Multiphase Sequence Analysis

Thomas Collas

1 Introduction

Many common notions such as lifecycle, adulthood, turning point or ratchet effect are based on the idea that various sets of sequences—especially individual careers—follow regular patterns defined by successive phases. Those phases that all cases encounter in the same order are usually defined by social moments—such as graduation, medical treatment, childbirth, job promotion, new research project, election as a congressman, Oscar nomination, etc.—or by calendar periods—e.g., minutes, months, years or decades. Such a conception of sub-level temporal structures nested in sequences and linked to one another is a cornerstone of life-course studies (Levy and The Pavie Team 2005). Nonetheless, it has rarely been discussed as a methodological parameter of sequence analysis.

Implementing the notion of phase can in particular deepen our understanding of how institutionalised narratives shape social processes (Abbott 2001). Dividing into phases also helps to reduce data complexity. In this chapter, I elaborate on these ideas by presenting visual and metric tools of multiphase sequence analysis (MPSA). Throughout the text, I will develop an example of two-phase sequences drawn from a study of the careers of participants in professional *pâtissier* (pastry cook) competitions in France.

In the first section, I present key properties of multiphase sequences. In the second section, I exemplify and discuss the implications of the division into phases. In the third section, I present tools to render multiphase sequences, introducing sliced representations. In the fourth section, I introduce a dissimilarity measure of multiphase sequences called multiphase optimal matching (MPOM).

T. Collas (✉)
F.R.S.-FNRS – Université de Louvain, Louvain-la-Neuve, Belgium
e-mail: thomas.collas@uclouvain.be

G. Ritschard, M. Studer (eds.), *Sequence Analysis and Related Approaches*,
Life Course Research and Social Policies 10,
https://doi.org/10.1007/978-3-319-95420-2_9

2 Sequences as Multiphase Structures

The pathbreaking proposition of sequence analysis was to study sequences as wholes (Abbott 1995), constructed as successions of atemporal timeslots.[1] The notion of phase adds an intermediary level to this binary hierarchical structure (sequences and time slots) by introducing discontinuities within these successions.

2.1 Characteristics of Multiphase Sequences

Lesnard and Kan (2011) provide an example of multiphase sequences in their analysis of workweeks. The whole sequences they study are each made up of 672 15-min time slots. Lesnard and Kan represent them as successions of 7 days of 96 15-min time slots with a two-state alphabet (work/non-work). For each sequence, seven successions (days) are nested in a larger succession (a week). This example of multiphase sequences is extremely regular: all sequences are made up of seven phases which contain exactly 96 time slots. It is possible to conceive of more diverse types of multiphase sequences.

First, the phases may be of uneven length within sequences and from one sequence to another. For instance, in France, whenever a new government is formed, many recruits enter the ministerial *cabinets*. Their job careers can be divided into three phases: their positions before entering the *cabinets* (for example, by separating private and public sectors), their positions within the *cabinets* (by differentiating *cabinets* and hierarchical positions) and their positions after they leave these *cabinets* (with the same alphabet as in the first phase). The lengths of the different phases differ because they are not defined according to their duration but with respect to common events that may happen at different moments in a career. Take the first phase for example: some enter a *cabinet* after a 20-year long career whereas others have only worked for three or four years before being recruited. More broadly, the division into phases often implies length differences even for sequences of equal length.

Secondly, some phases may be empty, i.e., of length zero. For example, *pâtissier* competitions suppose a division of careers into two phases: a phase in which competitors are employed as a junior or apprentice *pâtissier*, followed by a phase in which they are a fully-fledged *pâtissier*. Nonetheless, some competitors may quit the trade just after their apprentice period or enter the trade directly as senior workers.

These characteristics lead to a formal definition. Multiphase sequences S_1 and S_2 are successions of n phases such as $S_1 = (\zeta_1^{S_1}, \zeta_2^{S_1}, \ldots, \zeta_p^{S_1}, \ldots, \zeta_n^{S_1})$ and $S_2 = (\zeta_1^{S_2}, \zeta_2^{S_2}, \ldots, \zeta_p^{S_2}, \ldots, \zeta_n^{S_2})$, where $\zeta_p^{S_1}$ is phase p in sequence S_1 and $\zeta_p^{S_2}$ is phase p in sequence S_2. The length of each phase varies from 0 to the length of the sequence in which it is nested. If $n = 1$, we encounter the usual definition of sequences as continuous series of time slots.

[1] A time slot is atemporal since a single state is observed from the beginning to the end of it.

2.2 *Two Formal Properties of Phases and Two Methodological Assumptions*

Two formal properties of phases are captured by the "turning point—trajectories" model theorised by Abbott (2001, p. 253). In this model, turning points differentiate consistent episodes (trajectories) and link these episodes to the previous and following ones into a larger narrative. The first property is relative consistency: the division of a sequence into phases relies on the assumption that the succession of states within each phase is both consistent and different from successions within other phases.[2] The second property is processual location: a phase is defined by its position within a sequence, hence by its position relative to other phases. Returning to the example of the workweek from Lesnard and Kan (2011), Friday is regarded as a consistent period for work scheduling and its position within the workweek immediately preceding Saturday distinguishes it from other days.

Two crucial methodological assumptions follow these formal properties.

First, as sequences, phases are approached as sites (or locations) of narratives: as successions of time slots each containing one state, phases and sequences are constructed as compartments for modelling narratives.[3] To assert that a phase is relatively consistent and located in a sequence does not imply any assumption about the narrative it contains. "Stage" models in the traditions of Piaget or Parsons (for a synthesis, see Levy and The Pavie Team 2005) define the content of both typical sequences and typical phases within these sequences (the so-called stages). These models assume that each phase is the location of a single and specific kind of narrative (for instance, a certain behavioural development or a type of activity). The notion of phase discussed here helps one to appreciate the relevance of stage models but is not bound to these models. While comparing sequences, one assumes that there are patterns, i.e., types of narratives, to be discovered. While comparing multiphase sequences, one assumes that phase-structured patterns—types of narratives including types of sub-narratives—are to be discovered. As for sequence analysis at large, the only assumption about the content of these sub-narratives lies in the alphabets that are used and in the definition of the boundaries of the set of sequences. These elements limit the universe of possible narratives. Identifying these consistent sub-narratives is a different question from dividing into phases, just as identifying consistent narratives is a different question from delimiting a population of sequences.

Second, distinct phases are dissociated and should not be compared. In the workweek example, it would make little sense to compare one sequence's Tuesday to another's Sunday. Similarly, it does not seem relevant to directly compare contests

[2] As Abbott (2001) and Cornwell (2015, p. 94) indicate, this consistency can be conceived as a set of stable relationships between states as modelled by Markov chain models. A turning point is defined by Abbott (2001, p. 247) as a transition separating stable probability regimes.

[3] By methodological construction and to illustrate the notion of site, a narrative cannot be contained in a time slot, which can only contain one state.

in which apprentice *pâtissiers* compete with contests where senior *pâtissiers* compete. This incommensurability of phases has major implications for the comparison of multiphase sequences. I shall return to this point after discussing the practical operation of dividing into phases.

3 Division into Phases: Reference Frame, Alphabet(s) and Phase-Structure

This section starts with an example of division into phases before distinguishing three crucial aspects of this operation.

3.1 A First Hint: The Extended Example

Careers of participants in *pâtissier* competitions in France offer a case of two-phase sequences. The data are drawn from results of 2060 professional competitions. These competitions consist in making or presenting decorative sculptures (out of sugar, chocolate or ice), cakes or plated desserts before juries of peers.[4]

Here I present data from 777 *pâtissiers* who participated in two to 21 competitions and whose careers of participations in competitions began before 2002 and ended after 1990. Each time slot is a participation in a competition. Each participation is defined by an age category (apprentice, junior, senior), a rank (1st, 2nd, etc.) and a type of competition (preceded or not by screening contests).[5]

To be compared, these careers have been divided into two phases with respect to age categories since some contests are for apprentice and junior competitors only, others for senior competitors. This is a case of institutional division of careers into phases that is verified in the data. Indeed, returns to apprentice and junior competitions from senior competitions are unusual: amongst the competitors who participated in at least two competitions and at least one senior competition, only 7.6% participated in an apprentice or junior competition after competing in a senior competition. Interviews with *pâtissiers* within a larger research project made it clear

[4]For a detailed account, see Collas (2015).

[5]Data were gathered from archives and trade press collections. 2060 rankings covering the period 1933–2012 were coded. *Le Journal du pâtissier* (published since 1978), which is the main source, mentions competitions that are organised in different areas in France, while the other sources mention mainly competitions taking place in Paris. Beginning (before 2002) and end (after 1990) dates were chosen because of source heterogeneity and in order to limit right and left censoring. A first rank in the "*Un des meilleurs ouvriers de France*" competition was not kept since it was always gained at the end of a competitor's career. R software (R Core Team 2014) and the TraMineR package (Gabadinho et al. 2011) were used to visualise sequences, to compute OM-distances, to extract sets of representative sequences and to compute other sequence-related metrics.

that first participation in a senior competition is represented as an entry into a phase of evaluation which is not based on age or scholarship, but on the fact of being identified as a *pâtissier*. Thus, I postulate that participations in competitions preceding participation for the first time in a senior competition are not comparable with subsequent participations.

The dissociated phases are, first, the *ante*-senior phase—including only participations in apprentice and junior competitions and ending with the last participation before a participation in a senior competition or with the last participation if the competitor has never participated in a senior contest—and, second, the senior phase—which begins with the first participation in a senior competition and can include any type of participation. Each phase is defined by its relative position within the sequence and by its postulated internal consistency.

Since many competitors participate only in one type of competition (senior ones on the one hand, junior and apprentice ones on the other), a large fraction of sequences include a phase that does not contain any participation (an empty phase, i.e., of null length): these sequences are made up of participations in competitions during only one phase. The senior phase includes participations in 84.4% of the sequences, the *ante*-senior phase in 44.7% of the sequences.

Division into phases impacts the sequences' states alphabet(s). In that example, division leads to a shorter alphabet and, as a consequence, reduced data complexity: age categories are not taken into account in the definition of each participation since phases already bear this age dimension.

Two dimensions are used to define the alphabet. First, competitions preceded by screening contests are regarded as distinct. Amongst these competitions, two are singular and thus isolated as distinct states in the alphabet: a national plated dessert competition named *Championnat de France du dessert* (CFD) (labelled "C") and the oldest competition preceded by screening contests still organised today named *Un des meilleurs ouvriers de France* (One of the best craftsmen in France exam, labelled "M"). This competition is for senior competitors only. Several other competitions are gathered under the "S" label (for Screening contests).[6] Other types of competitions are labelled "W" (for Without screening contests). Second, each participation is defined by the rank awarded, in three categories: 1st rank (labelled "L", for laureate), 2nd or 3rd rank ("P", for podium) and 4th rank or below ("O", for off-the-podium). For example, a state "LC" indicates a first rank in the *Championnat de France du dessert*.

In order to include long pauses between two participations in the analysis, a state named "4Y" (for four years) was created to indicate every period lasting more than four years and less than eight years between two successive participations. The alphabet contains eleven possible states (Table 1). One, related to a senior competition (OM), is observed only during the senior phase.

[6] *Meilleur Apprenti de France* (Best Apprentice in France), *Meilleur Apprenti du Monde* (Best Apprentice in the World), *Coupe du Monde de la Pâtisserie*, *Grand prix international de la chocolaterie*, World Chocolate Master.

Table 1 States alphabet

Competition	Rank	State
Competition without screening contests	1st	LW
	2nd or 3rd	PW
	4th or below	OW
Championnat de France du dessert, CFD	1st	LC
	2nd or 3rd	PC
	4th or below	OC
Competition preceded by screening contests (other than CFD or MOF)	1st	LS
	2nd or 3rd	PS
	4th or below	OS
Un des meilleurs ouvriers de France, MOF (only in Phase 2)	Unranked finalist	OM
Four years pause between two successive participations		4Y

3.2 Three Aspects of Division into Phases

This case sheds light on three aspects of division into phases.

The first one is the reference frame of the division. Here, the reference frame is endogenous: two phases of participations in competitions are dissociated according to a characteristic of participations in competitions (the first participation in a senior competition). In other cases, the reference frame is exogenous: an external dimension is introduced in relation to a research question. For example, careers of participants in competitions could be divided according to job positions. Taking another example, in many systems, academic careers are structured by a limited number of phases (lecturer, assistant professor, associate professor, etc.). The dissociation of phases of academic activities (e.g., publications) according to these successive academic ranks helps to explore how institutionalised episodes impact scientific outputs. Such a division into phases is also a way to reduce data complexity: what can be regarded as two distinct channels (Pollock 2007; Gauthier et al. 2010)—academic ranks and publications—are reduced to one channel cut into successive phases.

The second aspect is the definition of the alphabet(s). Here, the phase division reduces by a half the number of possible states with a very limited loss of information regarding the age categories of each participation (only the above mentioned 7.6% of senior competitors are affected). But the division into phases can also accompany a definition of several alphabets. Since phases are regarded as consistent episodes within sequences, the number of relevant states for each phase may be quite low, especially when the division is endogenous. Such is the case for the careers in ministerial *cabinets* mentioned above. The types of possible positions during the *cabinet* phase are both more limited and more specific than before and after this phase. As a consequence, two alphabets may be defined. By delimiting a relevant universe of possible states for each phase, the plurality of alphabets reduces data complexity.

Determining the number and level of phases is a third key aspect to phase division. The case developed here is simple: only two successive phases are dissociated. As illustrated by several examples mentioned, this number can be higher, but the number of assumptions about the structure of the sequences rises accordingly. We can also envision sub-phases nested within phases. A tennis match is such a two-level nested structure: points are clustered within games and games within sets. Careers including gradations within ranks (2nd class, 1st class, etc.) such as academic careers in France present such a nested structure.

4 Rendering Multiphase Sequences

Two types of graphical representations help to render multiphase sequences. The event-aligned variety of simple alignment representations is suited for two-phase sequences. Sliced representations, introduced here, generalise the logic of event-aligned representations to *n*-phase sequences.

4.1 Simple Alignment on a Specific Event

Drawing on previous studies (Blanchard 2010; Giudici and Gauthier 2009), Colombi and Paye (2014) discuss a visualisation method that transforms the usual left- or right-aligned representation suited for one-phase sequences into a representation aligned on a specific event. This event is regarded as a turning point between two phases: "After synchronisation, each sequence (e.g., series of job positions) is positioned according to an event that takes place in a particular moment for each individual (e.g., childbirth)" (Colombi and Paye 2014, p. 250).

Two steps are followed. First, a relative time axis aligned on the specific event under study is introduced: the temporal scale is negative for states observed before this event, positive for states observed after. The format of each time slot is preserved but the time axis is distorted so as to preserve a social timing according to a supposed turning point. Second, blank time slots are inserted at the beginning and end of each sequence. As a result, the length of every sequence is equal to the length of the longest observed period preceding the studied event summed with the length of the longest observed period following this event.

Taking a toy example, if *A* marries at age 25 and *B* at age 28, their job sequences from 22 to 31 are left-aligned and event-aligned (on marriage date) as shown in Table 2 (E stands for employment, U for unemployment, m on the axis stands for marriage).

Table 2 Different alignments of sequences

Left-aligned sequences, age axis

A	U	U	U	U	E	E	E	E	U	U
B	U	U	E	E	E	E	E	E	E	U
Axis	22	23	24	25	26	27	28	29	30	31

Event-aligned sequences, event-relative axis

A				U	U	U	U	E	E	E	E	U	U
B	U	U	E	E	E	E	E	E	E	U			
Axis	−6	−5	−4	−3	−2	−1	m	1	2	3	4	5	6

Multiphase sequences, left-aligned sliced axis

A	U	U	U				U	E	E	E	E	U	U
B	U	U	E	E	E	E	E	E	E	U			
Axis	s1	s2	s3	s4	s5	s6	m1	m2	m3	m4	d1	d2	d3
Phase	Singlehood						Marriage				Divorce		

4.2 Multiple Alignment by Sliced Representation

Event-alignment is convenient for keeping the continuous representation of left-aligned sequences but such a technique is limited to two-phase sequences. Sliced representations help to render n-phase sequences. In the previous example, if one wants to include divorce in the reference frame of division and if A gets divorced at 29, the representation implies three dissociated phases (see bottom subtable in Table 2).

As for event-aligned representations, the temporal scale is relative, but the time axis is indexed on phases, not on a single event. In this example, the axis is left-aligned for each phase. The origin point is the first time slot of each phase. In case of a right-alignment, the origin point is the highest possible length for a given phase. A right-alignment for certain phases and a left-alignment for others can be envisioned. In that sense, the event-aligned representation is a special case of sliced representations.

Figure 1 shows three representations of the same sample of five sequences of participations in competitions. Division into phases underlines the proximity of the last two sequences during the senior phase. While the event-aligned representation highlights the sequences' continuity as the usual left-aligned or right-aligned representations of sequences, sliced representation focuses on the different successions within each phase, not on the sequels and aftermath of a supposed turning point.

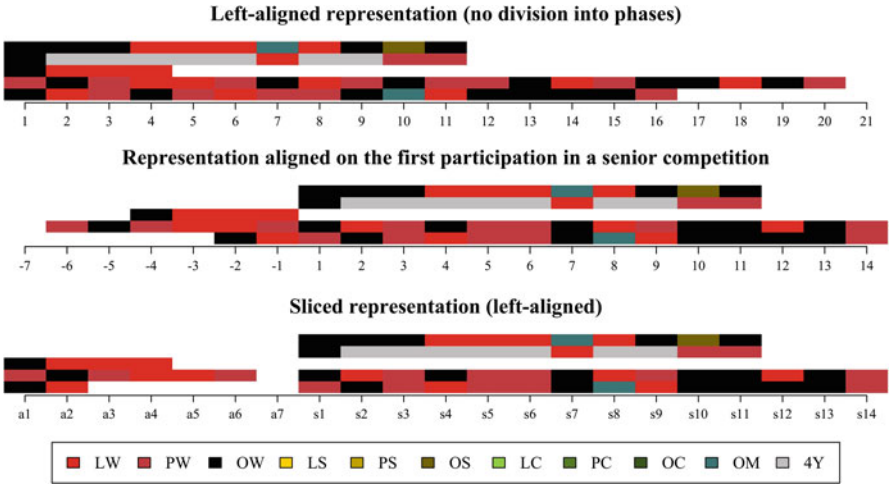

Fig. 1 Three representations of the same careers of participations in competitions (see explanation of state labels in Table 1)

5 Measure and Interpretation of Pairwise Distances Between Multiphase Sequences: Multiphase Optimal Matching

Besides visualisation, dissimilarity measures are commonly used tools for comparing sequences. When measuring dissimilarities between multiphase sequences, phases are regarded both as dissociated incommensurable episodes and as sites of narratives. These are the basic principles of multiphase optimal matching (MPOM) introduced here. This section focuses on optimal matching (OM) for three reasons: it is a seminal and widespread dissimilarity measure in the social sciences; the cost definition operations that OM implies have consequences for MPSA; the principles of OM are adapted to the example under study (this point is discussed below). Nevertheless, MPOM's analytical logic can be extended to other dissimilarity measures when they are applied to multiphase sequences.[7]

5.1 Analytical Logic

MPOM's analytical logic is twofold. First, pairwise distances between sequences are measured with respect to equivalent phases. Time slots belonging to Phase P_1 in Sequence S_1 are only compared with time slots belonging to Phase P_1 in Sequence S_2. Second, each phase is regarded as an ordered set of time slots. Equivalent phases

[7]For a review of dissimilarity measures between sequences, see Studer and Ritschard (2016).

are compared with the three basic operations of OM (Abbott and Forrest 1986): substitution, insertion and deletion. Costs assigned to these operations are defined for each phase.[8]

For each pair of sequences, MPOM measures a distance per phase (distance between S_1 and S_2 on Phase P_1, distance between S_1 and S_2 on Phase P_2, etc.) through OM operations and then a distance between sequences by summing distances between equivalent phases. The matrix of pairwise distances between sequences is the sum of the matrices of pairwise distances per phase. Thus, the contribution of each phase to the distance between two sequences depends on the differences in state composition of this phase from one sequence to another and on its maximal length (the longer a phase is, the heavier its impact on pairwise distance can be, due to the number of insertions and/or deletions necessary to edit one phase into another).

MPOM-measure is a fractal generalisation of OM measure. In the case of a one-level phase division, each phase is approached as OM approaches a sequence. In the case of a higher-level phase division (with multiple levels of phases, phases nested within phases), the same operation is reproduced at each level. Thus, an n-level MPOM-distance implies nested sums. For instance, a two-level MPOM-distance (see the tennis match example above) is a sum of one-level MPOM-distances.

Regarding empty phases, if Phase P_1 is of length $l_1^P = 0$ in S_1 and of length $l_2^P \geq 1$ in S_2, the impact of emptiness on the distance between S_1 and S_2 is equal to the cost of insertion and deletion ($indel$) multiplied by the length of the longest version of Phase P_1 (here l_2). More broadly, the impact of a length difference between two sequences is equal to this difference multiplied by the indel ($|l_2 - l_1| \times indel$).[9]

Thus, MPOM rests on two methodological principles that can be applied to other dissimilarity measures: dissociation of phases and combination of phase pairwise-distances into sequence pairwise-distances.

5.2 *MPOM Applied to Careers of Participants in 'Pâtissier' Competitions*

Turning to careers of participants in *pâtissier* competitions, optimal matching measure offers an appropriate tool for searching for regular patterns in these data for two reasons. First, these sequences are characterised by a high level of instability from one participation to another regarding ranking (one may rank first, then ninth,

[8]This involves a multiplication of cost-setting operations which may seem dubious since many criticisms of OM have focused on cost-setting operations (Abbott and Tsay 2000). Constant or data-driven substitution costs may be relevant in some cases.

[9]Since division into phases often implies length differences, pairwise distances between phases can be standardised with respect to the maximal possible distance for each phase, which I do not do here precisely in order to take length differences into account.

Table 3 Substitution costs

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
1 - LW	0	1.8	1.9	1.95	1.85	1.7	1.7	1.95	1.85	1.7	2
2 - PW	1.8	0	1.7	1.99	1.95	1.92	1.92	1.99	1.95	1.92	2
3 - OW	1.9	1.7	0	2	1.98	1.95	1.95	2	1.98	1.95	2
4 - LC	1.95	1.99	2	0	1.8	1.9	1.9	1	1.8	1.9	2
5 - PC	1.85	1.95	1.98	1.8	0	1.7	1.7	1.8	1	1.7	2
6 - OC	1.7	1.92	1.95	1.9	1.7	0	1	1.9	1.7	1	2
7 - OM	1.7	1.92	1.95	1.9	1.7	1	0	1.9	1.7	1	2
8 - LS	1.95	1.99	2	1	1.8	1.9	1.9	0	1.8	1.9	2
9 - PS	1.85	1.95	1.98	1.8	1	1.7	1.7	1.8	0	1.7	2
10 - OS	1.7	1.92	1.95	1.9	1.7	1	1	1.9	1.7	0	2
11 - 4Y	2	2	2	2	2	2	2	2	2	2	0

then third, etc.); insertion/deletion operations reduce distance due to lags. The second reason is relations between states. A first rank is closer to a second rank than to a tenth rank. Multiple substitution costs lie in the postulate of distance instead of strict difference between states.

So as to limit the number of assumptions, the same substitution costs between states are used for both phases. These costs have been set manually with respect to states' formal closeness: competitions with screening contests are closer to one another than to competitions without screening contests; a first rank is closer to a second rank than to a fourth rank; a second or third rank is closer to a fourth rank than to a first rank (due to the singular situation of ranking first). Minimum substitution cost between two distinct states is 1, maximum cost is 2. Minimum cost is used for a substitution between two states indicating the same position in two competitions preceded by screening contests (e.g., CFD and World Cup). Other substitution costs are set between 1.7 and 2 (Table 3). Several trials intended to take into account the unequal length of sequences without making it the first criterion of distance lead me to set insertion-deletion cost (*indel*) at 1.35 for both phases.[10]

[10]This indel is higher than the minimum substitution cost for distinct states (1) and lower than the minimum substitution cost for states differing in terms of both ranking and screening (1.7). For example, it is less costly to edit one sequence (ABC) into another sequence which is one time slot shorter and otherwise identical (AB) than to edit one sequence into another same-length sequence which is different only for the last time slot (ABC into ABD) only if states C and D are similar with respect to ranking or anterior screening contests.

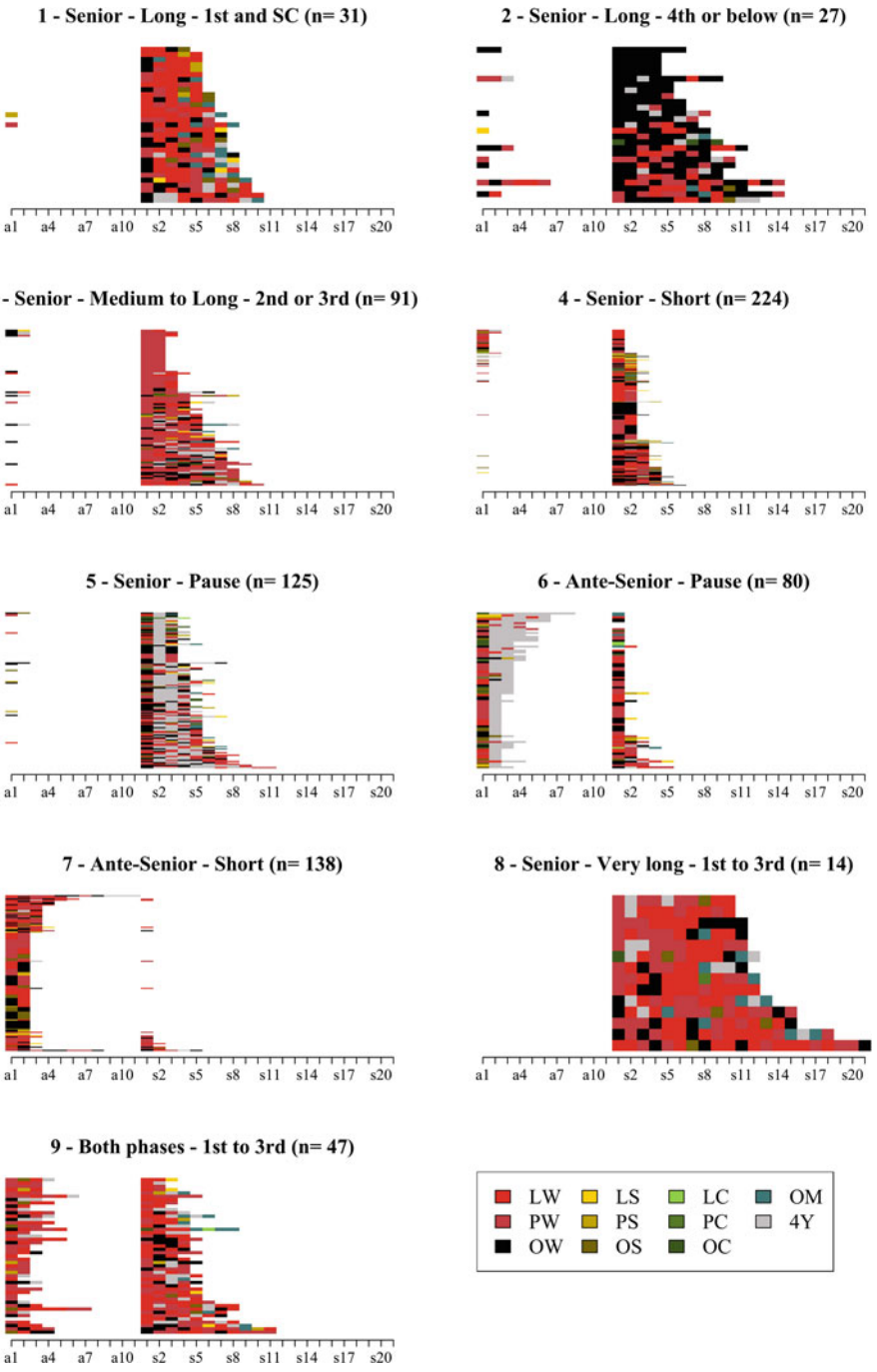

Fig. 2 MDS Sequence Index Plot for each cluster (sorted according to the first factor of multidimensional scaling following Piccarreta and Lior 2010). SC = competitions preceded by screening contests. (See explanation of state labels in Table 1)

Table 4 Sets of representative sequences for each cluster

Cluster	*Ante*-Senior	Senior
1		LW/3-PW/1-LW/1
		OW/1-PW/1-LW/1-OS/1-LW/1
		OW/1-LW/3-PS/1
		PW/1-LW/2-OW/1-OM/1
		LW/1-OW/1-LW/1-OM/1-LW/1-OS/1
		LW/1-PW/1-LW/1-PW/1-OS/1-OM/1
		OW/1-LW/1-OW/1-PW/1-OS/1-LW/1-OS/1
2		OW/5
3		PW/3
4		OW/1-LW/1
5		OW/1-4Y/2-PW/1
6	PW/1-4Y/2	PW/1
7	OW/1-PW/1	
8		PW/1-4Y/1-LW/3-PW/1-LW/2-PW/1-LW/1
		PW/1-LW/1-PW/3-LW/1-OW/1-OM/1-LW/2-OW/1
		PW/1-4Y/2-PW/3-LW/3-PW/2
		PW/2-LW/1-OW/1-LW/1-LW/1-LW/1-PC/1-LW/1-PW/2-OM/1
		PW/1-LW/1-OW/1-PW/2-LW/2-OM/1-4Y/2-OW/1
		OW/1-4Y/1-LW/1-PW/1-LW/1-PW/2-LW/2-PW/1-OM/1-4Y/1-OM/1
9	PW/1	LW/1-PW/1-OW/1-LW/1
	PW/1-LW/	LW/1-4Y/1
	1-PW/1	
	LW/2	PW/1-LW/1-PW/1-LW/1-LS/1
	PW/3	PW/2-LW/1-PW/1-LW/1

N.B.: In each cluster, the distance of at least one sequence out of two from one of the representative sequences is inferior to 30% of the maximum distance in the cluster (for details on the centrality criterion, see Gabadinho and Ritschard 2013). Sequences are sorted by representativeness. Each state is followed by its number of successive occurrences

A comparison between several clustering methods invited me to opt for a nine-cluster Ward's (1963) partition (see Fig. 2 for sequence index plots and Table 4 for representative sequences).[11]

[11] Any distance-based clustering method could be used including the property-based and fuzzy methods addressed by Studer (2018) in this bundle. Here, using the R package WeightedCluster (Studer 2013), several algorithms have been compared for a division into two to ten classes: hierarchical cluster solutions named Ward, single, complete, average (UPGMA), McQuitty (WPGMA) and beta-flexible (flexible-UPGMA) (for a presentation, see Müllner 2013; Belbin et al. 1992) and non-hierarchical partition around medoids algorithm (PAM) (Rousseeuw and Kaufman 1990). PBC, HC, HG and ASW measures have been used to compare the quality of the different clustering solutions (Hennig and Liao 2010). Ward's nine-cluster solution was the most relevant regarding both quality measures and readability. Except for PAM, other algorithms tend to produce a partition between a very heterogenous group containing more than 80% of the cases and two to

Three key elements of interpretation arise: participation in senior competitions, length and tonality (the most frequent state or family of states within the sequence).

Cluster 1 (4% of sequences) gathers sequences mainly characterised by an empty *ante*-senior phase (only two sequences out of 31 do not start with a senior competition) and a long senior phase (mean length is 6.87 time slots) including mainly first ranks and participations in competitions preceded by screening contests (64.7% of participations). Cluster 2 (3.5% of sequences) gathers medium to long senior careers (mean length is 8.63) in which first ranks are rare (less than 11% of participations). Cluster 3 (11.7% of sequences) gathers short to medium length senior careers (mean length is 4.99), mainly characterised by second and third ranks (70% of participations). Cluster 4 (28.8% of sequences) gathers short length senior sequences with no shared tonality. Clusters 5 (16.1% of sequences) and 6 (10.3% of sequences) are defined by at least one four-year pause, respectively during the senior phase and during the *ante*-senior phase. Cluster 7 (17.8% of sequences) gathers sequences defined by a short *ante*-senior phase and an empty senior phase with no shared tonality. Cluster 8 (1.8% of sequences) is made up of very long senior careers (mean length is 13) including mainly first to third ranks (76% of participations). Sequences in Cluster 9 (6% of sequences) share a symmetrical intensity regarding participations in *ante*-senior and senior phases and a relatively low rate of 4th ranks or below.

How far does this clustering take phases into account? First, the Ward two-cluster solution separates Clusters 6 and 7 from the seven other clusters, that is to say careers first defined by *ante*-senior participations from careers first defined by senior participations. Second, Clusters 6 and 7, both characterised by *ante*-senior participations, are distinct from one another with respect to participation in senior competitions. Third, a quarter of the sequences counting one or more ante-senior participations are not clustered in Clusters 6 and 7. In other words, closeness does not only rest on the (non-)emptiness of phases, but also on phases' tonality. Fourth, when, as here, the reference frame of the division is endogenous, that division greatly simplifies the interpretation: once the phases mainly portrayed by each cluster have been identified, interpretation is primarily based on ranking.

5.3 MPOM Compared

MPOM takes cues from two other families of dissimilarity measures that assume a division of sequences into dissociated and incommensurable episodes.

First, MPOM generalises Hamming Distance (Hamming 1950) and Dynamic Hamming Distance (DHD) (see Lesnard 2008, 2014), which measure dissimilarities

nine easy to analyse but very small groups. Quality measures are higher for Ward compared to PAM. A nine-cluster solution is associated with the highest value of HG and HC indexes, the second highest value of PBC and the fourth highest value of ASW.

position-wise. DHD can be approached as a specification of MPOM-distance in which each time slot is a phase (of length 1) and in which substitution costs are derived from transition rates before and after each time slot. This specification is suited for sequences defined by a limited number of states, observed in each sequence and spanning long spells.[12]

Second, compared with Qualitative Harmonic Analysis (QHA) (Deville and Saporta 1983; Robette and Thibault 2008; Robette and Bry 2012), in which sequences are divided into periods that are modelled as bundles of states varying from one another in proportions of time spent, MPOM pays attention to the order of the states within phases within phases.

Thus, the main advantage of MPOM is to take into account a unit nested in a sequence that is more malleable than time slots—its length varies from 0 to the whole sequence's length—and that is studied as a time-ordered structure.

6 Conclusion

This chapter has introduced the idea of multiphase sequences and several tools to study them, especially sliced representations and a multiphase dissimilarity measure (called MPOM) the logic of which can be extended to other dissimilarity measures. Two general issues have been raised and invite further investigation in the development of MPSA.

At the beginning of this chapter, sequences were defined as hierarchical structures, as narratives including sub-narratives and nested into larger narratives. This definition is partly consistent with Dumont's (1980) perspective of hierarchies as nested entities. Any sequence (marital biographies, job careers, dances, etc.) can be approached as a fragment of larger social processes (Abbott 2016) including other fragments of social processes. The approach to sequences as hierarchical structures could be further developed by investigating the variety of relations between nested temporal structures. Hybridisations of network analysis and sequence analysis (Cornwell 2015) may be a possible way to study these relations as a multilevel issue (Lazega and Snijders 2016).[13]

[12]To preserve a division into phases in the comparison of workweeks, Lesnard and Kan (2011) wrap each phase into an atomic time slot, thus describing each phase by a single state. Their two-step method identifies types of narratives through DHD and clustering procedure and then assembles these types in week-sequences analysed with DHD and clustering. Compared to MPOM, this wrapping solution is suited for periods of identical time spread (such as hours or days). Its main limitation is that its second step is based on the heterogenous inputs of a clustering procedure.

[13]Regarding MPOM, the analysis of the articulation of distances between phases and distances between sequences could be further developed since two sequences can be identical along some phases and clearly different along others. A related question is the importance of specific phases in the definition of a whole sequence. Various theoretical perspectives assume that certain phases are more crucial than others (childhood in a whole life-course for example). Such assumptions can orientate the parameters of MPOM by differentiating phases' weights.

A methodological assumption of MPSA is to approach phases as sites of narratives nested within sequences, which are sites of larger narratives. This assumption differentiates phases and the narrative patterns that the division into phases makes it possible to unveil. That echoes other notions centred on temporal substructures, such as subsequence (Elzinga et al. 2008) or motif (Han 2014). There is an open field for research on sub-narratives, their typical position(s) within sequences and their overlaps.[14] MPOM assumes a division into phases prior to the identification of narrative patterns. Different alphabets are defined, thus determining what patterns can be identified. A comprehensive method would manage three different steps: identifying types of narratives, identifying phases, identifying patterns of relation between phases and narratives. In other words, the division into phases could be dynamically revised and preceded by a moment of identification of patterns for different fragments of sequences under study.[15]

These elements invite renewed exploration of the various interrelations and continuities between temporal structures, a major question that sequence analysts have already extensively explored.

Acknowledgements I thank the participants to the LaCoSA 2 conference and the editors and anonymous reviewers of this book for valuable suggestions. I thank Claire Lemercier, Laurent Lesnard, Etienne Ollion and Loretta Platts from whom I received insightful comments on earlier versions of this chapter. I am thankful to Richard Nice for a meticulous proofreading of my English.

References

Abbott, A. (1995). Sequence analysis: New methods for old ideas. *Annual Review of Sociology, 21*(1), 93–113.

Abbott, A. (2001). *Time matters: On theory and method*. Chicago: University of Chicago Press.

Abbott, A. (2016). *Processual sociology*. Chicago: University of Chicago Press.

Abbott, A., & Forrest, J. (1986). Optimal matching methods for historical sequences. *The Journal of Interdisciplinary History, 16*(3), 471–494.

Abbott, A., & Tsay, A. (2000). Sequence analysis and optimal matching methods in sociology: Review and prospect. *Sociological Methods & Research, 29*(1), 3–33.

Bakeman, R., & Gottman, J. M. (1997). *Observing interaction: An introduction to sequential analysis*. Cambridge: Cambridge university press.

Belbin, L., Faith, D. P., & Milligan, G. W. (1992). A comparison of two approaches to beta-flexible clustering. *Multivariate Behavioral Research, 27*(3), 417–433.

[14]Regarding overlaps, my purpose was focused on clear-cut phases, divided from the single viewpoint of contractual events (a divorce, a recruitment, etc.), which is simplistic. For instance, the idea that a divorce begins with a contract in couples' histories is debatable. In most cases, the identification of phases as relevant locations of narratives is a challenging issue since many narratives overlap. Nevertheless, the notion of overlap is premised upon the idea of typical locations of narratives. Identifying those typical locations is thus a preliminary step to the study of overlaps.

[15]This identification could rely on tools for detecting consistency within segments, such as the stationarity test proposed by Bakeman and Gottman (1997).

Blanchard, P. (2010). Analyse séquentielle et carrières militantes. Research report, Institut d'études politiques et internationales.

Collas, T. (2015, Unpublished). *La pâte et le décor: Considération et formes professionnelles dans le monde des pâtissiers*. Ph.D. thesis, Sciences Po Paris.

Colombi, D., & Paye, S. (2014). Synchronising sequences: An analytic approach to explore relationships between events and temporal patterns. In P. Blanchard, F. Bühlmann, & J.-A. Gauthier (Eds.), *Advances in sequence analysis: Theory, method, applications* (pp. 249–264). Cham: Springer.

Cornwell, B. (2015). *Social sequence analysis: Methods and applications*. Cambridge: Cambridge University Press.

Deville, J.-C., & Saporta, G. (1983). Correspondence analysis, with an extension towards nominal time series. *Journal of Econometrics, 22*(1–2), 169–189.

Dumont, L. (1980). *Homo hierarchicus: The caste system and its implications*. Chicago: University of Chicago Press.

Elzinga, C., Rahmann, S., & Wang, H. (2008). Algorithms for subsequence combinatorics. *Theoretical Computer Science, 409*(3), 394–404.

Gabadinho, A., & Ritschard, G. (2013). Searching for typical life trajectories applied to childbirth histories. In R. Levy & E. Widmer (Eds.), *Gendered life courses-Between individualization and standardization. A European approach applied to Switzerland* (pp. 287–312). Vienna: LIT.

Gabadinho, A., Ritschard, G., Mueller, N. S., & Studer, M. (2011). Analyzing and visualizing state sequences in R with TraMineR. *Journal of Statistical Software, 40*(4), 1–37.

Gauthier, J.-A., Widmer, E., Bucher, P., & Notredame, C. (2010). Multichannel sequence analysis applied to social science data. *Sociological Methodology, 40*(1), 1–38.

Giudici, F., & Gauthier, J.-A. (2009). Différenciation des trajectoires professionnelles liée à la transition à la parentalité en Suisse. *Revue Suisse de Sociologie, 35*(2), 253–278.

Hamming, R. W. (1950). Error detecting and error correcting codes. *Bell Labs Technical Journal, 29*(2), 147–160.

Han, S.-K. (2014). Motif of sequence, motif in sequence. In P. Blanchard, F. Bühlmann, & J.-A. Gauthier (Eds.), *Advances in sequence analysis: Theory, method, applications* (pp. 21–38). Cham: Springer.

Hennig, C., & Liao, T. F. (2010). Comparing latent class and dissimilarity based clustering for mixed type variables with application to social stratification. Research Report 308, Department of Statistical Science, Department of Sociology, University of Illinois.

Lazega, E., & Snijders, T. (Eds.) (2016). *Multilevel network analysis for the social sciences*. New York: Springer.

Lesnard, L. (2008). Off-scheduling within dual-earner couples: An unequal and negative externality for family time. *American Journal of Sociology, 114*(2), 447–490.

Lesnard, L. (2014). Using optimal matching analysis in sociology: Cost setting and sociology of time. In P. Blanchard, F. Bühlmann, & J.-A. Gauthier (Eds.), *Advances in sequence analysis: Theory, method, applications* (pp. 39–50). Cham: Springer.

Lesnard, L., & Kan, M. Y. (2011). Investigating scheduling of work: A two-stage optimal matching analysis of workdays and workweeks. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 174*(2), 349–368.

Levy, R., & The Pavie Team. (2005). Why look at life courses in an interdisciplinary perspective. In R. Levy, P. Ghisletta, J.-M. Le Goff, D. Spini, & E. Widmer (Eds.), *Towards an interdisciplinary perspective on the life course* (pp. 3–32). Amsterdam-Boston: Elsevier.

Müllner, D. (2013). Fastcluster: Fast hierarchical, agglomerative clustering routines for R and Python. *Journal of Statistical Software, 53*(9), 1–18.

Piccarreta, R., & Lior, O. (2010). Exploring sequences: A graphical tool based on multi-dimensional scaling. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 173*(1), 165–184.

Pollock, G. (2007). Holistic trajectories: A study of combined employment, housing and family careers by using multiple-sequence analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 170*(1), 167–183.

R Core Team. (2014). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.

Robette, N., & Bry, X. (2012). Harpoon or bait? A comparison of various metrics in fishing for sequence patterns. *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique, 116*(1), 5–24.

Robette, N., & Thibault, N. (2008). L'analyse exploratoire de trajectoires professionnelles: Analyse harmonique qualitative ou appariement optimal? *Population, 63*(4), 621–646.

Rousseeuw, P. J., & Kaufman, L. (1990). *Finding groups in data*. New York: Wiley.

Studer, M. (2013). WeightedCluster library manual: A practical guide to creating typologies of trajectories in the social sciences with R. *LIVES Working Papers* (24).

Studer, M. (2018). Divisive property-based and fuzzy clustering for sequence analysis. In G. Ritschard & M. Studer (Eds.), *Sequence analysis and related approaches: Innovative methods and applications*. Cham: Springer (this volume).

Studer, M., & Ritschard, G. (2016). What matters in differences between life trajectories: A comparative review of sequence dissimilarity measures. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 179*(2), 481–511.

Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association, 58*(301), 236–244.