Cost Setting in Optimal Matching to Uncover Contemporaneous Socio-Temporal Patterns

Laurent Lesnard

SciencesPo. Paris
OSC / CNRS
Observatoire sociologique du changement

Cost Setting in Optimal Matching to Uncover Contemporaneous Socio-Temporal Patterns

Laurent Lesnard

Notes & Documents

n° 2009-03 Juin 2009

Résumé :

Cet article traite de la question des effets des coûts sur les types de régularités temporelles que les Méthodes d'Appariement Optimal (MAO) permettent de mettre au jour en sciences sociales. L'équilibre entre les coûts d'insertion et suppression (indel) et de substitution détermine le type de régularité temporelle. Alors que les insertions-suppressions privilégient les états codés identiquement à leur timing, les substitutions respectent le timing des événements au prix de leur simplification lorsqu'ils sont différents. Plus le ratio du coût de substitution sur le coût d'insertion-suppression est faible, plus les MAO sont portées vers la distance de Hamming où seules les substitutions sont utilisées. Plus il est élevé, plus les MAO s'approchent de la distance de Levenshtein II qui consiste à trouver la sous-séquence commune la plus longue. Quand le timing des séquences est de toute première importance, les opérations de substitution doivent être privilégiées aux insertions-suppressions et leurs coûts déterminés avec soin. Idéalement, les coûts de substitution devraient varier avec le temps de manière à mieux prendre en compte le timing des séquences étudiées. Comme les opérations d'insertion-suppression déforment le temps, donc le timing des séquences, il est suggéré de n'utiliser que des substitutions avec des coûts qui varient avec le temps inversement proportionnels aux fréquences de transitions toutes les fois que le timing des séquences est central pour l'analyse. Variante des MAO proche de la distance de Hamming, le Dynamic Hamming Matching est appliqué à la question des horaires de travail en France en 1985 et 1999 (N = 7 908) et comparé à trois variantes des MAO (Hamming et Levenshtein I et II). Conformément à ce que l'on pouvait attendre, les deux variantes de Hamming apparaissent meilleures, en termes d'entropie, pour identifier les types de journées de travail que les deux distances de Levenshtein.

Pour citer ce document :

Lesnard, Laurent (2009). « Cost Setting in Optimal Matching to Uncover Contemporaneous Socio-Temporal Patterns », Notes & Documents, 2009-03, Paris, OSC, Sciences Po/CNRS.

Pour une version électronique de ce document de travail et des autres numéros des Notes & Documents de l'OSC, voir le site web de l'OSC : http://osc.sciences-po.fr/publication/pub_n&d.htm

Abstract:

This article addresses the question of the effects of cost setting on the kind of temporal patterns Optimal Matching (OM) can uncover when applied to social science data. It is argued that the balance between indel (insertion and deletion) and substitution costs determines what kind of socio-temporal pattern can be brought to light. Insertion and deletion operations favor identically coded states irrespective of their locations whereas substitutions ones focus on contemporaneous similarities. The lower the ratio of substitution to indel costs, the closer OM is to the Hamming distance where only substitutions are used. The higher this ratio, the closer OM is to the Levenshtein II distance, which amounts to finding the longest common subsequence. When the timing of sequences is crucial, substitutions should be favored over indels and their costs should be carefully fixed. Ideally, substitution costs should vary with time to better take into account the timing of the sequences studied. As indels warp time, hence the timing of sequences, it is suggested to use only substitution operations with time-dependent costs inversely proportional to transition frequencies whenever the timing of sequences is central. This OM variant, coined Dynamic Hamming Matching, is applied to the question of the scheduling of paid work where timing is critical (1985 and 1999 French time-use surveys, N = 7908) along with three classical OM variants (Hamming and Levenshtein I and II). As expected, the two Hamming dissimilarity measures fare better to identify patterns of workday schedules, as measured by entropy, than the two Levenshtein ones.

Readers wishing to cite this document are asked to use the following form of words:

Lesnard, Laurent (2009). "Cost Setting in Optimal Matching to Uncover Contemporaneous Socio-Temporal Patterns", Notes & Documents, 2009-03, Paris, OSC, Sciences Po/CNRS.

For an on-line version of this working paper and others in the series, please visit the OSC website at: http://osc.sciences-po.fr/publication/pub_n&d.htm

1. Introduction

Dynamic statistical models appeared in the social sciences at the dawn of the 1980s. In the first review dedicated to these models, Nancy Tuma, Michael Hannan and Lyle Groeneveld (1979) enjoined social scientists to incorporate these new tools made available by the development of personal computers. In view of the widespread use and growing sophistication of dynamic regressions and other duration models this "dynamic model turning point" can be considered as successful. Even though statistical models are not always used in a true hypothesis testing perspective but also very often as "descriptive tools" (Abbott 1998) their greater explanatory power rely on additional assumptions which make them also more fragile. It is long known that, in order to be faithful to facts, simplification should be progressive (Simiand 1922, p. 48). However, until recently, applying this precept on sequence data proved challenging, as it required expertise in emerging methods only available in exotic statistical packages or programs. This was all the more unfortunate as dynamic models often rely on strong assumptions on causality and on the order of observed events (Bocquier 2006) and as a consequence, describing sequences before any causal analysis is attempted is essential (Abbott 1990).

So far, two kinds of statistical descriptive methods have been used to describe sequence data. The first one is related to the geometric data analysis (GDA) paradigm. GDA is particularly prominent in France where there is a long tradition, if not a "French school", of building empirical typologies of sequences using these techniques (Deville et Saporta 1980 ; Deville 1982 ; Degenne, Lebeaux et Mounier 1996). However, as the crux of these methods is multiple correspondence analysis (for a comprehensive presentation of MCA, see Le Roux et Rouanet 2004) they do not take advantage of the extra information contained in the ordering of events. Optimal Matching (called OM in the rest of this paper), introduced into the social sciences approximately at the same time by Andrew Abbott and colleagues (Abbott et Forrest 1986 ; Abbott et Hrycak 1990 ; Abbott 1995), is a family of descriptive methods adapted to sequences that make full use of the ordered dimension of longitudinal data.

In OM, the degree of dissimilarity between two sequences is determined by the least number of weighted 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's output is a dissimilarity matrix between all sequences that must be combined with cluster analysis, multidimensional scaling, or any other data reduction procedure handling dissimilarity objects. In the ancestor of OM, the Levenshtein distance (Levenshtein 1966 [1965]), the three basic operations are given equal weights: each operation cost one unit.

In theory, the choice of a cost system determines how sequences are matched, hence how sequence similarity is defined. In the social sciences, most early OM adopters claimed that results were little affected by changes in the relative weights of the three basic operations (for a review see

Abbott et Tsay 2000). OM detractors in this field have been interpreting this as a sign, not of robustness, but—often mistaking OM for a model—of weakness (Levine 2000). However, OM is a quite flexible family of methods that have been used in numerous fields to capture different kind of patterns depending on the material and question: computer science, coding theory, speech recognition, bird songs studies, gas chromatography, geology, human depth perception, biology, etc. And of course nowin the social sciences.

As underlined by Abbott (2000), "pattern search algorithms in general do not assume anything about the way the data are generated […but] They rather make assumptions about the kinds of patterns we expect to see". For instance, it is well known that when substitution operations are not allowed, or, this is exactly the same, when their cost is equal to or 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 et Liberman 1983). But exactly which kind of patterns go with which combination of costs remains nonetheless to be explored in the social sciences.

As a result, it seems that there are two ways of using OM in the social sciences, either to "fish […] for patterns" (Abbott 1990), that it to say to explore sequence data without any strong assumptions about the kind of patterns they may contain, or to find specific temporal patterns previously found and/or predicted by theory. As OM is used in the social sciences to uncover temporal patterns, the need to have precise ideas about the kinds of patterns looked for is not as pressing as it can be in some other fields, as for instance biology. However, this does not mean that social scientists can avoid reflecting on the relationships between edit operations and their costs and the kind of patterns that they can brought to light. Not knowing what kind of pattern a dataset conceal is one thing; disregarding how different parameterizations of OM lead to the uncovering of different sorts of patterns is another.

Sequences in the social sciences are not made of amino acids but express successions of social states or events[1]. The timing of event is often crucial in the social sciences as very often what matters is not only the events but when they occur. In this regard, it would be better to speak of *episodes* instead of events, that is of events occurring at *specific* moments within sequences[2]. Events coded identically but happening at distinct moments will be generally considered in the social sciences as different: "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

[1] An event is "something that is happening" (Merriam Webster) and can be represented by a change of state. States and events can be considered as different formulations of social processes see for instance the reply of Andrew Abbott to Lawrence Wu (2000).

[2] The *Merriam Webster* dictionary defines an episode as "an event that is distinctive and separate although part of a larger series".

the immediate change, that matters." (Abbott 1990). Whether OM is used as a sequence data mining tool or as a technique to capture different kinds of temporal patterns, more consideration should be paid to the link between costs and temporal patterns.

This article aims at addressing this concern. First I look into the consequences of the basic edit operations on the kind of temporal pattern that can be uncovered. Then I examine how it is possible to improve substitution costs in order to better capture the timing of sequences. Lastly, I contend that only substitution operations with time-dependent costs inversely proportional to transition frequencies should be used whenever the timing of sequences is central. This OM variant, coined Dynamic Hamming Matching, is applied to the question of the scheduling of paid work in France and compared with the three historical OM parameterizations.

2. Costs and temporal patterns

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 the social sciences as *Optimal Matching* comes in fact 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 and 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 strings 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 OM[3]. Levenshtein also suggested using only insertion and deletion operations to match strings. These two Levenshtein distances are usually considered as a refinement 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. Consequently, OM refers to the more general solution proposed by Levenshtein to the problem of sequence comparison and encompasses two particular cases: where the comparison is restricted to either substitution or insertion-deletion operations (see Table 1).

[3] Kruskal suggests a substitution penalty equal to 2, arguing that if the substitution cost is greater than 2 then "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)

Table 1 — The three historical OM variants and their costs

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

OM techniques were born in computer sciences and were subsequently imported into other scientific fields, especially biology. As OM was imported into the 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) all refer to biology to assess the use of OM in the 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 that were so, several of the fundamental biological operations involved in these transformations, such as swaps and larger transpositions, would be missing (Abbott 2000).

In actuality, 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 (Durbin *et al.* 1998). To achieve this, the key process is homology, where information about structure and/or function of sequences already known by experimentation is transferred to sequences with which significant similarities are found. In biology, indel and substitution operation do not have substantive meaning. Costs, however, are defined according to biological theories.

Substitution costs usually reflect evolutionary preferences for certain evolutions over others[4]. 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 processes involving the introduction of some unimportant residues between related alignments. However, even though it is also possible to turn the question of setting insertion and deletion costs into probability estimation, in practice this possibility is often disregarded and indel costs are usually set empirically relatively to substitution costs (Durbin *et al.* 1998, p. 16-17 and 44-45). Therefore, OM's three edit operations have no particular meaning in biology. They are just abstract operations used to align sequences. The key of the successful transposition of OM into the biological field rest on costs which are

[4] 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, which means that 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.

interpreted and defined according to biological theories. Social scientists should therefore not be too worried about the substantive meaning of edit operations[5] but should rather focus on cost setting.

In the social sciences, when an event is inserted or deleted, it is also time that is either added or removed. Indel operations warp time so as to align identically coded events. On the other hand, substituting an event by another preserves the timing of the sequences but at the cost of approximating an event by another one. In summary, insertion and deletion operations preserve events but distort time while substitution operations do just the opposite, *i.e.* they conserve time but alter events. As a result, OM applied to sequences of social events is a combination of accelerations/decelerations to match subsequences of identically coded events and of event approximations when the flow of time is normal (see Table 2). The expression "normal flow of time" has been used here to emphasize that once time has been warped, co-occurrences of events do not mean that these events are necessarily contemporaneous, unless time is accelerated then decelerated so that the respective time-scales of both sequences coincide again.

Table 2 — Edit operations and sequences of social events

	Insertion-Deletion	Substitution
Preserved	Events	Time
Altered	Time	Events

The warping of time by indel operations is a well-known feature of OM in the speech recognition field, which shares with the social sciences some of their concern with time[6]. While time warping is a valued feature in this field where it "has no intrinsic meaning and can be freely distorted" (Kruskal et Liberman 1983) this question is more problematic in the social sciences. Indeed, time warping means 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 I and II distances, neither the nature of the events suppressed nor their locations in the sequence are considered as relevant. As a consequence, time warping destroys the temporal links between sequences, their *contemporaneity*. To insert time to identify unemployment spells of approximately equal length suggests that the events themselves and their order are more important

[5] Some authors (Levine 2000 ; Wu 2000 ; Elzinga 2003) expressed concerns about the sociological meaning of the three basic operations of OM, some arguing that the legitimacy of OM in biology was stemming from the theoretical relevance of the three edit operations.

[6] 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, and (3) build 'average' sequences. 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. Both indel and compression-expansion operations are used in speech recognition. The former are used in order to recover interpolated or deleted sounds (e.g. "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 sections 6 and 7). It is interesting to note that, as in biology, it is through costs that OM is fine-tuned in order to suit the requirements of the analysis.

than *when* they occur (e.g. in a mass unemployment or a full employment period); thus events lose their indexicality.

When sequences are put together in order to be analyzed by means of sequence analysis techniques, it is assumed that they are ordered according to a common time scale and that the aim of sequence analysis is to study the thus implicitly defined *calendar*. A calendar is not necessarily an institutionalized system of division of time as the year, the month, the hour, but can be defined as any relevant social system of division of time as for instance the calendar of footsteps of the Ilmington dances (Abbott et Forrest 1986), the calendar of the German musician careers (Abbott et Hrycak 1990) or the calendar of lynching (Stovel 2001). The term calendar is used here to emphasize that the aim of applying sequence analysis on social science data is to uncover socio-temporal regularities. This term refers to the precursory work of Durkheim on time (Durkheim 1912): "The calendar expresses the rhythm of collective activities, while at the same time its function is to assure their regularities". Calendars reveal the rhythm(s) of collective life but at the same time help individuals to anticipate, plan and orient themselves. Calendars can be more or less structured, institutionalized, recognized by actors, etc., but as long as there is some sort of collective activities there is a calendar.

As a consequence, what time-warping and contemporaneity mean depend on the nature of the calendar implied by putting sequences together. Contemporaneity does not refer exclusively to the common period of time in which sequences may unfold. For example in a panel of individuals followed over a period of years, trajectories involve age and period effects. But with such data, other types of sequences can be defined. For instance, Brendan Halpin (Halpin 2008), using the British Household Panel Survey, studied the six-year monthly labor market histories of women who had a birth at the end of the second year, classified into full-time and part-time employment, unemployment and non-employment. In this case, even if the time unit is still months, the calendar studied is defined by the cohort of women who became mothers at the end of the second year, whatever this year is. Even if trajectories are not anymore located in the same historical time, time warping is still an issue as the aim of the analysis is to identify different temporal patterns of labor market attachment after entry into motherhood: whether women get back to work six months or two years after giving birth to their first child matters for the analysis.

As shown by this example, the effect of time warping also depends on how sequences are arranged and coded. Coding states amounts to defining the social space in which unfold the series of states studied. With OM, social sequences are indeed not considered as "the list of successive realizations of an underlying stochastic process" (Abbott 1990) but as social processes unfolding in interactional fields governed by rules and regularities (Abbott 1997). Consequently, the kind of temporal patterns that can be uncovered using OM depends first on the state space defined in this coding stage. The kind of temporal patterns that can be identified by OM and as a result whether or not time-warping is a desirable feature primarily depend on the definition and constitution of the social field studied. In history, OM was applied to identify patterns of folk dances (Abbott et Forrest) or musicians careers (Abbott et Hrycak). In the field of stratification analysis, OM has been used to

identify intragenerational mobility patterns (Halpin et Chan 1998); in time-use analysis it has been applied to examine daily lifestyles (see for instance Saint Pol 2006 ; Lesnard 2008)[7].

The kind of temporal patterns that can be brought to light with OM can be located on a scale (see Figure 1) ranging from the number of identical states identically located in the sequences (Hamming distance, see Table 1) to the longest common subsequences irrespectively of their location in the sequences using only indel transformations (Levensthein II distance, see Table 1). When all the states have the same substitution cost, setting indel costs to a value smaller than or equal to twice the cost of a substitution amounts to finding the longest common subsequences wherever their locations in the sequences. When one insertion and one deletion cost more than one substitution, as for instance in the Levenshtein I distance (see Table 1), then both kinds of operation are used and it is not anymore the longest common subsequences which are found but the longest quasi-common subsequences. A quasi-common subsequence has some states not aligned in between two series of common states. Using more than one substitution cost allows even more flexibility in the balance between identical subsequences and very similar subsequences as it gives the possibility to define what kind of quasi-common subsequence is acceptable or not. States with high substitution costs, that is, higher than one insertion and one deletion cannot be part of the longest quasi-common subsequences whereas states which substitution costs are lower than two indels can be.

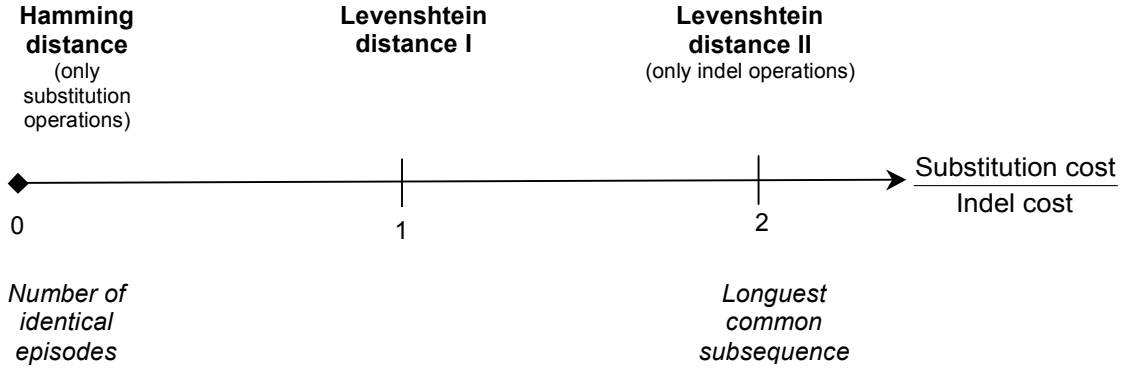

Figure 1 — Ratio of substitution to indel costs and kinds of pattern captured by OM

The balance between indel and substitution operations will focus the analysis towards temporal patterns located between two polar ideal-types, one where the timing of events is less important than their order (the Levenshtein II pole) and the other where the timing of events is crucial (the Hamming pole). Using only indel transformations makes it possible to identify long common subsequences whereas using only substitution operations amounts to measuring the degree of contemporaneity of sequences. In their review, Abbott and Tsay (2000) underline that indel costs are most of the time set empirically once substitution costs are defined, either empirically or theoretically.

[7] For a review of the different uses of OM in the social sciences, see Abbott and Tsay (2000).

As a consequence, setting substitution costs so that they adequately capture contemporaneous similarities is the major challenge social scientists are facing, whether or not indel operations are also used.

3. Improving substitution costs to capture contemporaneous similarities

When the timing of event is crucial, insertion and deletion operations warp time and smooth out a great part of the social structure of sequences. This is the case in the time use field where the timing of everyday activities is decisive. But it is can also the case in other fields such depending on the research questions, as for instance life course research where the timing of the different stages analyzed is very often critical (Aisenbrey et Fasang 2007). Preserving the timing of sequences comes at the expense of distorting episodes whenever they are different. Indeed, substitution costs reflect the penalty of replacing a state by another one: the higher the penalty, the more different states are. Substitution costs should then be interpreted as the likelihood that two different episodes are contemporaneously close i.e. that they belong to the same trajectory pattern even though they are different. In this respect, it seems better to allow substitution costs to vary with time in order to improve the extraction of the social structuring of the timing of events. Time-independent substitution costs amount to assuming that the likelihood that two different episodes are contemporaneously close is time-constant, which is a strong assumption.

Yet, once sequences are time warped by indel operations, their respective time scales do not coincide anymore and time-varying substitution costs cannot really be used unless a choice is made regarding which date of the two sequences should be considered. The simplest way to implement time-varying substitution costs is to keep sequences always in sync by using only substitution operations, which is possible only when sequences are of equal length. When no indel operations are used, matching is based on the identical parts of the two sequences and on the time-varying degree of proximity of the differing episodes.

Table 3 — Time-dependent substitution costs: an example

	Low rate of unemployment			High rate of unemployment		
	1	2	3	4	5	6
i	E	E	E	U	U	E
j	U	U	E	E	E	E
k	E	E	E	E	E	E

If we consider two sequences describing work stability with two states, employed (E) and unemployed (U), then using time-varying substitution costs makes it possible to define unemployment spells as being closer to employment ones when the unemployment rate is high. For example, if the employment rate is low at the beginning of the period studied (t = 1,2) but high after, then the distance between *j* and *k* will be higher than the one between *i* and *k* because being

unemployed at a time of full employment is more atypical than when unemployment is widespread (Table 3). Such time-varying substitution costs also mean that the distance between *i* and *j* will be higher than the one between *k* and *i* because even if they have more events in common (they both experience unemployment), these events occur at different dates with different rates of unemployment: when unemployment is low, being unemployed is more atypical than at times of mass unemployment. Of course, if the unemployment rate were stable throughout the period studied, then using time-varying substitution costs would be irrelevant.

In this example, substitution costs are defined according to the rate of unemployment, which can be calculated from the same data. Such a method to derive substitution costs becomes problematic for sequences with three or more states[8]. A solution to take into account the timing of sequences is to use the series of transition matrices that describe the transitions between all states between two consecutive dates. A transition matrix is a macro representation of individual trajectories between all the different states between two consecutives dates. The strength of the flux between two different *states*, measured by transitions, can be used as an indicator of how close two different *events* are. A low transition rate between two states means that, at that particular moment, these two states are not connected hence that they can be considered as being part of two distinct trajectories. On the contrary, a high transition rate between two states can be interpreted as a change of state within a single trajectory.

For example, in the 1999 French Time Use survey, 22% of the respondents started to work between 8:00 and 8:10 but only 3% between 10:40 and 10:50. Conversely, only 78% of those not at work at 8:00 were still not working at 8:10 whereas 97% of the non-workers at 10:40 did also not work 10 minutes later. In the vocabulary of markov chain analysis, between 10:30 and 10:40, work and non-work are very close to being two *absorbing* states, that is, two states from which it is impossible to leave, suggesting that these two states belong to two different processes. If in two workdays considered at 8 AM, one has work but not the other, then even if these two episodes are different, they are however likely to belong to the same type of workday.

As a result, the cost for substituting work for non work should reflect that even though episodes are different, empirical evidence at hand suggest that at that particular moment in time, they are likely to be two slightly shifted variants of the same type of workday. On the contrary, because transitions are very low between 10:30 and 10:40, the states work and non-work found in two sequences should be considered as very different at that time. Whereas it is hard to tell around 8 AM if two persons, one working and not the other, have different work schedules, it is easier at 10:40.

[8] When there are only two states, the contemporaneous proximity can be derived indifferently from either the rate of unemployment ($p_t(U)$) or the rate of employment ($p_t(E)$) since $p_t(U) = 1 - p_t(E)$.

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 is clearly different. But the difference in the absolute number of hours that separate them can be either increased or lessened by collective rhythms. For instance, the social difference between one hour of work from 4 PM to 5 PM (9-to-5 workday) and another from 7 PM to 8 PM (evening work) is larger than the absolute number of hours that separate 4 PM to 7 PM[9].

4. Dynamic Hamming Matching

The solution suggested in this paper is (1) to use time-varying substitution costs inversely proportional to transition rates (2) to only use substitution operations[10]. When all sequences have the same length[11], and the sample and coding are defined so as to uncover contemporaneous similarities, then is it possible to use only substitution operations with costs derived from transitions. Temporal distortions are avoided since indel operations are not used. This method is no longer based on optimality principles, precisely because it is the search of logic optimality that causes time warping. In this regard, the OM variant suggested here can be seen as an extension of the Hamming distance with substitution costs derived from the series of transition matrices describing the sequences. Sample weights can be used to estimate the transitions matrices so that the survey design can be to a certain extent integrated in OM[12].

---

[9] Before turning to the solution proposed in this paper to the question of substitution cost setting, it seems necessary at this point of the article to address the issue of the software implementation of OM. It should be clear that importing directly into the social sciences programs that were designed in other fields is delicate. While it is no longer maintained but still available, it is worth mentioning the program designed by Andrew Abbott, *Optimize*. Only time-invariant substitution costs can be used and indel costs are determined relatively to them according to a scale factor. A sequence module is available in the *TDA* package, a freeware developed by Götz Rohwer and Ulrich Pötter of the University of Bochum originally to apply event history models. By default indel and substitution costs are respectively set to 1 and 2 but can be set to other values. Indel costs can be set using a single value, a user-defined matrix or a linear indel function cost with two parameters. Transition frequencies or any user-defined matrix can also be used as substitution costs. A set of *Stata* ado files proposing roughly the same functionalities have been recently released (Brzinsky-Fay, Kohler et Luniak 2006). More recently, a *R* library, TraMineR, brings sequence analysis, including optimal matching, to *R*.

[10] It could be possible to use indel operations by using dynamic costs defined relatively to substitution ones; for instance the middle of the distribution of substitution costs.

[11] In the social sciences, sequential materials are collected by means of survey and consequently are not necessarily of equal length. For instance, in a survey with retrospective questions on family and work biographies carried out on a representative sample of the population with age ranging from 18 to 65, family and work sequences are of different length. Analyzing with OM social sequences of uneven length seems highly problematic: what kind of regularities can be obtained out of sequences so varied in their completeness? Of course OM handles such sequences, but in a very cursory way, through insertions; the quality of such extrapolation then depends on insertion costs, in particular whether or not they vary with time. In the above example, the only solution available to analyze sequences of equal length would consist in focusing on partial biographies, between 18 and 30 for instance (transition to adulthood). It would amount to exclude respondents younger than 30 (incomplete biography) and to truncate the other sequences over that age.

[12] Sample 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

The fact that substitution costs are derived from transitions between states and are 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 an individual and sequential synthetic measure of those relationships. This sequence comparison method is basically turning transition matrices into inter-individual differences.

This variant can appear similar to the common practice of setting substitution costs using information about transitions (Abbott et Forrest 1986 ; Abbott 2000). If this strategy has indeed already been used, substitution costs are usually time independent, *i.e.* they are derived from a global transition matrix collapsing all the couples of dates, thus disregarding the intra-sequences variability.

When the sequences have all the same length it is suggested to estimate the $\{p_{ab,t}\}$, the proximity of two states occurring at the same time, 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 $t+1$: $p(X_t = b|X_{t-1} = a)$[13], $p(X_{t+1} = b|X_t = a)$, $p(X_t = a|X_{t-1} = b)$, $p(X_{t+1} = a|X_t = b)$, where X_t is a random variable describing the occurrence (event) of the t^{th} episode of a sequence. In other words, a diachronic distance is substituted for a synchronic one. From a probabilistic point of view the higher the probability of transition between the two *states* before and after t, the closer the two *episodes*. One possible way to do this is simply to define the substitution cost function as[14]:

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

account. The matching procedure in itself, *i.e.*, the comparison of pair of sequences does not require any weights; it is by definition a one to one procedure. However, sample weights should be turned on to interpret results, for instance, if cluster analysis is used, the size of the clusters obtained must be weighted.

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

[14] This formula is a quite straightforward adaptation of the one used in TDA to implement transition-based substitution costs (Rohwer et Pötter 2005, p. 496-497). It 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 : $$s_T(a,b) = \begin{cases} 4 - 2\left[p(X_T = a|X_{T-1} = b) + p(X_T = b|X_{T-1} = a)\right] & \text{if } a \neq b \\ 0 & \text{otherwise} \end{cases}$$

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 *episodes a* and *b* at t (with a lower bound of 0). Indeed, high transitions mean that many individuals have just changed from *a* to *b* or from *b* to *a*, or that they are about to do so. In statistical terms, the probability at t that *a* and *b* belong to the same trajectory is high. On the contrary, low transitions mean that these two states are not connected around t, that, from a probabilistic viewpoint, they belong to two different types of trajectories. Thus, substitution costs depend on time and are derived from the transitions observed in the sample studied. As transition rates necessary imply two consecutive dates while dissimilarity is only needed for a single date, it seems better to smooth a little bit substitution costs by taking into account the two transitions immediately before *and* after the date of interest rather than only the one before *or* after.

Other implementations of this type of transition-based substitution costs are possible. More transitions before and after the date of interest could have been taken into account. It would have even been possible to use all the transitions before and/or after t in order to smooth more substitution costs[15]. However, as the aim of DHM is precisely to uncover temporal patterns, smoothing should never be too strong. Overall, the more dates considered in the calculation of such substitution costs, the more timing is smoothed. However, the effect of the number of dates ultimately depends on both the time unit and the timing of the variations of the process studied. If daily activities were observed minute by minute instead of every ten minutes, it might have been necessary to use more dates before and after t. On the contrary, if daily activities were only observed every hour, then using more dates would have certainly smoothed out most of the temporal variations. The question of the correspondence between the time unit and the variations of the phenomenon measured is however unlikely to appear in practice as the time units of longitudinal data are very often scaled to the temporal variations of the process of interest.

Before turning to the application of this method to the scheduling of paid work, it is worth noting that transitions from *a* to *b* as well as from *b* to *a* are used to estimate the degree of proximity of the states *a* and *b*. Deriving substitution costs from transition does not imply that substitution costs are conceived in terms of transitions. Substitutions are *diachronic* in essence whereas transition are by definition *synchronic*. In the example of job stability, it means that both those who become unemployed and those who find a job are taken into account to assess the proximity of the states employment an unemployment. A sequence with three employment spells followed by three unemployment spells can be considered as quite similar as another one with three unemployment spells and three of employment at $t = 3$ if the substitution cost is low, but the total distance will be nonetheless quite high as they never coincide: “The fact that there is a temporal or linear logic (that

[15] In this case, rather than assigning equal weights to past and/or future transitions, decreasing weights with the temporal distance of transitions from t could be used. For instance, it might be interesting to use exponentially decreasing weights similar to those used in the exponential smoothing technique in time-series analysis.

certain states are disproportionately likely to follow or precede other specific states) is a feature of the longitudinal nature of the trajectory rather than of the state space" (Halpin 2008).

5. 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. Dynamic Hamming Matching is nonetheless not bind to this question and to this kind of data but can be applied to any social sequence dataset where timing is essential. For instance it has been successfully applied to life course data to identify trajectories to old age security in West Germany (Aisenbrey et Fasang 2007). It has also been applied to more complex time-use sequences to describe jointly the work schedules of dual-earner couples with the help of four states (Lesnard 2008) or the scheduling of work over the week with short sequences (seven days) made of five states (Lesnard et Saint Pol 2009). The simplicity of the analysis of work schedules where sequences are just made of zeroes (not at work) and ones (at work) is intentional and aims at exploring how Dynamic Hamming Matching fares on an ideal-typical problem. To do so, DHM will be compared to the three classical OM variants described in Table 3.

Work schedules have been usually reduced to either durations (the number of hours of work) or categorical indicators (e.g. day vs. night work). In order to distinguish night work from work schedules shifted in the afternoon/evening or in the morning, precise criteria are required. Despite the fact that these criteria can be based on *a priori* knowledge, they require setting threshold and as such, necessarily entail some arbitrariness. As a result, the scheduling of work is most of the time reduced to simplistic and rigid dichotomies, eg. day vs. night work, which makes it difficult to study work schedules with the necessary level of details. Indeed, when the entire distribution of work hours over the day is taken into account, it appears that if night work remained stable in the US since the 1970s, work before 9 AM and after 5 PM increased significantly (Hamermesh 1999). This trend can be linked to the growth of the service sector where many occupations have work hours at the fringes of the of the 9-to-5 workday (Presser 2003). These low-skilled occupations also tend to work fewer hours than in the past (Gershuny 2000), yet, short workdays do not necessarily go hand in hand with shifted schedules. If previous studies gave some very useful first elements on the correlation between work schedules and occupation, only a detailed typology of workdays can give more insights on this issue. As the timing of work is more important for the analysis than the number of hours of work, OM variants close to the Hamming pole on Figure 1 should in theory give better results.

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 the "stylized questions" on time directly asking respondents to give average estimates of the time they spend doing some pre-defined activities, in time use surveys information on time is collected in diaries in which respondents describe, with their own words, the

sequence of activities they did on a specific day. These descriptions are then coded according to a nomenclature of activities. 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 et Sullivan 1998). The last two French time use surveys (1985-86 and 1998-99) used here were done in person by the French Institute of Statistics (INSEE) over the course of a year[16] and had high response rates[17]. In the 1985-86 survey, one respondent was selected among household members ages 15 and over using the Kish method. When the respondent had a partner, he or she was also interviewed. In the 1999 survey, all household members over 15 year old were interviewed. In both surveys, respondents were asked to describe their activities over the course of one day, imposed by interviewers so that all the days of the week were represented equally. As the aim of the analysis is to describe workdays, the information about daily activities contained in the diaries of these two surveys has been drastically reduced to two activities: work vs. non work. Diaries of both surveys cover 24 hours (midnight to midnight), with minor differences in precision[18], and as a result all sequences have the same length (144 10-minute spells) and are day-synchronized.

Four OM analyses were conducted on the two samples merged (N = 7,908)[19]:

- Hamming
- Dynamic Hamming
- Levenshtein I
- Levenshtein II

The four dissimilarity matrices were analyzed with the flexible beta cluster algorithm, also known as flexible WPGMA (Weighted Pair Group using arithMetic Averages), proposed by Lance and Williams (1967), one of the most efficient method in presence of noise and outliers (Milligan 1980; Milligan 1981; Milligan 1989). The same settings have been used ($\beta = -0.5$) for the four dissimilarity matrices. Following Rohwer and Pötter (2005, p. 468-470), entropy (Shannon's H) is used to compare the homogeneity of state distribution in the four typologies. If p_{tj} is the proportion of individuals who are in state j at t, then entropy at time *t* can be defined as:

---

[16] 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).

[17] 65% for the 1985-86 French TUS, and 80% for the 1998-99 one.

[18] The 1985-86 and 1998-99 surveys have respectively 5- and 10-minute time slots: comparability can be an issue but an unpublished methodological study (Alain Chenu, personal communication) suggests 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.

[19] DHM is available in *SAS* as a macro, in *Stata* as a plug-in (see the author's web page), and in *R* in the TraMineR library. All the OM analyses were carried out in *R* with TraMineR. Detailed results will only be provided for the Dynamic Hamming Matching typology.

$$H_t = -\sum_{j=1}^{q} p_{tj} \ln(p_{tj}) \quad (2)$$

H_t is bounded by 0 and ln(q), values reached respectively when all individuals are in the same state and when individuals are equally distributed among the q states. Therefore, the lower H_t, the higher the homogeneity of state distribution at t. In other words, low entropy values signal that all the individuals considered are in the same state (work for instance) at the same time. As a result, entropy can be used as a measure of contemporaneous similarity of the four typologies. Entropy is by no means an absolute quality index as it obviously favors high degree of contemporaneous similarities. It is used here only to see whether or not Dynamic Hamming Matching captures this kind of temporal pattern better than Levenshtein I and II dissimilarity measures.

Table 4 — Average entropy (twelve-cluster solutions)[20]

	H	%
Hamming	0.2121	30.60
Dynamic Hamming	0.2172	31.33
Levenshtein I	0.2183	31.50
Levenshtein II	0.2182	31.48
Whole sample	0.4000	57.70

Note. The first column shows the absolute values of entropy (weighted averages over cluster and time for the typologies) and the second, values of entropy relative to the maximum possible value (ln(2)). The lower entropy, the higher homogeneity.

The entropy figures corresponding to each of the four typologies (see Table 4) were obtained in the following way. First, entropy was derived from (2) for each time slot and for each of the twelve groups of a given typology[21]. To get an entropy indicator for a group of a given typology, these 144 entropy figures were then averaged (simple mean). At this stage each of the twelve groups of the four typologies is characterized by an average entropy. In order to obtain a single figure for each of the four typologies, these twelve entropy measures were finally weighted by the size of their respective clusters and averaged. These successive averages are likely to be responsible for smoothing out most of the differences in entropy between the four typologies. However, even if differences are small, the two Hamming dissimilarity measures have indeed the lowest entropy values.

[20] 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 inter-group distance for the last steps in the grouping process can give some guidelines as a spike reveals that two dissimilar clusters have just been joined. The graph (not shown) for Dynamic Hamming Matching suggests that an eight-class scheme is the most acceptable synthetic representation of the structure of the data. Other spikes 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. A twelve-class classification was finally adopted after close inspection of the shape and relevance of all the cluster solutions between fifteen and eight.

[21] The seqstatd command of the *R* library TraMineR was used.

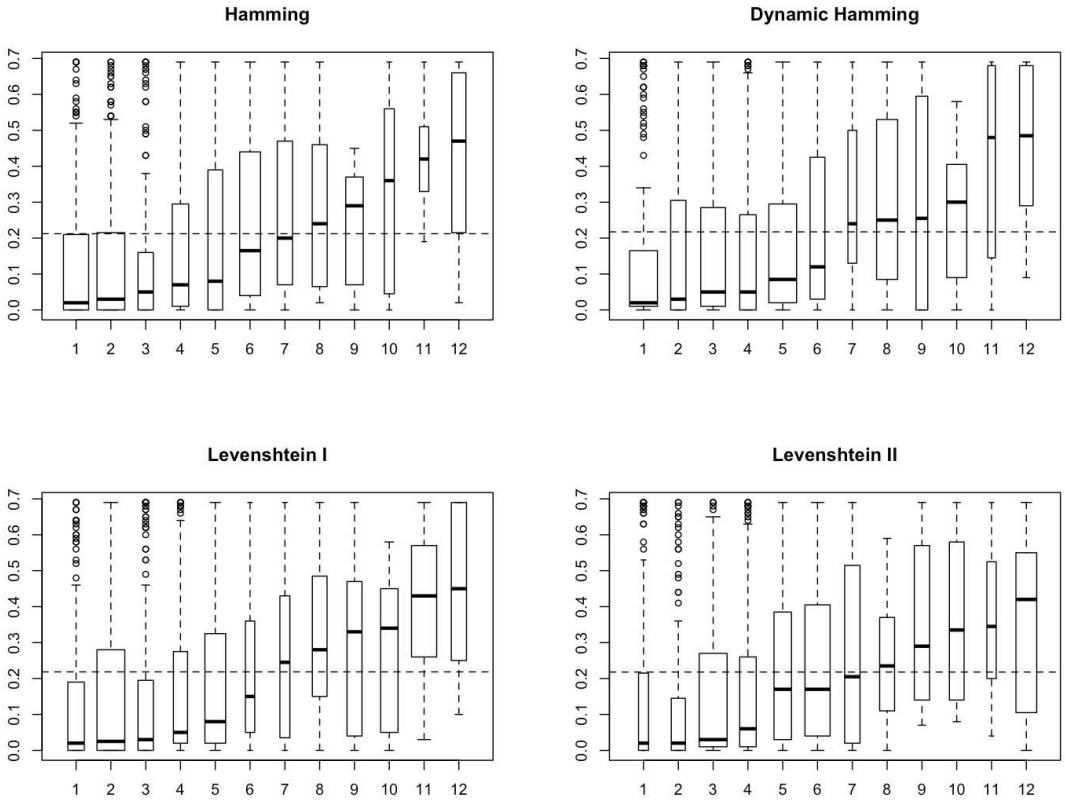

Figure 2. Entropy distribution for the four twelve-cluster typologies

Note: The dotted lines indicates the average entropy value (cf. Table 4). Box widths are proportional to the square root of the size of each cluster.

The inspection of the four series of twelve boxplots of the 144 entropy values (Figure 2) gives a better picture of the differences between the four OM variants. The two hamming dissimilarity measures keep entropy at very low levels for about half of the clusters whereas entropy figures are low for only four of the twelve clusters derived from the Levenshtein II one. It seems that the better entropy efficiency of the Hamming dissimilarities for a larger number of clusters comes at the expense of a two or three small clusters with very high entropy values. This explains why on the average the four cluster solutions are about the same. It should also be noted that the low-value entropy clusters of Levenshtein II are smaller than the Hamming ones or even the Levenshtein I ones. As Levenshtein II does not favor contemporaneity, it is just because the data contain highly synchronized workdays that they are nonetheless picked up by this OM variant. However, only perfectly synchronized workdays are grouped together and even if the four techniques can identify the same highly synchronous patterns, their relative size is very different. For instance, with Levenshtein II, quite synchronized workdays but of different lengths will end up in two different clusters, just because parameterization favors identically coded events, here work duration, over their timing. The Levenshtein I cluster solution is in between these two patterns: it has both more low- and high-entropy groups than Levenshtein II but less than Hamming and Dynamic Hamming.

Homogeneity of state distribution can also be assessed visually by plotting 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 (or chronogram) and the latter is an individual tempogram or index plot. Both kinds of tempograms help to interpret and assess visually the quality of sequence classifications. The gradient and the height of the curve of aggregate tempograms indicate how homogeneous clusters are: the steeper 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 last two clusters, which clearly lack homogeneity, the overall quality of the Dynamic Hamming Matching taxonomy assessed visually with aggregate tempograms appears quite satisfactory (see Figure 3). Individual tempograms (see Figure 4) confirm these impressions and measures, showing that most clusters contain very similar sequences. Tempograms of the two Levenshtein typologies (not shown) look less homogeneous, confirming previous findings.

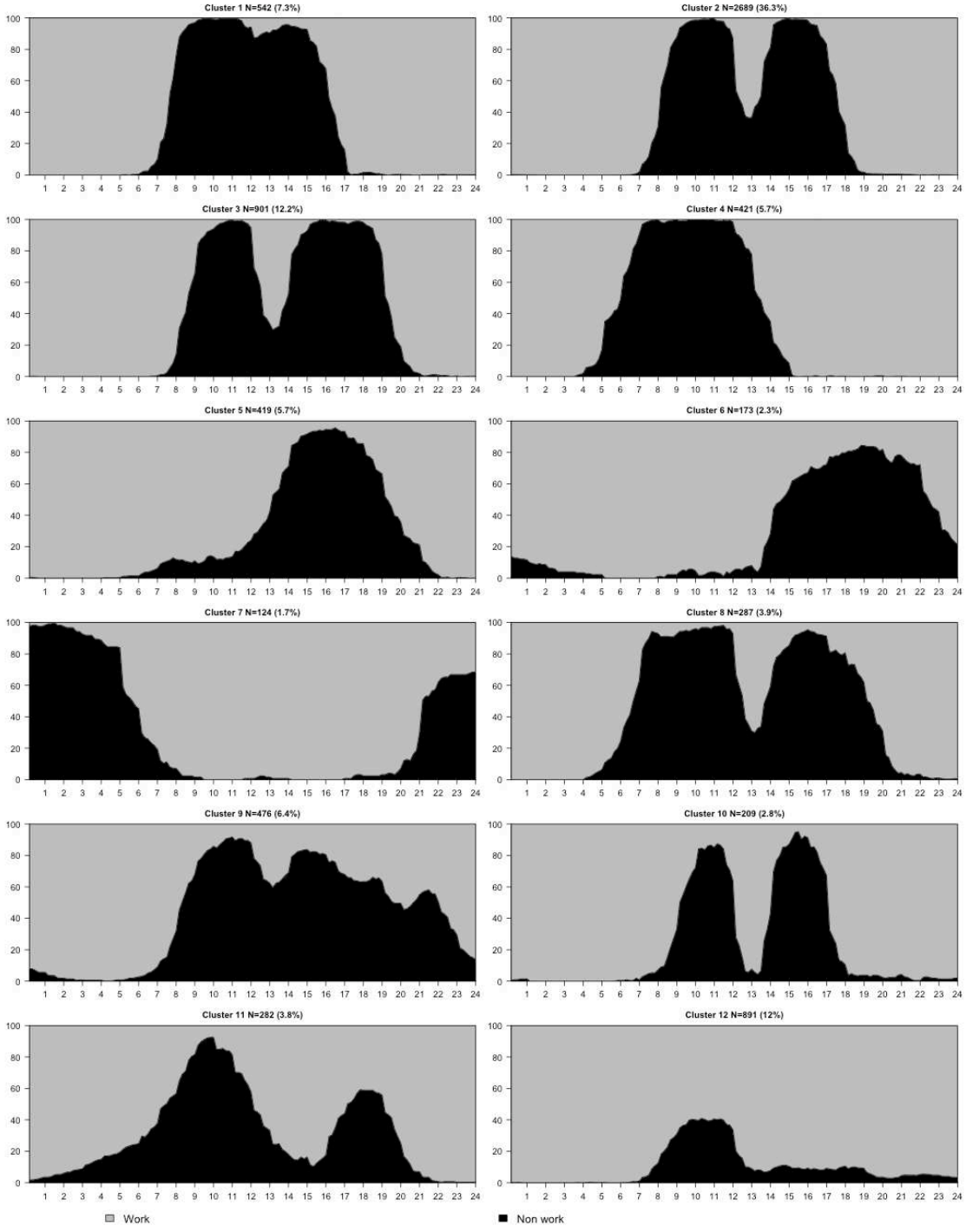

Figure 3 — Aggregate tempograms for the Dynamic Hamming Matching typology

Note. Cluster id numbers are different from Figure 2.

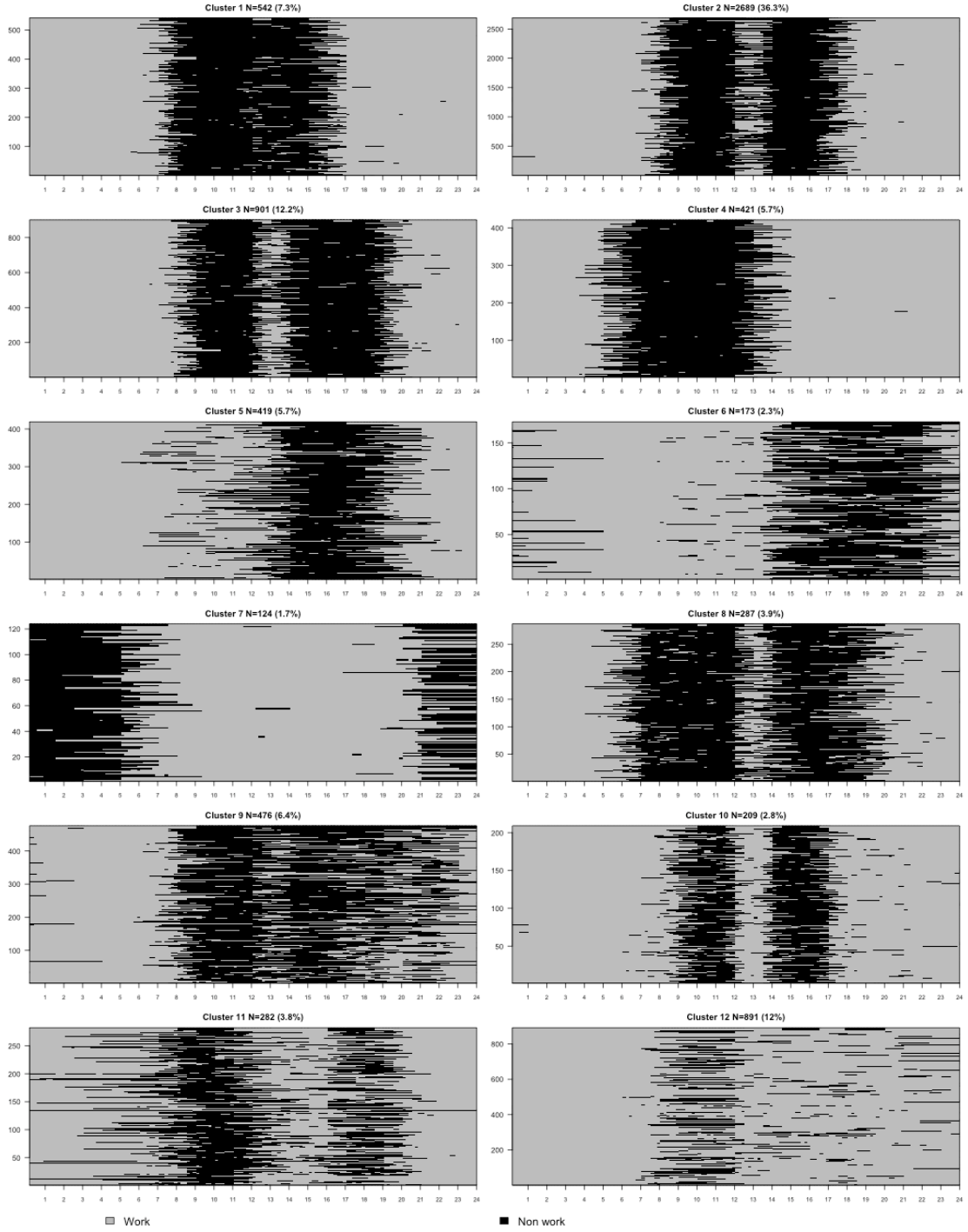

Figure 4. Individual tempograms for the Dynamic Hamming Matching typology

Note. Cluster id numbers are different from Figure 2.

But more importantly, the other typologies are less interpretable. In the case of the Levenshtein typologies, it is certainly because the social structuring of the timing is partially blurred by indel operations. It is the opposite for Hamming, which is so good at spotting contemporaneous similarities that it tends to group sequences together not because they are alike but just because they are very dissimilar from the very synchronized ones. Work schedules can be described roughly by two simple indicators: the number of work hours and the time of the day corresponding to the middle of the workday (mid-workday), which gives a very rudimentary indication of the scheduling of work within the day. With the help of Table 5 and of aggregate and individual tempograms, the Dynamic Hamming Matching clusters can be easily labeled and interpreted. The first three clusters consist of the *9 to 5 workdays* 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 of "shifted work schedules". Work schedules located at the margin of the 9 to 5 work day have increased in France as it was found for the US with visual estimates (Hamermesh 2002).

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

		1985-86			1998-99			
	Type of work day	Size (%)	Mid-work day	Duration	Size (%)	Mid-work day	Duration	Average entropy
	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	.1390
2	9 to 5	38.17	12:53	8:17	33.88	12:57	8:23	.1720
3	10 to 7	10.69	14:01	9:09	14.03	14:03	9:39	.1872
	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	.1381
5	In the afternoon	5.40	15:32	6:46	6.43	15:24	6:43	.2812
6	In the evening	2.08	17:02	7:20	2.49	17:20	7:04	.3383
7	In the night	1.66		7:38	1.57		7:56	.2394
	Long	**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	.2899
9	10 to 7 spreading in the evening	5.59	14:38	10:18	7.52	14:46	10:58	.4321
	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	.2327
11	Fragmented full time	3.46	12:15	8:06	4.22	12:11	7:20	.4343
12	Very short work day	13.32	12:52	2:14	10.54	13:31	2:41	.2483
	Total	**100.00**		**7:32**	**100.00**		**7:58**	**.2172**

Long workdays come in two flavors: either in a long version of the standard workday, *i.e.* beginning earlier and ending later than the 9 to 5 workdays, or in a long version of the 10 to 7 ones, *i.e.* ending later than 7 PM. Other patterns of workdays are less clear and are generally made up of short and/or fragmented workdays. Fragmented means that work schedules have at least two distinct work periods separated by considerable time. The best example of this is supermarket cashiers 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 (Prunier-Poulmaire 2000). Fragmented part-time workdays are often concentrated around the lunch break, *i.e.* at the end of the morning and the beginning of the afternoon. Fragmented full-time workdays are fragmented workdays *par excellence*. Indeed, although their duration average eight hours, they are made of two distinct but highly variable work periods separated by several hours. In this case, mid-work day is a very poor indicator of the scheduling of work. Finally, the last cluster groups very short workdays together. Since all days with at least a 10-minute work spell have been considered as workdays, this last cluster collects in fact very short work days without having to *a priori* define a minimum work time.

6. Conclusion

Up to now, OM has been mainly used in the social sciences as a kind of sequence data mining tool capable of uncovering socio-temporal patterns. There is nothing wrong with this kind of use but even if OM can be used without any specific expectations on the kind of socio-temporal patterns buried in data, it seems crucial to know what kind of patterns can be uncovered with OM and how those different patterns are linked to cost setting. Indel operations warp time in order to match identically coded states but occurring at different moments in their respective sequen ces. Substitutions do the opposite as substituting one event by another preserve their location in their respective sequences but entails approximation. As a result, the kind of socio-temporal patterns that can be brought to light by OM vary with costs and range from finding the longest common subsequences irrespective of their locations, when indel costs are low relatively to substitution ones (Levenshtein II), to identifying contemporaneous similarities, when indel costs are high relatively to substitution ones (Hamming). The flexibility offered by OM is even greater when more than one substitution costs are used and when costs vary with time.

Two consequences can be drawn on. First, that if OM can be used as a sequence data mining tool, different combinations of costs should be used in order to explore the different types of temporal patterns concealed in data. In this respect, the Levenshtein I dissimilarity measure might represent a good starting point, as it combines limited time-warping with neutral substitution costs. In a way, Levenshtein I plays a similar role in OM than the uniform prior distribution in Bayesian statistics. Second, if OM is used to measure specific similarities, then costs should chosen accordingly. Of course in any case, coding is likely to play a major part in the kind of temporal patterns that can be uncovered. This step is as crucial as parameterizing correctly OM given that socio-temporal patterns are captured within the bounds laid out by the different states chosen. If no difference is made between two states playing a fundamental part in the trajectories studied, then it will be hard to get something out of OM, whatever costs are chosen.

The greatest challenge social scientists are facing to apply OM is to find sensible ways to determine substitution costs to capture adequately contemporaneous similarities. This issue is even more prominent when the timing of the sequences studied is of primary importance, as it can be in time-use studies, but also in other field of social sciences, as for instance for life-course research.

Indeed, using indel operations amounts to voluntarily adding noise to the phenomenon under study and should be seldom used whenever the timing of events is considered as crucial for the analysis. Dynamic Hamming Matching (DHM), which only use substitution operations with time-varying costs derived from the series of transition matrices, has been specially designed for this purpose. Indeed, 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 describing a set of sequences can also be seen as the macro description of these collective rhythms. With substitution costs inversely proportional to empirical transition frequencies, 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 trajectories. On the contrary, high transition probabilities between two states may reveal changes in a single trajectory. Deriving substitution costs from transition matrices amounts to disaggregating and connecting this macro information on collective rhythms.

Dynamic Hamming Matching was applied to study the timing of paid work and compared to the three classical OM variants. The four dissimilarity matrices were analyzed using flexible WPGMA. Despite the fact that DHM only uses substitution operations, differences in timing can appear within clusters. Indeed, as OM is only the first stage of the analysis and is supplemented by cluster analysis, giving the priority to contemporaneous similarities do not totally prevent from finding other kinds of patterns. But the cluster analysis stage is far from removing all the effects of cost setting. In terms of the homogeneity of state distribution (entropy), Dynamic Hamming Matching fared better than the two Levenshtein dissimilarity measures. The different types of workday are also more interpretable because information on the timing of sequences is not blurred by indel operations.

As the goal of this article was to introduce the method and its rationale, it was not possible to push any further the methodological comparison of those four methods. It is however a much needed next step. OM is still quite new to the social sciences and therefore requires abundant critical use, replication, and validation (Levine 2000). Different ways of describing social patterns must be systematically compared using different kinds of data. In this regard, future methodological work should not be restricted to OM but should consider other forms of sequence analysis techniques but also alternative methods such as multiple correspondence analysis and direct cluster analysis.

REFERENCES

Abbott, Andrew. 1990. « Conceptions of time and events in social science methods ». *Historical methods* 23:140-150.

—. 1995. « Sequence analysis: new methods for old ideas ». *Annual Review of Sociology* 21:93-113.

—. 1997. « Of Time and Space: The Contemporary Relevance of the Chicago School ». *Social Forces* 75:1149-1182.

—. 1998. « The causal devolution ». *Sociological Methods and Research* 27:148-181.

—. 2000. « Reply to Levine and Wu ». *Sociological Methods and Research* 29:65-76.

Abbott, Andrew and John Forrest. 1986. « Optimal matching methods for historical sequences ». *Journal of Interdisciplinary History* 16:471-494.

Abbott, Andrew and Alexandra Hrycak. 1990. « Measuring resemblance in sequence analysis: an optimal matching analysis of musicians careers ». *American Journal of Sociology* 96:144-185.

Abbott, Andrew and Angela Tsay. 2000. « Sequence analysis and optimal matching methods in sociology ». *Sociological Methods and Research* 29:3-33.

Aisenbrey, Silke and Anette E. Fasang. 2007. « Beyond Optimal Matching: The 'Second Wave' of Sequence Analysis ». *CIQLE Working Paper*. New Haven: Yale University.

Bocquier, Philippe. 2006. « Les effets peuvent-ils précéder les causes ? Traitement des intentions et des anticipations », p. 239-259 in Philippe Antoine and Éva Lelièvre. *États flous et trajectoires complexes*. Paris: Ined.

Brzinsky-Fay, Christian, Ulrich Kohler and Magdalena Luniak. 2006. « Sequence analysis with Stata ». *Stata Journal* 6:435-460.

Degenne, Alain, Marie-Odile Lebeaux and Lise Mounier. 1996. « Typologies d'itinéraires comme instrument d'analyse du marché du travail », p. 27-42 in Alain Degenne, Michèle Mansuy, Gérard Podevin and Patrick Werquin. *Typologie des marchés du travail. Suivi et parcours. 3e journées d'étude Céreq/Cérétim/Lasmas-Institut du Longitudinal. L'analyse longitudinale du marché du travail. Rennes, 23 et 24 mai 1996*. Marseille: Céreq, Document Céreq No. 115.

Deville, Jean-Claude. 1982. « Analyse des données chronologiques qualitatives, comment analyser les calendriers ? ». *Annales de l'I.N.S.E.E.* 45:45-104.

Deville, Jean-Claude and Gilbert Saporta. 1980. « Analyse harmonique qualitative », p. 375-389 in Edwin Diday. *Data Analysis and Informatics. Proceedings of the International Symposium on Data Analysis and Informatics*. Amsterdam, New York: North-Holland.

Durbin, Richard, Sean R. Eddy, Anders Krogh and Graeme Mitchison. 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.

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

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

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

Halpin, Brendan. 2008. « Optimal Matching Analysis and Life Course Data: The Importance of Duration ». *Department of Sociology Working Paper Series*: University of Limerick.

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

Hamermesh, Daniel S. 1999. « The Timing of Work Over Time ». *The Economic Journal* 109:37-66.

—. 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.

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

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

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

Le Roux, Brigitte and Henri Rouanet. 2004. *Geometric Data Analysis. From Correspondence Analysis to Structured Data Analysis*. Dordrecht: Kluwer Academic Publishers.

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

Lesnard, Laurent and Thibaut de Saint Pol. 2009. « Patterns of Workweek Schedules in France ». *Social Indicators Research* 93:171-176.

Levenshtein, Vladimir I. 1966 [1965]. « Binary codes capable of correcting deletions, insertions, and reversals ». *Soviet Physics Doklady* 10:707-710.

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

Presser, Harriet B. 2003. *Working in a 24/7 Economy: Challenges for American Families*. New York: Russel Sage Foundation.

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 », p. 33-62 in F. Thomas Juster and Frank P. Stafford. *Time, Goods, and Well-Being*. Ann Arbor: University of Michigan Press.

Rohwer, Götz and Ulrich Pötter. 2005. « TDA User's Manual », p. 468-470: Ruhr-Universität Bochum.

Saint Pol, Thibaut de. 2006. « Le dîner des Français : un synchronisme alimentaire qui se maintient ». *Économie et Statistique* 400:45-69.

Simiand, François. 1922. *Statistique et expérience. Remarques de méthode*. Paris: Éditions Marcel Rivière.

Stovel, Katherine. 2001. « Local Sequential Patterns: The Structure of Lynching in the Deep South, 1882-1930 ». *Social Forces* 79:843-880.

Tuma, Nancy Brandon, Michael T. Hannan and Lyle P. Groeneveld. 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.