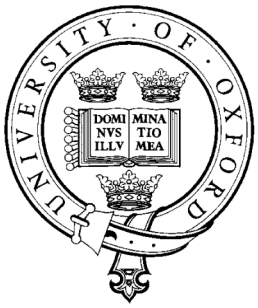
# Two-Stage Optimal Matching Analysis of Workdays and Workweeks

Laurent Lesnard, Man Yee Kan




*Sociology Working Papers*
*Paper Number 2009-04*

# Two-Stage Optimal Matching Analysis of Workdays and Workweeks

Laurent Lesnard
Man Yee Kan

***Department of Sociology***

*University of Oxford*
*Manor Road*
*Oxford OX1 3UQ*
*www.sociology.ox.ac.uk/swp.html*

# Two-Stage Optimal Matching Analysis of Workdays and Workweeks

**Laurent Lesnard**[i]

**Man Yee Kan**[ii]

[i]Sciences Po, France

[ii]Department of Sociology
University of Oxford

**Addresses for correspondence:**

Laurent Lesnard, Sciences Po, Observatoire sociologique du changement, 27 rue Saint Guillaume 75007, Paris, France. Email: laurent.lesnard@sciences-po.fr

Man Yee Kan, Department of Sociology, University of Oxford, Manor Road, Oxford, OX1 3UQ, UK. Email: man-yee.kan@sociology.ox.ac.uk

This project was supported by a Postdoctoral Fellowship from the British Academy and a British Academy/CNRS Joint Project Grant. Earlier versions of this paper were presented at the IATUR 2008 conference, the ISA Research Committee 33 2008 conference, and seminars in Oxford and Surrey. We are grateful to the participants for their helpful comments.

# Two-Stage Optimal Matching Analysis of Workdays and Workweeks

**Summary.** We apply Optimal Matching (OM) at two stages for the analysis of workdays and workweeks using data from the UK 2000 Time Use Survey. We only employ substitutions but no insertion or deletion when calculating the distance matrix between sequences. The costs are defined according to the transitional frequencies of events at a given time. Our study demonstrates how OM can be adapted to the number of periodicities and theoretical concerns of the topic by adjusting its costs and parameters. There are 7 main types of workweeks in the UK and standard workweeks account for only 1 in 4 workweeks.

*Keywords*: Cost; Optimal Matching; Time Use; Two-Stage Optimal Matching; Work Schedule; Work Time

## 1. Introduction

Work time is an important dimension of quality of life and social stratification. Previous studies showed that the average work time has been on the decline whereas that on leisure has increased since the 1960s in industrialised countries (Dumazedier 1967, Robinson and Godbey 1999, Gershuny 2000). Yet recent research has shown that the decrease in work time has been the most prevalent among low-earning and low-status workers but has increased in the case of higher-grade professionals and managers. The two opposite trends in work time have progressively reversed the social class-work time gradient so that at the beginning of the 2000s, high skilled workers have longer work hours than unskilled ones (Gershuny 2000). Relatively few studies, however, have focused on the scheduling of work time, although it is obvious that not only the total amount of work time but also its scheduling are very significant to one's social and family life. In this article, we introduce a statistical technique that is particularly useful for analysing the scheduling of time use. We call this technique "two-stage optimal matching" (2SOM). We employ it for the analysis of 7-day diary data from the UK 2000 Time Use Survey to define a typology of workweeks.

A handful of studies have explored the scheduling of everyday activities using time diary data. Time diary data are usually collected by respondents' records of their activities at every 10 minute or 15 minute time slots. Unlike questionnaire stylised data, these data not only provide information on the amount of time spent on daily activities but also the *scheduling* of these activities. However, before the advent of Optimal Matching Analysis (OM), these rich data, especially their scheduling dimension, were not thoroughly analysed. For example, the first graphical representations of the timing of daily activities can be found in Szalai (1972). Wilson (1998) first applied OM to time diary data to explore the timing of daily activities. Lesnard (2004) introduced an advanced version of OM, Dynamic Hamming Matching, DHM, which is particularly adapted to the analysis of time use data. Using this sequence analysis variant, researchers have identified a variety of workdays in France (Lesnard 2006a) and in Belgium (Glorieux et al. 2008). The workdays include "shifted" (morning, evening or night shifts), "fragmented" (two short work spells

where there is a long break between them), "short" (a short spell of work episode), standard (9-to-5 work hours), and "long" (a long spell of work episode). In France, standard workdays increasingly gave way to non-standard work schedules between the 1980s and the 1990s. Lesnard (2008) suggests that the expansion of non standard workdays is one of the key factors to explain the increase in desynchronization of work time for dual-earner couples.

Most studies on the scheduling of work have been confined to one day diary data despite the fact that work is very likely organized according to longer time frames - one week to the least. Average workweek time is regularly measured and reported in official labour statistics. The number of weekly work hours is also a topic of frequent debates and negotiations among policy makers, trade unions and academics. Nevertheless, relatively little attention has been given to scheduling of those work hours over the week. Admittedly, the analysis of workweek schedules has been restricted by the lack of suitable data (most time use surveys only collected day-long rather than week-long diaries). Furthermore, analysing the scheduling of work over the week is methodologically more challenging than focusing on workdays because it requires taking into account both of the scheduling of work hours within the day and that of workdays over the week.

Following the recent Eurostat guidelines on collecting time use data, some recent national time use surveys (e.g. in UK 2000, France 1999, Belgium 1999 and Finland 2000) have collected 7-day working time data using the "workweek grid" method (Robinson et al. 2002). The week-long time use data provide great opportunities for researchers to investigate patterns of work schedules, but so far very few studies have fully exploited the strength of these using advanced statistical techniques. For example, when examining individuals' workweeks and the synchronicity of work time of dual earner couples, Chenu and Robinson (2002) adopted a series of indicators and numeric indexes such as the length of workweek and the amount of work during weekends to analyse individual workweeks rather than a systematic tool of estimating the distance among different types of workweeks and that between partners' work time.

## 2. Optimal Matching and cost setting

Methods for describing sequential data (e.g. data concerning lifecycle events, and career trajectories) have been available to social scientists for more than three decades. Among these methods, the most popular one is certainly Optimal Matching Analysis (OM), introduced to the social sciences by Andrew Abbott and his colleagues in the 1980s[1].
OM is basically a distance measure adapted to sequence data in which dissimilarity between two sequences is given by the minimal total cost to match them (Kruskal 1983, Durbin et al. 1998). Three transformations are allowed to transform one sequence into the other one: insertion, deletion, and substitution. The total cost of any matching is the sum of the weighted number of transformations required. The lowest matching cost is used as the measure of the dissimilarity between two sequences. As a result, the kind and the number of transformations used depend on the relative cost of

[1] Abbott and Forrest (1986) and Abbott and Hrycak (1990). For early OM applications, see Abbott and Tsay (1990).

insertion-deletion and substitution. Insertion and deletion, commonly called indel, are completely symmetrical in OM and therefore are given the same cost.

Despite OM has been used for more than three decades, users are often uncertain about how to set costs and how this might affect results (Wu 2000). As Stovel et al (1996) put it, "The assignment of transformation costs haunts all optimal matching analyses". Statistical software in the early 1980s were not yet well adapted for analysing sequence data, and empirical data analyses based on them usually took hours or even longer to complete. Hence it was difficult to test how different values of cost might lead to different results. Considerable progress has been made on this issue with the developments in relevant statistical packages and programs. Furthermore, social scientists have increasingly been employing, experimenting with, and reflecting on optimal matching.

In social science research, sequence data are usually concerned with events and time, which determine how the two kinds of operations (i.e. indel and substitutions) are used in OM. Indels warp time in order to match identically coded but remote events while substitutions focus on the comparison of contemporaneous but distinct events (Lesnard and Saint Pol 2006). The succession of indels and substitutions for matching two sequences can be seen as a series of acceleration and deceleration to match identically coded events, as well as a couple of replacements of states by one another. The ratio of indel cost to substitution one determines whether it is preferable to simplify time or events when comparing pairs of sequences. In fact, what is now commonly known as Optimal Matching was originally a refinement suggested by Vladimir Levenshtein (1966) to improve the similarity measure introduced by Richard Hamming (1950), who measured similarity by the number of identical contemporaneous tokens (see Table 1).

**Table 1 – The three historical Optimal Matching distances**

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

When no substitution operations are used, whether because they are not allowed as in the Levenshtein II distance or because their cost is greater than twice the indel one,[2] then OM is equivalent to finding the longest common subsequence (Kruskal 1983)[3], Time warps, string edits, and macromolecules: the theory and practice of sequence comparison, 1-44}. When no indel operations are used (because they are not allowed by definition, e.g. Hamming distance or their cost is greater than the substitution one),

[2] If the cost for substituting two events is higher than that for one insertion and one deletion, then substitutions are never used (Kruskal 1983).

[3] The dissimilarity measure suggested by Elzinga (2003) does not belong to the Optimal Matching family. However, if it had to be located on this scale, it would be on the far right of it.

OM amounts to counting the number of dissimilar contemporaneous events[4]. As a result, the ratio of indel cost to substitution one determines the kind of similarity that OM will be most sensitive to (see Figure 1). If the timing of events is not too important, then users should favor costs close to Levenshtein II. On the contrary, high indel costs should be used when timing is important for the analysis. Users should set the same cost for indel and substitution if they want to use both kinds of operations in a more or less balanced way (Lesnard 2009a).

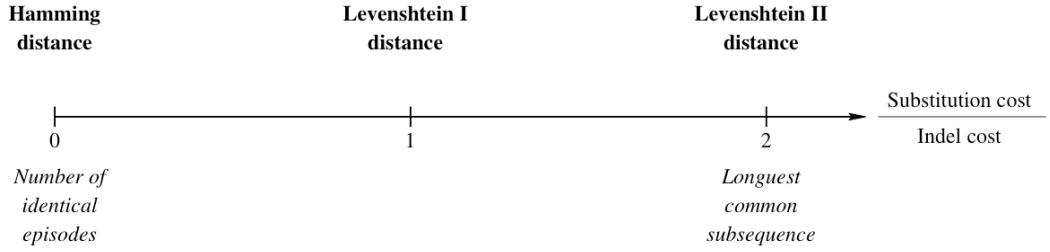



**Figure 1 - Patterns corresponding to costs in OM**

Unlike the three historical OM variants described in Table 1, it is common, especially in social research, to define a substitution cost for each pair of states[5]. This approach makes it possible to have some substitution costs higher as well as some lower than the indel cost. When it is higher, the two different events will not be substituted but the algorithm will try to shift the two sequences so as to find identical but shifted subsequences. When it is lower, the two different events will be substituted and be kept locally in synchronization. However it should be noted that defining a substitution matrix amounts to deciding a priori that some pairs of states are closer to one another than others. This a priori knowledge can be informed by theory (for instance, see Halpin and Chan 1998), hypotheses (for an illustrative example, see Stovel et al. 1996), previous findings or even the sequences themselves[6].

When adapting OM to various kinds of data and research questions, the flexibility offered by substitution costs is even greater than in the case of defining a pairwise substitution matrix. For example, Lesnard (2004, 2009a) recommends a variant of Hamming matching called Dynamic Hamming Matching, in which only substitution operations with time-varying costs inversely proportional to the series of transition matrices are used, to be applied to the case where sequences are of equal length and that preserving the timing of event is crucial. Indel costs are usually set relatively to substitution ones, but a notable exception is the OMv dissimilarity measure (introduced by Halpin 2008) in which the indel cost is inversely proportional to the spell length so as to take the duration of an event spell into account.

The versatility of OM is also illustrated when it is applied to multiple domain sequences, i.e. sequences unfolding in several distinct, but theoretically interdependent, domains such as family, employment and housing careers (Pollock 2007). Although it is possible to conduct sequences analyses for each of the domains,

[4] When indel operations are allowed but are penalized by a very high cost, they are used only when sequences are not of equal length. In such cases, the resulting dissimilarity is likely to be defined above all by the length of sequences.

[5] A recent exception is Brzinsky-Fay (2007), who set indel cost to 1 and substitution cost to 2.

[6] Frequencies of transitions between states have often been used as an inverse measure of distance between states (e.g. Abbott and Hrycak 1990).

a simpler way is to define substitution matrices for each and then combine them (Stovel et al. 1996, Pollock 2007), a method which is called Multiple Sequence Analysis (MSA). In Gary Pollock's (2007) study, the substitution cost concerning employment statuses (e.g. employed or self-employed) as w*ell as* housing tenure statuses (e.g. owning with a mortgage or owning outright) is considered to be equal to the sum of the substitution costs concerning respectively employment statuses and housing tenure statuses. Hence the multiple domain issue can be dealt with by multiple OM analyses or by combining multiple substitution matrices for a single OM analysis[7].

As suggested by Andrew Abbott (2000), the hypotheses on which OM rests are not about how data are generated[8] but on the kinds of patterns that users expect to see prior to the analysis. In standard OM, the ratio of indel cost to substitution one is a kind of slider with which users can choose among infinite combinations of patterns, ranging from the number of similar contemporaneous events to the longest common subsequence. The importance attached to the timing of sequences decreases as the cursor moves along. Focusing the analysis on a certain type of pattern does not imply, as unfairly criticized by Levine (2000) and Wu (2000), that OM will create ex nihilo this particular pattern. It just implies that the pattern being looked for will be easier to be identified from the data if it exists. Using multiple substitution costs enables researchers to tune OM finely for particularly nature of data and research questions. But, as illustrated by Multiple Sequence Analysis, the flexibility offered by OM is actually greater than that by setting a matrix of pairwise substitution costs. In what follows, we use workweeks as an example to demonstrate how OM can be adapted to cope with long sequences and multiple periodicities.

## 3. Analysing workdays and workweeks

The data come from the UK Time Use Survey (UKTUS) 2000-01, carried out by the Office for National Statistics from June 2000 to September 2001 (Ipsos-RSL and Statistics 2000). The sample is nationally representative of about 6,400 households in the UK. The response rate was 60%. All individuals aged 8 or above in the households were requested to fill in individual questionnaires and diaries. In addition to the traditional 2 single day diaries (one in the week, one in the weekend), the UKTUS also collected 7-day workweek grid diaries. In each day of the workweek grid diaries, the time is divided into 96 15-minute slots. Respondents were requested to indicate their work or study episodes by drawing a line across the start and the end of each of their work or study episode. They were also instructed to exclude travelling time and meal breaks in their work or study time.

The design of the diaries, however, does not enable users to distinguish between work and study spells. In order to build a typology of workweeks for the present study, only respondents in employment (part-time or full-time), as reported in the household questionnaires, were selected. Of the 9,823 respondents who filled in the week grid diaries, 4,944 were in employment and recorded work time on at least one of the 7

[7] In the case of multiple OM analyses, domains are assumed to be independent (correlations may only be identified from the results). When combining multiple substitution matrices for a single OM analysis, domains are assumed to be interdependent.

[8] That is, the aim of the analysis is not to provide a model of underlying processes, but rather to describe the data.

days. In the workweeks, there are 21,122 workdays in which respondents recorded at least one work episode[9].

Investigating workweek patterns involves analysing two nested periodicities: days within weeks and hours within days. At the level of the day, the focus should be placed on the scheduling of work hours. At the level of the week, it is rather which days are scheduled for work that is of interest. Although it is possible to apply OM directly to the 672 15-minute time slots of the workweek grids, it will be more appropriate to take account of these two nested periodicities in the analysis as workers are likely to schedule their work time at two stages in real life. The issue of intra-day work time variations over the week is similar to that of seasonality in time-series analysis. As the main goal of time-series analysis is to model trends (e.g. trends of the unemployment rate), seasonality is often considered not directly relevant and therefore only controlled by modelling it separately. In the case of scheduling of work hours within a day, intra-day variation is a much more important issue than the case of seasonality in time-series analysis, so it should be analysed separately.

Our approach is to apply OM in two steps, a method we call two-stage optimal matching (2SOM). At the first stage, OM is applied to the 96 15-minute time slots to define typologies of workdays. The sample is consisted of day-long diaries recording at least 15 minutes of work time (there are repeated records from respondents who had more than one work day in the week). 21,122 workdays are derived from the original 4,944 workweeks. These sequences are made of two states: work and non work. Cluster analysis is applied to the resulting dissimilarity matrix to produce a typology of workdays. At the second stage, OM is employed to analyse 7-day weeks, which were made of the types of workdays identified in the first stage. The states include the types of workdays (e.g. standard, long and so on) and "rest", a category to take into account the days with no work at all.

As mentioned in the previous section, the costs at the two stages should be set according to the importance of timing for the analysis. Focussing on timing is certainly crucial for the first stage, which is concerned with the scheduling of work time during the day. The parameters used should be those on the Hamming distance pole in Figure 1. We will use a variant of Hamming distance, Dynamic Hamming Matching, DHM (Lesnard 2004), which has been applied to the analysis of work schedules in recent studies (Glorieux et al. 2008, Lesnard 2006a, Lesnard 2006b, Lesnard 2008). In DHM, only substitution operations are used. In order to make the costs sensitive to the timing of sequences, their values are defined to be varying with time and inversely proportional to transition frequencies between pairs of states at a particular time. The rationale for DHM is that transition frequencies between states reveal their relative distances at a given t. A high frequency of transitions between any two states at t indicates that many individuals switch states at the time and therefore the likelihood that these two states belong to the same type of trajectory at that time is high. As a result, the distance between these two states is considered to be short. On the contrary, a low frequency of transitions suggests that the two states belong to two different types of trajectory and hence their distance is considered to be long at t. DHM fits well with the requirements of the first stage of the analysis. At the second

[9] As a consequence, full-time students, some of whom might have a part-time job, were not included in the analysis. This is likely to decrease the number of part-time workweeks.

stage, timing is crucial too. For example, working on an evening shift is likely to have different implications for social life on different days of the week (e.g. Saturday vs. Monday). Thus DHM is also an appropriate parameterization for the second stage. We hence analyse workweeks with two-stage Dynamic Hamming matching. However, it should be noted that different OM parameterizations could have been used at the different stages of the analyses. Figure 2 summarizes the analyses we have conducted.

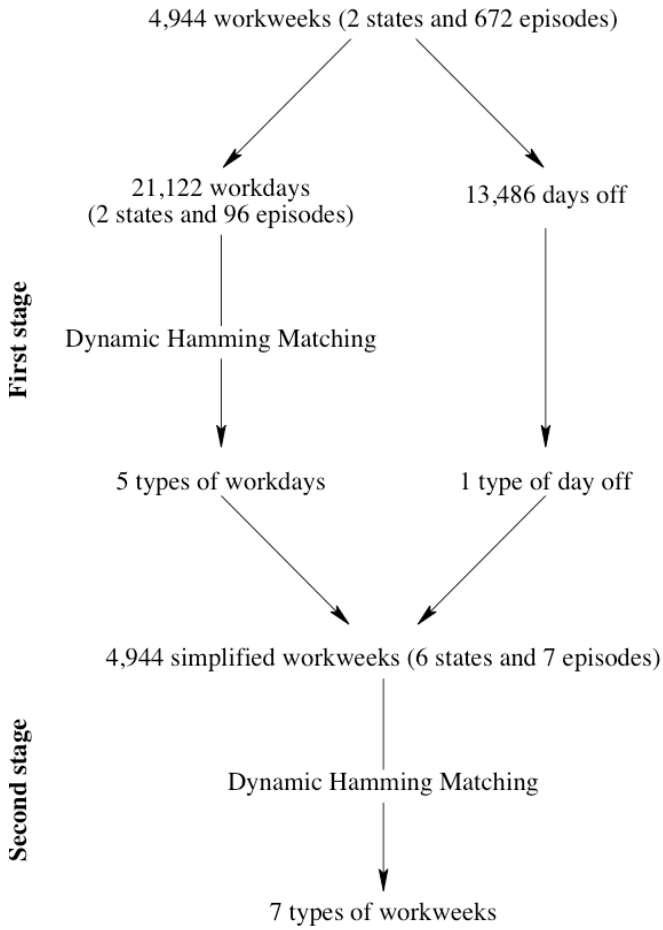



**Figure 2 - Synoptic representation of Two-Stage Dynamic Hamming Matching applied to workweeks.**

At each of the two stages of the analysis, we employ the method beta-flexible (Belbin et al. 1992, Milligan 1989) to calculate the distance between groups (as opposed to the original elements that is called linkage in cluster analysis). Beta-flexible has proved more robust to recover structure in presence of outliers and noise than Ward's or other classical linkages (Milligan 1980, Milligan 1981). We conduct the sequence analysis by using the seqcomp plugin in Stata[10]. We used SAS for the cluster analysis because the beta-flexible linkage is not implemented in Stata[11].

Results

[10] This plugin is available free of charge at http://laurent.lesnard.free.fr.

[11] Now it is possible to conduct all the analyses with the TraMineR library in *R* (Gabadinho et al. 2008), which implements Dynamic Hamming Matching since the version 1.4 released on August 6, 2009.

## 4.1. First stage

At the first stage of analysis, we focused on the 34, 608 days in the 4,944 weeks from the sample[12]. We applied Dynamic Hamming Matching (DHM) to the 21,122 days with at least one work spell (61% of the days). Nevertheless the sample was too big for the 4 Gb memory limit imposed by 32-bit systems of our computers. We thus split these 21,122 days into two sub-samples and conducted two analyses separately[13]. Agglomerative hierarchical clustering (linkage: beta-flexible with ß = - 0.3) is applied to the two distance matrices produced by DHM. There are no definitive criteria to determine the number of clusters, but the "elbow criterion" usually gives interesting starting points[14]. In both samples, the first significant spike in the intergroup distance occurs for the seven-group partition, suggesting that very dissimilar groups have just been joined and hence there are at least eight types of workdays in the data. Another smaller spike is observed in the 9-cluster solution of the second sample. We examined and compared visual representations of the clusters and the summary statistics (including average total worktime, medians of the start, the middle and the end time of the workday) of the different partitions between 11 and 8 groups.

[12] 1,078 out of the 34,608 days (3.11%) have missing values. Visual inspection reveals that these missing values appear to be coding errors caused by a single data coder, who coded "missing" instead of "zeroes" for work spells during work days and for all values during non-work days. Once these false missing values were replaced with zeroes, no further missing value is found.

[13] Clustering a 21,122 x 21,122 matrix with SAS 9.1 on a MS Windows 32-bit computer was not possible. The sample was randomized before being split.

[14] An elbow, or a spike, in the intergroup distance indicates that the two very dissimilar clusters have been merged. In such a case, the cluster solution just before this merging should be considered rather than the one just after.

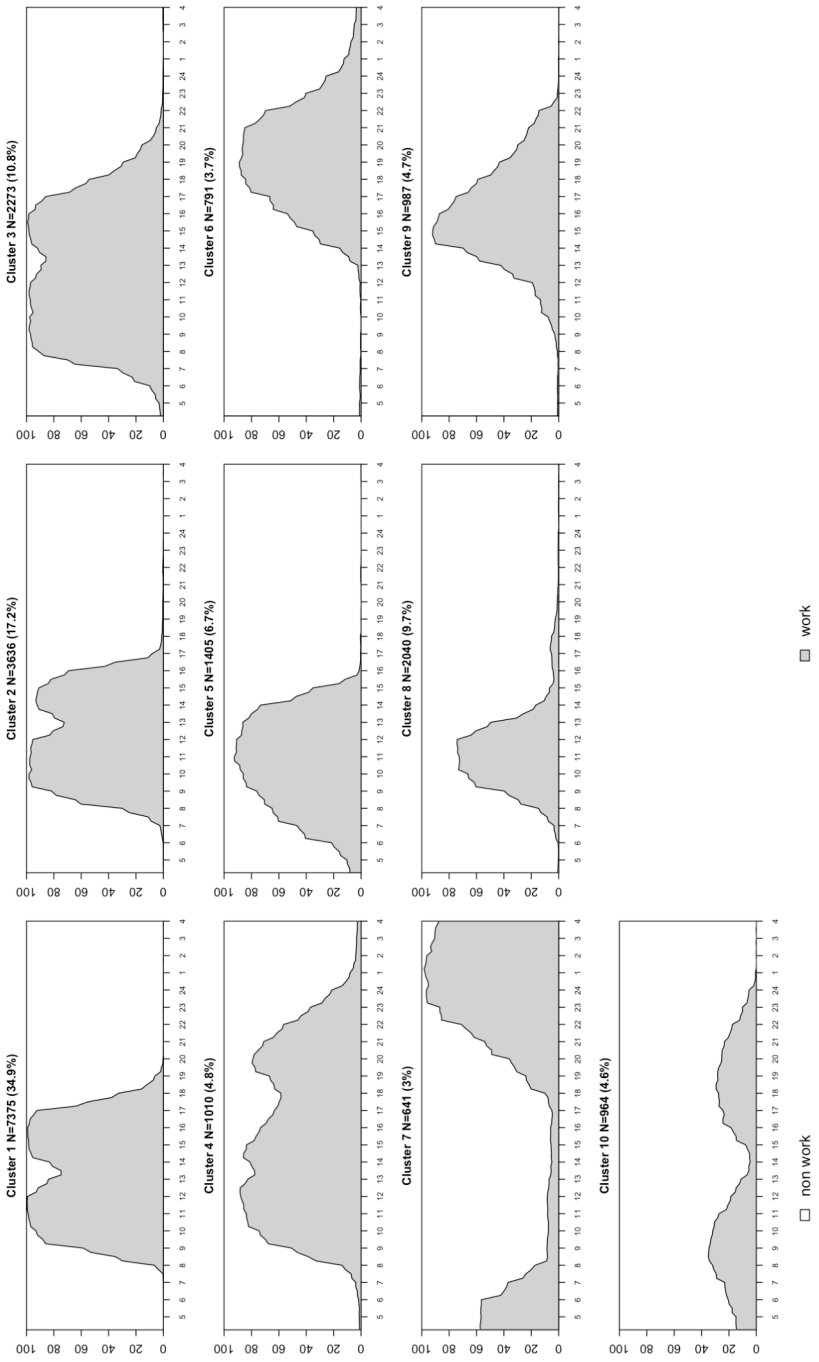



**Figure 3 - Typology of workdays**

Finally, we adopted the 10-cluster solution as it is the most succinct one that incorporates all major types of part-time workdays, which are important characteristics of the UK workdays, and all the other major categories. We moved on to match identical clusters in the typologies from the two samples based on tempograms[15] and summary statistics. This step has been straightforward as the different types of workdays identified from the two randomized sub-samples are highly similar to each other, as well as to typologies found from previous studies

[15] A tempogram is a graphical representation of the state distribution for each time-slot.

(Glorieux et al. 2008, Lesnard 2006a, Lesnard 2009b). The final typology is summarized in Table 2 and represented graphically in Figure 3.

**Table 2 – Types of workdays**

| | Name | | Size (%) | |
|---|---|---|---|---|
| 1 | Standard | 9 to 5 | 34.92 | 52.13 |
| | | 8 to 4 | 17.21 | |
| 2 | Long | long | 10.76 | 15.54 |
| | | long day and evening | 4.78 | |
| 3 | Shift | morning shift | 6.65 | 13.42 |
| | | evening shift | 3.74 | |
| | | night shift | 3.03 | |
| 4 | Part-time | part-time morning | 9.66 | 14.33 |
| | | part-time afternoon | 4.67 | |
| 5 | Short | short atypical | 4.56 | 4.56 |

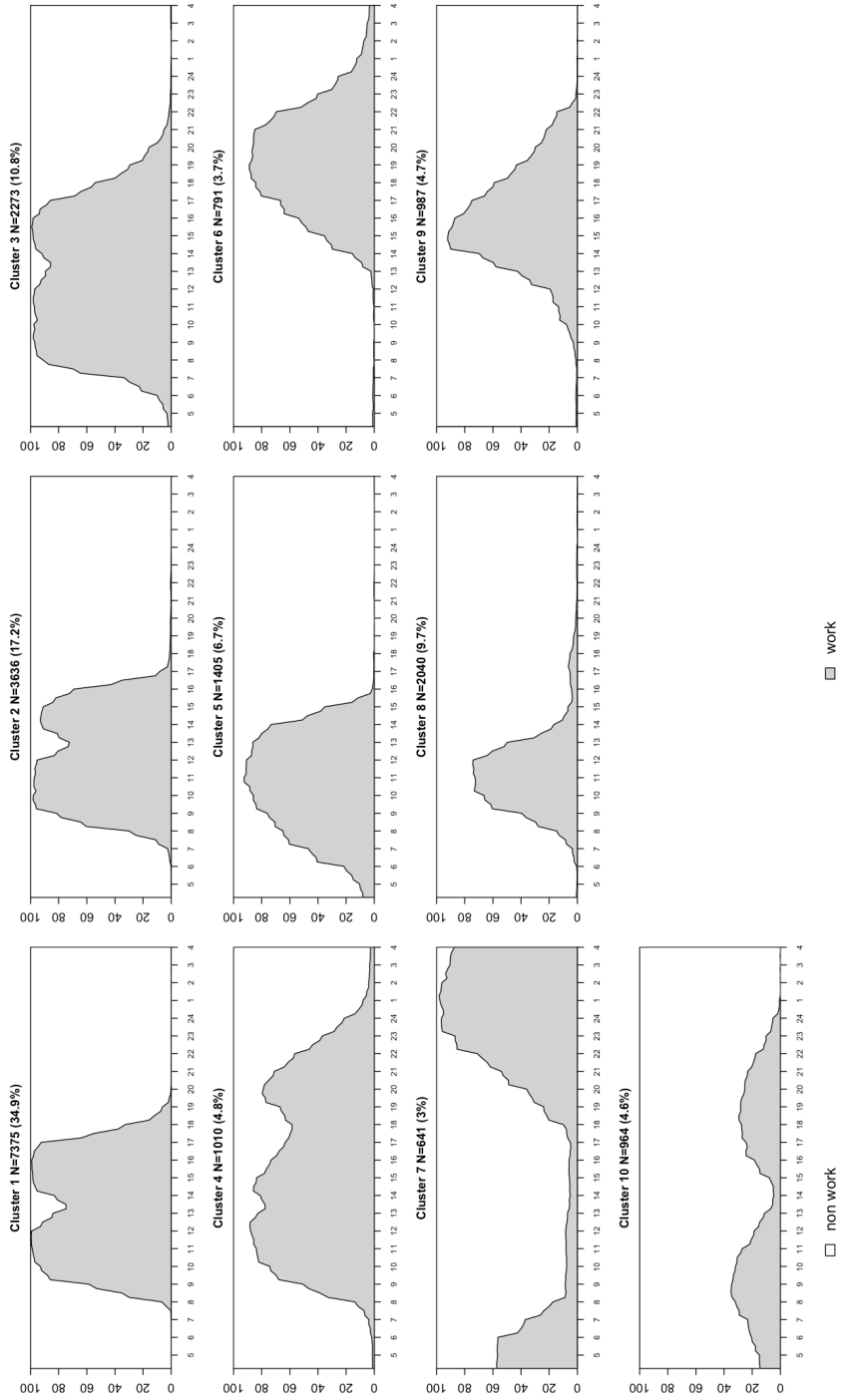



**Figure 3 – Tempograms of the typology of workdays**

The first two types are standard workdays (Type 1, Clusters 1 and 2). The first one, which is also the most common type of workday in the UK, is the traditional 9-to-5 workday. Workdays of this group start at around 9 am and end at around 5 pm. The second type of standard workdays is a variant of the 9-to-5 one, but the starting time and the end time are both one hour earlier - we hence call it the 8-to-4 workday. Although being the most common type of work schedules, standard workdays (9-to-5 and 8-to-4) account for just over half of the workdays in the UK (52.1%). Other types of workdays deviate from the standard workdays in two main ways: length and schedule. Type 2 of workdays, long workdays (Clusters 3 and 4), have distinctly long

total work time (over 10 hours). There are minor differences between them: the former is a longer version of the 9-to-5 workdays and the latter is characterized by evening work in the workplace or at home. Long workdays constitute 15.5% of total workdays. There are three groups of shorter workdays. Two of them are part-time workdays (Type 4, Clusters 8 and 9), which make up 14.3% of the total workdays. They have the major work spell in the morning or in the evening. The third one, s*hort* workdays (Type 5, Cluster 10) contains very short total work time and is characterized by multiple, short, and staggered work spells. Short workdays represent 4.6% of total workdays. The final type, shift workdays (Type 3, Clusters 5, 6 and 7), depart from the standard ones in their schedules. There are three types of shifts: morning, evening, and night, which add up to 13.4% of the total workdays. The total work time of these shift workdays is more or less the same as the standard ones. However, most of the work on these days is carried out before 9 am or after 5 pm. Interestingly, morning shift is the most common among the three types of shift workdays, whereas night shift is the least common form.

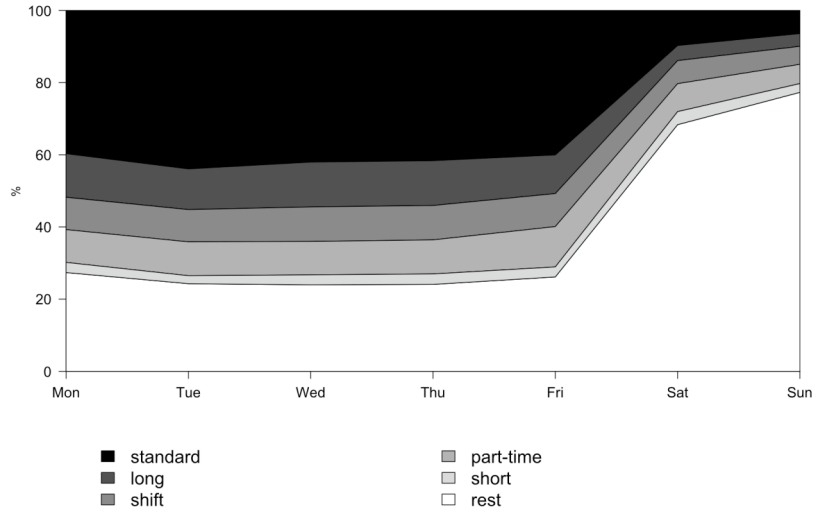



**Figure 4 – Simplified Workweeks**

To build simplified workweeks, we employ a 5-category typology as described in the above paragraph and in Table 2[16]. Furthermore, we add the category rest days to take account of the days that contain no work spell. Therefore every week contains 7 episodes (Monday to Sunday) and 6 states (5 types of workdays and 1 type of rest days). The visual representation of these simplified workweeks is given in Figure 4. The proportion of each of the 5 types of workdays remain more or less stable during weekdays. Standard workdays make up about 60% of the weekdays. As expected, the results are very different during weekends. Work is uncommon on Saturdays and Sundays. It is worth mentioning that the proportion of non standard workdays (shift, part-time and short) is much higher on weekends than on weekdays. That is, weekend work is atypical, and the types of workdays on weekends are more likely to be atypical as well.

[16] Previous studies on French workweeks also supported the simplified groupings for different types of standard workdays, shift workdays and part-time workdays respectively. The simplification can be justified by the proximity in workers' backgrounds and characteristics among the subgroups (For more detail, see Lesnard 2006a, Lesnard 2009b).

## 4.2. Second stage

To build a typology of workweeks, we run DHM on these simplified workweeks. In Figure 5, the intergroup distances for the series of nested partitions indicate that there are at least five types of workweeks. Two other, smaller, spikes occur at the 10-cluster and 18-cluster solutions. The results suggest that we should examine the series of partitions ranging from 18 to 5.

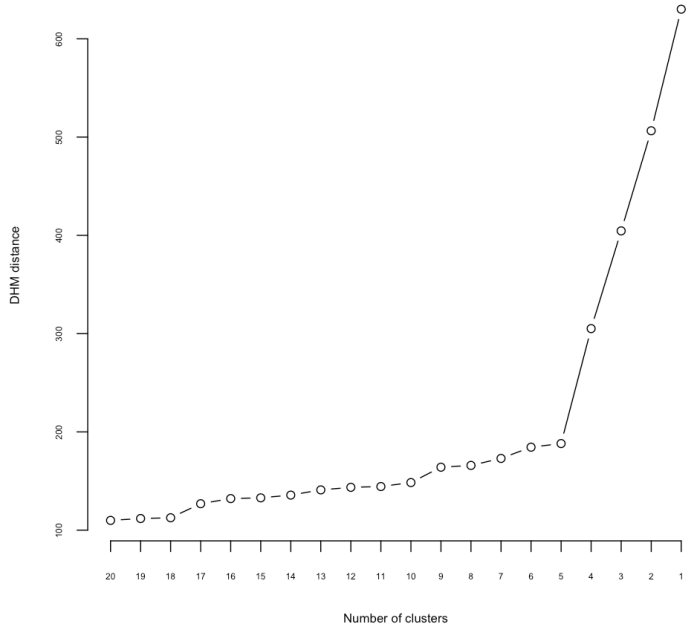



**Figure 5 - Cluster solutions and DHM distance (beta-flexible linkage)**

We first tried to reduce the number of groups starting from the 18-group solution by the algorithm results. However, some of the groupings suggested did not seem appropriate[17]. We therefore decided to reduce the number of groups manually based on descriptive statistics and visual representations of the cluster solutions. Cluster analysis is basically an algorithm used to produce nested sets of clusters. As an algorithmic method, it is based on the repetition of a finite sequence of simple and intuitive instructions. In the beginning, each element is a separate cluster and on each step the two "closer" clusters are merged[18]. This sequence of instructions could be done manually, though it is much more efficient to run it by computers. However, efficiency comes at the expense of rigid rules which may not be entirely satisfactory for the last steps of the grouping. For the last few steps of the agglomerative clustering, successive groupings suggested by the algorithm are not theoretically or empirically satisfactory. We therefore conduct the grouping by linking clusters that are similar in theoretically important characteristics (e.g. number of workdays, proportion of work done on weekends, and total work hours). To adopt results

[17] For instance, the first grouping suggested by the algorithm is to combine clusters 5 and 16.

[18] At first, each element forms a separate cluster. The original dissimilarity matrix determines the distances among the clusters at this stage. However, when two elements are combined to form a new cluster, the original matrix cannot be referenced to directly because it does not provide information about the distance between the newly formed cluster and the rest of the original elements. At this stage, we conduct the linkage manually.

different from those suggested by the algorithm, however, one has to provide convincing arguments and justifications.

**Table 3 - Descriptive statistics of the eighteen-cluster solution produced by beta-flexible clustering**

| Original cluster id | New cluster id | Size | Work time | Number of days off | Number of work days | Proportion of work on Sat. | Proportion of work on Sun. | Proportion of work on weekend | Proportion of full Saturdays off | Proportion of full Sundays off |
|---|---|---|---|---|---|---|---|---|---|---|
| 2 | 1 | 26.10 | 42.24 | 1.76 | 5.24 | 4.97 | 3.19 | 4.08 | 80.43 | 86.90 |
| 8 | 2 | 4.39 | 49.18 | 1.64 | 5.36 | 10.37 | 8.23 | 9.30 | 59.49 | 72.31 |
| 10 | 2 | 3.62 | 60.53 | 1.21 | 5.79 | 40.81 | 30.21 | 35.51 | 6.83 | 25.47 |
| 13 | 2 | 5.40 | 46.37 | 1.95 | 5.05 | 7.63 | 5.13 | 6.38 | 72.50 | 80.00 |
| 14 | 2 | 4.57 | 57.20 | 1.59 | 5.41 | 10.00 | 6.06 | 8.03 | 62.56 | 81.28 |
| 16 | 2 | 2.00 | 46.84 | 1.55 | 5.45 | 8.82 | 5.27 | 7.05 | 56.18 | 74.16 |
| 6 | 3 | 4.03 | 31.78 | 2.23 | 4.77 | 13.86 | 9.60 | 11.73 | 53.63 | 65.92 |
| 7 | 3 | 5.67 | 42.94 | 1.51 | 5.49 | 15.89 | 14.87 | 15.38 | 45.63 | 51.98 |
| 15 | 3 | 2.41 | 35.08 | 2.11 | 4.89 | 3.32 | 4.53 | 3.92 | 88.79 | 84.11 |
| 9 | 4 | 4.90 | 35.14 | 1.61 | 5.39 | 16.26 | 13.10 | 14.68 | 38.99 | 51.38 |
| 12 | 4 | 2.05 | 23.34 | 1.98 | 5.02 | 6.81 | 6.42 | 6.62 | 57.14 | 73.63 |
| 11 | 5 | 6.86 | 20.39 | 2.30 | 4.70 | 5.93 | 3.87 | 4.90 | 73.77 | 81.31 |
| 17 | 5 | 2.27 | 23.12 | 3.03 | 3.97 | 5.13 | 3.19 | 4.16 | 75.25 | 88.12 |
| 4 | 6 | 5.83 | 36.76 | 2.32 | 4.68 | 12.56 | 6.99 | 9.78 | 59.07 | 77.61 |
| 5 | 6 | 4.54 | 31.15 | 3.08 | 3.92 | 10.10 | 3.44 | 6.77 | 66.83 | 87.13 |
| 1 | 7 | 7.24 | 13.06 | 5.11 | 1.89 | 9.15 | 7.08 | 8.12 | 67.70 | 74.53 |
| 3 | 7 | 6.41 | 30.89 | 2.85 | 4.15 | 11.40 | 8.39 | 9.89 | 63.16 | 70.53 |
| 18 | 7 | 1.71 | 23.08 | 3.78 | 3.22 | 14.39 | 15.75 | 15.07 | 57.89 | 52.63 |

Figure 6 shows the eighteen-cluster solution. Table 3 displays descriptive statistics on the eighteen clusters and how we have reorganised them. Cluster 2 represents the standard 9-to-5 Monday-to-Friday workweek (standard workweek). We combine Clusters 8, 10, 13, 14, and 16 into a single category, long workweek, because they are all characterized by long work hours over the week (the average workweek time is longer than 45 hours, and in three of them it is over 48 hours, i.e., the maximum limit suggested by the European Working Time Directives). We then group Clusters 6, 7, and 15 together because they all consist of shift hours on workdays (shift workweeks). Clusters 9 and 12 form another category, alternate workweeks, in the new typology. Unlike other types of workweeks, they are not composed of a uniform type of workdays. Instead the types of workdays vary over the week (e.g. standard hours on Monday, shift hours on Tuesday, part-time work on Wednesday and so on).

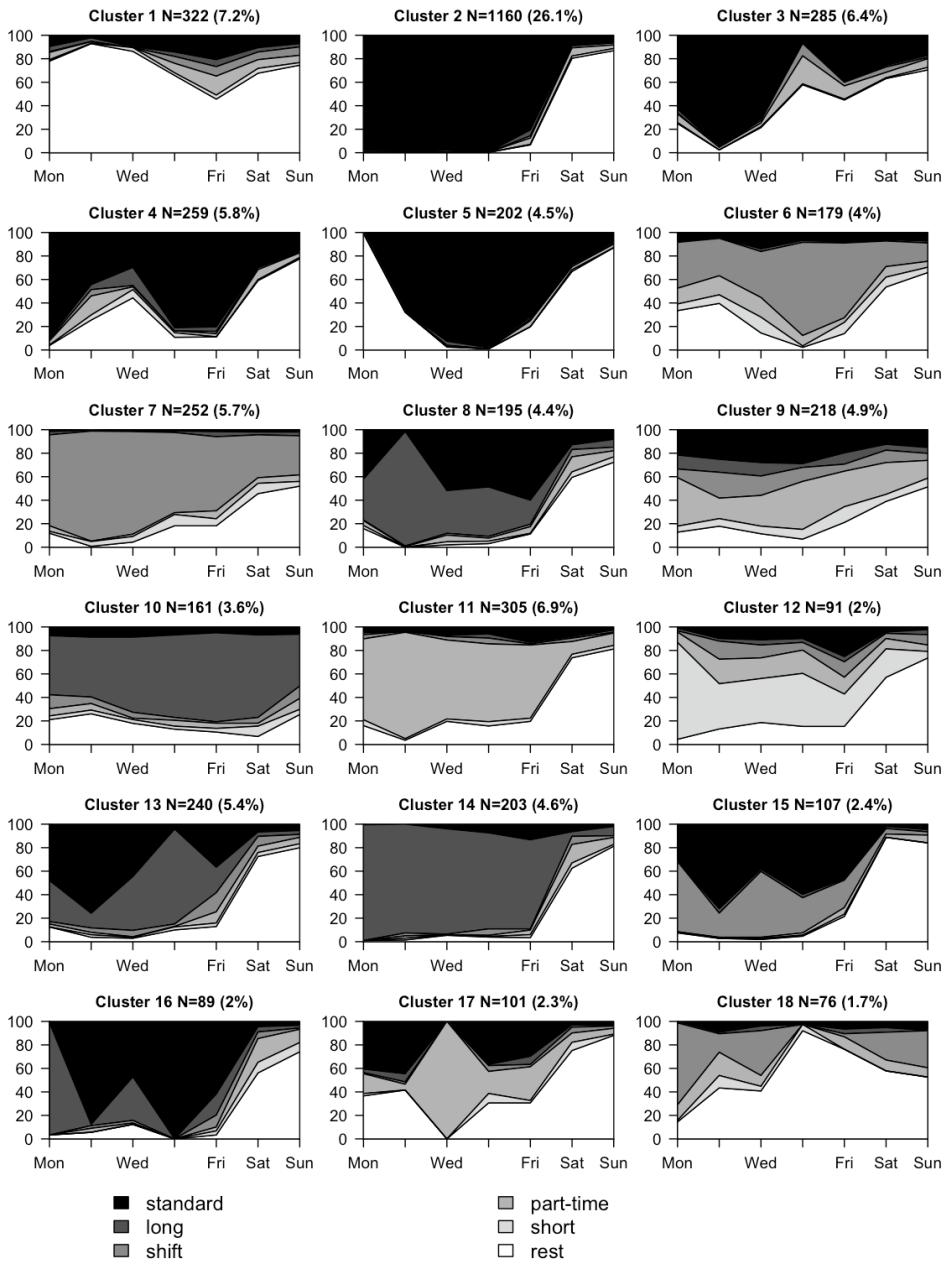



**Figure 6 - The eighteen-cluster solution produced by flexible-beta clustering**

The rest of the clusters are variants of short, part-time workweeks. First, for clusters 11 and 17, individuals usually work on weekdays but they tend to have part-time work hours on each of the workdays.(part workday workweek). Another type of part-time work is found in clusters 4 and 5, where respondents tend to take one or half a day off during weekdays, but they are likely to work at standard hours during their workdays (standard workday part-time). The final type of short workweek is made of very few workdays in a week (clusters 1, 3, and 18, short workweek).

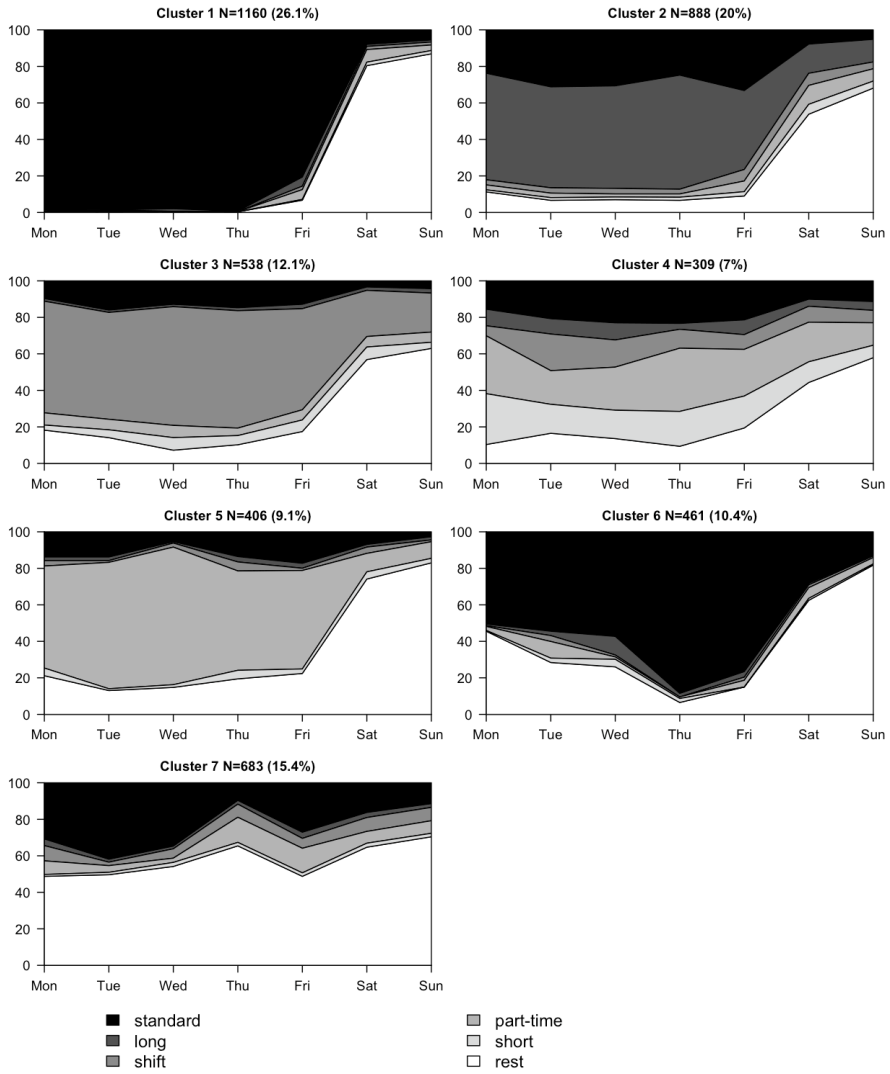



**Figure 7 – Tempograms of the final typology of workweeks**

Descriptions of the 7-group typology of workweeks are provided in Figure 7 and Table 4. Type A, the standard workweek, is composed of five standard workdays from Monday to Friday and the average work time is 42.2 hours. Although being the most common type of workweeks, they only account for about one fourth (26%) of the total workweeks. Another common type of workweeks are long workweeks (Type B), which make up one fifth of the total workweeks. Long workweeks are composed of one or more long workdays. The average work hours are ten hours longer than the standard workweeks. In addition, they deviate from the standard workweeks in the timing of work: 13% of them contain weekend work (c.f. 4% for standard workweeks).

**Table 4 - Descriptive statistics of the final typology of workweeks**

| | Name | Size | Work time | Number of days off | Number of work days | Proportion of work on Sat. | Proportion of work on Sun. | Proportion of work on weekend | Proportion of full Saturdays off | Proportion of full Sundays off |
|---|---|---|---|---|---|---|---|---|---|---|
| A | Standard | 26.1 | 42.2 | 1.8 | 5.2 | 5 | 3.2 | 4.1 | 80.4 | 86.9 |
| B | Long | 20 | 52.1 | 1.6 | 5.4 | 14.9 | 10.6 | 12.7 | 53.8 | 68.1 |
| C | Shift | 12.1 | 37.7 | 1.9 | 5.1 | 12.7 | 11.1 | 11.9 | 56.9 | 63 |
| D | Alternate | 7 | 31.7 | 1.7 | 5.3 | 13.5 | 11.1 | 12.3 | 44.3 | 57.9 |
| E | Part-time I | 9.1 | 21.1 | 2.5 | 4.5 | 5.7 | 3.7 | 4.7 | 74.1 | 83 |
| F | Part-time II | 10.4 | 34.3 | 2.7 | 4.3 | 11.5 | 5.4 | 8.5 | 62.5 | 81.8 |
| G | Short | 15.4 | 21.6 | 4 | 3 | 10.7 | 8.6 | 9.6 | 64.7 | 70.4 |

Type C, shift workweeks, constitute 12% of the total workweeks. Like long workweeks, they have a high proportion of weekend work (12%). The average work time is shorter than the standard workweek (37.7 hours). From the cluster (Cluster 4, Figure 7), we see that in the case where Saturday or Sunday is a workday, a rest day will take place between Monday and Friday. In other words, shift workweeks are not only characterized by shift hours of work, but also a shift of the days off from weekends to weekdays.

Alternate workweeks, Type D, are made of more than one type of workdays. There is a high proportion of part-time workdays and hence the average weekly work time is considerably short (31.7 hours). Weekend work is also common in this type of workweeks (12.3%). Accordingly, days off tend to shift towards Monday to Friday.
It is well documented that part time work rate is relatively high in the UK compared with other developed countries. In the present study, we have identified two main types of part-time workweeks, which together form about one fifth of the total workweeks. Type E, the standard workday part time, is similar with the standard workweek but have about one more day off (2.5 c.f. 1.8). Type F, part workday part-time, is characterized by part workweek (2.7 days off on average) and part-time work hours during workdays. Both Types E and F of part-time workweeks have shorter total work hours than the standard workweeks (the figures being 21.1 and 34.3 respectively). Furthermore, weekend work is more common than the standard workweek, especially working on Saturdays. These results are consistent with previous studies on work time trends, which show that long and short workdays are both increasingly common in economically advanced societies (Gershuny, 2000). Our typology goes further to identify the distribution of workdays over the week, and that working on weekends is a key characteristic of the work time trends.
Finally, Type G, short workweeks, which represent 15% of the workweeks, are composed of only three work days. Nevertheless, standard work hours are usually involved during workdays. The average weekly work hours are short (21.6 hours). In

fact, short workweeks can also be defined as a variant of part-time work in the UK. Similarly with Types E and F, there is a high proportion of weekend work (9.6%).

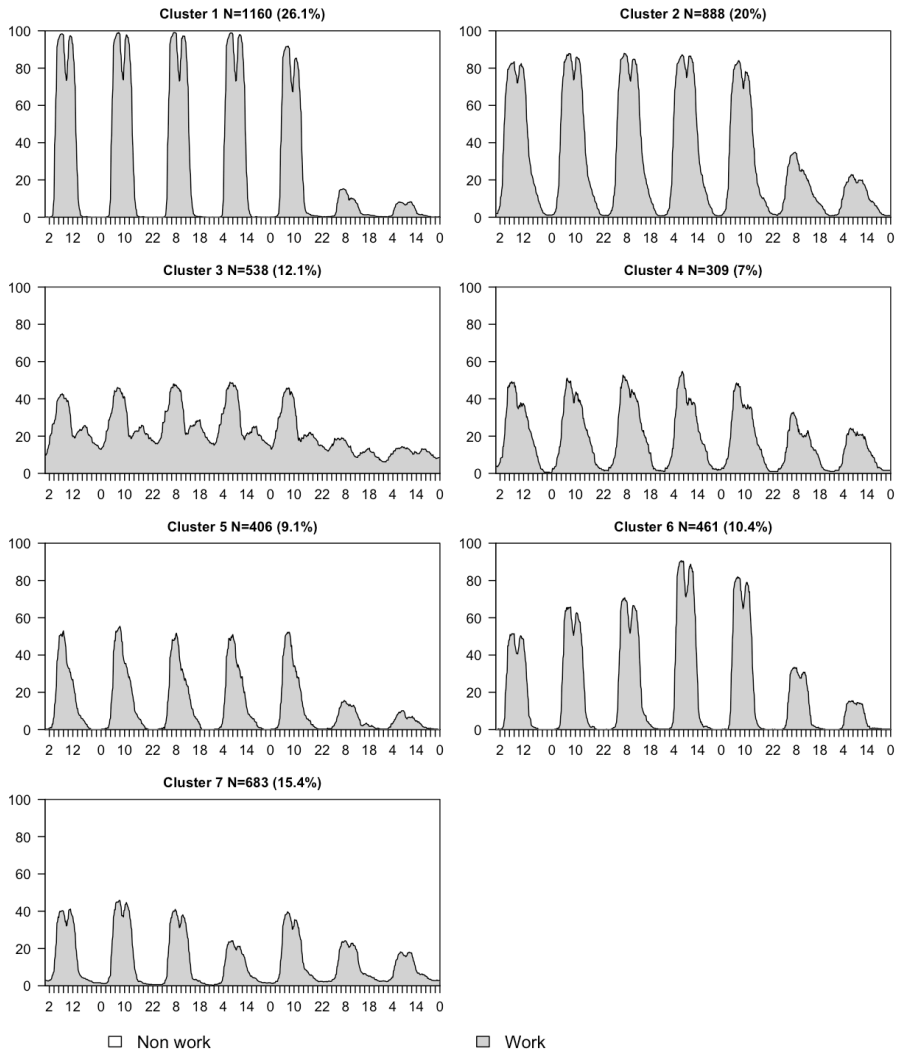



**Figure 8 – Tempograms of the final typology of workweeks (15-minute time slots)**

To assess the effectiveness of two-stage optimal matching, we represent the typology using the 672 15-minute episodes in Figure 8. As can been seen, the seven types are very distinct from one another, suggesting that the first step of 2SOM is effective. The strength of 2SOM lies in making the interpretation of results easier. The results would have been much more difficult to interpret if we had not obtained the first stage results. It should be noted that, however, applying OM directly on the 672 15-minute episodes yields slightly different results from 2SOM[19].

19 Given the fact that the typologies are difficult to interpret, it is not possible to describe precisely the difference in the results of one-stage OM and 2SOM. Overall, the difference should not be significant because the first-stage 2SOM is guided by theories and empirical findings of previous studies. The additional figure is available upon request to the authors.

## 4.3. Advantages of analysing both workdays and workweeks

As can be seen from Figure 7, workweeks are usually dominated by one type of workdays: for example, *long* workweeks are composed of mostly *long* workdays, and *shift* workdays are common in *shift* workweeks. This suggests that work is not randomly scheduled over days and weeks, but is instead highly temporally structured. Previous research demonstrated that work schedules mostly reflect the preferences of the employers rather than those of the employees (Lesnard 2008, Golden 2001).

This lack of variation in the types of workdays within a workweek gives some confidence to researchers who only have day-long time use data: that analysing how work is organised at the level of the day is likely to give good insights into how work is scheduled over a longer period. Nevertheless, our findings also show that there is one major drawback from this approach: the overall proportion of atypical or non-standard workweeks will be underestimated if the figures are generalized from the analysis of workdays alone. It is because standard workdays, though not being the dominant types of workdays, occur also in *long*, *shift* and all types of *part-time* workweeks. In other words, observing a standard workday in a sample is not a good predictor of whether or not the rest of the week will be made of only standard workdays. In contrast, a non standard workday is an excellent predictor of atypical workweeks.

Thus, researchers will overestimate the proportion of *standard* workweeks based on the number and proportion of *standard* workdays in their samples. In this study, 52% of wordays are *standard* (31% when both workdays and rest days are taken into account), but standard workweeks only account for 26% of the total workweeks. On the other hand, *long* workdays represent 16% of the workdays but 20% of the workweeks are *long*. In sum, the proportion of workweeks will be more accurately represented should the analyses be conducted at both the day- and the week- levels rather than solely at the day-level.

# 4. Conclusion

We have demonstrated that two-stage optimal matching (2SOM) can be usefully applied to the analysis of workweeks. Our study surpasses past ones by providing important insights into the schedule of work hours within workdays and the structure of work days over the week. We have identified 7 types of workweeks and more varieties of part-time work in the UK.

Methodologically, this study has contributed to the on-going reflections and discussions on how costs should be set in OM. We suggest that costs should be defined in accordance with the kind of patterns that researchers expect to see or consider theoretically interesting. For example, we have defined costs based on transitional frequencies of two states at a given time in this study. Setting the costs based on theoretically concerns will help to emphasize and capture certain patterns effectively. But it does not necessarily imply that the results will be completely changed should a different cost is chosen. In cluster analysis, there are often stable

clusters, which exist in the outputs however the parameters are set[20]. But the proportion of the stable clusters may vary with the cost. That is, non core cases of a cluster are prone to move to another cluster if the algorithm changes. In this regard, to stress a certain kind of patterns in OM by defining the ratio of indel cost to substitution ones is tantamount to adjusting the sensitivity of border cases to the stable clusters. When indel costs are low compared to substitution ones, the distance will favour the number of identical events regardless of their location in the sequences. When they are high, similarity of the border cases will be estimated according to their existing positions in the sequences.

In the case of the scheduling of work, it is essential to compare sequences based on their local similarity (otherwise the schedule of events itself will be altered). Furthermore, we employ time-varying substitution costs because transition frequencies provides significant empirical and theoretical information on sequence proximity. To apply OM on other research topics, we recommend users define the costs based on theories and previous findings. When no previous reference is available, researchers may adopt neutral costs (i.e. close to Levenshtein I) so as to favour neither local nor remote similarity.

OM is a versatile technique that can easily take into account two periodicities in its analysis. Although it is possible to apply OM directly to the 672 episodes, it is more appropriate to focus on each of the nested periodicities so that the patterns found are clearer and easier to be identified at each stage. The two-stage optimal matching (2SOM) for cases of nested periodicities is analogous to noise filtering or seasonal adjustment in time series analysis. The first stage of OM, in which analyses are guided by theories and previous findings, acts as a form of noise filtering.

Finally, we have a suggestion for researchers who intend to apply 2SOM on other workweek data. The UK 2000 time use data used in this study, unlike other survey data such as the France Time Use Study 1998-1999, do not contain information about whether the respondents considered the workweeks filled in to be a "normal" week. Consequently our sample is likely to contain respondents who were on vacation or had some usual work patterns. In future research, where possible, researchers should restrict their analyses to workweeks that respondents indicated as normal.

[20] In OM, stable clusters are composed of identical or almost identical sequences, which will end up being grouped together in no regard to the costs, since costs are, by definition, only used when sequences are dissimilar.

## References


Abbott, A. (2000) Reply to levine and wu. *Sociological Methods and Research*, **29**, 65-76.

Abbott, A. and Forrest, J. (1986) Optimal matching methods for historical sequences. *Journal of Interdisciplinary History*, **16**, 471-494.

Abbott, A. and Hrycak, A. (1990) Measuring resemblance in sequence analysis: An optimal matching analysis of musicians careers. *American Journal of Sociology*, **96**, 144-185.

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

Brzinsky-Fay, C. (2007) Lost in transition? Labour market entry sequences
of school leavers in europe. *European Sociological Review*, **23**, 409-422.

Chenu, A. and Robinson, J.P. (2002) Synchronicity in the work schedules of working couples. *Monthly Labor Review*, **125**, 55-63.

Dumazedier, J. (1967) *Toward a society of leisure?* New York: Free Press.

Durbin, R. *et al*. (1998) *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge (UK), New York: Cambridge University Press.

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

Gabadinho, A. *et al*. (2008) Mining sequence data in r with the traminer package: A user's guide.

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

Glorieux, I., Mestdag, I. and Minnen, J. (2008) The coming of the 24-hour economy? Changing work schedules in belgium between 1966 and 1999. *Time & Society*, **17**, 63-83.

Golden, L. (2001) Flexible work schedules: Which workers get them? *American Behavioral Scientist*, **44**, 1157-1178.

Halpin, B. (2008) Optimal matching analysis and life course data: The importance of duration. *Department of Sociology Working Paper Series*, **WP2008-01**,

Halpin, B. and Chan, T.W. (1998) Class careers as sequences: An optimal matching analysis of work-life histories. *European Sociological Review*, **14**, 111-130.

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

Ipsos-Rsl, Statistics, O.F.N. (2000) United kingdom time use survey.

Kruskal, J.B. (1983) An overview of sequence comparison. *Time warps, string edits, and macromolecules: the theory and practice of sequence comparison*, 1-44.

Lesnard, L. (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, L. (2006a) Flexibilité des horaires de travail et inégalités sociales. In *Données Sociales - La société française,* (ed Insee), pp. 371-378. Paris: Insee.

Lesnard, L. (2006b) Flexibilité et concordance des horaires de travail dans le couple. In *Données Sociales - La société française,* (ed Insee), pp. 379-384. Paris: Insee.

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

Lesnard, L. (2009a) Cost setting in optimal matching to uncover contemporaneous socio-temporal patterns. *Notes & Documents*,

Lesnard, L. (2009b) *La famille désarticulée. Les nouvelles contraintes de l'emploi du temps*. Paris: PUF.

Lesnard, L. and Saint Pol, T.d. (2006) Introduction aux méthodes d'appariement optimal (optimal matching analysis). *Bulletin de Méthodologie Sociologique*, **90**, 5-25.

Levenshtein, V.I. (1966) Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, **10**, 707-710.

Levine, J.H. (2000) But what have you done for us lately?: Commentary on abbot and tsay. *Sociological Methods and Research*, **29**, 34-40.

Milligan, G.W. (1980) An examination of the effect of six types of error perturbation on fifteen clustering algorithms. *Psychometrika*, **45**, 325-342.

Milligan, G.W. (1981) A monte carlo study of thirty internal criterion measures for cluster analysis. *Psychometrika*, **46**, 187-199.

Milligan, G.W. (1989) A study of the beta-flexible clustering method. *Multivariate Behavioral Research*, **24**, 163-176.

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

Robinson, J.P., Chenu, A. and Alvarez, A.S. (2002) Measuring the complexity of hours at work: The weekly work grid. *Monthly Labor Review*, **125**, 44-54.

Robinson, J.P. and Godbey, G. (1999) *Time For Life. The Surprising Ways Americans Use Their Time*. University Park: Pennsylvania State University Press.

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

Stovel, K., Savage, M. and Bearman, P. (1996) Ascription into achievement: Models of career systems at lloyds bank, 1890-1970. *American Journal of Sociology*, **107**, 358-399.

Szalai, A. (Ed.) (1972) *The Use of Time. Daily Activities of Urban and Suburban Populations in Twelve Countries* Mouton, The Hague, Paris.

Wilson, W.C. (1998) Activity pattern analysis by means of sequence-alignment methods. *Environment and Planning A*, **30**, 1017-1038.

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