LIVES
Swiss National Centre of Competence in Research

FNSNF
Swiss National Science Foundation

Collas, T. (2016)

Multiphase Optimal Matching: An Application to Careers of Participation in Pâtissiers' Competitions

in G. Ritschard & M. Studer (eds), Proceedings of the International Conference on Sequence Analysis and Related Methods, Lausanne, June 8-10, 2016, pp 571-588.

UNIVERSITÉ DE GENÈVE

FONDATION POUR L'UNIVERSITÉ DE LAUSANNE

UNIVERSITÉ DE GENÈVE
GENEVA SCHOOL OF SOCIAL SCIENCES

UNIL | Université de Lausanne
Institut des sciences sociales

Multiphase Optimal Matching: An Application to Careers of Participation in *Pâtissiers'* Competitions

Thomas Collas

Center for the Sociology of Organizations (Sciences Po), 19. rue Amélie, 75007 Paris
thomas.collas@sciencespo.fr

Preliminary and incomplete draft for the International conference on sequence analysis and related methods (LaCoSA 2), Lausanne, June 8th-10th 2016.

Please do not quote or circulate.

Many social sciences theories (*e.g.* Becker, 1963 ; Wilensky, 1964 ; Lang and Lang, 1984 ; for a synthesis on 'universal narratives' see Abbott, 1992) and many common notions (lifecycle, adulthood, turning point, ratchet effect, etc.) are premised upon the idea that some sets of processes follow regular patterns defined by successive phases. A phase can be defined as a moment in a sequence, a succession of episodes (time units in a sequence) that can be regarded as singular compared to other successions of episodes (other phases) in the same sequence[1].

Optimal matching analysis (OMA) is commonly used to reveal these successive phases. In this talk, I use OMA to compare sequences defined by two or more phases through what I call multiphase optimal matching (MPOM).

In a first section, I define the main properties of multiphase sequences and the main features of MPOM. In a second section, I present a simple example of MPOM to compare careers of participation in *pâtissiers*' competitions in France[2].

[1] Such is also the basic postulate of Qualitative Harmonic Analysis (Deville, 1974). For an application of QHA and a comparison with Optimal Matching Analysis, see Robette and Thibault (2008) and Robette and Bry (2012).

[2] I keep the french word *pâtissiers* since 'baker' refers to bread making – to *boulanger* in french –, 'pastry chef' or 'pastry cook' are relevant to describe *pâtissiers*' restaurant work and 'confectioners' refers to candies making.

1 Multiphase sequences and multiphase optimal matching (MPOM)

The basic principle of MPOM is to compare sequences according to equivalent phases. Phase is made a relevant unit of comparison. Sequence is both regarded as a succession of phases and as a coherent unit compared to other sequences.

1.1 Multiphase sequences: four properties

The main property of multiphase sequences is that they are defined before the analysis as composed by two or more phases that are supposed to be common to the whole set of sequences under study. For example, sequences synchronization (Blanchard, 2010; Giudici and Gauthier, 2009; Colombi and Paye, 2014) is a visualization technique suited for two-phase sequences, that is to say sequences in which one event is *postulated* as a turning point (marriage, employment, entry into an association, etc.). This event thus divides the sequence between a first phase (episodes before the turning point) and a second one (episodes occurring after the turning point). As Denis Colombi and Simon Paye (2014) have underlined, this turning point can be endogenous or exogenous. In the first case, the turning point is a transition between two states belonging to the sequence alphabet. In the second case, the turning point is defined on another channel. For example, an occupational career can be divided between two phases when married people are compared: before and after marriage. Following Andrew Abbott (1997), synchronization can be defined as a presumed division into two trajectories, *i.e.* phases characterized by distinct 'regimes of probability'. This is the second property of multiphase sequences.

Obviously, there can be more than *one* turning point in what appears to be a single sequence (*e.g.* an individual career), which is then divided into three or more phases. Synchronization can be generalized to n phases. Every sequence in a set can be defined as a succession of n phases such as $A = (\zeta_1^A, \zeta_2^A, ..., \zeta_p^A, ..., \zeta_n^A)$ and $B = (\zeta_1^B, \zeta_2^B, ..., \zeta_p^B, ..., \zeta_n^B)$, where ζ_p^A is the phase p in sequence A and ζ_p^B is the phase p in sequence B. Take tennis matches for example. To visualize matches as sequences of games (each time unit is a game, each state in the alphabet is the number of points in a game), each last point of a set can be regarded as a turning point. Thus, each match can be divided into five phases (the highest possible number of sets) to indicate that it is not only a sequence of games but, above all, a sequence of sets of games. This toy example points out two properties of multiphase sequences. First, equivalent phases often vary in length from one sequence to another. There can be six games in the first set of match A and ten games in the first set of match B. The end of the set announces in each case the second set. Even if complete sequences are of equal length, synchronization generally implies differences in phase lengths. If not, each turning point is met at the same time for the whole set of sequences. Such is the case if turning points are ages. Second, the length of a phase can be equal to zero. This phase is thus considered as an empty one. This can occur if two hypothetic idiosyncratic turning points are observed at the same time or if a supposed turning point is not yet observed, never observed or immediately observed at the beginning

of the sequence. Match A can come to an end after three sets, while match B can last five sets. Within match A, sets 4 and 5 are empty phases.

This applies to the former example: occupational careers can be divided into more than two phases with respect to marriage. A first phases can be 'being single', a second one 'being married', a third one 'being divorced or widowed', a fourth one 'being remarried' and so on. Each of these phases necessarily follows the previous ones: 'being remarried' implies celibacy, marriage and divorce or widowhood. The length of these different phases can vary largely and can obviously be equal to zero, for example for people who never get married, for people who never get divorced or for people who get remarried during the same time unit (a month for example) they get divorced (two assumed turning points are met at the same time)[3].

Let us summarize the four properties underlined:

1. Multiphase structure is not a result but a *postulate*, even if it can result from previous analysis.
2. Each phase is regarded as an *hypothetic trajectory*.
3. Equivalent phases often *differ in length*.
4. Phase length can be *equal to zero*. This does not affect the relevance of a constant multiphase definition for all sequences under study.

1.2 MPOM: a short definition

Let us now consider that we do not only want to visualize a set of n-phase sequences, but to compare them through optimal matching.

The main concern is to keep the multiphase structure. As Laurent Lesnard (2008, p. 463) has stressed, insertion/deletion operations 'loosen the connection of processes with their temporal scale'. Dynamic Hamming Matching (DHM), the method suggested by the author (for details, see Lesnard, 2014), rests on two conditions to avoid such a loosening. First, each time unit is regarded as incommensurable with other time units: if A's workday is compared to B's one, what A is doing (working/not working) at 7PM can only be compared to what B is doing at the exact same time. Then, only substitution operations are used and insertion/deletion operations are excluded. Second, the cost of substitution for a state to another depends on the observed rate of transition between the two states at the time unit in question. Then, a cost of substitution between each pair of states is defined for each time unit. This solution, which does not preserve only one synchronization point but *each time unit synchronization*, is highly suited for sequences defined by a limited number of states, observed in each sequence and spanning over long periods, that is to say sequences varying from one another in timing and duration.

MPOM loosens the 'connection of [sequences] with their temporal scale' (Lesnard, 2008, p. 463) only within phases. In other words, phases, not time units, are regarded as incommensurable. Three features can be underlined:

[3] Another simple example of multiphase sequence is a French academic career at University with turning points such as becoming *Maître de conférences* and becoming *Professeur des universités*.

- The three basic operations of OM (Abbott and Forrest, 1986) – substitution, insertion, deletion – are kept and used to compare equivalent phases from one sequence to another (ζ_p^A to ζ_p^B). Length variations resulting from synchronization induce several complications. While the 'core program' (Gauthier and al., 2014. p. 5) of sequence analysis 'mostly refers to sequences of equal length, with age as a time axis and year as a time unit' (Ibid.), analysts of sequences of unequal length are often mostly preoccupied, since the seminal study by Andrew Abbott and John Forrest (1986), by normalizing sequences with respect to length or by minimizing distance due to length (Abbott and Hrycak 1990; Stovel and al., 1996; Stovel and Bolan, 2004). As Philippe Blanchard (2010. p. 57) has stressed, taking length differences into account implies a cautious arbitration between substitution costs and insertion/deletion costs in the search of their most relevant relative definition regarding the sequences under study.
- Substitution costs as well as insertion/deletion cost (indel) can be defined for each phase. That involves a multiplication of cost-setting operations that may seem dubious, since many criticisms addressed to OMA focused on cost-setting operations (see the widely cited debate between Wu, 2000 and Abbott and Tsay, 2000). Transition-rates based substitution costs can be relevant in some cases and can be defined for each phase. If a pair of states can be observed in any phase, transition between these states can be defined as a combination between intra-phase transition rates and intra-sequence transition rates.
- For each pair of sequences, the aim of the operation is to compute a distance per phase and then to compute an inter-sequences distance. The simplest way to do so is to sum OM-distances per phase, the resulting distance matrix being the sum of phase-distance matrices (see example below). Thus, each phase contribution to the full distance depends on differences in states-composition and in length. Another way to compute a total distance is to standardize OM-distances per phase with respect to maximal possible distance for each phase.

2 A two-phase example: competitors' careers as ante-senior and senior phases

I used MPOM while preparing my PhD thesis on *pâtissiers*' work in France since the 1970s. I was studying careers of participation in *pâtissiers*' competitions in France, which mostly consist in sugar or chocolate sculptures competitions.

I was fishing for regular patterns, especially for cumulative patterns in terms of ranking. Cumulative models (DiPrete and Eirich, 2006), when applied to competitions, predict a close relation between length and structure (for a detailed account, see Collas, 2015): short careers' patterns should mainly follow a 'succession of failures' pattern, long careers should even follow 'successions of victories' patterns – through indirect screening of candidates (Menger, 2009) – or progressive pattern – for example through learning-by-doing due to imperfect information on the competition during the first participations.

Despite of numerous cost setting operations, OMA appeared as the best solution to search for regular patterns

- The first reason was *instability*: the sequences are characterized by a high instability from one time unit to another concerning ranking (one can rank first, then ninth, then third, etc.) ; insertion/deletion operations reduce distance due to some lags.
- The second reason was *relations between states*: a first rank is closer to a second rank than to a tenth rank. OM algorithm lies in the postulate of distance between states (through substitution costs) instead of strict difference between them.

2.1 Data

Data were gathered from 2060 rankings (120 competitions, recurring or not), mainly from trade press, but also from organizers' archives for some major competitions. Data cover the period 1953-2012. Due to source heterogeneity[4], I focused on careers beginning after 1985 and ending before 2007 and thus kept 3927 names out of 6264[5]. I used OM only to compare participation careers counting two participations or more, that is 1258 careers[6]. Career length varies from 2 participations to 21 participations.

Time unit is a participation in a competition: the time axis is thus a process one. Careers are defined with 24 states (Table 1, p. 6).

Three dimensions have been taken into account to define the states:

- Recurring competitions preceded by pre-selection contests are regarded as distinct. That is the case of *Championnat de France du dessert* (CFD, a national 'dessert on plate' competition) for apprentice and for senior competitors, of *Meilleur Apprenti de France* (Best Apprentice in France) and *Meilleur Apprentice du Monde* (Best Apprentice in the World) – these two are put in a same category, BA –, of '*Un des meilleurs ouvriers de France*' (MOF) competition[7] and, finally, of senior international competitions[8] with pre-selection contests that I put in a same category: CM. Competitions which do not verify these two conditions (recurrence and pre-selections) are categorized R2 and regarded as equivalent.
- Every state is defined by the rank occupied, in three categories: 1^{st} rank (L), 2^{nd} or 3^{rd} rank (P) and 4^{th} rank or beyond (H).

[4] *Journal du pâtissier* (published since 1978) mentions competitions organized in different areas in France, while the other sources mention mainly competitions taking place in Paris.

[5] The mean duration of career counting two participation or more and ending after 1980 is 5 years.

[6] 1263 careers counted two participations or more but I have not kept a first rank at the '*Un des meilleurs ouvriers de France*' (MOF) competition if it was gained at the end of a career since 98 % of participation careers including this rank end with this rank. Then, it allowed me to compare MOF laureates' careers with other participation careers.

[7] This can be translated as 'one of the best craftsmen in France' but the usual category is simply the french expression for 'best craftsman in France'.

[8] *Coupe du Monde* (in Lyon), *Grand prix international de la chocolaterie*, World Chocolate Maser.

Table 1. States definition

State	Competition	Rank	Category
L-CFDAp	CFD	1^{st}	Apprentice
P-CFDAp	CFD	2^{nd} or 3^{rd}	Apprentice
H-CFDAp	CFD	4^{th} or beyond	Apprentice
L-BA	BA	1^{st}	Apprentice
P-BA	BA	2^{nd} or 3^{rd}	Apprentice
H-BA	BA	4^{th} or beyond	Apprentice
L-R2Ap	R2	1^{st}	Apprentice
P-R2Ap	R2	2^{nd} or 3^{rd}	Apprentice
H-R2Ap	R2	4^{th} or beyond	Apprentice
L-R2Ju	R2	1^{st}	Junior
P-R2Ju	R2	2^{nd} or 3^{rd}	Junior
H-R2Ju	R2	4^{th} or beyond	Junior
L-CFDSe	CFD	1^{st}	Senior
P-CFDSe	CFD	2^{nd} or 3^{rd}	Senior
H-CFDSe	CFD	4^{th} or beyond	Senior
L-CM	CM	1^{st}	Senior
P-CM	CM	2^{nd} or 3^{rd}	Senior
H-CM	CM	4^{th} or beyond	Senior
L-MOF	MOF	1^{st}	Senior
H-MOF	MOF	unranked finalist	Senior
L-R2Se	R2	1^{st}	Senior
P-R2Se	R2	2^{nd} or 3^{rd}	Senior
H-R2Se	R2	4^{th} or beyond	Senior
4years	4 years void between two participations		

- Every state is defined by the age category of contestants, in three categories again: Senior (Se), Junior (Ju) or Apprentice (Ap).

A state named '4years' was added for every period lasting more than 4 years and less than 8 years between two successive participations.

These sequential data were characterized by high heterogeneity[9], high instability[10] and high length variance.

2.2 First senior competition as an assumed turning point between ante-senior phase and senior phase

In addition to that, it seemed dubious to compare these sequences from start to finish. Some competitions are for apprentices and junior competitors only. Furthermore,

[9] Amongst 1258 careers, we count 794 different patterns (mean occurrence is 1.56). Each sequence of length 6 or more ($N = 174$) is unique.

[10] Mean complexity (Gabadinho and al., 2010) is 0.43. This metric varies from 0 to 1 and is based on the number of transitions and the on longitudinal entropy.

a first analysis pointed out a low rotation between these competitions and senior ones : the first participation in a senior competition implies no later participation in junior or apprentice competitions, except for 6 % of senior contestants[11]. That first analysis also stressed that an important part of individual careers did not include any senior competition (44 % of the 557 sequences that include a junior or apprentice competition) and that a more important part did not include any junior or apprentice competition (69 % of the 1017 sequences that include a senior competition).

Based on that analysis and on interviews with *pâtissiers* ($N = 58$), I assumed that the first participation in a senior competition represents entry into a career of evaluation which is not based on age or scholarship, but on the single fact of being identified as a *pâtissier*.

The sequences compared are defined as two-phase sequences verifying the aforementioned properties with the first participation in a senior competition as a supposed turning point. States observed before the first participation in a senior competition were regarded as incommensurable with states observed after that participation. There are an *ante*-senior phase (defined solely by participations in apprentice and junior competitions) and a senior phase (that can include any state in Table 1). The first phase is defined by 218 distinct patterns (mean occurrence is 2.6) and the second phase by 504 patterns (mean occurrence is 2.02).

Figure 1 shows a sample of 30 phases following that synchronization[12].

2.3 MPOM as a sum of OM-distances per phase

MPOM preserves synchronization while computing distance between sequences.

In this example, I define MPOM distance between two sequences as a sum of OM-distances per phase. Let us formalize it.

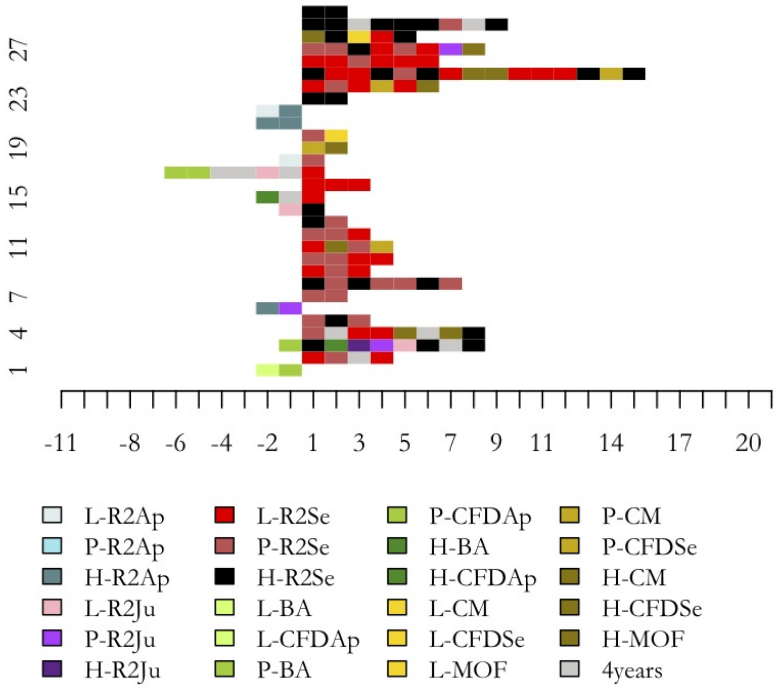
Following the presentation of OM distance by Cees Elzinga (2014. p. 55), let S_1 and S_2 denote two n-phases sequences and ζ denotes the set of n postulated phases. For a phase ζ_p in $\{\zeta_1, \zeta_2, ..., \zeta_p, ..., \zeta_n\}$ over the alphabet of states $\Sigma_p = \{\lambda_p, a_p, b_p, ...\}$ in ζ_p (with λ denoting the empty state), let $t_{\zeta_p^{S_1}, \zeta_p^{S_2}} = t_1 ... t_l$ denotes a series of admissible phase edits. For any pair of equivalent phases from one sequence to another, many distinct series of edits that transform $\zeta_p^{S_1}$ (the phase p in sequence S_1) into $\zeta_p^{S_2}$ (the phase p in sequence S_2) may exist and I write $T(\zeta_p^{S_1}, \zeta_p^{S_2})$ to denote the set of such edit-series. Furthermore, to each edit t_i, a nonnegative cost of weight $c(t_i)$ is assigned and the cost of an edit-series $C(t_{\zeta_p^{S_1}, \zeta_p^{S_2}})$ equals the sum on the edits involved. The OM-distance per phase is defined as the minimum of the costs of the edit-series in $T(\zeta_p^{S_1}, \zeta_p^{S_2})$.

Thus, MPOM-distance $d_{MPOM}(S_1, S_2)$ between sequences S_1 and S_2 is defined as follow :

[11] 22 % of the 311 competitors who participate in senior and apprentice or junior competitions, participate in junior or apprentice ones after a first participation in a senior one.

[12] R software (R Core Team, 2015) and TraMineR R package (Gabadinho and al., 2011) were used to visualize sequences, to compute OM-distances, to extract sets of representative sequences and to compute other metrics related to sequences (length and complexity).

Fig. 1. Random sample of 30 sequences



N.B.: Due to a high number of states, visualization is a bit tricky, diversity has been impoverished to improve legibility: some states, which are rare states and for which substitution costs are low, are visualized with the same colors

$$d_{MPOM}(S_1,S_2) = \sum_{p=1}^{n} min\left\{C(t_{\zeta_p^{S_1},\zeta_p^{S_2}}) : t_{\zeta_p^{S_1},\zeta_p^{S_2}} \in T(\zeta_p^{S_1},\zeta_p^{S_2})\right\} \quad (1)$$

2.4 Definition of substitution and insertion/deletion costs

In trying to uncover cumulative patterns, substitution costs based on transition rates would be of little help. Substitution costs between states have been defined with respect to their formal closeness: competitions with pre-selection contests are closer to one another than to competitions without pre-selection contests; a first rank is closer to a second rank than to a fourth; a second or a third rank is closer to a fourth rank than to a first rank (due to the singular position of ranking first).

Minimum substitution cost between two distinct states is 1 and maximum cost is 2. 80 % of substitution costs are superior to 1.6, which is the minimum substitution cost between two states differing by rank and by pre-selections. Due to the division into two phases, only 5 % of the costs differences between distinct states are due to a difference in age categories, 95 % are due to a difference in rankings and in types

Table 2. Substitution costs for senior phase

	L-BA	P-BA	H-BA	L-R2Ap	P-R2Ap	H-R2Ap	L-R2Ju	P-R2Ju	H-R2Ju	L-CFDSe	P-CFDSe	H-CFDSe	L-CM	P-CM	H-CM	L-MOF	H-MOF	L-R2Se	P-R2Se	H-R2Se	4years
L-BA	0	1.6	1.81	1.88	1.93	1.95	1.92	1.97	1.99	1.05	1.65	1.86	1.05	1.65	1.86	1.05	1.65	1.93	1.98	2	2
P-BA	1.6	0	1.6	1.84	1.92	1.94	1.88	1.96	1.98	1.65	1.05	1.65	1.65	1.05	1.65	1.65	1.05	1.9	1.97	1.99	2
H-BA	1.81	1.6	0	1.6	1.9	1.93	1.64	1.94	1.97	1.86	1.65	1.05	1.86	1.65	1.05	1.86	1.65	1.65	1.95	1.98	2
L-R2Ap	1.88	1.84	1.6	0	1.88	1.91	1.04	1.92	1.95	1.93	1.9	1.65	1.93	1.9	1.65	1.93	1.9	1.05	1.93	1.96	2
P-R2Ap	1.93	1.92	1.9	1.88	0	1.6	1.92	1.04	1.64	1.98	1.97	1.95	1.98	1.97	1.95	1.98	1.97	1.93	1.05	1.65	2
H-R2Ap	1.95	1.94	1.93	1.91	1.6	0	1.95	1.64	1.04	2	1.99	1.98	2	1.99	1.98	2	1.99	1.96	1.65	1.05	2
L-R2Ju	1.92	1.88	1.64	1.04	1.92	1.95	0	1.88	1.91	1.92	1.88	1.64	1.92	1.88	1.64	1.92	1.88	1.04	1.92	1.95	2
P-R2Ju	1.97	1.96	1.94	1.92	1.04	1.64	1.88	0	1.6	1.97	1.96	1.94	1.97	1.96	1.94	1.97	1.96	1.92	1.04	1.64	2
H-R2Ju	1.99	1.98	1.97	1.95	1.64	1.04	1.91	1.6	0	1.99	1.98	1.97	1.99	1.98	1.97	1.99	1.98	1.95	1.64	1.04	2
L-CFDSe	1.05	1.65	1.86	1.93	1.98	2	1.92	1.97	1.99	0	1.6	1.81	1	1.6	1.81	1	1.6	1.88	1.93	1.95	2
P-CFDSe	1.65	1.05	1.65	1.9	1.97	1.99	1.88	1.96	1.98	1.6	0	1.6	1.6	1	1.6	1.6	1	1.84	1.92	1.94	2
H-CFDSe	1.86	1.65	1.05	1.65	1.95	1.98	1.64	1.94	1.97	1.81	1.6	0	1.81	1.6	1	1.81	1.6	1.6	1.9	1.93	2
L-CM	1.05	1.65	1.86	1.93	1.98	2	1.92	1.97	1.99	1	1.6	1.81	0	1.6	1.81	1	1.6	1.88	1.93	1.95	2
P-CM	1.65	1.05	1.65	1.9	1.97	1.99	1.88	1.96	1.98	1.6	1	1.6	1.6	0	1.6	1.6	1	1.84	1.92	1.94	2
H-CM	1.86	1.65	1.05	1.65	1.95	1.98	1.64	1.94	1.97	1.81	1.6	1.81	1.6	1.6	0	1.81	1.6	1.6	1.9	1.93	2
L-MOF	1.05	1.65	1.86	1.93	1.98	2	1.92	1.97	1.99	1	1.6	1.81	1	1.6	1.81	0	1.6	1.88	1.93	1.95	2
H-MOF	1.65	1.05	1.65	1.9	1.97	1.99	1.88	1.96	1.98	1.6	1	1.6	1.6	1	1.6	1.6	0	1.84	1.92	1.94	2
L-R2Se	1.93	1.9	1.65	1.05	1.93	1.96	1.04	1.92	1.95	1.88	1.84	1.6	1.88	1.84	1.6	1.88	1.84	0	1.88	1.91	2
P-R2Se	1.98	1.97	1.95	1.93	1.05	1.65	1.92	1.04	1.64	1.93	1.92	1.9	1.93	1.92	1.9	1.93	1.92	1.88	0	1.6	2
H-R2Se	2	1.99	1.98	1.96	1.65	1.05	1.95	1.64	1.04	1.95	1.94	1.93	1.95	1.94	1.93	1.95	1.94	1.91	1.6	0	2
4years	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	0

Table 3. Substitution costs for ante-senior phase

	L-CFDAp	P-CFDAp	H-CFDAp	L-BA	P-BA	H-BA	L-R2Ap	P-R2Ap	H-R2Ap	L-R2Ju	P-R2Ju	H-R2Ju	4years
L-CFDAp	0	1.6	1.81	1	1.6	1.81	1.88	1.93	1.95	1.93	1.98	2	2
P-CFDAp	1.6	0	1.6	1.6	1	1.6	1.84	1.92	1.94	1.9	1.97	1.99	2
H-CFDAp	1.81	1.6	0	1.81	1.6	1	1.6	1.9	1.93	1.65	1.95	1.98	2
L-BA	1	1.6	1.81	0	1.6	1.81	1.88	1.93	1.95	1.93	1.98	2	2
P-BA	1.6	1	1.6	1.6	0	1.6	1.84	1.92	1.94	1.9	1.97	1.99	2
H-BA	1.81	1.6	1	1.81	1.6	0	1.6	1.9	1.93	1.65	1.95	1.98	2
L-R2Ap	1.88	1.84	1.6	1.88	1.84	1.6	0	1.88	1.91	1.05	1.93	1.96	2
P-R2Ap	1.93	1.92	1.9	1.93	1.92	1.9	1.88	0	1.6	1.93	1.05	1.65	2
H-R2Ap	1.95	1.94	1.93	1.95	1.94	1.93	1.91	1.6	0	1.96	1.65	1.05	2
L-R2Ju	1.93	1.9	1.65	1.93	1.9	1.65	1.05	1.93	1.96	0	1.88	1.91	2
P-R2Ju	1.98	1.97	1.95	1.98	1.97	1.95	1.93	1.05	1.65	1.88	0	1.6	2
H-R2Ju	2	1.99	1.98	2	1.99	1.98	1.96	1.65	1.05	1.91	1.6	0	2
4years	2	2	2	2	2	2	2	2	2	2	2	2	0

of competition. The only difference in the definition of substitution costs between the two phases (see Tables 2 and 3, pp. 9-10) is related to age categories: substitution costs between junior and apprentice competitions are higher during ante-senior phase than during senior phase.

Insertion-deletion cost (indel) is set to 1.3 for both phases, between minimum substitution cost for distinct states (1) and minimum substitution cost for states differing both in terms of ranking and pre-selection (1.6). In other words, it is less costly to turn a sequence ABC into AB than to turn ABC into ABD only if C and D are similar in ranking or in pre-selection contests organization. This indel definition takes into account the unequal length of sequences without making it the first criterion of distance between sequences.

After a comparison between several clustering methods[13], the outcome of Ward (1963) algorithm appeared as the most suitable one. The dendrogram produced indicates a 9-clusters solution nested in a 3-clusters solution, itself nested in a 2-clusters solution. I comment on the 9-clusters solution from the figures (2 to 4) and tables (4 to 7) presented below.

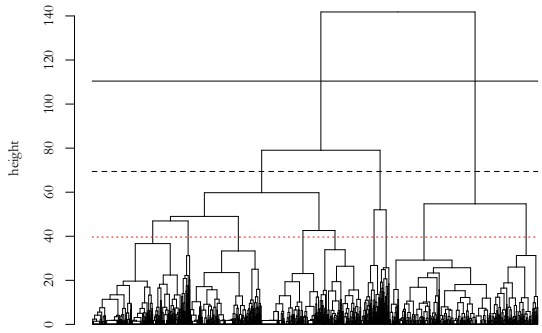
2.5 Two-phase structure as a key to interpreting clustering

I briefly describe each cluster. Three key elements of interpretation arose: participation in senior competitions, length and 'tonality' (that is the most often reached rank). Regarding high heterogeneity of data, clustering is quite noisy.

Cluster 1 gathers sequences characterized by a short but not empty ante-senior phase. Cluster 2 is defined by at least one 4-years void before the first participation in a senior competition. Sequences in cluster 3 share a symmetrical intensity regarding participation in ante-senior and senior phases. The senior phase is mainly characterized by 1^{st} to 3^{rd} ranks. Cluster 4 gathers short to medium length senior sequences mainly characterized by podium positions (2^{d} or 3^{rd} ranks). Cluster 5 gathers short to medium length senior careers too, but mainly characterized by 1^{st} ranks. Cluster 6 gathers short failure senior careers. Cluster 7 gathers senior careers characterized

[13] Single, Complete, UPGMA, WPGMA, Ward and Beta-flexible (Maechler and al., 2016).

Fig. 2. Dendrogram



by a 4-years void after the first participation in a senior competition, that is mainly due to participation in specific competitions that are organized only every 4 years (the golden bricks on sequence index plots). Cluster 8 gathers mean to long senior careers of failures. Cluster 9 gathers long senior careers of success and is the closest to what can be called a cumulative pattern. Except for cluster 9, this clustering does not reveal a sharp connection between length and career structure.

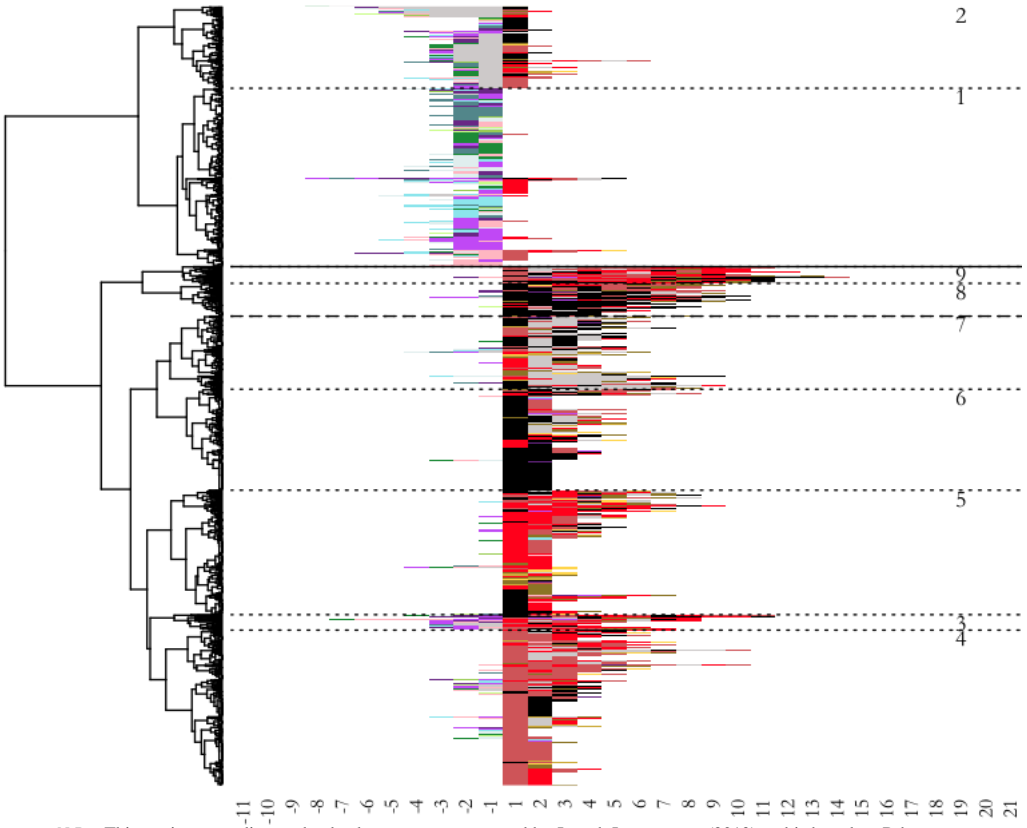
How far does this clustering take synchronization into account ? Four points can be stressed:

- First, the 2-clusters solution separates clusters 1 and 2 from the seven other clusters. In other words, the 2-clusters solution separates careers firstly defined by ante-senior participations from careers firstly defined by senior participations.
- Second, clusters 1 and 2, both characterized by ante-senior participations, are distinct from one another with respect to participation in senior competitions.
- Third, a quarter of sequences counting one or more ante-senior participations are not clustered in clusters 1 and 2. In other words, closeness does not only rest on the (non-)emptiness of phases, but also on phases' composition (what I called tonality).
- Fourth, when, as here, synchronization is endogenous, multiphase structure simplifies greatly the interpretation. Once the phase mainly portrayed by each cluster has been highlighted, clustering's interpretation is based primary on ranking.

Conclusion

As is often the case, the structure of this paper is the exact reverse of the research story. What I have called MPOM from page 1 was first an *ad hoc* method suited to compare a set of careers of participation in *pâtissiers*' competitions. It was a kind of

Fig. 3. Sorted Sequence Index Plot



N.B. : This sorting according to the dendrogram was suggested by Joseph Larmarange (2013) and is based on R heatmap function (R Core Team, 2015). Line types refer to the three different clustering solutions kept (plain line for 2-clusters, dashed for 3-clusters and dotted for 9-clusters).

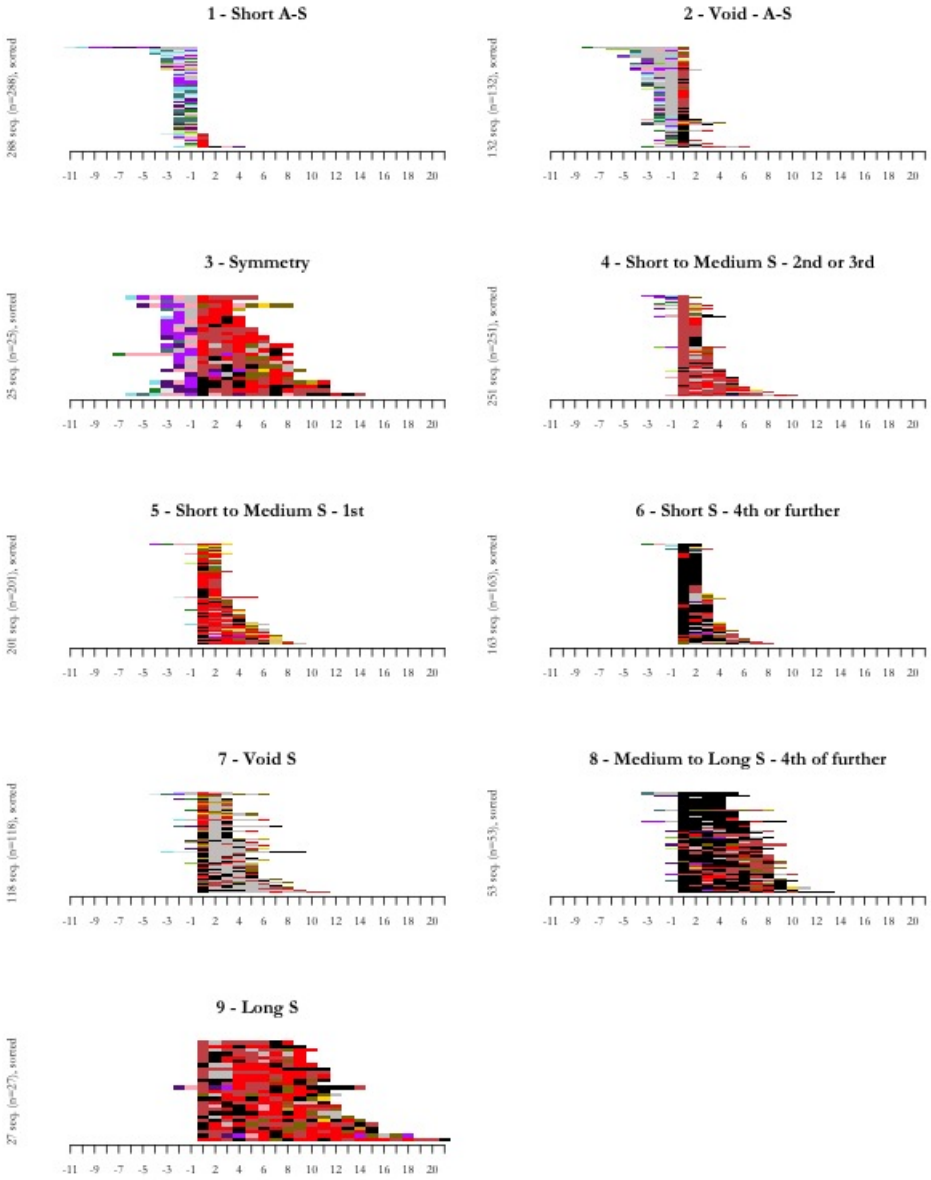
bricolage, a simple way of keeping synchronization while computing distances. That brought me to figure out what synchronization means. Let us underline two features.

First, synchronization means making phases a main character in the story. It is not at all a result, but a postulate made at the beginning of the study. Identifying what phases appear as the most important to explain differences can be regarded as a result. On this point, relations between phases and sequences could be investigated further.

Second, synchronization obviously means distorting the time axis to replace it with an odd sliced one, especially when the number of phases is superior to two. The aim of this calendar timing slaughter is to preserve an assumed social timing with phases regarded as the coherent trajectories within a career[14]. MPOM can be an

[14]For sure, these are two heavy sociological postulates: something is a career – often because different events imply the same human being – and some others things are trajectories

Fig. 4. MDS Sequence Index Plot for each cluster



N.B : Following Piccarretta and Lior (2010), sequences are sorted according to their score on the first factor derived by applying multidimensional scaling (MDS) to the dissimilarity matrix.
A-S = Anter-Senior ; S = Senior.

Table 4. Mean state distribution by cluster within Ante-Senior phases for which length is superior to zero (in length percentage)

	1		**2**		**3**		**4**		**5**		**6**		**7**		**8**		**9**	
$N =$	288		132		25		53		19		9		19		11		1	
Mean/St. dev.	$\bar{X}$	σ	$\bar{X}$	σ	$\bar{X}$	σ	$\bar{X}$	σ	$\bar{X}$	σ	$\bar{X}$	σ	$\bar{X}$	σ	$\bar{X}$	σ	$\bar{X}$	σ
L-R2Ap	13	22.9	3.2	14.8	1	5	10.4	28.3	4.4	13.4	3.7	11.1	10.1	21.3				
P-R2Ap	14.2	25.7	6.4	17.5	4.3	10.5	10.7	28.1	5.3	22.9	11.1	33.3	12.3	21.4				
H-R2Ap	15.6	27.5	7.4	19.9	6.7	14.4	3.8	16.6	8.8	21.1	11.1	33.3	10.1	21.3	16.7	32.5		
L-R2Ju	13.8	26.4	9	20.5	17.6	21.8	21.1	35.7	17.1	33.4	37	48.4	2.6	11.5			50	
P-R2Ju	17.7	28	11	22.3	31.7	28.1	17.6	30.4	22.4	41.6	5.6	16.7	15.4	33.8	15.2	34.5		
H-R2Ju	9.5	21	6.8	20.2	17.8	26.6	8.5	18.1					21.1	38.4	36.4	50.5	50	
L-BA	2.3	9.7	0.3	2.9			3.8	16.6	7.9	25.1					9.1	30.2		
P-BA	3.3	12.6	0.4	3.6			4.4	17	10.5	31.5			10.5	31.5	9.1	30.2		
H-BA	8.2	19	6.9	19.6	1.6	5.7	5.7	21.2	11.8	31.6	25.9	43.4	9.6	25.6				
L-CFDAp	0.2	2.9	0.3	2.9			1.9	13.7	5.3	22.9								
P-CFDAp			1.1	9.2					5.3	22.9								
H-CFDAp	1	10.2	2.3	8.5			2.8	15.2										
4years	1.1	6.3	44.9	25.4	19.3	26.2	9.4	19.7	1.3	5.7	5.6	16.7	8.3	17.3	13.6	24.5		

Table 5. Mean state distribution by cluster within Senior phases for which length is superior to zero (in length percentage)

	1		**2**		**3**		**4**		**5**		**6**		**7**		**8**		**9**	
$N =$	49		130		25		251		201		163		118		53		27	
Mean/St. dev.	$\bar{X}$	σ	$\bar{X}$	σ	$\bar{X}$	σ	$\bar{X}$	σ	$\bar{X}$	σ	$\bar{X}$	σ	$\bar{X}$	σ	$\bar{X}$	σ	$\bar{X}$	σ
L-R2Se	51	45.3	16.7	33.8	39.3	24	13.1	18.8	44.4	26.8	7.6	15.4	11.7	16.9	5.8	8.9	32.8	11.3
P-R2Se	36.3	45	29.7	43.2	29.7	22	60.2	25.9	14.6	18.5	13	20.6	11	14.6	16.4	17.7	28	16.2
H-R2Se	3.6	11.6	41.8	47.1	13	14.5	11.9	19.4	14.4	19.6	61.9	30.1	22.9	22.1	65.7	24.6	20.7	13.6
L-CM	1.1	5.5	0.7	4.5	0.4	2.2	0.5	3.9	3.5	10.7	1.4	7.2	0.8	4	0.2	1.5	0.3	1.7
P-CM			0.9	6	1.4	5.2	1.6	7.5	1.9	8.5	0.5	4.7	2.4	10.6	0.2	1.7	0.2	1.3
H-CM			0.3	2.9	1.1	3.7	2.2	8.9	6.5	15.5	1.4	5.6	0.9	5	1.8	6.4	3	4.1
L-CFDSe			1.7	12.5	0.5	2.5	0.4	4.5	1	6.2	0.4	2.9						
P-CFDSe			0.4	4.4			0.1	2.1	0.7	5.1	0.1	1.6	1.1	4.8			0.3	1.6
H-CFDSe	2	14.3	2.3	15.1	0.5	2.5	0.7	4.8	2.6	10.5	1.7	6.5	3.8	13.4	1	5.5	0.3	1.6
L-MOF															0.2	1.2	0.3	1.4
H-MOF	0.5	3.6	3.7	17.9	5.9	10.9	0.5	3.7	3.6	10.4	1.5	6.3	4.1	10.6	1.3	3.7	4.1	5.1
L-R2Ap							0.2	3.2	0.2	3.5								
P-R2Ap	0.5	3.6					0.2	3.2	0.7	5.5	0.3	3.9						
H-R2Ap											0.4	4.1						
L-R2Ju	2.6	10.5			1.1	3.7	0.3	3.3	0.7	5.3	0.2	2	0.2	2.3	0.5	2.4	1.1	3.6
P-R2Ju	1	7.1			1.4	4	0.6	5.5	0.4	3.1	1.2	6.9	0.1	1.5	0.6	2.6	0.9	3.4
H-R2Ju	0.5	3.6			0.7	2.6	0.3	3.6	0.7	4.9	0.8	6.1	0.5	3	0.9	4.1	0.5	1.7
L-BA							0.1	1.6										
P-BA							0.1	1.6	0.1	1.4								
H-BA									0.2	2.4	0.2	2.6			0.2	1.7		
4years	0.9	4.5	1.9	7.8	5	9.6	6.9	13.4	3.8	9	7.4	12.9	40.5	13.8	5	8	7.5	8.6

alternative to multiple-sequence analysis (Pollock, 2007 ; Gauthier and al., 2010) if one channel appears to be tightly structured along phases. MPOM can be combined with this definition of multi-channel distance if more than two channels are taken into account. For example, individual sequences can be defined as a combination

within this career – often because we consider that some events are more challenging than others for this same human being.

Table 6. Some features of the 9 clusters

Cluster	Frequency		Mean value of three metrics		
			Distance	*Complexity*	*Length*
1 - Short Ante-Senior	288	22.9 %	4.31	0.41	2.67
2 - Void Ante-Senior	132	10.5 %	5.51	0.53	4.05
3 - Symmetry	25	2.0 %	12.66	0.61	10.20
4 - Short to Medium Senior - 2^d or 3^{rd}	251	20.0 %	4.56	0.41	3.59
5 - Short to Medium Senior - 1^{st}	201	16.0 %	4.96	0.46	3.56
6 - Short Senior - 4^{th} or further	163	13.0 %	3.44	0.33	3.11
7 - Void Senior	118	9.4 %	6.30	0.50	4.97
8 - Mean to Long Senior - 4^{th} or further	53	4.2 %	7.26	0.34	7.51
9 - Long Senior	27	2.1 %	12.84	0.54	12.07

Table 7. Sets of representative sequences for each cluster

Cluster	**Ante-senior phase**	**Senior phase**
1	(H-R2Ap,1)-(P-R2Ju,1)	
2	(P-R2Ju,1)-(4years,1)	(H-R2Se,1)
3	(P-R2Ju,1)-(4years,1)	(P-R2Se,1)-(L-R2Se,2)
	(H-R2Ap,1)-(4years,1)-(P-R2Ju,1)	(L-R2Se,1)-(P-R2Se,2)-(L-R2Se,1)-(4years,1)-(L-R2Se,1)-(H-R2Se,1)
	(P-R2Ju,1)-(L-R2Ju,1)-(H-R2Ju,1)	(L-R2Se,1)-(P-R2Se,4)
4		(P-R2Se,2)
5		(L-R2Se,2)
6		(H-R2Se,2)
7		(H-R2Se,1)-(4years,2)-(H-R2Se,1)
8		(H-R2Se,5)
9		(P-R2Se,1)-(H-R2Se,1)-(P-R2Se,1)-(L-R2Se,1)-(P-R2Se,2)-(L-R2Se,2)-(H-R2Se,1)-(P-R2Se,1)
		(P-R2Se,1)-(H-R2Se,1)-(L-R2Se,1)-(H-R2Se,1)-(P-R2Se,1)-(4years,1)-(L-R2Se,1)-(H-CM,1)-(L-R2Se,1)
		(P-R2Se,1)-(L-R2Se,1)-(P-R2Se,3)-(L-R2Se,1)-(H-R2Se,1)-(H-MOF,1)-(L-R2Se,2)-(H-R2Se,1)
		(P-R2Se,1)-(L-R2Se,2)-(H-R2Se,1)-(L-R2Se,2)-(H-R2Se,3)
		(P-R2Se,1)-(4years,2)-(L-R2Se,3)-(P-R2Se,1)-(L-R2Se,2)-(P-R2Se,1)-(L-R2Se,1)
		(P-R2Se,1)-(4years,2)-(P-R2Se,3)-(L-R2Se,3)-(P-R2Se,2)
		(P-R2Se,1)-(L-R2Se,1)-(H-R2Se,1)-(P-R2Se,2)-(L-R2Se,2)-(H-MOF,1)-(4years,2)-(H-R2Se,1)
		(P-R2Se,1)-(4years,1)-(P-R2Se,4)-(H-R2Se,1)-(L-R2Se,1)
		(H-R2Se,2)-(L-R2Se,2)-(H-R2Se,1)-(L-R2Se,1)-(P-R2Se,1)-(L-R2Se,1)-(H-R2Se,1)-(H-CM,1)-(L-CM,1)

N.B.: The centrality criterion has been used to extract these sets of representatives (Gabadinho and Ritschard, 2013). Coverage threshold is 50 % and neighborhood radius is 30 %. In other words, in each cluster, the distance of at least one sequence out of two to one of the representative sequences is inferior to 30 % of the maximum distance within each cluster. Sequences are sorted by representativeness. This presentation of sequences follows Abbott and Hrycak (1990), what has been named State-Permanence Sequence (SPS) format (Aassve and al., 2007): each state is followed by its number of successive occurrences.

of employment and residential sequences divided into phases defined by a marital sequence.

MPOM is thus a simple by-product of OMA, a by-product which is heuristic if sequences are defined by a set of phases and which takes into account the succession of states within phases. That is a major difference with other methods making the postulate of a phase division, such as QHA (Deville, 1974).

References

Aassve, A., Billari, F. Piccarreta, R. (2007). Strings of Adulthood: A Sequence Analysis of Young British Women's Work-Family Trajectories. *European Journal of Population*, 23-3. 369-388.

Abbott, A. D. (1992). What Do Cases Do ? Some Notes on Activity in Sociological Analysis. In Becker, H. S., & Ragin, C. C. (eds.). *What is a Case ? Exploring the Foundations of Social Inquiry* (pp. 53-82), Cambridge: Cambridge university.

Abbott, A. D. (1997). On the concept of turning point. *Comparative Social Research*,16. 85-106.

Abbott, A. D., & Forrest, J.. (1986). Optimal Matching Methods for Historical Sequences. *The Journal of Interdisciplinary History*, 16. 471-494.

Abbott, A. D., & Hrycak, A.. (1990). Measuring Resemblance in Sequence Data: An Optimal Matching Analysis of Musicians' Careers. *American Journal of Sociology*, 96-1. 144-185.

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

Becker, H. S. (1963). *Outsiders: Studies in the Sociology of Deviance*. New York: The Free Press.

Blanchard, Ph. (2010). Analyse séquentielle et carrières militantes. Research Report. http://hal.archives-ouvertes.fr/hal-0047619. Accessed 10 January 2015.

Collas, T. (2015). La pâte et le décor: Considération et formes professionnelles dans le monde des pâtissiers. PhD Thesis. Sciences Po.

Colombi, D., & Paye, S. (2014). Synchronising Sequences: An Analytic Approach to Explore Relationships Between Events and Temporal Patterns. In Blanchard, Ph., Bhlmann, F., Gauthier, J.-A. (Eds.). *Advances in Sequence Analysis: Theory, Method, Applications* (pp. 249-264), New York Heidelberg Dordrecht London: Springer.

Deville, J.-C. (1974). Méthodes statistiques et numériques de l'analyse harmonique. *Annales de l'Insee*, 15. 3-101.

DiPrete, T. A., & Eirich, G. M. (2006). Cumulative Advantage as a Mechanism for Inequality: A Review of Theoretical and Empirical Developments. *Annual Review of Sociology*, 32. 271-297.

Elzinga, C. H. (2014). Distance, Similarity and Sequence Comparison. In Blanchard, Ph., Bhlmann, F., Gauthier, J.-A. (Eds.). *Advances in Sequence Analysis: Theory, Method, Applications* (pp. 51-73), New York Heidelberg Dordrecht London: Springer.

Gabadinho, A. & Ritschard, G. (2013). Searching for typical life trajectories applied to child birth histories In R Lévy & E. Widmer (eds.), *Gendered Life Courses* (pp. 287-312). Vienna: LIT.

Gabadinho, A., Ritschard, G., Mller, N. S., Studer, M. (2011). Analyzing and Visualizing State Sequences in R with TraMineR. *Journal of Statistical Software*, 40(4). 1-37.

Gabadinho, A., Ritschard, G., Studer, M., Mller, N.S. (2010). "Indice de complexité pour le tri et la comparaison de séquences catégorielles", In *Extraction et gestion des connaissances (EGC 2010), Revue des nouvelles technologies de l'information RNTI*. E-19. 61-66.

Gauthier, J.-A., Bhlmann F., Blanchard Ph. (2014). Introduction: Sequence Analysis in 2014. In Blanchard, Ph., Bhlmann, F., Gauthier, J.-A. (Eds.). *Advances in Sequence Analysis: Theory, Method, Applications* (pp. 1-17), New York Heidelberg Dordrecht London: Springer.

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

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

Lang, G., & Lang, K. (1988). Recognition and Renown: the Survival of Artistic Reputation. *American journal of sociology*, 94. 79-109.

Larmarange, J. (2013). Représenter un tapis de séquences ordonnées selon un dendrogramme. http://joseph.larmarange.net/?Representer-un-tapis-de-sequences. Last access: Feb. 2nd 2015.

Lesnard, L. (2008). Off-Scheduling within Dual-Earner Couples : An Unequal and Negative Externality for Family Time. *American Journal of Sociology*, 114. 447-490.

Lesnard, L. (2014). Using Optimal Matching Analysis in Sociology: Cost Setting and Sociology of Time. In Blanchard, Ph., Bhlmann, F., Gauthier, J.-A. (Eds.). *Advances in Sequence Analysis: Theory, Method, Applications* (pp. 39-50), New York Heidelberg Dordrecht London: Springer.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., Hornik, K.(2016). cluster: Cluster Analysis Basics and Extensions. R package version 2.0.4.

Menger, P.-M. (2009). *Le travail créateur: S'accomplir dans l'incertain*. Paris: Gallimard-Seuil.

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

Pollock, G. (2007). Holistic Trajectories: a Study of Combined Employment, Housing and Family Careers by Using Multiple-Sequence Analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 170. 167-183.

R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria.

Robette, N., & Bry, X. (2012). Harpoon or Bait? A Comparison of Various Metrics in Fishing for Sequence Patterns. *Bulletin de Mthodologie Sociologique*, 116. 524.

Robette, N., & Thibault, N. (2008). Analyse harmonique qualitative ou méthodes d'appariement optimal?. *Population*, 63(4). 621646.

Stovel, K., & Bolan, M. (2004). Residential Trajectories: Using Optimal Alignment to Reveal the Structure of Residential Mobility. *Sociological Methods & Research*, 4. 559-598.

Stovel, K., Savage, M., Bearman, P. (1996). Ascription into Achievement: Models of Career Systems at Lloyds Bank, 1890-1970. *American Journal of Sociology*, 102-2. 358-399.

Ward, J. H. Jr. (1963). Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*, 58. 23644.

Wilensky, H. L. (1964). The Professionalization of Everyone?. *American Journal of Sociology*, 70. 137-158.

Wu, L. L. (2000). Some Comments on Sequence Analysis and Optimal Matching Methods in Sociology: Review and Prospect. *Sociological Methods & Research*, 29. 41-64.