Optimal Matching and Social Sciences

Laurent Lesnard

Optimal Matching and Social Sciences

Laurent Lesnard[1]

This working paper is a reflection on the conditions required to use optimal matching (OM) in social sciences. Despite its striking success in biology, optimal matching was not invented to solve biological questions but computer science ones: OM is a family of distance concepts originating in information and coding theory were it is known under various names among which Hamming, and Levenshtein distance. As a consequence, the success of this method in biology has nothing to do with the alleged similarity of the way it operates with biological processes but with choices of parameters in accordance with the kind of materials *and* questions biologists are facing. As materials and questions differ in social sciences, it is not possible to import OM directly from biology. The very basic fact that sequences of social events are not made of biological matter but of events and time is crucial for the adaptation of OM: insertion and deletion operations warp time and are to be avoided if information regarding the social regulation of the timing of event is to be fully recovered. A formulation of substitution costs taking advantage of the social structuration of time is proposed for sequences sharing the same calendar: dynamic substitution costs can be derived from the series of transition matrices describing social sub-rhythms. An application to the question of the scheduling of work is proposed: using data from the 1985-86 and 1998-99 French time-use surveys, twelve types of workdays are uncovered. Their interpretability and quality, assessed visually through aggregate and individual tempograms, and box plots, seem satisfactory.

Ce document de travail se veut une réflexion sur les conditions d'utilisation des méthodes d'appariement optimal en sciences sociales. En dépit de ses succès retentissant en biologie, l'appariement optimal n'a pas été inventé pour résoudre des problèmes en biologie mais en informatique et en théorie du codage où il est connu notamment sous les noms de distance de Hamming et de Levenshtein. Par conséquent, le succès de cette méthode en biologie n'a rien à voir avec la possible ressemblance de son mode opératoire avec des processus biologiques mais provient de choix de paramètres cohérents avec le type de matériel et de questions auxquels les biologistes font face. Parce qu'en sciences sociales les séquences sont composées d'événements et de temps, les succès obtenus en biologie ne peuvent être importés directement. En particulier, les opérations d'insertion et de suppression distordent l'échelle temporelle et doivent être évitées dès lors que l'objectif de l'analyse est de repérer la régulation sociale du timing des événements. Une formulation des coûts de substitution tirant profit de la structure sociale du temps est proposée pour les séquences qui partagent le même calendrier : des coûts de substitution dynamiques peuvent être dérivés de la série de matrices de transitions qui décrit les sous rythmes. Une application à la question des rythmes de travail est proposée : douze types de journées de travail sont identifiés dans les enquêtes sur l'emploi du temps réalisées par l'Insee en 1985-86 et 1998-99. Leurs interprétabilité et qualité, visuellement évaluées au travers de chronogrammes agrégés et individuels et de diagrammes à surfaces, apparaît satisfaisante.

[1] Observatoire sociologique du changement (Sciences-po & CNRS) and Laboratoire de sociologie quantitative (Crest – Insee). Please address correspondence to laurent.lesnard@sciences-po.fr.

Introduction

Dynamic statistical models appeared in social sciences at the dawn of the 1980s: in the first review dedicated to dynamic models, Nancy Tuma, Michael Hannan and Lyle Groeneveld (1979) were enjoining social scientists to incorporate these new tools made available by the development of personal computers. This "dynamic turning point" has been successful in view of the widespread use of dynamic regressions and other duration models as well as their growing sophistication. According to Andrew Abbott (1998), this success can be attributed to a large extent to the first generation of users who was also successful in turning this methodological novelty into academic power, but also to the "causal devolution" in social sciences and the consubstantial devaluation of description in favor of modeling.

However, rather than being used in a true hypothesis testing perspective, statistical models are generally used as descriptive tools (Abbott, 1998). This is not completely absurd since statistical models and descriptive tools both aim at representing data in the best possible way. Both are *statistical abstractions*[2], but with different degrees in abstraction: models are more abstract as a result of additional assumptions made[3]. However, this greater simplification power is fragile for it relies completely on those assumptions, sometimes untestable and often untested.

Consequently, the problem of modeling is not very different from the more general issue of articulating statistical abstractions with the real: in order to be faithful to facts, simplification must be progressive[4]. In this respect, statistical abstractions developed for the analysis of sequences are limited to models resting on hypotheses which are all the more strong that no basic descriptive methods adapted to this kind of data are equally available. A new tool, Optimal Matching (called OM in the rest of this paper), introduced in social sciences by Andrew Abbott and other authors (Abbott, 1984; Abbott and Forest, 1986; Abbott and Hrycak, 1990) is a good candidate to describing sequences of social events. However, this new descriptive method has not yet been adopted by social scientists.

In OM, the degree of dissimilarity between two sequences is determined by the least number of edit operations that are necessary to turn one sequence into the other (i.e. to match the two sequences). Three kinds of edit operations are generally used: insertion, deletion, and substitution. OM will be introduced in more details later, at present it is only necessary to understand that sequences are manipulated, transformed, altered, with the help of a certain number of basic operations, in order to assess their degree of similarity.

In the ancestor of OM, the Levenshtein distance (Levenshtein, 1966 [1965]), the three basic operations are given equal weights: each operation costs one unit. In theory, the choice of a cost system determines the matching procedure and to a certain extent the results obtained. In social sciences, most OM users claim that results are little affected by changes in the relative weights of the three basic operations (for a review see Abbott and Tsay, 2000). OM detractors have been interpreting this as a

[2] This expression is used by François Simiand in his seminal book on the use of statistics in social sciences (1922).

[3] The affinity between regression analysis and correspondence analysis is better seen when the geometric dimension of regression is taken into account. For more details about this subject, see Rouanet *et al.* (2002) and Le Roux and Rouanet (2004).

[4] "Ne nous lassons donc pas de répéter que, pour avoir chance de ne pas se prendre à des représentations inexactes et par suite à des coïncidences fortuites ou trompeuses, notre expérimentation statistique doit toujours s'appliquer à saisir, d'abord, dans son allure propre le fait étudié, à le saisir dans la *succession de ses phases*, dans la *décomposition de ses parties* si c'est le cas ; et si elle en simplifie ensuite l'expression, comme il est peut-être utile ou nécessaire pour la recherche même, si elle en laisse tomber telles ou telles particularités pour n'en retenir que certaines autres, elle doit savoir qu'elle fait cette élimination et pourquoi et avec quelles conséquences possibles sur les résultats ultérieurs. " (Simiand, 1922, p. 48).

sign, not of robustness, but, often mistaking OM for a model, of weakness (Levine, 2000). There has also been concern about the sociological meaning of the three basic operations of OM (Levine, 2000; Wu, 2000; Elzinga, 2003): it has been argued that the legitimacy of OM in biology was stemming from the theoretical relevance of the three edit operations. In sociology, these operations seem loosely linked to theory and the choice of a particular cost system appears arbitrary. As a set of descriptive tools, however, it seems natural that the optimal matching analysis of sequences with different cost settings yields roughly the same results (Abbott, 2000). Nonetheless, if OM is not to be limited to exploratory analysis, then attention should be paid to these differences, hence to the empirical and sociological consequences of cost settings, or in other words to the sociological meaning and consequences of sequence transformations in OM.

This working paper tries to address this issue and is organized as such. It is meant to be a methodological as well as a theoretical reflection on the use of OM in social sciences but neither an exhaustive review of the different use of OM in social sciences nor an in-depth technical presentation of OM[5]. However, OM is first presented in a non-technical way but with sufficient details to grasp its functioning. As biology is often referred to when the use of OM in social sciences is assessed the specificity of the use of OM in this scientific discipline is emphasized. In order to understand the specificity of sequences in social sciences, the second part of this paper focuses on time and on the epistemological consequences of OM in this respect. A method is proposed taking into account these specificities. Finally, an application of this method to work schedules is presented.

Optimal Matching and its use in biology

A short history and non-technical presentation of Optimal Matching

Optimal Matching is a family of dissimilarity measures between sequences derived from the distance originally proposed in the field of information theory and computer science by Vladimir Levenshtein (1966 [1965]). What is known in biology, and now in social sciences, under the name of *sequence analysis* is in fact coming from research on *coding theory* and *string editing*. Coding theory refers to the body of research dealing with the reception of coded information through noisy channels such as radio or telegraph. Strings are basic components of computer science and the indispensable 'find' or 'replace' functions of text processing software are probably the most obvious implementation of such algorithms.

The Levenshtein or edit distance between two sequences (or string in the computer science vocabulary) is given by the smallest number of operations needed to turn one sequence into the other (i.e. to match them). The different edit operations allowed, insertion, deletion, or substitution, are penalized by a cost, which is equal to one in the original version of OM6. Levenshtein also suggested using only insertion and deletion operations to match strings. These two Levenshtein distances are usually considered as an improvement of the distance proposed by Richard Hamming (1950). The Hamming distance between two sequences is the number of substitutions required to change one sequence into the other. As a result, and contrary to the Levenshtein distance, the Hamming distance can only be applied to sequences of equal length. It is interesting to note that the Hamming distance is related to the Manhattan distance, or L1 distance and, as a consequence, is not a Euclidean metric.

[5] A review of recent applications of sequence analysis in social sciences can be found in Abbott and Tsay (2000). The standard text on sequence analysis in computer sciences and computational biology is Sankoff and Kruskal (1983). A more recent reference is Durbin *et al.* (1998).

[6] Kruskal suggests a substitution penalty at least equal to 2, arguing that if the substitution cost is greater than 2 than "it is always shorter for a listing to use a deletion-insertion pair in place of a substitution, and if [it is equal to 2] it is as short" (1983, p. 18).

Consequently, OM refers to the more general solution proposed by Levenshtein to the problem of sequence comparison and encompasses two particular cases: when comparison is restricted to either substitution or insertion-deletion operations (see Table 1)

	Operations used	
	Substitution	**Insertion and deletion**
Hamming	Yes (cost=1)	No
Levenshtein I (OM)	Yes (cost=1)	Yes (cost=1)
Levenshtein II	No	Yes (cost=1)

Table 1 – OM and the Hamming and Levenshtein distances

For example, the Levenshtein I distance between the sequences *S1* and *S2* (see Table 2) is 2. The matching process can be represented in a matrix where horizontal, vertical, and diagonal movements correspond to the three edit operations (respectively an insertion, a deletion and a substitution) and each cell contains the cumulated minimum cost to reach it (see Figure 1). The optimal path is represented by connected circles. The Hamming distance between *S1* and *S2* is simpler to calculate: as there is no *episode*[7] common to the two sequences, the distance is 4, that is to say the length of the sequences.

	Episodes			
	0	**1**	**2**	**3**
S1	A	B	C	D
S2	D	A	B	C

Table 2 – Two basic sequences of equal length

		-1	0	1	2	3
			A	B	C	D
-1		0	1	2	3	4
0	D	1	1	2	3	3
1	A	2	1	2	3	4
2	B	3	2	1	2	3
3	C	4	3	2	1	2

Figure 1 – Matrix representation of the Levenshtein distance calculation between *S1* and *S2*

The matrix representation of the matching process helps to understand how the algorithm works. OM is by definition an optimization problem: all the possible combinations of edit operations to match two sequences must be considered in order to identify the most efficient solution. This problem can be solved recursively by dynamic programming and is based on the fact that there are only three possibilities to attain a cell: from the left, the top or the diagonal. Each of these three directions corresponds to an edit operation: if the sequence to match is in the columns of the matrix (as in Figure

[7] The i^{th} episode is understood here as the i^{th} component of a sequence. Therefore, an episode has the same location in all the sequences.

1), then a horizontal movement represents the insertion of the corresponding column element after the corresponding row element, a vertical movement is the deletion of the corresponding row element and a diagonal movement is either a cost free movement if the elements of the corresponding row and column are identical or a substitution if not. Each cell contains the minimum cumulative cost to reach it from one of these three possibilities: the top left cell contains 0 and is the starting point whereas the bottom right cell contains the dissimilarity measure of the two sequences.

The correspondence of the horizontal and vertical movements with insertion and deletion operation is reversed when the target sequence is not located in the first row of the matrix but in the first column: insertion and deletion operations are symmetrical and this is why their costs are always identical in OM. This symmetry can also be seen when the target sequence is still located in the first row but this time is not *S1* but *S2* (see Figure 2): the matching matrix is in this case the transposed version of the previous matrix. Insertion and deletions are symmetrical operations and are often jointly referred as *indel*[8] or as *gaps* in biology.

		-1	0	1	2	3
			D	A	B	C
-1		0	1	2	3	4
0	A	1	1	1	2	3
1	B	2	2	2	1	2
2	C	3	3	3	2	1
3	D	4	3	4	3	2

Figure 2 – Matrix representation of the Levenshtein distance calculation between *S2* and *S1*

In theory, the choice of a cost system determines how sequences are matched and to a certain extent the dissimilarities obtained. When substitution operations are not allowed, or, this is exactly the same, when their cost is strictly greater than the cost of an insertion and a deletion, then the Levenshtein distance between two sequences is equivalent to finding their longest common subsequence, whatever their location in the two sequences (Kruskal, 1983, p. 30). On the contrary, using only substitution operations will focus the analysis on finding contemporaneous similarities. OM is a quite flexible family of methods that have been used in numerous fields: computer science, coding theory, speech recognition, bird songs studies, gas chromatography, geology, human depth perception, biology, etc. And of course now social sciences. There is no room here to present, even broadly, what is the meaning of the edit operations and how costs are chosen in all these fields[9]. We preferred to focus on biology given the role this discipline is playing in the assessment of the relevance of OM in social sciences[10].

Before focusing on how weights are determined in biology, it is worth noting that as OM is a kind of correlation coefficient for sequences, the output is a gigantic dissimilarity matrix between all sequences (individuals, and not variables): OM must be combined with cluster analysis, multidimensional scaling, or any other data reduction procedure handling dissimilarity objects. As a

[8] Indel is an acronym formed by the beginnings of *in*sertion and *del*etion and is therefore designating jointly insertion and deletion operations.

[9] For an overview, see Sankoff and Kruskal (1983).

[10] Speech recognition will be also roughly presented at the beginning of the section on OM and social sciences.

consequence, OM's output is always accessed indirectly, most of the time in social sciences through the former technique. This issue will be addressed at the end of this paper.

Optimal Matching and biology

OM techniques were born in computer sciences and were subsequently imported into other scientific fields, among which biology. As OM was imported into social sciences through biology, this scientific field is the *de facto* reference in terms of its integration into pre-existing theories. Indeed, Levine (2000), Wu (2000), and Elzinga (2003) refer to biology to assess the use of OM in social sciences and claim that in biology the edit operations used in OM are linked to chemical properties and transformations of sequences of DNA, RNA and proteins. It can be said here and now that if it were so, several of the fundamental biological operations involved in these transformations, such as swaps and larger transpositions, would be missing (Abbott 2000).

Sequence analysis is used in biology as an approximation to avoid costly and lengthy experimentations. This is not to say that sequence analysis is a computational reproduction of biological experimentations but it is precisely the opposite, a way to solve the question of the identification of the structure and/or functions of DNA or proteins without what is considered as the most reliable way to do so: experimentation (see Durbin *et al.* 1998). To achieve this, the key process is *homology*: information about structure and/or function of sequences *already known by experimentation* is transferred to sequences with which significant similarities are found. Consequently, biological theories are not central in the use of sequence analysis in this field: "most of the problems in computational sequence analysis are essentially statistical" (Durbin *et al.* 1998, p. 1).

OM is one of the tools that have been used and developed in biology to identify these similarities: it is basically an adaptation of the Levenshtein distance to these problems. Therefore, the three edit operations, insertion, deletion and substitution, have nothing to do with biology but, once their relative costs were given some thoughts, were considered as not completely absurd and above all produced results. What separate the Levenshtein I distance from OM, a family of dissimilarities, are the relative costs of the edit operations, called *scoring model* in biology. Consequently, the essence of OM is not in the three edit operations but in the way they are used and combined through cost settings to analyze biological sequences.

The theoretical congruence of OM with biological theory is therefore not as advanced as some have claimed. However, as Elzinga (2003) suggests, "oftentimes, there is a plausible theory or credible hypothesis about the probability that such a set of operations really took place or could have taken place in the course of evolution". As the goal of the analysis is to identify similarities between new sequences and experimentally known sequences, the main difficulty computational biologists are facing is to discern "significant similarities between anciently divergent sequences amidst a chaos of random mutation, natural selection, and genetic drift" (Durbin *et al.*, 1998, p. 1). Consequently, substitution costs must reflect evolutionary preferences for certain evolutions over others. A low substitution cost between two states in an alignment means that under some phylogenetic assumptions the two sequences are probably related. As a result, substitution matrices are above all a question of probability estimation: the main task of computational biologists is to constitute a good sample of confirmed alignments but also of alignments which are plausible under certain phylogenetic assumptions in order to estimate these probabilities.

This is a quite complex operation in practice given that protein sequences come in family and other problems of the same sort. Constituting these matrices requires considerable work and is an essential step in using sequence analysis. The PAM matrices, developed in 1978 by Dayhoff *et al.* are derived from alignments between proteins experimentally or hypothetically related, especially regarding the percentage of accepted mutations (PAM is the acronym of Point Accepted Mutation). Matrices for greater evolutionary distances are extrapolated from this matrix by simply raising it to the power of the evolutionary distance researched. The BLOSUM series of matrices (Henikoff and Henikoff 1992) has

been developed according to the same principles but with a more elaborated treatment of the differences between short time and longer term evolutionary distance.

Computational biologists believe that indel costs should reflect the probability of inserting a gap in a sequence, possibly depending on the kind of "residue" (event) inserted. Insertion and deletion operations are mainly used in biology to take into account possible evolutionary process involving the introduction of some unimportant residues between related alignments. However, although it is also possible to turn the question of the determination of insertion and deletion costs into probability estimation, these costs are often disregarded (Durbin *et al.*, 1998, pp. 16-17 and 44-45).

Elzinga is therefore right when she claims that edit operations are linked to evolutionary hypotheses. However, only substitution cost matrices are given some theoretical attention whereas indel costs are almost always chosen on a complete empirical basis. Furthermore, substitution matrices are not the exact product of chemical or phylogenetic models: theory intervenes mainly in the constitution of the samples of alignments, confirmed or hypothetical, which are used to estimate substitution probabilities. Hence substitution costs are not theoretically determined one by one: theory is just used to provide guidelines to estimate probabilities. The way theory is used is quite interesting: the exact nature of the relations between sequences and their phylogenetically plausible mutations does not need to be perfectly known. These relations are uncovered during the stage of probability estimation and used as a yardstick to distinguish between insignificant (gaps) and significant evolutionary changes in other samples. Contrary to stochastic modeling, the interest is not in a single evolutionary scenario, true on average: all the complexity of the evolutionary change is taken into account and summarized in substitution matrices. In other words, the parameters used by computational biologists are derived from descriptive statistics, of course judiciously chosen.

Elzinga's claim that specifying a cost function is to use a model (2003) is consequently not true, at least in biology: substitution matrices are not generated *ex nihilo* from a pure chemical model but are based on frequencies observed in a particular sample of sequences. The theory intervenes only in the constitution of that sample in a very minimal way: if it was possible to build the substitution matrix out of the chemical and evolutionary properties of biological sequences, OM would simply not be necessary. It is because biological theory is not that advanced that the only solution is to gather hypothetically related sequences and infer probabilities about how they are related. OM is not used in a traditional way in biology since these descriptive statistics are used to detect new similarities in new samples: this is not to say that this is modeling, but a more complex and unusual way to describe data given the particularity of the questions asked and material used in this discipline: in biology, OM is parameterized with descriptive statistics to produce new descriptive statistics.

Consequently, OM in biology is neither a reproduction of the bio-chemical phenomena of interest nor are its parameters derived from standard modeling strategies: OM is used in biology as a descriptive tool. Of course OM is somewhat more elaborated than an arithmetical mean and requires more care to simplify with sufficient accuracy the biological materials. OM was not invented to answer biological questions but to address issues interesting coding theory and computer science. Biologists successfully used this statistical abstraction because they managed to parameterize it to fit the kind of data and problems they were facing. Social sciences share with other sciences, among which biology and other "hard" sciences, the fact that they resort to abstractions to simplify with accuracy an otherwise too complex material (Simiand 1922, pp. 29-30 in particular). As Simiand remarks, the problem is not the abstractions, but to use them in adequacy to the material analyzed: the use of OM in social sciences should be evaluated according to the same principle.

Optimal Matching and social sciences

The role and consequences of edit operations in social sciences

This brief outline of the conditions under which OM is used in biology is emphasizing the key role played by the costs of the three edit operations. This is through the relative penalties associated with these operations that a method elaborated in a totally different scientific field, namely computer science, was adapted to the requirements of biology. OM will only go beyond the marginality of an exotic exploratory technique by clarifying the meaning these operations and costs have for social sciences. As sequences in social sciences are not made of amino acids but express successions of social events, it means that the coding of events, and time are central in this process, and that the question of the relevance of OM in social sciences should be reformulated as whether or not OM represents with a sufficient degree of faithfulness sequences of social events.

Indeed, events are the fabric of sequences in social sciences but are not given, as amino acids are, but constructed and coded by social scientists. As a result, the meaning of the three edit operations depends primarily on the way sequences are constructed, hence on the indispensable preliminary work of constituting the object of the analysis. There is obviously no unique answer to this question, as there are different substitution matrices in biology according to the kind of sequence analyzed and of questions asked. Only scientific debates can contribute to the establishment of guidelines relevant for certain kinds of analysis, career analysis or time-use analysis for instance. This calls for the highest scientific standards in terms of argumentation and clarification of all the details of the analysis, and of sharing algorithms and other data management procedures, an especially crucial point given that this family of methods is not yet widely available in standard statistical packages[11].

	Insertion-Deletion	**Substitution**
Preserved	Events	Time
Altered	Time	Events

Table 3 – Edit operations and sequences of social events

Second, the matter of sequences in social sciences is also *time*. As a consequence, the very fact of manipulating sequences to assess their similarity means for social sciences that OM is based on *manipulations of time*: inserting or deleting an event is also warping the timing of the processes analyzed in order to identify sub-sequences of identically coded events. On the contrary, substituting an event by another means that the timing is preserved but that an event is approximated by another. In summary, insertion and deletion operations preserve the events but distort time while substitution operations just do the opposite, i.e. they conserve time but alter events. As a result, OM with sequences of social events is a combination of accelerations/decelerations to match identical subsequences of events and of events approximations when the flow of time is normal (see Table 3). Note that the expression of "normal flow of time" has been used here: once time has been warped, co-occurrences of events do not mean that these events are necessarily contemporaneous, unless time was accelerated then decelerated to that the calendars of both sequences coincide again.

The warping of time by indel operations has also been studied in the speech recognition field, which shares with social sciences some of their concern with time. In this field, OM is used to: 1. measure the variability of compression-expansion between two sequences 2. determine the degree of resemblance of two sequences independently of differences in compression-expansion 3. build

[11] The program designed by Andrew Abbott, *Optimize*, is no longer maintained but is still available on the author's web page at the University of Chicago. A sequence module is available in the *TDA* package, a freeware developped by Goetz Rohwer and Ulrich Poetter of the University of Bochum to apply event history models.

'average' sequences (Kruskal and Liberman, 1983). In this context, indel operations can be used to compress and expand time so that different delivery speeds of the same words can be taken into account[12] (see Table 4). Warping time is in this field absolutely necessary and can be seen as multiple re-synchronizations of the time scales of two sequences. Time is freely warped here because it is used only as an ordering support, and is in this respect quite similar to residues in biological sequences.

Compression-expansion	Deletion-insertion
Compress 2 units into 1	Delete 1 unit
Expand 1 unit into 2	Insert 1 unit

Table 4 – Correspondence between time warping and indel operations (reproduced from Kruskal and Liberman, 1983)

The question of the use of indel operations to analyze sequences of social events can be reformulated as whether or not it is legitimate to distort time. Warping time means here that events coded identically but occurring at different moments are considered as almost perfectly equivalent except for the weighted number of episodes that separate them. In the Levenshtein distance, indel weights are all equals to 1: time is considered as a linear dimension and neither the nature of the events suppressed nor their location in the sequence are considered as relevant. This would be a rather strong assumption, quite contrary to the shift from causes to events that characterizes OM as "a particular value of [a variable] may have no absolute meaning independent of time […] A given value may acquire significance because it is the first reversal of a long, steady fall, or because it initiates a long steady state. In either case, it is the general temporal context, not the immediate change, that matters." (Abbott, 1990).

Of course when the sequences studied do not share the same time scale, warping time is not really a problem[13]. But when they do, warping time destroys the temporal links between sequences, their *contemporaneity*. Inserting time so that unemployment spells of approximately equal length can be identified means that the events are of importance, not when they occur: events lose their indexicality[14]. Consequently, the use of indel operations with sequences of social events can have undesirable consequences and should be avoided whenever the timing of events is crucial.

In social sciences, sequence analysis is used to "fishing for patterns" (Abbott, 2000), to take into account the complexity of sequences and as such partakes of the break with causes to focus on events. As a descriptive technique, OM should be able to discriminate between events pertaining to different rhythms and events whose cadence is close: the goal of OM in social sciences is ultimately to identify sub-rhythms of social processes[15]. In career analysis, OM has indeed been used to identify different trajectory patterns (see for instance Halpin and Chan, 1998) and in time-use analysis, to locate different daily routines (see for instance Lesnard, 2004 or Saint Pol, 2005). Given that indel operations are

[12] In fact both indel and compression-expansion operations are used in speech recognition. The former are used in order to recover interpolated or deleted sounds (eg. "probably" may be pronounced "prob'ly", etc.) whereas the latter are used to synchronize identical sub-sequences. The difference between these two very similar operations, both implemented by indel operations, lies in their respective costs (more details can be found in Kruskal and Liberman, 1983, especially in the sections 6 and 7): once again, it is through costs that OM can be fine-tuned in order to suit the requirements of the analysis.

[13] Analyzing sequences with different calendars is looking for unvarying patterns, rules which are valid for different historical periods. In other words, the property of indexicality of time is disregarded to focus on transhistorical properties.

[14] On indexicality, see Abbott (1999). Being unemployed in a time of mass unemployment is likely to be a different experience than in a time of full employment.

[15] Consequently, OM is compatible with the theory of time sketched out by Abbott (1999): when mainly substitution operations are used, OM respects indexicality and enables "multiple times" to be identified. This theme is developped in the next section.

warping time and also in that case are blurring the temporal links between individual sequences, indel operations make the identification of sub-rhythms harder and should thence be seldom and carefully used.

Consequently, the question of the legitimacy of OM in social sciences can be reformulated as: what is the meaning of substitution operations in social sciences and above all how to use them, *i.e.* how to choose their costs, in order to identify patterns of chain of social events? Substituting one event for another can be seen as altering one element in a chain of social events: for instance replacing an unemployment spell by a part-time work event. With the exception of preserving the respective time scales of sequences, such operation has no particular meaning in social sciences: it is an abstract operation, in this respect not very different from calculating the arithmetic mean of a series of numbers, used only in order to assess the degree of similarity of sequences. In such a perspective, it does not matter if substitution operations can be or not related to specific social processes: substitution operations are just some of the building blocks of the abstract process of assessing the degree of similarity between sequences.

This is also true in biology: as Abbott made clear, indel and substitution operations have nothing to do with actual biological processes. In the matching process of two biological sequences, completely evolutionary unrelated elements can be substituted with one another, but with a high penalty if the substitution matrix is well defined[16]. As a result, substitution operations are not used in biology as an equivalent to evolutionary transformations of proteins or DNA but are *interpreted* as such only when their costs are low. Substitution operations *per se* are not used as functional equivalent of evolutionary processes. Substitution costs are.

Whereas indel costs should be defined as a function of the temporal proximity of identically coded events, substitution costs should represent the closeness of two different events at a particular position in their respective sequences. As biologists use OM to infer biological properties from known sequences, they want this closeness to be related to evolutionary processes, and they interpret and estimate substitution costs accordingly by using evolutionary evidence and hypotheses. In social sciences, the aim is to identify diverse groups of sequences, *i.e.* multiple sub-rhythms: substitution costs should be interpreted in terms of sub-rhythms and estimated accordingly. As a sub-rhythm is an ideal-typical sequence of social events, the chances of having a group of identical sequences are infinitesimal. Consequently, substitution costs should be low when two events belong to the same sub-rhythm and high when they do not.

Furthermore, substitution costs should depend on time, *i.e.* on the location of events in the sequences compared. Fixed substitution costs mean that the differences between sub-rhythms are constant and expressed once and for all by oppositions between certain events. Unless this fixity is pursued, a variable and time dependent definition of the closeness of sub-rhythm seems preferable. Time-dependent substitution costs mean a considerable increase in the number of parameters to be determined.

Having defined the general properties substitution costs should fulfill, their exact formulation remains to be specified. Since biology went quite far in the explicitation of the probabilistic foundations of OM, it can be useful to have a look at them at this point of the discussion. We use here the general probabilistic model proposed by Durbin *et al.* (1998). If we consider two events *a* and *b* occurring in two sequences at the same time *t*, then the substitution cost function at that time should be of the form[17]

[16] In that case it might be preferable to suppress the event or to insert another one: the exact outcome will depend on the relative cost structure.

[17] See Durbin *et al.*, 1998, pp. 14-15 for more mathematical details. We have just added a temporal reference to the probabilistic framework they proposed.

$$s_t(a,b) = \frac{p_{ab,t}}{q_{a,t}q_{b,t}}$$

Where $p_{ab,t}$ is the probability of observing jointly the events *a* and *b* at *t* and $q_{a,t}$ is the probability of observing *a* at *t*. In other word, substitution costs should reflect the likelihood, or proximity, of two events occurring at the same time. In biology, the joint probability of *a* and *b*, $p_{ab,t}$, is interpreted as the evolutionary plausibility of their relationships, and is accordingly estimated using a sample of hypothetically and/or confirmed alignments. This probability is divided by the product of the individual probabilities of occurrence of the two events separately in order to take into account the possibility of observing the same pair by accident even though they are in fact independent. The ratio used as the substitution cost is therefore the net probability of the hypothetical evolutionary association between *a* and *b*.

Substitution costs based on transition matrices

A complete and detailed theory is not necessary: otherwise an *ad hoc* full-blown mathematical model could be developed, in biology as well as in social sciences. What is needed is a principle to generate those values, a generative principle consistent with social theory. Biology uses chemical properties as well as evolutionary theory. Social sciences need a social theory of time to interpret and determine substitution costs.

The fact that time is social is almost a truism (Abbott 1999). Émile Durkheim was the first social scientist to throw light on the links between time and society. In the book he wrote on religion (1912), Durkheim demonstrated how the calendar of undifferentiated societies was structured by collective life and religion: the crucial days of the calendar of the Aborigines were also celebrations, i.e. intensive collective moments. On the contrary, profane days were undifferentiated and solitary moments. Calendars reveal the rhythm(s) of collective life but at the same time help individuals to anticipate, plan and orient themselves daily in society. This double dimension of time has been condensed by Durkheim in the formula[18]: “The calendar expresses the rhythm of collective activities, while at the same time its function is to assure their regularities” (Durkheim, 1912).

As a consequence, “quantitatively equal periods of time are rendered socially unequal and unequal periods are socially equalized” (Sorokin and Merton, 1937). In other words, time is not purely quantitative because it is socially differentiated: the different social symbols used to represent time (calendars and clocks) should not be confused with time itself (Elias, 1992). The main channel of this socially differentiation is collective rhythms: it is what the entire society do that differentiate the continuous flow of events[19].

The statistical translation of “collective rhythms” is “transition matrices”. Indeed, a transition matrix describes trajectories between all the different states between two dates. A transition matrix is a synthetic representation of individual sequences at a certain moment. Transition matrices are the macro representations of micro phenomenon: distances between states are social but trajectories are individual. The strength of the flows between states, measured by transitions, is an indication of the different sub-rhythms that punctuate social life: a low transition rate between two states mean that these two states are at that particular moment not communicating hence that they are socially distinct sub-

[18] The translation has been taken from the first english translation of the book: Émile Durkheim, *Elementary forms of religious life*, New York and London, 1926.

[19] Although they constitute an interesting contribution to the growing academic debate about sequence analysis, the pure axiomatic approach proposed by Elzinga (2003) and Dijkstra and Taris (1995) is of little relevance for social scientists. Indeed, one of their premises, namely that the goal is “to find a representation of the sequences and their similarities that is free of sociological or historical theory – one that just relies on the basic properties of a sequence” (Elzinga, 2003: 7) clearly reveals the disconnection of this kind of purely theoretical solution with what sequence analysis is, in biology as well as in social sciences.

rhythms; on the contrary, when there are many transition between two states, it means that a change in a social sub-rhythm has been spotted and that these two states belong to it.

Consequently, substitution costs should be inversely proportional to transition rates. This sociological interpretation helps to understand why the quite common empirical practice of setting substitution costs using information about transitions yields good results and is "wise" according to Abbott (2000: 4). This strategy has indeed already been used successfully (see for instance Abbott and Forrest, 1986) and is one of the cost strategies proposed by TDA, one of the few statistical packages with OM capabilities available to date. However, substitution costs were temporally fixed, *i.e.* they were derived from a global transition matrix between all states built by merging the different episodes to build a global transition matrix disregarding the intra-sequences variability. This Markovian approach that only takes into account transitions and not their dates is very different from what we suggest here.

When the sequences have all the same length and that they share the same calendar, for instance in a career analysis all the individuals belong to the same cohort, we propose to estimate the $p_{ab,t}$ by the series of conditional probabilities describing the transitions between the *states* *a* and *b* considered between the dates *t*-1 and *t*, and *t* and *t*+1: $p(X_t = b|X_{t-1} = a)$[20], $p(X_{t+1} = b|X_t = a)$, $p(X_t = a|X_{t-1} = b)$, $p(X_{t+1} = a|X_t = b)$, where X_t is a random variable describing the occurrence (event) of the t^{th} episode of a sequence. In other words, we propose to substitute a diachronic for a synchronic distance. From a probabilistic point of view the higher the probability of transition between the two *states* before and after *t*, the closer the two *events*. One possible way to do this is simply to define the substitution cost function as[21]:

$$s_t(a,b) = \begin{cases} 4 - \left[p(X_t = a|X_{t-1} = b) + p(X_t = b|X_{t-1} = a) + p(X_{t+1} = a|X_t = b) + p(X_{t+1} = b|X_t = a)\right] & \text{if } a \neq b \\ 0 & \text{otherwise} \end{cases}$$

The higher the transitions between the states *a* and *b* and between *t*-1 and *t*, and between *t* and *t*+1 (with an upper bound of 4), the lower the substitution cost between the two events *a* and *b* at *t* (with a lower bound of 0). Indeed, high transitions mean that a lot of changes between these two states have just occurred and/or are about to occur, in other words that these states are statistically close. On the contrary, low transitions mean that these two states are from a probabilistic viewpoint very dissimilar. Thus, substitution costs depend on time and are derived from the transitions observed in the sample studied. It is possible to use only substitution operations with such costs when sequences have equal length. In that case, there is no more 'optimality' in the sense that the path followed to match pairs of sequences is simply the diagonal: it is an extension of the Hamming distance with substitution costs derived from the series of transition matrices describing the sequences.

Contrary to biology, it is not possible to constitute a sample of sequences to estimate these probabilities: the interdependence relationships between sequences of social events are not fixed and the goal of the analysis is not to identify plausible mutations but simply to describe these relationships for the sample analyzed. Of course, the generalizability of such a parameterized OM depends on the representativeness of samples analyzed: with representative samples, results can be generalized to the

---

[20] It is formally the probability of reaching the state *b* at time *t* conditionally to being in the state *a* at time *t*-1.

[21] The above formula is valid on the interval $]1,T[$, where*T* is the length of the sequences. The bounding formula are in this case simply:

If $t = 1$, then: $$s_1(a,b) = \begin{cases} 4 - 2\left[p(X_2 = a|X_1 = b) + p(X_2 = b|X_1 = a)\right] & \text{if } a \neq b \\ 0 & \text{otherwise} \end{cases}$$

If $t = T$, then: $$s_T(a,b) = \begin{cases} 4 - 2\left[p(X_T = a|X_{T-1} = b) + p(X_T = b|X_{T-1} = a)\right] & \text{if } a \neq b \\ 0 & \text{otherwise} \end{cases}$$

entire sampled population for they depend on probability estimates. In this regard, weights can be used to estimate the transitions matrices so that the survey design can be to a certain extent integrated in OM[22].

As in biology, the exact nature of the relations between sequences does not need to be known but is uncovered during the calculation of substitution costs. The fact that substitution costs are derived from transitions between states and used to compare events could appear in this regard as a kind of circularity. In fact, there is indeed some circularity here but this is not a problem since description is the only goal of the analysis: the output of OM, a distance matrix between sequences, is indeed just a new way of presenting the underlying series of transition matrices. However, whereas a series of transition matrices represent just *macro* relationships without connection with one another, the OM presentation proposed here is *individual* and is a synthetic measure of those relationships. This sequence comparison method indeed is basically turning transition matrices into inter-individual differences. The price of this individualization of collective rhythms is that the sequential differences between individuals are collapsed into a highly synthetic figure, the dissimilarity measure.

This is the reason why additional methods are required if some of this information is to be partially recovered. In this respect, cluster analysis seems more adapted than multidimensional scaling. Indeed, the goal of this second stage of the analysis is to reveal the underlying temporal regularities that have generated the distance matrix: the goal of cluster analysis is precisely to reveal the different groups hiding behind distances. Consequently, homogeneous groups identified with cluster analysis are also the temporal patterns social scientists are looking for. Classically, one of the issues of using cluster analysis to identify groups, what is called here temporal patterns, is the homogeneity of clusters: homogeneity measures should always accompany the labels used to describe classes. But the main issue is certainly which clustering method should be used?

Cluster analysis is not a particularly well known statistical discipline and though its principles are ancient, is not well implemented in standard statistical packages. Of course they all contain the historical methods such as single, complete, average or Ward algorithm. The Ward method is often considered as the best method available certainly because of its proximity with mainstream statistics: the Ward clustering method is indeed based on variance maximization/minimization. However, this method is far from being the best clustering algorithm. Although the Ward criterion perform well with well structured data sets it tends to join clusters with a small number of observations, is strongly biased toward producing equal size clusters, and is also very sensitive to noise and outliers (Milligan 1980 and 1981). The Ward agglomeration strategy is adapted to Euclidean distances (L_2) and when the clusters to be recovered have been generated from multivariate normal mixture, have equal spherical covariance matrices and sampling probabilities. These assumptions are very strong in the case of OM, and in particular with the method proposed, given that the Hamming distance is closely related to the Manhattan distance (L_1).

The flexible beta method, also known as flexible WPGMA (Weighted Pair Group using arithMetic Averages), proposed by Lance and Williams (1967) is much better to use with empirical data (Milligan, 1989): when noise and outliers are present, flexible WPGMA outperforms all the other algorithms, including Ward's. Flexible UPGMA (Unweighted Pair Group using arithMetic Averages), proposed by Belbin, Faith and Milligan (1992) is even better. Flexible WPGMA is available in SAS at least since the version 6 but not flexible UPGMA[23]. Stata 9 and SPSS 14 do not feature either, whereas Clustan Graphics, a statistical package specialized in cluster analysis, is in this respect no better than SAS. The

[22] Weights should only be used to calculate transition matrices, and consequently substitution costs: instead of counting the number of transitions, it is simply the weighted number of transitions which should be taken into account. The matching procedure in itself, *i.e.* the comparison of pair of sequences does not require any weights: it is by defintion a one to one procedure. However, weights should be used to interpret results, for instance, if cluster analysis is used, the size of the clusters obtained must be weighted.

[23] Flexible WPGMA is called "Flexible-Beta Method" in SAS and in ClustanGraphics.

SAS implementation of flexible WPGMA will be used in the second part of this paper in the absence of more efficient algorithms available in standard statistical packages[24].

Sequences with different length and/or disconnected calendars

The question of the length and of the calendars of sequences is a major scientific question. What is at stake here is the scientific legitimacy of comparing sequences with unequal length and/or completely unrelated calendars. Let us consider a hypothetical situation where retrospectively collected life courses are submitted to OM. Since the sample is not a cohort, the age of respondents, and, as a result, sequences' length, vary greatly[25]

Convincing sociological arguments are required to justify such a comparison. If transition to adulthood is of interest, then it seems bold to compare trajectories so varied in their completeness. In other words it seems crucial to work on sequences with roughly, if not exactly, the same length. If a generational sample, in other words if sequences have the same calendar, is chosen, then it would be even possible to see how those transitions relate to socio-historical changes (unemployment rate, female labor participation rate, economic growth, higher education prevalence, etc.). If the sample is constituted on a retrospective basis, then different cohorts can be compared and trans-generational similarities and dissimilarities appear.

In both cases, sequences have the same calendar. Indeed, the definition of a period of observation and the coding of events create in that case a common calendar that is precisely the subject of the analysis: the transition to adulthood calendar. It is a calendar in its own right because previous work emphasized how socially regulated is the timing of the entry into adulthood. It is however different from the calendar we use daily life as it does not exist in a symbolic form. In other words it is a kind of hidden social calendar that exists objectively but less subjectively (in comparison with the clock for instance) that sequence analysis can uncover.

However, the cohort sample presents another advantage: the transition to adulthood calendar is in that case also synchronized with what is happening in the rest of society. Transition to adulthood is a process involving three major social fields: school, economics, and family. When a cohort sample is considered then it becomes possible to establish a clear link between the process studied and the characteristics of these social fields. To see how the changes occurring within these fields interrelate with transition to adulthood it would be necessary to mix together a finite number of cohort samples of sufficient size. With a sample mixing too many different cohorts, the relations between the calendar and social structure is blurred and only strong structural regularities can appear.

This example helps to clarify further the use of sequence analysis in social sciences. The goal pursued is to throw light on temporal patterns: in other words to identify social calendars of some sort, in all their complexity and their variations. As a consequence the structure of the sample must be in accordance with this goal. Events should also be coded so as to facilitate the uncovering of the kind of temporal patterns researched. Ultimately, the interpretation of results should take into account these two crucial parameters.

When all sequences have the same length, and that the sample and the coding are defined so as to uncover a certain calendar then it is possible to use only substitution operations with costs derived from transitions. Temporal distortions of the processes are avoided since indel operations are not used. This

[24] We discovered since then that the statistical language R features both methods.

[25] Quite paradoxically, the example first proposed by Dijkstra and Taris (1995) and reused by Elzinga (2003) is finally quite close to such a situation despite the fact that they used a survey where a cohort is followed longitudinally but decided to represent only transitions between different states so that the length of the sequences is the number of transitions (see Dijkstra and Taris, 1995: 223), and are not identical and proportional to the number of waves of the survey. The authors acknowledge this high variation in the sequence length and they even draw the attention of readers on this aspect of their data, considering it as a particularly challenging test of the methods they propose (Elzinga, 2003: 17).

method is no longer based on optimality principles, precisely because it is the research of logic optimality that causes temporal warping. Events coding identically but occurring at different moment are considered not as identical events that are shifted but as different events because they are shifted. This solution satisfies the principle of social structuration of time, the kind of questions asked by social scientists and the nature of the data at their disposal

When sequences have unequal lengths and the period of time considered is not too wide in comparison with the unit of analysis, some indexicality subsists but the different sequences are not perfectly synchronous. As a consequence, it is not possible to use time-varying substitution costs derived from transitions matrices as it was proposed above and thus the only solution is to calculate a single transition matrix which will retain some of the social structuration of the underlying time scale. Indel operations are particularly useful here – and absolutely necessary when sequences do not have all the same length – as they can help to re-synchronize the different sequences. But they can also increase their desynchronization.

The question of the costs of indel operations is in this case quite difficult to solve. If indel operations are used together with transition-based substitution costs it seems wise to set indel costs to the middle of the distribution of substitution costs so that a time shift is privileged to a substitution when transitions are low. Another solution would be to use once again information from the sample on the relative weight of the states where an event is to be inserted or deleted. It seems that this question is as problematic in social sciences as it is in biology where most of the time indel costs are chosen on a complete empirical basis. Even more, perhaps, given that there are some theoretical justification to the insertion and deletion of residues in biology whereas the necessity of warping time seems less assured.

An application to the daily scheduling of paid work

Contrary to the order required by communication, it is through the question of the scheduling of paid work within the day that the theoretical considerations that have been proposed first were in fact elaborated. Work time is indeed difficult to summarize and is usually reduced either to durations (the number of hours worked) or to indicators (*e.g.* night work). In order to distinguish night work from work schedules shifted in the afternoon/evening or in the morning, precise criteria are required. These criteria are based on *a priori* knowledge but also on arbitrariness. As a result, the scheduling of work is most of the time reduced to the dichotomy day vs. night work.

The lack of adequate tools to describe the scheduling of work time is becoming critical with the rise of dual-earner couples in most developed countries. Indeed, if individual work schedules cannot be described, so are the “family work days” and the problem of desynchronization some spouses face (Nock and Kingston, 1984). The consequences for daily life, and in particular for childcare, of a major social change remain unknown because tools to describe sequences of daily events are missing.

The number of hours worked as well as their scheduling are crucial economic parameters for firms in societies with economic organizations based on the division of labor (Moore, 1967). It has been demonstrated that the number of work hours are related to social position, this relation evolving with economic changes (Gershuny, 2000). Work time is socially structured and its rhythms can be legitimately studied and uncovered with the modified Hamming method proposed in this paper.

Data and coding

Information on work time can be collected using various methodologies, but it has been proven that the time diary approach produces far better estimates than any other method (Robinson, 1985). Indeed, contrary to “stylized questions” on time asking directly to respondents to give average estimate of the time they are spending doing some pre-defined activities, information on time is collected in time diary

surveys through respondents' description with their own words of the sequence of activities they did a specific day. Unfortunately, this sequential information on daily life is usually reduced to aggregate durations (time budgets) despite the wealth of sociological information they contain, in particular on the sequencing of daily life (Gershuny and Sullivan, 1998).

The two last French time use surveys (1985-86 and 1998-99) used here were done in person by the French Institute of Statistics (INSEE) over a year[26] and had high response rates (64.7% and 80%). The modified Hamming distance has been applied on the two samples merged so that the evolution between 1986 and 1998 can be easily taken into account. Diaries of both surveys cover 24 hours (1 am to midnight), with minor differences in precision[27], and as a result all sequences have the same length and are perfectly synchronized[28].

We have implemented ourselves the modified Hamming distance we proposed in SAS as a macro and in Stata as a plug-in. Both are available from the author.

Taxonomy of work days

There is no absolute and rigid rule to decide how many clusters are necessary to give a synthetic but faithful representation of the data analyzed. However, considering the flexible WPGMA height for the last steps in the grouping process can give some guiding elements as a jump reveals that two dissimilar clusters have just been joined. Figure 1 suggests that an eight-class scheme is the most acceptable synthetic representation of the structure of the data. Other jumps are occurring when the number of classes is reduced from eleven to ten, and from fifteen to fourteen. The right number of classes is therefore between thirteen and eleven. We adopted a twelve-class classification after a close inspection of the shape and relevance of clusters for various numbers of classes between fifteen and eight.

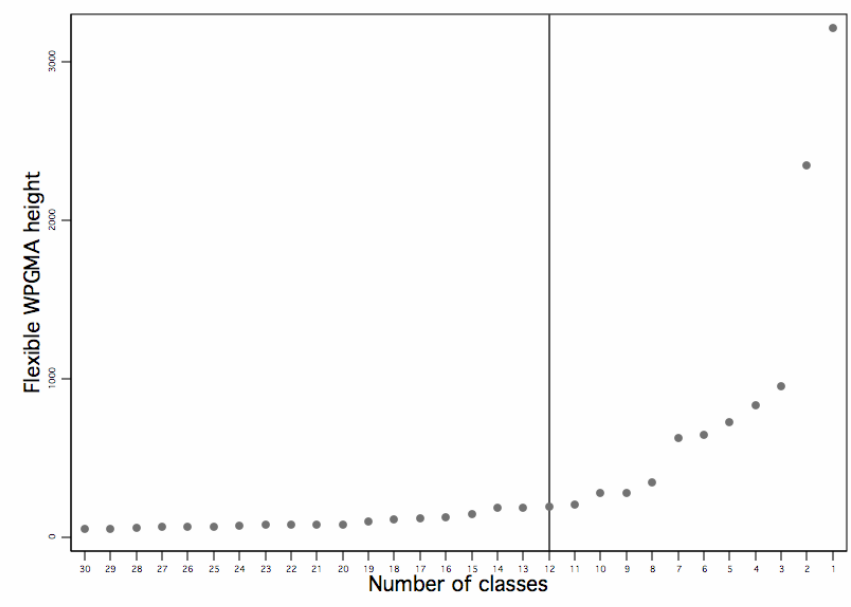

Figure 3 – Number of classes and flexible WPGMA height

[26] With the exception of summer and Christmas holidays. A year is a small observation window with respect to the pace of changes in the use of time (on changes in the use of time since the 1960s, see Gershuny, 2000).

[27] The 1985-86 and 1998-99 surveys have respectively 5- and 10-minute time slots: comparability can be an issue but unpublished methodological studies (Alain Chenu, personal communication) suggest that problems are likely to be minor and limited to very specific sequences of activities (clearing the table vanishes in having meal for instance). Work time should not be too affected by this methodological difference.

[28] They are synchronized with regard to the calendar unit "day".

In this particular example, work schedules can be described roughly by two simple indicators:

- the number of work hours;
- The time of the day corresponding to the middle of workday (mid-workday), which gives a very rudimentary indication on the scheduling of work within the day.

With the help of Table 5 and of visual representations of clusters that will be presented later, clusters can be easily labeled and interpreted. The first three clusters consist of the 9 to 5 work day and of two variants, one slightly shifted to the left in the morning, the other slightly shifted to the right but also markedly longer. Another group of clusters consists of shifted schedules: in the morning, in the afternoon, in the evening and in the night. As a result, we see that night work, the only shifted work schedule usually taken into account, is only the tip of the iceberg "shifted work schedules". Work schedules located at the margin of the 9 to 5 work day have increased in France: a similar result, though not based on a classification but on visual estimates, has also been found for the US (Hamermesh, 2002).

		1985-86			1998-99		
	Type of work day	**Size (%)**	**Mid-work day**	**Duration**	**Size (%)**	**Mid-work day**	**Duration**
	Standard	**56,45**	**12:59**	**8:26**	**54,71**	**13:06**	**8:43**
1	8 to 4	7,60	12:00	8:14	6,79	11:53	8:22
2	9 to 5	38,17	12:53	8:17	33,88	12:57	8:23
3	10 to 7	10,69	14:01	9:09	14,03	14:03	9:39
	Shifted	**14,41**		**7:16**	**16,55**		**7:16**
4	In the morning	5,26	9:44	7:39	6,07	9:45	7:44
5	In the afternoon	5,40	15:32	6:46	6,43	15:24	6:43
6	In the evening	2,08	17:02	7:20	2,49	17:20	7:04
7	In the night	1,66		7:38	1,57		7:56
	Long work day	**9,12**	**13:57**	**10:29**	**11,60**	**14:06**	**11:02**
8	Long 9 to 5	3,53	12:54	10:47	4,08	12:53	11:08
9	10 to 7 spreading in the evening	5,59	14:38	10:18	7,52	14:46	10:58
	Other	**20,02**	**12:50**	**3:45**	**17,14**	**13:11**	**4:13**
10	Fragmented part-time	3,23	13:21	3:50	2,38	13:28	5:33
11	Fragmented full time	3,46	12:15	8:06	4,22	12:11	7:20
12	Very short work day	13,32	12:52	2:14	10,54	13:31	2:41
	Total	100,00		7:32	100,00		7:58

Table 5 – Basic characteristics of the classification (averages in hours:minutes per day).

Longer work days come in two flavors: either in a long version of the standard work day, *i.e.* beginning earlier and ending later than the 9 to 5, or in a long version of the 10 to 7, *i.e.* ending later than 7 pm. Other patterns of work days are less clear and are generally made of short and/or fragmented work days. By fragmented we mean that work schedules have at least two distinct work periods separated by considerable time: the best example is supermarket cashiers (Prunier-Poulmaire, 2000) who are asked to work only during peak shopping periods, *i.e.* during the 9 to 5 workers' lunch break and after the 9 to 5 work day. Fragmented part-time work days are concentrated mostly around the lunch break, *i.e.* at the end of the morning and the beginning of the afternoon. Fragmented full-time work days are fragmented work day *par excellence*: although their duration is on average of eight hours, they are made of two distinct work periods separated by several hours. In this case, mid-work day is a very poor indicator of the scheduling of work. Eventually, in the last cluster are gathered very short work days: since all days with at least a 10-minute work spell have been considered as work days, this last cluster collects in fact the very short work days without our defining *a priori* and unavoidably arbitrarily a minimum work time.

Quality

One possibility to assess the quality of clusters is to display the distribution of work durations and mid-workdays, for instance in box-plots. This solution is however not too satisfying as it relies in the first place on the relevance of the indicators used. As we have already seen, mid-workday is a very rough indicator of the scheduling of work and can be tricky. The problem of using variance and other standard statistical analysis tools to assess the quality of the clusters takes us back to the problem of defining synthetic and faithful indicators of sequences: if it was possible to design relevant indicators, there would be no need to do sequence analysis.

A natural visual representation of clusters made of similar sequences is to plot for each episode the proportion of sequences in the cluster that are in the different states. An alternative is to stack all individual sequences horizontally. The former is an aggregate tempogram and the latter is an individual tempogram. Both kinds of tempogram help to interpret but also to assess visually the quality of sequence classifications. The gradient and the height of the curve of an aggregate tempogram indicate how homogeneous clusters are: the steepest and the higher, the more homogenous clusters are. If individual sequences are represented in individual tempograms by colored sub-segments then it is possible to assess the quality of clusters by the homogeneity of the different patches of color.

With the exception of the two last clusters which clearly lack homogeneity, the overall quality of the taxonomy measured through duration box plots (see Figure 4) and also to a lesser extent through mid-workday box plots (see Figure 5) is satisfactory: overall, boxes are small and distinct from one another. As expected, the two last clusters are the less homogeneous in terms of mid-workdays: this indicator is particularly inappropriate to describe fragmented work hours. Clusters appear also remarkably homogeneous depicted by aggregate tempograms (see Figure 6).

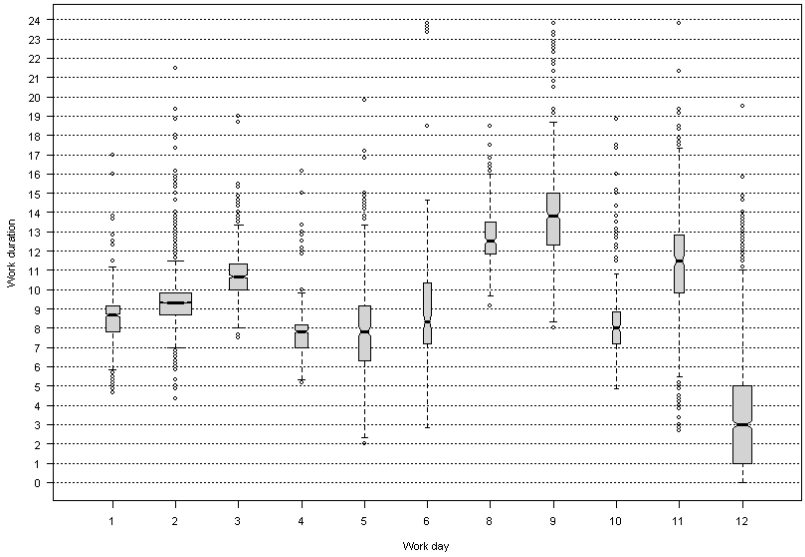

Figure 4 – Boxplot of clusters' work durations (boxes' width are proportional to the size of clusters).

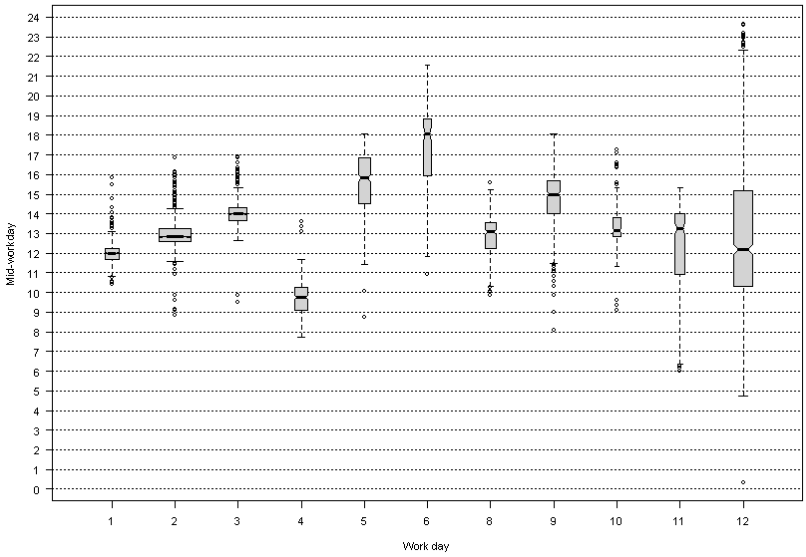

Figure 5 – Boxplot of clusters' mid-workdays (boxes' width are proportional to the size of clusters).

Individual tempograms (see Figure 7) confirm these impressions and measures: most clusters contain very similar sequences. The homogeneity of the first four clusters is quite impressive and corresponds in fact to traditional work schedules[29]: the three variants of 9 to 5 (standard workers) as well as the work schedules shifted in the morning and also in the night (shift workers) correspond to the industrial organization of work (Fordist). Indeed, fragmented work schedules are here to satisfy the new temporal requirements of the service industry (shop and services opening hours) and are by definition less socially structured. In other words, the lack of homogeneity found in some clusters is to a large extent not due to a defect in the method proposed but on the contrary to crucial social phenomena: work schedules' variability is increasing.

[29] For more details, see Lesnard (2006).

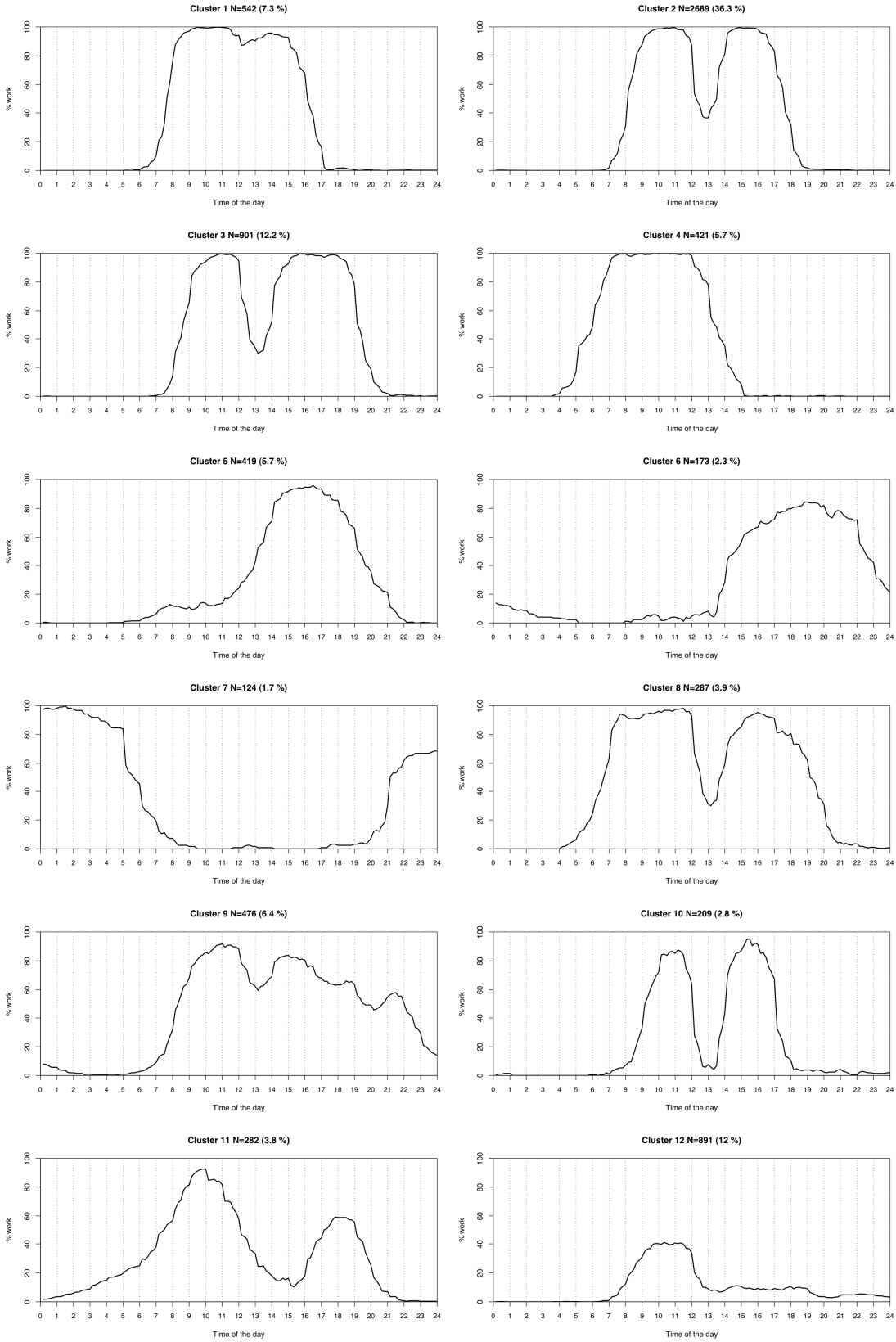

Figure 6 – Aggregate tempogram of the classification of work days

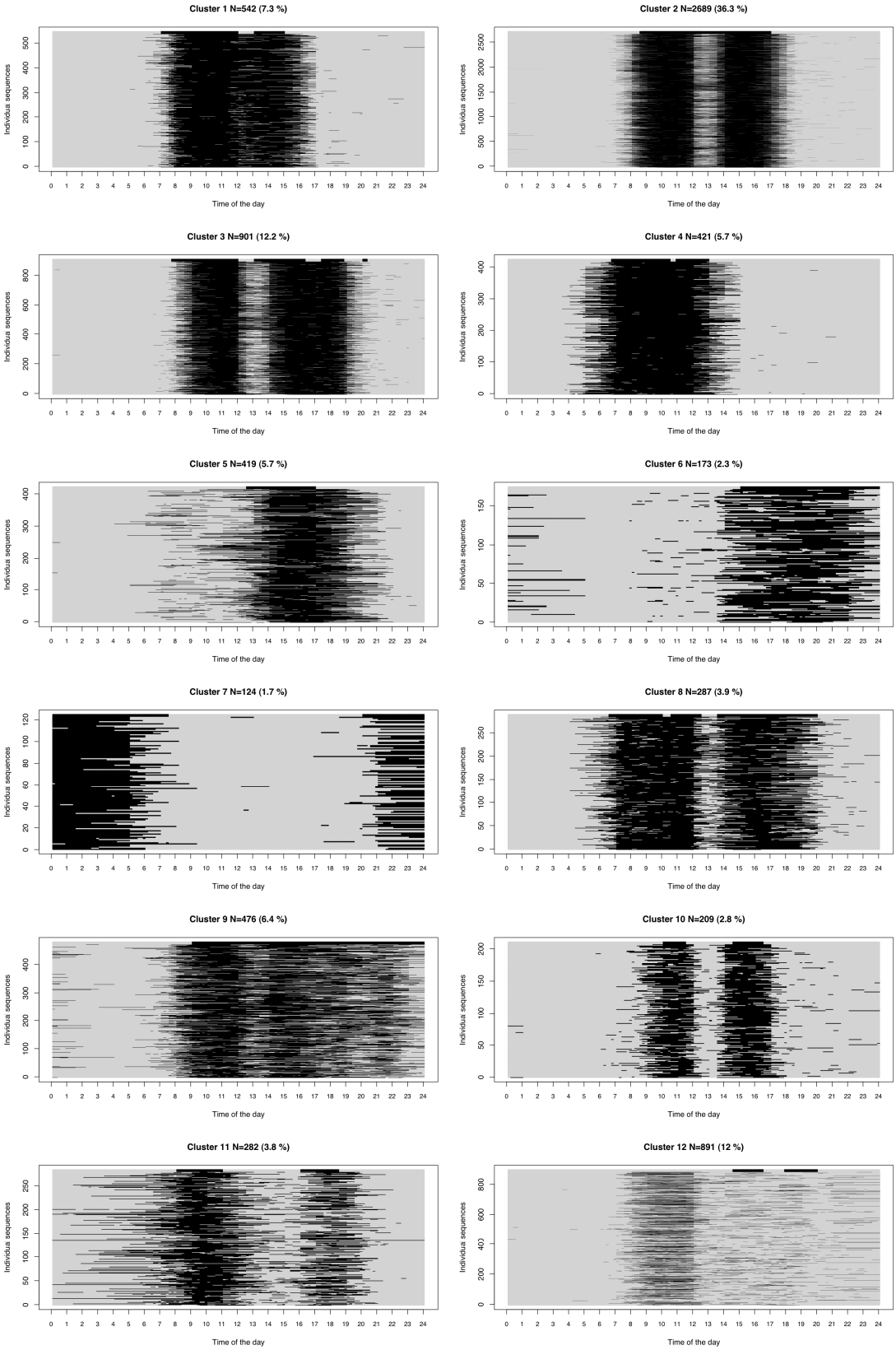

Figure 7 – Individual tempogram: individual sequences are represented horizontally. Black indicates work spells and light gray non-work spells.

Despite only substitution operations are used and because OM is only the first stage of the analysis and is supplemented by cluster analysis, these differences in timing appear in the results. Indeed, as

collective rhythm is the basis of the measure of similarity between sequences atypical rhythms are easily found because they are opposed to major sub-rhythms: temporal shifts are crucial components of sequences and disregarding them result in a loss of fundamental information on the dynamic aspect of the processes studied.

Conclusion

As in biology, the three elementary operations used in OM are of little theoretical relevance in social sciences: it is costs, their interpretation and determination, which are central. Indeed, the success of OM in the biological field does not rely on any resemblance between insertion, deletion, and substitution operations, and bio-chemical processes: substitutions are interpreted as plausible evolutionary changes only when substitution costs are low. In other words, costs help biologists to distinguish between evolutionary changes and random mutations: "pattern search algorithms in general do not assume anything about the way the data are generated. (They rather make assumptions about the kinds of patterns we expect to see.)" (Abbott, 2000: 3).

Evolutionary changes in biology, social rhythms in social sciences: the aim of the analysis is not to detect plausible evolutionary changes but, as sequences are not made of biological matter but of events and time, to cast light on social rhythms, on the social structuration of the timing of events. Indeed, time is socially structured: the continuous flow of events is differentiated by collective rhythms, by what a part or the whole society is doing. Calendars used nowadays are objectified social symbols of former social rhythms, marked by religion and strong temporal symmetry (Durkheim, 1912; Zerubavel, 1985). However, these calendars have lost their original connection with social rhythms with the transformation of collective rhythms following social differentiation: modern time is plural and not as institutionalized as the collective rhythms fossilized in calendars. As a result the time of contemporary societies is harder to analyze and uncovering this plural structuration, these social rhythms, is ultimately what is at stake in the analysis of sequences of social events, whether OM or event history models are used.

If OM should be used in social sciences in order to uncover social calendars, then costs should be set in order to distinguish sequences belonging to identical or different collective rhythms. Since the nature of the elements of sequences are not given by nature but decided, social scientists have more freedom than biologists. Indeed, prior to sequence analysis *per se*, states have to be defined and this step is as crucial as parameterizing correctly OM for collective rhythms are measured within the bounds laid out by the different states chosen. If no difference is made between two states playing a fundamental part in the differentiation of time then it will be hard to get something out of OM, whatever costs are chosen.

Another parameter must be taken into account if social rhythms are to be uncovered: the effects indel and substitution operations have on sequences in social sciences. Indel operations warp time in order to match identically coded states but occurring at different moments in their respective sequences. Substitutions do the opposite: substituting two events is warping them in order to conserve their co-occurrence. One major consequence of the social structuration of time is that the timing of events is not random but on the contrary reflects the social rhythms analyzed. It is not because two events are coded identically that they are socially equivalent: a one-hour work spell in the middle of the afternoon vs. one at the beginning of the night are clearly different. But this difference is only partly due to the absolute number of hours that separate them: this pure numerical difference is indeed altered by collective rhythms: the social difference between one hour of work from 4 pm to 5 pm and another from 7 pm to 8 pm is larger than the absolute number of hours. Therefore, using indel operations amounts to voluntarily adding noise to the phenomenon under study and should be used with extreme caution.

In biology, costs are not coming from theoretical models (otherwise these model could be used directly) but are derived from a sample of confirmed and/or hypothetical alignments. How theory is used is particularly interesting here: relationships within the set of alignments used as a yardstick are not perfectly known (otherwise, once again, OM would not be necessary) but are synthesized into costs to be used to analyze other samples. The substitution costs proposed here partakes of the same logic: a complete social and historical model (Elzinga, 2003) is not necessary if principles to derive substitution costs capable of distinguishing social flows can be established. As collective rhythms are behind the social differentiation of time, they should be central in the definition of substitution costs. The series of transition matrices associated with a sample provides precisely an aggregate description of collective flows between the states defined in the analysis. With substitution costs inversely proportional to empirical transition probabilities low transition flows mean high substitution costs: when two states are disconnected in terms of transition probabilities, they will be considered as belonging to two distinct social rhythms. On the contrary, high transition probabilities between two states may reveal changes in a single rhythm.

It is argued here that substitution operations, with costs inversely proportional to transition probabilities, should be used alone whenever it is possible, *i.e.* when sequences are contemporaneous and of equal length. When the sequences under study are not contemporaneous, it is not possible to only use substitution operations nor to set their costs as inversely proportional to transition probabilities. The object of the analysis is also changed: as time loses its indexicality, it is only average social calendars, trans-historical regularities, that can be uncovered. In such a case, the series of transition matrices lost most of their meaning and should not be used to set substitution costs. However the average Markovian transition matrix can be used as a description of some of the trans-historical regularities analyzed. When sequences' length varies, indel operations have to be used. Once again the goal of the analysis is also at the same time transformed and the legitimacy of the comparison itself is at stake.

Deriving substitution costs from transition matrices amounts to individually connect this aggregate information on collective rhythms: with such substitution costs, OM is basically a way to individualize and connect collective transition matrices. However, this connection is synthesized by single measures – dissimilarity coefficients – and information on the sequential nature of these different rhythms is also disappearing at the same time. Cluster analysis recovers most of this information and is therefore a crucial step of OM. As the underlying distance measure is unlikely to be Euclidean, the Ward algorithm should not be used and all the more so as new techniques such as flexible WPGMA and UPGMA have been proven far superior to recover information on the structure of data in presence of outliers and noise.

The method proposed in this paper has been applied to the timing of paid work. As all sequences have the same length (144 10-minute time slots), only substitution have been used with costs inversely proportional to transition probabilities. The dissimilarity matrix produced by this modified Hamming distance was then analyzed using flexible WPGMA. The quality and interpretability of the classification of workdays obtained suggest that OM is not only an exploratory tool but also a powerful method to identify social rhythms when parameters are chosen accordingly.

Since OM is new in social sciences, considerable work needs to be done in order to demonstrate the reliability and interest of this method. Results must be replicated and validated: in other words abundant critical use of OM is needed (Levine, 2000). However, this task is not facilitated by the computer power required by this method but also by the lack of programs proposing this method[30]. It

[30] Besides Optimize, a program supervised by Abbott but no longer maintained, and TDA, OM is not implemented in any statistical packages intended for social scientists. Numerous OM packages are available in biology but are most of the time almost impossible to use in social sciences because of the dramatic differences in the aim of the analysis and in the nature of sequences, as it should be clear to the reader now.

goes without saying that standard statistical packages do not feature OM: a new method is by definition difficult to use and therefore to validate.

Nonetheless, this is more than a catch-22. Statistical software do not feature all statistical methods equally: if regression and inferential statistics are well implemented, multidimensional descriptive methods are lagging far behind. Geometric data analysis and cluster analysis are two striking examples. Although correspondence analysis is theoretically well established and has long proven empirically its worth, its implementation in major statistical packages such as SPSS or Stata is indigent. Whereas it has never been easier to run a complex duration regression full of untestable and untested causal hypotheses, performing a basic multivariate correspondence analysis with supplementary variables is purely and simply impossible in the current versions of SPSS and Stata. Cluster analysis is treated somewhat better but it seems that statistical software companies believe that cluster analysis is a finished or frozen research project with no new algorithms or techniques since the 1970s: major improvements such as flexible WPGMA and UPGMA, proposed in the 1980s, are missing. It is also during the 1980s that OM was introduced in social sciences: OM has been around for more than twenty years now and is still ignored by standard statistical packages.

The indigent situation of the implementation of multivariate descriptive methods is obviously related to what Abbott calls the causal devolution (1998): the advent of a new generation of social scientists with strong quantitative skills and taste for new methods corresponds to the diffusion of personal computers and of the first statistical packages. The dominant academic positions they acquired oriented on a long-term basis the teaching of statistics in social sciences, but also indirectly what kind of statistical procedures are implemented in statistical packages. Whereas the latest regression models are widely available, social scientists who need to use cluster analysis either have to resign to using 20-year old methods (something unthinkable for econometricians) or to try to find if an obscure specialized package is available. The plurality of statistical methods as reflected in the statistics literature is far from being respected in statistical packages and the marginality of OM in social sciences is doubly affected by this phenomenon since it relies on other multivariate descriptive procedures. If standard multivariate descriptive procedures are still not well implemented, OM is unlikely to be featured in the next version of Stata, not to mention SPSS: this calls for a better explicitation of the sequences of treatment and procedures OM users are using, and for sharing programs when users designed their own computer solutions. The growing number of articles using this method evidences that the unprecedented insights on sequences offered by OM outweigh the huge difficulties to apply it.

References

Abbott, Andrew (1990). “Conceptions of time and events in social science methods”, *Historical methods*, 23, 140-150.

Abbott, Andrew (1995). “Sequence analysis: new methods for old ideas”, *Annual Review of Sociology*, 21, 93-113.

Abbott, Andrew (1998). “The causal devolution”, *Sociological Methods and Research*, 27, 148-181.

Abbott, Andrew (1999). “Temporality and process in social life”, in Kalleberg, Ragnvald and Engelstad, Frederik (eds), *Social time and social change*, Oslo, Scandinavian University Press, 1999.

Abbott, Andrew (2000). “Reply to Levine and Wu”, *Sociological Methods and Research*, 29, 65-76.

Abbott, Andrew and Forrest, John (1986). “Optimal matching methods for historical sequences”, *Journal of Interdisciplinary History*, 16, 471-494.

Abbott, Andrew and Hrycak, Alexandra (1990). “Measuring resemblance in sequence analysis: an optimal matching analysis of musicians”, *American Journal of Sociology*, 96, 144-185.

Abbott, Andrew and Tsay, Angela (2000). "Sequence analysis and optimal matching methods in sociology", *Sociological Methods and Research*, 29, 3-33.

Belbin, Lee and Faith, Dan and Milligan, Glenn W. (1992). "A Comparison of Two Approaches to Beta-Flexible Clustering", *Multivariate Behavioral Research*, 27, 417-433

Dayhoff, Margaret O., Schwartz, Robert M. and Orcutt, B. C. (1978). "A model of evolutionary change in proteins", in Dayhoff, M. O. (ed.) *Atlas of protein sequence and structure*, Washington, National Biomedical Research Foundation 5, 345-352.

Dijkstra, Wil and Taris, Toon (1995). "Measuring the agreement between sequences", *Sociological Methods and Research*, 24, 214-231.

Durbin, Richard, Eddy, Sean R., Krogh, Anders and Mitchison, Graeme (1998). *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*, Cambridge (UK), New York, Cambridge University Press.

Durkheim, Émile (1912). *Les formes élémentaires de la vie religieuse*, Paris, Alcan.

Elias, Norbert (1992). Time *: an essay*, Oxford, Basil Blackwel.

Elzinga, Cees H. (2003). "Sequence similarity: a nonaligning technique", *Sociological Methods and Research*, 32, 3-29.

Gershuny, Jonathan and Sullivan, Oriel (1998). "The sociological uses of time-use diary analysis", *European Sociological Review*, 14, 69-85.

Gershuny, Jonathan (2000). *Changing Times: Work and Leisure in Postindustrial Society*, Oxford: Oxford University Press.

Halpin, Brendan and Chan, Tak Wing (1998). "Class careers as sequences: an optimal matching analysis of work-life histories", *European Sociological Review*, 14, 111-130.

Hamermesh, Daniel S. (2002). "Timing, Togetherness, and Time Windfalls", *Journal of Population Economics*, 15, 601-623

Hamming, Richard W. (1950). "Error-detecting and error-correcting codes", *Bell System Technical Journal*, 29, 147-160.

Henikoff, Steven and Henikoff, Jorja G. (1992). "Amino acid substitution matrices from protein blocks", *Proceedings of the National Academy of Sciences of the USA*, 89, 10915-10919.

Kruskal, Joseph B. (1983). "An overview of sequence comparison", *in* Sankoff, David and Kruskal, Joseph B. (ed.) *Time warps, string edits, and macromolecules: the theory and practice of sequence comparison*, Reading, MA, Addison-Wesley, 1-44.

Kruskal, Joseph B. and Liberman, Mark (1983). "The symmetric time-warping problem: from continuous to discrete", *in* Sankoff, David and Kruskal, Joseph B. (ed.) *Time warps, string edits, and macromolecules: the theory and practice of sequence comparison*, Reading, MA, Addison-Wesley 125-161.

Lance, Godfrey N. and Williams, W. T. (1967). "A General Theory of Classification Sorting Strategies. 1. Hierarchical Systems", *Computer Journal*, 9, 373-380.

Le Roux, Brigitte and Rouanet, Henry (2004). *Geometric data analysis: from correspondence analysis to structured data analysis*, Dordrecht, Boston and London, Kluwer.

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, 63-88.

Lesnard, Laurent (2006). "Flexibilité des horaires de travail et inégalités sociales", *in* Insee, *Données Sociales 2006*.

Levenshtein, Vladimir I. (1966). "Binary codes capable of correcting deletions, insertions, and reversals", *Soviet Physics Doklady*, 10, 707-710. Originally published in Russian in *Doklady Akademii Nauk SSSR*, 163 (4): 845-848, 1965.

Levine, Joel H. (2000). "But what have you done for us lately? Commentary on Abbot and Tsay", *Sociological Methods and Research*, 29, 34-40.

Milligan, Glenn W. (1980). "An Examination of the Effect of Six Types of Error Perturbation on Fifteen Clustering Algorithms", *Psychometrika*, 45, 325-342.

Milligan, Glenn W. (1981). “A Monte Carlo Study of Thirty Internal Criterion Measures for Cluster Analysis”, *Psychometrika*, 46, 187-199

Moore, Wilbert E. (1963). *Man, time, and society*, New York and London, Wiley and sons.

Nock, Steven L. and Kingston, Paul W. (1984). “The Family Work Day”, *Journal of Marriage and the Family*, 46, 333-343.

Prunier-Poulmaire, Sophie (2000). “Flexibilité assistée par ordinateur. Les caissières d'hypermarché”, *Actes de la recherche en sciences sociales*, 134, 29-65.

Robinson, John P. (1985). “The validity and reliability of diaries versus alternative time use measures”, *in* Juster, F. Thomas and Stafford, Frank P. (ed.) *Time, Goods, and Well-Being*, Ann Arbor: University of Michigan Press.

Rouanet, Henri, Lebaron, Frédéric, Le Hay, Vivane, Ackermann, Werner, and Le Roux, Brigitte (2002). “Régression et analyse géométrique des données: réflexions et suggestions”, *Mathématique et Sciences Humaines*, 160, 13-45.

Saint Pol, Thibaut de (2005). “Le dîner des Français : étude séquentielle d’un emploi du temps”, *Document de travail du CREST*, 2005-19.

Sankoff, David and Kruskal, Joseph B. (ed.) (1983). *Time warps, string edits, and macromolecules: the theory and practice of sequence comparison*, Reading, MA, Addison-Wesley.

Simiand, François (1922). *Statistique et expérience. Remarques de méthode*, Paris, Marcel Rivière.

Sorokin, Pitrim A. and Merton, Robert K. (1937). “Social Time: A Methodological and Functional Analysis”, *American Journal of Sociology*, 42, 615-629

Thompson, Edward P. (1967). “Time, Work-Discipline, and Industrial Capitalism”, *Past and Present*, 38, 56-97.

Tuma, Nancy B., Hannan Michael T. and Groeneveld, Lyle P. (1979). “Dynamic analysis of event histories”, *American Journal of Sociology*, 84, 820-854.

Wu, Lawrence L. (2000). “Some comments on "Sequences analysis and optimal matching methods in sociology: review and prospects"”, *Sociological Research and Methods*, 29, 41-64.

Zerubavel, Eviatar (1985). *Hidden Rhythms: Schedules and Calendars in Social Life*, Los Angeles, University of California Press.