substitutions in order to preserve timing and avoid the temporal distortions of alignment (Lesnard 2010); modifying the costs, such as giving a negative cost to matches so that long common patterns are rewarded by a smaller distance; making ICs vary along time, for example according to transition matrices that shift with the context (Lesnard 2010) or by adjusting SCs locally in the sequence (Hollister 2009); sophisticating the calculation of SCs from the data at hand, for example by optimizing the ratio of matches (Gauthier et al. 2010); instead of aligning sequences, ignoring non common codes and calculating the size of the largest nonsuccessive common subsequence (Elzinga 2003). These options have spread less than the standard model, whose emergence resulted from several intertwined reasons: alternative options make the calculus more complex statistically and heavier for computers; they have not been implemented on standard packages; they limit the control of the analyst on the algorithm; and, the improvement they bring to the method has not been tremendous enough to convince sequence analysts to invest time in them yet. With time, when more scholars will join the SA commmunity, bringing in more diverse empirical fields and more diverse theoretical concerns, these alternatives might get more success.

6. Step-by-step implementation of sequence analysis

This section summarizes the main steps of standard SA, as they are usually implemented before and/or after the application of the OM algorithm (fig. 1). We use as an illustration an application to the members of the main French organization committed to the AIDS cause, *Aides*, based on data collected among present and past members of this organization (Broqua and Fillieule 2001. The project adopted the perspective of symbolic interactionnism in order to understand how activists engage in *Aides*, abide by it and eventually leave it. At each stage of a career (Becker 1966), attitudes and behaviours are determined by past attitudes and behaviours, and in turn condition future possibilities, thus resituating the periods of commitment in the entire life cycle (Fillieule 2010). Such an perspective on careers had been implemented before by means of monographs and life interviews, not yet through statistics.

a. Collecting sequence data

Sequence data may be collected from administrative or organisational archives (Stovel et al. 1996; Lemercier 2005; Buton, Lemercier and Mariot [forthcoming]), from the aggregation of ethnographic studies (Forrest and Abbott 1990), from national family/household panels (Billari and Piccarretta 2005), from one-shot retrospective surveys (Robette 2010), or from the aggregation of separate inventories (Abbott and DeViney 1992; Stovel 2001). Any sources of diachronic data may be relevant, as long as they be systematic enough over one period, over one given age-range or any other time window. Remembering dates and setting up a biographical calendar being a demanding task, both intellectually and emotionally, for respondents, survey designs should ensure that the questionnaire makes the best out of their memory and compliance. The "life history calendar" (LHC) is helpful in this respect. Instead of a long and sometimes redundant list of questions, the respondent is invited to fill in a table-like calendar, with years or months as columns and life domains as lines. Combined with the use of colours, drawings and symbols, the LHC has been shown to improve the quantity and quality of the answers (Freedman et al. 1988; Glasner and van der Vaart 2007).

We use a retrospective longitudinal self-administered mail survey of members of Aides. Thanks to the access to the member directory granted to us by the organization, we sent our questionnaire to all the 1,969 members of Aides Ile-de-France (central region around Paris) in 1998. The response rate was decent (25%, N = 502). We contacted not only current volunteers (response rate: 33%) but also former volunteers (r.r.: 20%). Their inclusion gives us a view which includes terminated, fully informed commitments, as well as unfinished, statistically "censored" trajectories. We had the opportunity to test the reliability of our sample with data from Aides' archives. All representativity tests on sex, birth year, date of entry and date of exit give satisfying results: monovariate distributions are highly tied between survey sample and population and the hierarchy of bivariate correlations are similar enough in the sample and in the population. We did not use the LHC, but response rates to questions asking for calendar-years were quite high and of good quality.

b. Coding sequences

A good alphabet is an alphabet that faithfully lists all possible states/events in the trajectory under study. No blank codes should remain, except missing values. Memory mistakes (surveys) or recording mistakes (archives) should be tracked and, if possible, corrected. If not possible, they should be interpreted: do they come from memory failures, refusals for moral or political reasons, misunderstanding of the questionnaire, lack of time, fear for confidentiality? Missing or uncertain

values with recurrent interpretations should be kept in the alphabet as "weak" (less reliable) codes, rather than be ignored. As in any coding, one given code should carry the same meaning for every individual, especially in the case of subjective states, such as questions of opinion or evaluation.

In the case of a fuzzy code, that is, when an individual is half *a* and half *b*, the analyst has the possibility to a create specific, fuzzy code *ab*, as long as its interpretation makes some sense inside the alphabet. The SC attributed to pair (*ab,c)* could be mean[cost(a,c),cost(b,c)]. No definitive solution has been proposed to nonordered multiple response variables, for example, if some employees hold two positions p_1 and p_2 at the same time. One may either drop the lowest position, or create a synthetic code $p_{1,2}$. SA with multiple simultaneous sequences has been implemented (MSA, i.e. multiple sequence analysis - see section 8) but it is applicable only if sequences made with second (third…) answers are fully known. In the case of selective second positions, the blank codes attributed to single-position employees do not allow the MSA option.

The career of members of Aides combines two kinds of information: being engaged/not yet/not anymore; the function and the degree of involvement, measured by the time spent for the cause. The alphabet is made of twelve codes: being less than eighteen years old; not being involved yet; being in relation with the organization but not active; being involved with a low/intermediate/high degree of involvement; being a manager; being a paid employee; having left temporarily; having left but keeping contacts in the organization/with no contacts; no answer. This alphabet is not fully hierarchical: sequences with the same codes may have a different order, resulting in rather complex combinatorics of states.

c. Exploring sequences

Several kinds of graphs help explore the data at hand. The year-by-year aggregated representation gives a first glimpse at proportions between elements of the alphabet. In the case of Aides (fig. 2), we notice the exponential development of the association, since its creation in 1984, the dominance of non managing, non paid positions (first three shades of blue) and the high proportion of exits from 1994 on (green). The next graph (fig. 3) represents individual sequences. A multidimensional scaling (MDS) procedure enables some raw order to emerge from otherwise hardly readable data. In spite of the theoretically high combinatorics, a limited number of sequences are really realized. Intense involvements in the organization (top part of the graph) and former members (bottom) oppose shorter, less committed involvements (middle). This elementary sorting through MDS gives a useful exploratory glimpse at biographical profiles (Piccarreta and Lior 2010). Yet it does not account for the order of states: based on a year-by-year (column-by-column) comparison of

sequences, MDS fails to express biographical similarities, that is, time-lagged common patterns. The same output, left-aligned with the year of entry in Aides (fig. 4), has the advantage of bringing together careers with similar beginnings. But it does not account properly for common patterns, such as two or three years of intense activity surrounded by more superficial periods, that would happen after time lags that vary across individuals. These exploratory graphs may nonetheless help (re)frame research hypotheses.

d. Performing the optimal matching analysis

Following the pragmatic approach to costs favoured in section 6, we test the different options and finally chose objective SCs, that is, costs that reflect "objective" transitions. In our case, we already had information about what being in and out, or holding this or that position meant from interviews with members and former members. We conceive of costs as a mix of the meaning actors give to their past, present and prospective statuses, belongings and actions, and of the objective knowledge the analyst and some actors share about these statuses, belongings and actions. This way, the algorithm will both account for the indigenous logics of activists and for our theoretical questions about activism.

Some authors consider that SCM should respect the triangular inequality: for any three costs ab, bc and bc, we should have: $ab \leq bc + ac$. This principle of a two-dimension metrics as not been fully demonstrated, but for the sake of consistence, we suggest that the alphabet be as close as possible to a scale. In the case of members of Aides, the degre of organisational involvement and the statuses do more or less generate a hierarchical scale. But the periods before and after engagement are outside of this scale. As a consequence, SCs are adjusted at best so as to mark the distance between being in and out Aides, and between the different statuses in. We start by the most simple pairs of codes, the ones that form a pseudo-scale: substitution of low by intermediate and intermediate by high investment costs 1, high investment by full-time or management position costs 2 (fig. 5). Jumping two or three of the steps of the scale costs the corresponding addition of one-step costs. Then, we set not born/minor and not engaged yet codes at even, low distance from in-the-organization codes, so that they do not weight too much in the analysis. Exit codes are set at higher distance from in-the-organization codes, because leaving is a more "costly" move than entering, especially for intensely engaged people. Unknown codes are set at null distance, so as to neutralize them as much as possible. The resulting SCM is necessarily approximative, as much as our theory to assess all costs, but it gives a meaningful account of which moves should make major differences between two sequences, and which ones should make minor differences.

e. Interpreting the distance matrix

OMA outputs an N*N dissimilarity (or distance) matrix *DM*. Like SCM, DM is symmetric along a main diagonal made of 0s. DM creates a structure inside the population of sequences that is formally similar to a network structure. Each sequence is more or less linked with all other sequences. Identical sequences have a distance of 0, they are at the same place in the structure. Pairs of sequences composed of very different states, and/or in very different proportions, and/or in a very different order, will be far away in the structure.

We now have to explore and synthesize this space. Several methods can be applied to DM, among which network analysis, cluster analysis, multidimensional scaling, and correspondence/factor analysis. Clustering is the dominant method in the SA literature, usually by means of a hierarchical, ascending model with square euclidean distances and Ward aggregation algorithm. The typology produced by clustering is easy to represent, interpret and to reuse as a dependent or independent variable. But one may use network analysis in order to represent graphically the whole population and to visualize the respective position of this or that individual, this or that cluster. Multidimensional scaling is a rough but straight way to introduce order into a graph of individual sequences (see fig. 3-4). Correspondence analysis heps unfold and describe the cleavages that structure the population of sequences. It also provides useful two-dimension graphs where clusters may be drawn around clouds of individuals. Besides, clusters may be described and/or explained through cross-tabulation and logistic regression models. Explanatory models may use variables both endogenous and heterogeneous to longitudinal data. The interpretation of biographical sequences typically relies on social and economic resources, ideological attitudes, socioeconomic status (SES), and so on.

In our example, cluster analysis applied to DM distinguishes seven groups of trajectories, represented in separate individual plots in figure 6. All clusters appear homogeneous with regard to the nature of states, their length and order, and the length of whole careers. SA does account for the richness of social time. At the same time, the heterogeneity between clusters is obviously high. The most intense involvements concentrate in two clusters, one with old, long careers (c3) and one with more recent, shorter ones (c6). Cluster 5 is composed of early members, some disengaged at survey time, some with very long engagement. At the opposite, c1 gathers short, lighter involvements, unended at survey date (1999). Clusters c2 and c4 are intermediate, mostly made of ended trajectories, either very short (c4) or longer with possibly a year of more intense activity (c2). A few careers with missing values compose the last cluster (c7). Cross-tabulations help understand the reasons why these clusters distinguish themselves from each other. Factors include the evolution of the organization's

recruitment policy, the kind of persons who volunteered to enter it, the public perception of the AIDS cause and of people living with AIDS (PLWA), the medical advances, the perception of homosexuals and homosexual activism, the creation of a multi-organizational field and the government's health policy and its consequence on funding.

Let us just point at one crucial factor of differentiation between Aides careers: generations (Fillieule and Blanchard, forthcoming). Some activists are young male homosexuals who feel close to PLWA through affective and cognitive proximity (c5). During the first two years of its existence, recruitment is through cooptation and mutual acquaintances, securing a very strong social and ideological collective identity. Gradually, the epidemic expands, more people volunteer and the increasing ideological heterogeneity produces a split in March 1987 and the subsequent departure of some of the founding members (c5 - bottom). Public authorities start getting involved against the epidemic in 1987, triggering the creation of other associations. From 1989 on, the movement opens up a new dynamic in homosexual activism.

During the second phase, from 1988 to the early 1990s, public authorities get more involved. Specialized public agencies are created. The public understands that not only homosexuals and drug users are affected. New associations criticize the inaction of public authorities. Act Up-Paris is the vanguard of the expression of a public word for PLWA and the demand for recognition of the homosexual identity. Simultaneously, the AIDS cause diversifies and generalizes, which translates into the arrival of heterosexual women, participating in a spirit of solidarity, motivated by their professional involvement in the (traditionally female) field of health and social services. Reinforcing existing cleavages that were latent during the initial phase of the mobilization and generating new lines of division, leading to a new wave of exits among volunteers (c2).

In the next three years, AIDS turns into a chronic affair. The management of AIDS is consolidated and professionalized, two to four-year full-time jobs are created (c3 - red lines). A vigorous growth provokes tensions, especially about the massive hiring of new employees with questions about the division of labor between volunteers and professionals (c2-c3). During the next phase, the cause of AIDS moves to the normality stage. New therapies (the HAART, proposed from 1996-1997) dramatically extend life expectancy, offering new perspectives to PLWA. Individuals affected refocus their attention on day-to-day concerns. HAART turn the public image of the epidemic from that of a fatal illness to that of a chronic disease. This results in a certain demobilization and in short, low-intensity involvements by younger activists with a lower affective proximity to AIDS (c4). Besides, long-term activists, often burnt-out, partly or completely disengage. With Aides getting involved into "parallel" causes (immigrants, prisoners, etc.) and in the issue of gay and lesbian civil rights, in the context of the large social movement of 1995, groups inside the organization compete

with each other (c4-c6). At the end of the 1990s, Aides is less cohesive but the major generational clash has been avoided, thanks to an efficient leadership and management, to an active training of newcomers and to the departure of militants with less organisational resources.

Besides the connection of biographical sequence patterns with the organisational and historical dynamics of the cause, two clusters deserve additional comments. C1, made of very homogeneous, two-to-six years, low-intensity commitments, is part of the generation of activists attracted by a "normalized" cause. But a large majority of them are still engaged: their trajectory is statistically right-censored. C7 gathers only nine sequences dominated by missing values. In spite of their formal incompleteness, c1 and above all c7 were rightly kept inside the analysis. The SA/clustering treatment demonstrates their specificity, instead of excluding them from the analysis.

7. Complementary tools and advanced issues

The seven principles presented in section 6 frame the standard OMA practices. To date, they have been used successfully in about a hundred empirical studies. The other tools presented above - raw and ordered graphs, clustering - are also part of the success of the method. By lack of space, let us present briefly here the other elements of the SA package.

Quantitative statistics specific to SA provide an indication about the empirical diversity of sequences. They may treat individual sequence *s*: *entropy* (inspired by Shannon's work on the English language) measures the diversity of states experienced in *s*; *complexity* depends on the quantity of non successive subsequences inside *s* and on the variance of their length (Elzinga 2010). One may also consider a set of sequences *S*: *ANOVA* applied to the distance matrix DM shows how heterogeneous the elements of *S* are (Studer et al. 2009); *entropy* measures how heterogeneous *S* is at each time point. These quantitative indicators fit to research hypotheses about measurable phenomenons, such as the degree of destandardization of modern life course or the historical diversification of trajectories of voters or of involvements in political organization. They are less useful in an interactionnist perspective.

The search for *prototypical sequences PS* is a powerful way to summarize a cluster of sequences. PSs may be either real, that is, extracted from the population at hand according to average features of the cluster (Stovel et al. 1996), or fictitious, that is, aggregations of dominant features of the cluster. They may be chosen automatically according to certain statistical parameters, such as the mean OM distance to other sequences (Aassve and Billari 2007), or manually from a visual inspection of the cluster. Figure 7 gives en example of PSs representing six clusters out of a ten-cluster typology

of activist careers in Aides. The typology is built on a three-dimension SA, adding the sociosexual career and the career relatively to AIDS to the calculation of OM distances. Each PS was chosen manually inside the real population so that the nature of the states, their length and their order, as well as the length of the career, represented at best each cluster. This graph is a relevant abstract of exhaustive sequence plots when the sample is large.

Sequence mining complements the SA package. It enables retrieving specific subsequences, defined by the user, or to select the sequences that discriminate a population or a subpopulation. Discriminating subsequences help identify PSs inside a cluster. As a whole, clusters resulting from a typology can be finely described by means of each cluster's individual graph, some sequence statistics, prototypical sequences and the list of frequent subsequences.

Practices of SA still differ on several aspects of the method. Regarding the general methodological strategy, authors are shared between putting the accent on theoretical rigour, on statistical consistence, on pragmatic sequence mining and on inference. Nonetheless, the dominant use of SA is a cautious practice with descriptive purpose, before some causal model, inference and validation statistics are applied. The clustering approach typically fits in this descriptive perspective, as well as the importance given to a proper coding. Meaningful results are considered a legitimation of SA, as much as more formal justifications, such as a full demonstration of the OM algorithm, the choice of the most relevant costs or the clustering options that make the best out of the distance matrix.

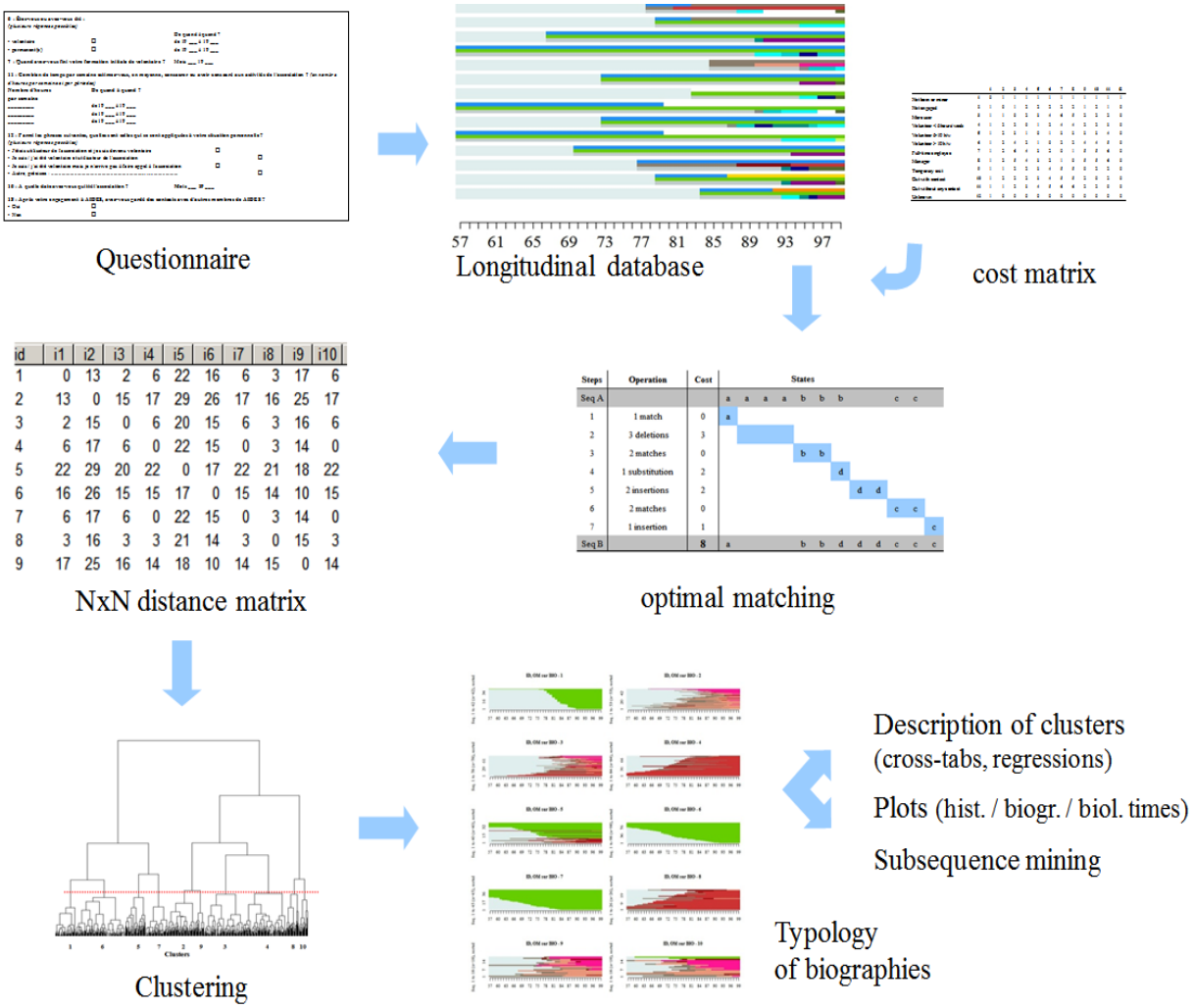
More could be done, by means of bootstrapping and simulation, in order to make the results more robust, that is, less sensible to some options taken during coding or treatment. Advanced issue also include multidimensional sequence analysis (e.g. Gauthier 2007; Gauthier et al. 2010; Pollock 2007; Blanchard and Fillieule, in progress) and visualization of sequences (e.g. Müller et al. 2008; Piccarreta and Lior 2010). The framing of time is still an open issue (should one focus on sequences of the same length? or should the length of sequences be standardized for OMA and reintroduced in the analysis subsequently?), as well as the optimal costs and the meaning of the OM metrics. Behind several of these topics, sequence analysts are shared on a larger alternative: is SA used as a pragmatic, somewhat blind mining technic, to which all means fit as long as they bring order into the population of sequences? or is it a theoretically loaded method, whose different steps need to be optimally formalized?

Let us conclude with a review of software adapted to SA for social sciences. Some generalist programs implement many of the functions presented in sections 6 and 7. The *TDA [Transition Data Analysis]* package (Rohwer, Poetter, 2007) was the first but is not maintained anymore. In *Stata*, the

ado-package SQ (Brzinsky-Fay et al. 2006) and the *seqcomp* plug-in (Lesnard and Kan 2009) are available. The *Trajectory Mining in R* package (*TraMineR* - Gabadinho et al. 2009) enables formatting sequence data, sequence plotting, sequence comparison, diverse sequence statistics and sequence mining. Other software bring added value to more specific tasks, like visualizing the optimal matching calculation (*Optimize*: Abbott 1997) or implementing "optimized" OM (*Saltt* package in *T-Coffee*: Notredame et al. 2000; Gauthier 2007). All treatments presented here were realized with *TraMineR*[1], which provides a complete set of tools, in a rich and rapidly evolving environment.

[1] For more information on TraMineR and empirical applications, see http://mephisto.unige.ch/traminer/. A special thank to Matthias Studer, who helped us realize figure 7.

Figure 1: Standard general design of sequence analysis



id	i1	i2	i3	i4	i5	i6	i7	i8	i9	i10
1	0	13	2	6	22	16	6	3	17	6
2	13	0	15	17	29	26	17	16	25	17
3	2	15	0	6	20	15	6	3	16	6
4	6	17	6	0	22	15	0	3	14	0
5	22	29	20	22	0	17	22	21	18	22
6	16	26	15	15	17	0	15	14	10	15
7	6	17	6	0	22	15	0	3	14	0
8	3	16	3	3	21	14	3	0	15	3
9	17	25	16	14	18	10	14	15	0	14

Steps	Operation	Cost	States
Seq A			a a a a b b b c c
1	1 match	0	a
2	3 deletions	3	
3	2 matches	0	b b
4	1 substitution	2	d
5	2 insertions	2	d d
6	2 matches	0	c c
7	1 insertion	1	c
Seq B		8	a b b d d d c c c

Figure 2: Careers in *Aides*. Yearly frequency plot, from 1980 to 1999 (N=502)

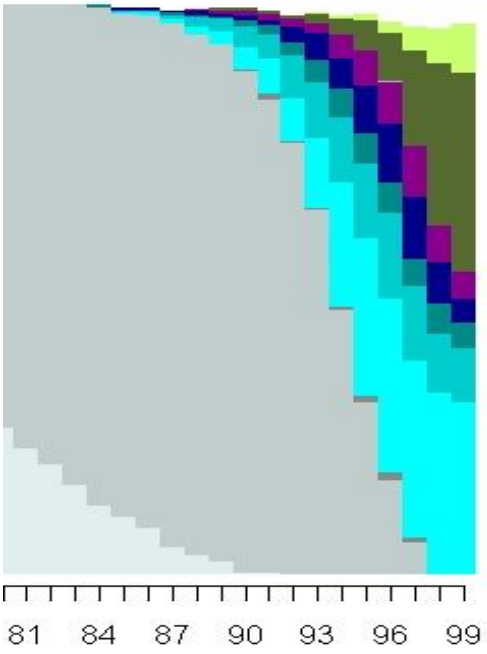


Figure 3: Careers in *Aides*. Individual sequences, ordered by multidimensional scaling, from 1980 to 1999 (N=502)

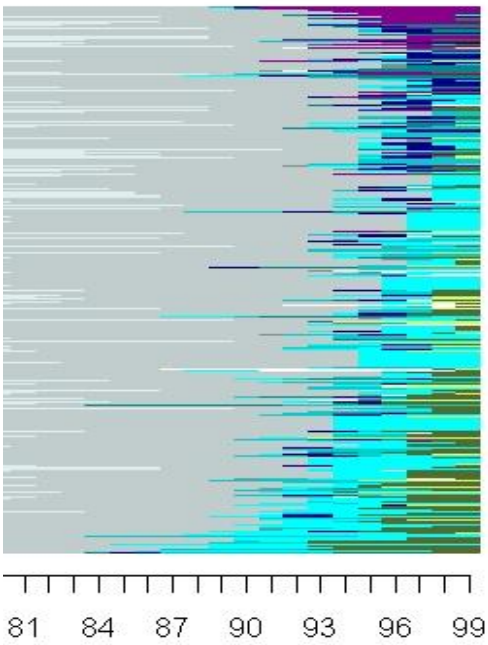


Figure 4: Careers in *Aides*. Ordered by multidimensional scaling, from entry in Aides (E) to year of survey (1999) (N=502)

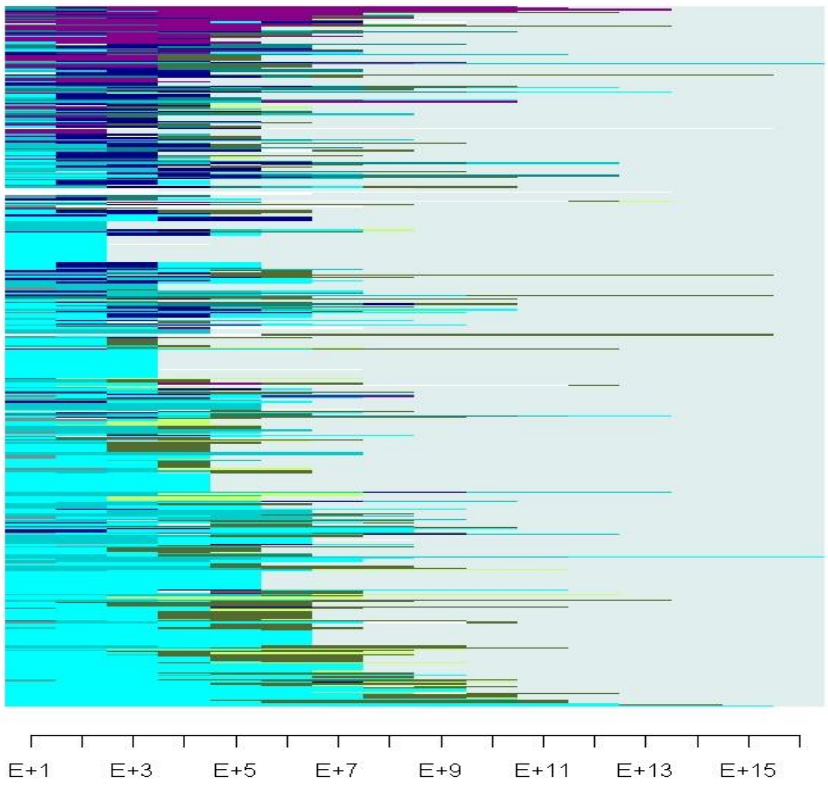


Figure 5: Substitution cost matrix based on objective knowledge about transitions

		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Not born or minor	**1**	0	1	1	1	1	1	1	1	1	1	1	1
Not engaged	**2**	1	0	1	2	2	2	2	2	1	2	1	0
Mere user	**3**	1	1	0	2	3	4	6	5	2	2	2	0
Volunteer < 6 hours/week	**4**	1	2	2	0	1	2	4	4	2	2	3	0
Volunteer 6-10 h/w	**5**	1	2	3	1	0	1	3	3	3	3	4	0
Volunteer > 10 h/w	**6**	1	2	4	2	1	0	2	2	4	4	5	0
Full-time employee	**7**	1	2	6	4	3	2	0	1	5	5	6	0
Manager	**8**	1	2	5	4	3	2	1	0	5	5	6	0
Temporary exit	**9**	1	1	2	2	3	4	5	5	0	2	2	0
Out with contact	**10**	1	2	2	2	3	4	5	5	2	0	2	0
Out without any contact	**11**	1	1	2	3	4	5	6	6	2	2	0	0
Unknown	**12**	1	0	0	0	0	0	0	0	0	0	0	0

Figure 6: Clusters of careers in Aides resulting from optimal matching analysis, ordered by MDS in historical time.

Objective substitution costs from -1 to 6, fixed insertion-deletion costs of 2, hierarchical ascending cluster analysis with Euclidean distance, aggregation with Ward algorithm

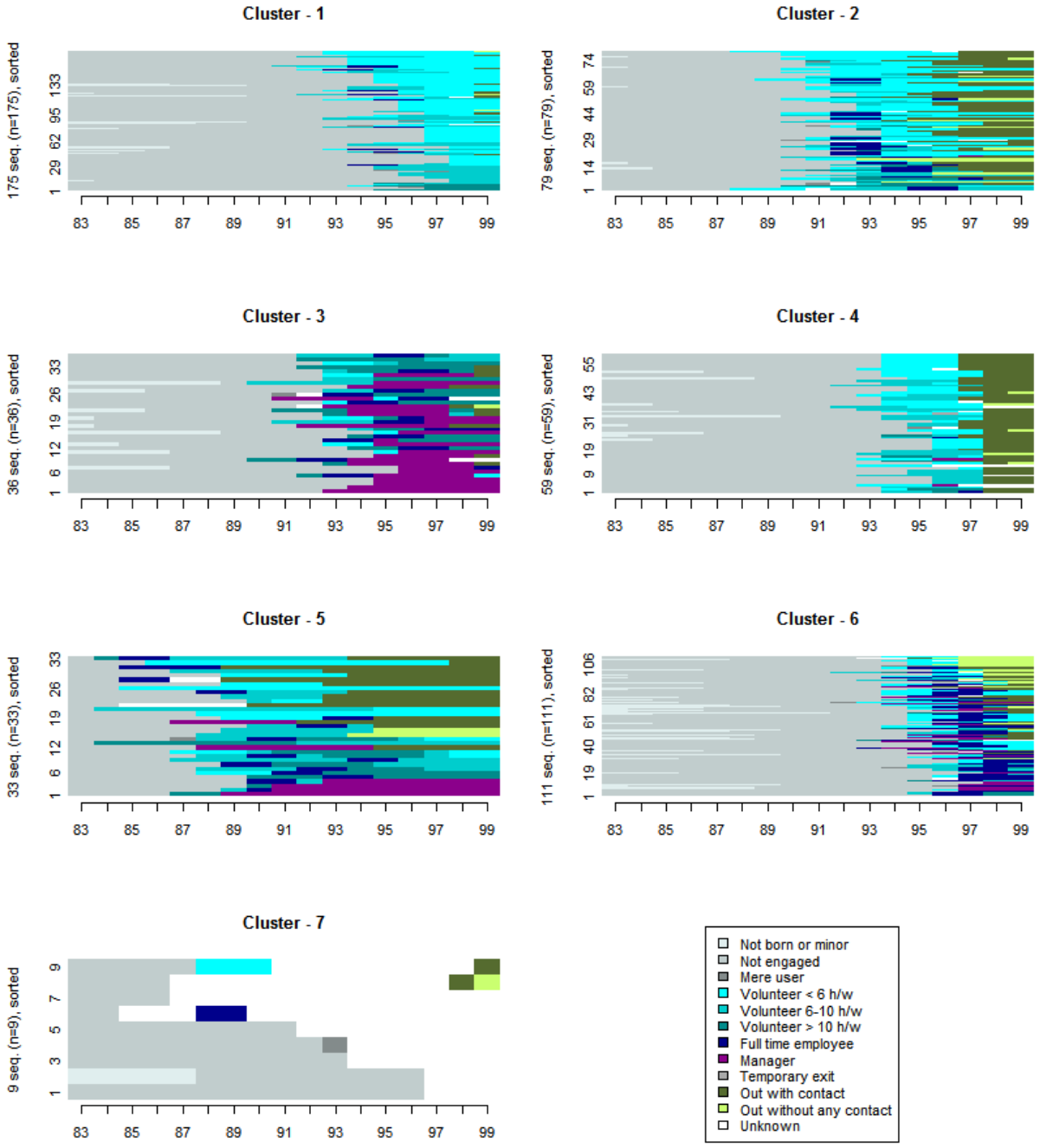
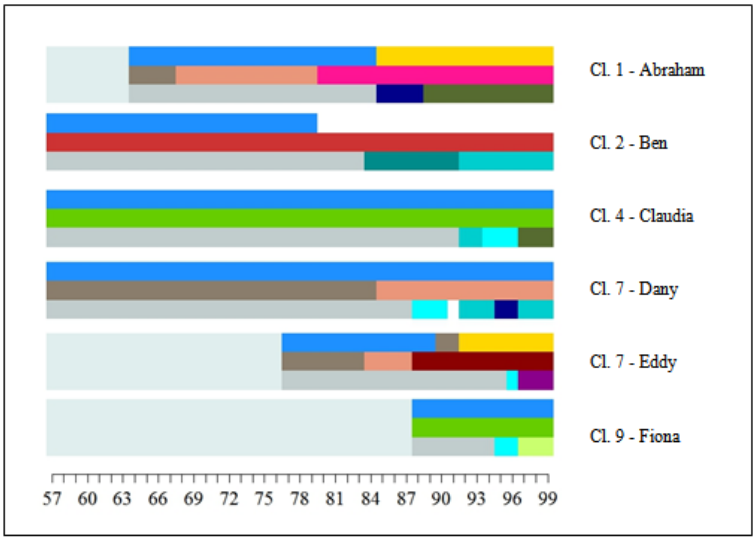
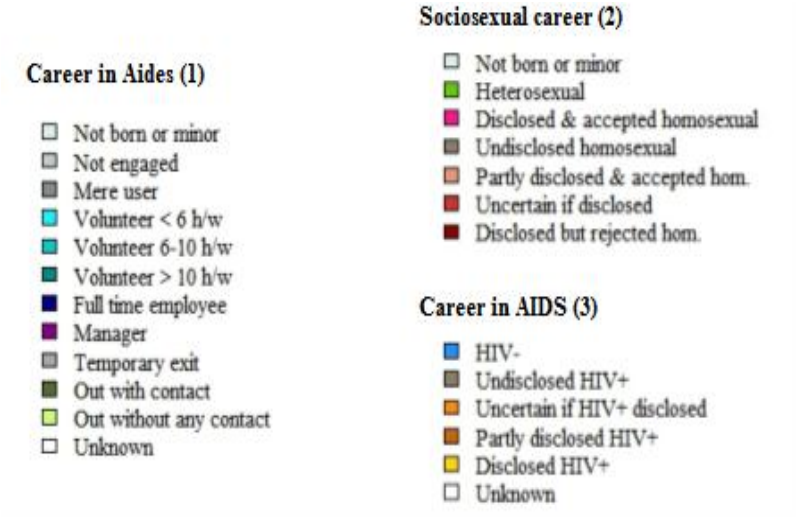


Figure 7: Prototypical activists of six clusters out of a ten-custer typology
Real activists chosen among clusters according to cluster-representative nature, length, and order of states in three-career biographies. Names are fictitious.



Legends



References

AASSVE Arnstein, BILLARI Francesco, and Raffaella PICCARRETA. 2007. Strings of Adulthood: A Sequence Analysis of Young British Women's Work-family Trajectories. *European Journal of Population* 23:369-88

ABBOTT Andrew. 1983. Sequences of Social Events: Concepts and Methods for the Analysis of Order in Social Processes. *Historical Methods* 16 (4): 129-147

ABBOTT Andrew. 1991. The Order of Professionalization. *Work and Occupations* 18: 355-384

ABBOTT Andrew. 1992. From causes to events: Note on Narrative positivism. *Sociological Methods and Research* 20: 428–455

ABBOTT Andrew. 1995. Sequence Analysis: New Methods for Old Ideas. *Annual Review of Sociology* 21: 93-113

ABBOTT Andrew. 1997. *Optimize*. http://home.uchicago.edu/~aabbott/om.html#optimize

ABBOTT Andrew. 2001. *Time matters: on Theory and Method*. Chicago: University of Chicago Press, 2001

ABBOTT Andrew and Emily BARMAN. 1997. Sequence Comparison via Alignment and Gibbs Sampling. *Sociological Methodology* 27: 47-87

ABBOTT Andrew and Stanley DEVINEY. 1992. The Welfare State as Transnational Event: Evidence from Sequences of Policy Adoption. *Social Science History*. 16 (2): 245-274

ABBOTT Andrew and John FORREST. 1986. Optimal Matching for Historical Sequences. *Journal of Interdisciplinary History*. 16: 471-494

ABBOTT Andrew and Alexandra HRYCAK. 1990. Measuring resemblance in sequence data : an optimal matching analysis of musicians' careers. *American journal of sociology* 96: 144-185

ABBOTT Andrew and Angela TSAY. 2000. Sequence Analysis and Optimal Matching Methods in Sociology. Review and Prospect. *Sociological Methods and Research* 29 (1): 3-33

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. *Sociological Methods & Research* 38(3): 420-462

BILLARI Francesco C. 2001. The Analysis of Early Life Course: Complex Descriptions of the Transition to Adulthood. *Journal of Population Research* 18 (2): 119-142

BLAIR-LOY M. 1999. Career Patterns of Executive Women in Finance: An Optimal Matching Analysis. *American Journal of Sociology* 104: 1346-97

BLANCHARD Philippe. 2005. Multidimensional biographies. Explaining disengagement through sequence analysis. ECPR General Conference, Budapest

BLANCHARD Philippe. 2010. *Analyse séquentielle et carrières militantes*. HAL-Open Archives: http://hal.archives-ouvertes.fr/hal-00476193/fr/

BLANCHARD Philippe and FILLIEULE Olivier. In progress. A Descriptive Approach to Multiple-Sequence Analysis.

BROQUA Christophe and Olivier FILLIEULE. 2001. *Trajectoires d'engagement : AIDES et Act Up*. Paris: Textuel

BRZINSKY-FAY Christian, Ulrich KOHLER and Magdalena LUNIAK. 2006. Sequence analysis with Stata. *The Stata Journal* 6(4): 435-460

BÜHLMANN Felix. 2008. The Corrosion of Career? Occupational Trajectories of Business Economists and Engineers in Switzerland. *European Sociological Review* 24 (5): 601-616

BÜHLMANN Felix. 2010. Routes into the British Service Class. Feeder Logics according to Gender and Occupational Group. *Sociology* 44(2): 195-212

BUTON François, LERMERCIER Claire and MARIOT Nicolas (under review). The Household Effect on Electoral Participation. A Multilevel Analysis of Voter Signatures from a French Polling Station (1982-2007).

DIJKSTRA Wil and Toon TARIS. 1995. Measuring the agreement between sequences. *Sociological methods and research* 24: 214-231

ELZINGA Cees H. 2003. Sequence Similarity: A Nonaligning Technique. *Sociological Methods and Research* 32: 3-29

ELZINGA Cees. 2007. *CHESA 2.1 User Manual*. Amsterdam: Frije Universiteit

ELZINGA Cees H. 2010. "Complexity of Categorical Time Series. Department of Social Science Research Methods. Amsterdam: Vrije Universiteit Amsterdam

FILLIEULE Olivier. 2010. Some Elements of an Interactionnist Approach to Political Disengagement. *Social Movement Studies* 9 (1): 1-15

FILLIEULE Olivier and Philippe BLANCHARD (forthcoming). Fighting Together. Assessing Continuity and Change in Social Movement Organizations Through the Study of Constituencies' Heterogeneity, in Kauppi Niilo (ed.) *The New Political Sociology*. ECPR Editions

FORREST John and Andrew ABBOTT. 1990. The Optimal Matching Method for Studying Anthropological Sequence Data. *Journal of Quantitative Anthropology* 2: 151-70

FREEDMAN Deborah, THORNTON Arland, CAMBURN Donald, ALWIN Duane and YOUNG-DEMARCO Linda. 1988. "The Life History Calendar: A Technique for Collecting Retrospective Data. *Sociological Methodology* 18: 37-68

GABADINHO Alexis, RITSCHARD Gilbert, STUDER Matthias and Nicolas MÜLLER. 2009. *Mining sequence data in R with the TraMineR package: A user's guide*. Department of Econometrics and Laboratory of Demography, University of Geneva, 2009

GAUTHIER Jacques-Antoine. 2007. *Empirical categorizations of social trajectories: a sequential view on the life course*. PhD dissertation, University of Lausanne

GAUTHIER Jacques-Antoine, WIDMER Eric, BUCHER Philip and Cédric NOTREDAME. 2009. How much does it cost? Optimization of costs in sequence analysis of social science data. *Sociological methods and research* 38 (1): 197-231

GAUTHIER Jacques-Antoine, WIDMER Eric, BUCHER Philip and Cédric NOTREDAME. 2010. Multichannel Sequence Analysis Applied to Social Science Data. *Sociological Methodology* 40 (1): 1-38

HALPIN Brendan and Tak Wing CHAN. 1998. Classs Careers as Sequences*:* An Optimal Matching Analysis of Work-Life. *European Sociological Review*, 14 (2): 111-130

HOLLISTER Matissa. 2009. Is Optimal Matching Suboptimal? *Sociological Methods and Research* 38: 235-64

LEMERCIER Claire. 2005. Les carrières des membres des institutions consulaires parisiennes au XIXe siècle. *Histoire et mesure* XX (1/2): 59-95

LESNARD Laurent. 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-earner couples. *Electronic International Journal of Time Use Research* 1 (1) : 60-84

LESNARD Laurent. 2010. Setting Cost in Optimal Matching to Uncover Contemporaneous Socio-Temporal Patterns. *Sociological Methods and Research* 38 389-419

LESNARD Laurent and Man Yee KAN. 2009. Two-Stage Optimal Matching Analysis of Workdays and Workweeks. Sociology Working Papers 2009-04, Department of Sociology, University of Oxford, http://hal.archives-ouvertes.fr/halshs-00435422

MACINDOE Heather and Andrew ABBOTT. 2004. Sequence Analysis and Optimal Matching Techniques for Social Science Data: 387-406 in *Handbook of Data Analysis*, edited by M. HARDY and A. BRYMAN. Thousand Oaks, CA: Sage

MÜLLER Nicolas, GABADINHO Alexis, RITSCHARD Gilbert, STUDER Matthias, "Extracting knowledge from life courses: clustering and visualization", *Computer Sciences Proceedings of DaWaK*, Turin, Italie, septembre 2008

NEEDLEMAN Saul and WUNSCH Christian. 1970. "A general method applicable to the search for similarities in the amino acid sequence of two proteins". *Journal of Molecular Biology* 48 (3): 443–53

PICCARRETA Raffaella and Orna LIOR. 2010. Exploring sequences: a graphical tool based on multi-dimensional scaling. *Journal of the Royal Statistical Society Series A* 173, part 1: 165-184

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*, 170, Part 1: 167-183

ROBETTE Nicolas. 2010. The diversity of pathways to adulthood in France: evidence from a holistic approach. *Advances in Life Course Research*

ROHWER Goetz and Ulrich POETTER. 2007. *Transition data analysis*, Ruhr University, Bochum, 2007, http://www.stat.ruhr-uni-bochum.de/tda.html

STOVEL Katherine. 2001. Local Sequential Patterns: The Structure of Lynching in the Deep South, 1882-1930", *Social Forces*, 79, 3, March 2001, pp. 843-880

STOVEL Katherine, SAVAGE Michael and Peter BEARMAN. 1996. Ascription into Achievement: Models of Career Systems at Lloyds Bank, 1890-1970. *The American Journal of Sociology* 102 (2): 358-399

STUDER Matthias, RITSCHARD Gilbert, GABADINHO Alexis & Nicolas MÜLLER. 2009. Analyse de dissimilarités par arbre d'induction. In Extraction et gestion des connaissances (EGC 2009), *Revue des nouvelles technologies de l'information RNTI* E-15 : 7-18.

WILSON Matthew and CASPER Gretchen. 2011. Bargaining within and across Crises. APSA Conference, Seattle