A 'global interdependence' approach to multidimensional sequence analysis

Nicolas Robette, Xavier Bry, Eva Lelièvre

A "Globally Interdependent Multiple Sequence Analysis" approach to uncover patterns of linked life courses

Authors: Nicolas Robette, Printemps (UVSQ-CNRS, UMR 8085)
Xavier Bry, I3M, Université Montpellier 2
Éva Lelièvre, INED

Contact :
nicolas.robette@uvsq.fr

Abstract:

While sequence analysis has now become a widespread approach in social sciences, several strategies have been developed to handle the specific issue of multidimensional sequences. These strategies have distinct characteristics, related to the way they explicitly emphasize multidimensionality, interdependence and parsimony. In this context, we introduce an original approach based on structural links between the dimensions, combining Optimal Matching Analysis (OMA), Multidimensional Scaling (MDS), canonical Partial Least Square (PLS) and Clustering, an approach we call "Globally Interdependent Multiple Sequence Analysis" (GIMSA). We then apply GIMSA to mother-daughter employment histories in France and discuss the value of this method.

1. INTRODUCTION

Since the mid-1970s, life course analysis has become a major field of interest in social sciences. Longitudinal micro-individual data - such as panels or retrospective surveys – have become more available and, at the same time, statistical methodology has undergone a profound evolution. In this context, event-history analysis, which can be viewed as adding a diachronic dimension to traditional regression models, rapidly became established as the dominant approach: it aims at modeling the duration in a given situation or the risk of experiencing a given event. However, during the last decade, a large corpus of more descriptive sequence analysis methods has been disseminated. Their main goal is to identify patterns and resemblances among sets of diverse sequences (made up of series of successive states), most often resulting in typologies of ideal-typical sequences. Nowadays, sequence analysis provides a powerful means to describe and better understand the unfolding of many social processes.

Most of the sequence analysis techniques currently used in social sciences are related either to algorithmic methods (Abbott and Tsay 2000) or to correspondence analysis methods (Grelet 2002)[1]. All have particularities, advantages and drawbacks, although they usually give quite similar results (Robette and Thibault 2008; Robette and Bry 2012). Optimal Matching Analysis (OMA) is by far the most widespread sequence analysis technique. It has the major conceptual advantage of jointly addressing the different temporal aspects of a sequence: the moment of a transition, the duration of a stage and the order within the sequence.

OMA was initially developed in molecular biology and introduced into the social sciences by Andrew Abbott in the 1980s (Abbott and Forrest 1986). The principle is based on the notion of similarity between pairs of sequences. The dissimilarity between two sequences is measured in terms of the "cost" of transforming one into the other. The transformation is performed using three types of basic operation: insertion (inserting an element into the

sequence), deletion (deleting an element) and substitution (replacing one element by another). A cost is associated with each operation; the distance between two sequences is thus defined as the minimum cost of the operations required to transform one into the other. Matching the entire set of sequences creates a matrix of pairwise distances which is then used to group together those that are most similar, e.g. using clustering techniques, and so obtain a typology[2].

Although OMA has been applied to a wide range of social objects (see for instance Abbott and Barman 1997 for pattern searching focusing on sequences of atypical types), it has also received some criticism (Wu 2000; Elzinga 2003); some innovative developments have emerged as a consequence (see Aisenbrey and Fasang 2010 for a review; see Biemann 2011 for a recent example).

While sequence analysis of work careers and social trajectories in general is now well established (Brzinsky-Fay and Kohler, 2010), notable previous developments have tended to explore the possibility of considering not only a single dimension of the life course (e.g. employment) but also of integrating combinations of statuses pertaining to other dimensions of the life-course such as housing and family, introducing multidimensional sequence analysis (Pollock, 2007). In this paper, we examine another aspect of life-course analysis: that of patterns of transmission from one generation to the next. While social mobility, for instance, is more usually studied in terms of cross-sectional comparison of the father's position at a given period compared to that of the son, we here consider the entire histories of both generations. More precisely, we study women’s involvement in paid employment for two successive generations marked, in France, by massive female entry into the labour force. Our aim here is to identify the way specific work profiles in one generation were followed by others in the next one, rather than the possible similarity of mothers’ and daughters work careers. Each generation’s involvement in paid activity has been shaped by their specific

historical context and the same profile is unlikely to recur. The relevant question is: what are the main female lineage patterns, in terms of school-to-work transition and employment history, which underlie the macro trends in work and family (Barrère-Maurisson, 1992).

To this end, we first review the methods already available for multidimensional sequence analysis, then we introduce a new approach based on structural links between the parents' and children's sequences, combining Optimal Matching Analysis (OMA), Multidimensional Scaling (MDS), canonical Partial Least Square (PLS) and Clustering, an approach we call "Globally Interdependent Multiple Sequence Analysis" (GIMSA). We then apply GIMSA to mother-daughter employment histories in France and discuss its benefits for summarizing complex sequence data such as linked life courses.

2. HANDLING MULTIDIMENSIONAL SEQUENCES

In the early days of sequence analysis in the social sciences, the successive elements composing sequences were hard to simplify into a unique and limited set of states: methodological adjustments were needed to capture the diversity and complexity of individual social statuses. In other words, to make a detailed study of careers as sequences, various dimensions must be considered. For example, in their seminal article about the careers of musicians in Germany during the Baroque and Classical eras, Abbott and Hrycak (1990) combine position (e.g. vocalist or instrumentalist) and sphere (e.g. court or church). Likewise, Stovel et al.'s occupational status variable is a combination of position and branch size (Stovel et al. 1996), Blair-Loy combines job code and organization size (Blair-Loy 1999) and Han and Moen mix work status, organization and occupation (Han and Moen 1999).

Later, sequence analysis applications focused on life courses, and the need to handle simultaneously their various dimensions became a key concern. Conjugal, parental,

occupational and residential histories unfold in an interdependent way (Courgeau and Lelièvre 1992), and scholars explored the methodological repercussions of this theoretical construct. Pollock (2007) introduced "Multiple-Sequence Analysis" (MSA), which was subsequently systematized by Gauthier et al. (2010) and renamed "Multichannel Sequence Analysis" (MCSA)[3].

The various multidimensional sequence typology-building strategies found in the literature can be summarized into 4 groups. The first strategy consists in creating a new state variable which combines the simple states composing each dimension (Dijkstra and Taris 1995; Elzinga 2003; Aassve et al. 2007; Elzinga and Liefbroer 2007; Lesnard 2008; Chaloupkova 2010). For instance, in the case of conjugal and parental histories, possible combined states would be "single with no child", "married with no child", "married with one child", etc. This may quickly lead to a large alphabet, i.e. a set of very numerous states. Hence, in the case of 4 dimensions with 3 simple states in each, the combined variable would potentially have 3*3*3*3=81 states. Such an extended alphabet may be impractical when aiming to set substitution costs specifically tailored for each pair of combined states. However this drawback can easily be circumvented by setting a constant substitution cost or by using transition likelihood between the states[4] (Lesnard 2008).

The second strategy is a more refined approach to avoid the need for a large extended alphabet: it is based on combining the substitution costs of the various dimensions. For instance, the substitution of "single with no child" and "married with one child" will be equivalent to a combination of the substitution cost between "single" and "married" and the substitution cost between "no child" and "one child". A possible combination is the sum (or average) of the costs defined for each dimension (Stovel et al. 1996; Blair-Loy 1999; Pollock 2007; Gauthier et al. 2010; Salmela-Aro et al. 2011), e.g. the sum (or average) of the substitution cost between "single" and "married" and the substitution cost between "no child"

and “one child”. One can also imagine a more refined linear combination of the different dimensions (Abbott and Hrycak 1990), for instance by applying weights to these dimensions (Gauthier et al. 2010, p. 34). If there is a unique substitution cost for each dimension and it is identical between the dimensions, substituting two multidimensional states is equivalent to counting the number of dimensions which differ (Robette 2010). For example, replacing “single with no child” by “married with one child” will cost 2 while replacing “single with no child” by “married with no child” will cost 1. Moreover, this second strategy may be seen as a particular case of the first one, so substitution costs can be set simply and efficiently.

A third strategy consists in computing a dissimilarity matrix independently for each dimension and then summarizing them into a single distance matrix by linear combination (Han and Moen 1999; Blanchard 2005).

Lastly, distinct typologies of sequences can be built for each dimension and then compared (Blanchard 2005), e.g. with cross-tabulation[5].

These four strategies can be systematically compared and classified according to three criteria: multidimensionality, interdependence and parsimony (see table 1)[6]. With *multidimensionality* we refer to the fact that an approach where the contribution of each dimension to the overall results may or may not be explicit (i.e. unequivocal) and flexible (i.e. parameterizable by the analyst). The first strategy takes the multiple dimensions into account, as do the other three; still, by hiding dimensions in a single combined state variable, it is the only one which does not emphasize multidimensionality: for instance, it is not possible to have specific parameters for each dimension, to give more importance to a particular dimension by weighting it or to assess the impact of each dimension on the results.

With *parsimony* we refer to the fact that an approach may or may not lead to a limited and manageable number of ideal-types[7]: combining typologies produced for each dimension (as is done with strategy 4) may lead to an uncomfortably large number of clusters.

With *interdependence* we refer to the fact that the relationship between dimensions may be masked (strategy 3) or taken into account locally (as in strategies 2 and 3) – i.e. the focus is on the dependency between dimensions *at each point in time* – or globally (as in strategy 4) – i.e. the focus is on the dependence between dimensions through *sequences as wholes*.

Local interdependence implies an emphasis on contemporaneousness. Indeed, with strategies 1 and 2, dimensions associated with a given sequence are synchronized. Once defined at a given time point in a given sequence, a multidimensional situation "remains the same throughout the alignment procedure" (Gauthier et al. 2010, p.9): the various dimensions shift jointly. Technically, the synchronization of the dimensions means that they have to be defined with the same time window – i.e. the same starting and ending points (whether these are ages or dates) – and the same time clock (e.g. one year or one month by time point). This also implies the use of a unique dissimilarity measure (e.g. optimal matching).

Global interdependence releases the emphasis on contemporaneousness; it is the overall shape of a dimension which is related to the others. In strategy 4, pairwise comparisons are made separately for each dimension in a first step, and then the relationships between dimensions' patterns are examined. As distance matrices are computed independently for each dimension, distinct dissimilarity measures can be used, e.g. a metric focusing on timing (such as Hamming distance) for one dimension and a metric based on order (such as Longest Common Subsequence metric, see Elzinga 2008) for another dimension. Different time windows and clocks can also be used.

Strategy 3 also allows for different dissimilarity measures, time windows and clocks. However, by simply adding distances matrices, the relationships between dimensions are not adequately handled. For instance, with two-dimensional sequences, if two individuals have a dissimilarity of 1 for a dimension and are identical for the second, they will have the same global distance as two individuals who are identical on the first dimension and have a dissimilarity of 1 on the second.

It is important to keep in mind that these criteria – multidimensionality, parsimony and local/global interdependence – may or may not be desirable, depending on theoretical or data issues. There is no better strategy *per se* and the choice of one or other should be grounded on sociological and empirical criteria.

TABLE 1

Taxonomy Of Strategies To Handle Sequences With Various Dimensions

Strategy	*Multidimensionality*	*Parsimony*	*Interdependence*
Combining states **(1)**	No	Yes	Local
Combining costs **(2)**	Yes	Yes	Local
Combining distance matrices **(3)**	Yes	Yes	No
Combining typologies **(4)**	Yes	No	Global

Recently, a few sequence analysis papers took into account one of the key elements shaping life courses in Elder's paradigm (Giele and Elder, 1998), e.g. the fact that individual life courses are embedded into social relationships, which means that "linked lives" are studied. Most applications have dealt with the transmission of histories between parents and children (Falcon 2012; Liefbroer and Elzinga 2012; Robette et al. 2012; Fasang and Raab 2014), although a few papers have focused on husband and wife histories (Lesnard 2008; Robette et al. 2009; Lelièvre and Robette 2010).

To assess the strength of transmission between parents and children, Liefbroer and Elzinga (2012) left typologies to one side and analyzed the dissimilarities between the sequences of the relatives themselves. However, one might argue that in some cases, perfect similarity between the sequences of parents and children is not appropriate evidence of transmission processes. Indeed, the median age at parenthood, for instance, may be about 20 for the parents' generation and about 25 for that of their children. So, parents and their children may be viewed as having dissimilar sequences, strictly speaking, while they actually have perfectly equivalent histories given the structural changes to the historical context in which their lives unfolded. Moreover, a given amount of dissimilarity may have distinct reasons – e.g. a two year difference in age at marriage may be judged equivalent whether it be two years earlier or two years later, although it does not have the same meaning – which remains invisible in this "dissimilarity approach".

Some have adopted the second of the strategies presented earlier, i.e. the combination of substitution costs (Robette et al 2009; Fasang and Raab 2014). This strategy has the advantage of allowing the identification of contrasting patterns - i.e. groups of dyads in which parents' and children' sequences are distinct but often associated - in the case of an intergenerational transmission study (Fasang and Raab 2014). This strategy is especially appropriate when it is meaningful to synchronize the sequences of each dimension within a dyad e.g. compare the timing and pace of transitions of parents and children – or to characterize the situation of the dyad at a given point in time (*local interdependence*). This is not always the case, however. For instance, sometimes the various dimensions may differ substantially in nature. An alternative objective may be to study the relationship between the parents' overall career and their children's school-to-work transition. In this example, parents' career sequences could span over 45 years from ages 14 to 60 (with years as time units) and the alphabet (i.e. the set of possible states) would be based on an occupational classification; when children's school-

to-work sequences could stretch for only three years after leaving school (with months as time units) and the alphabet would then be composed of employment statuses (e.g. education, unemployment, part-time employment and full-time employment). Technically, the difference in sequence length can be handled with missing values. Nevertheless in this hypothetical example with such different time windows and clocks, this would become at the very least inelegant – even impractical - and would certainly obscure the results. More importantly, from a substantive point of view, “pasting” one sequence onto the other and locally aligning them is pointless here: what we want to examine is rather the overall shapes of the separate dimensions – and the patterns in them - and the relationship between these patterns (*global interdependence*)[8].

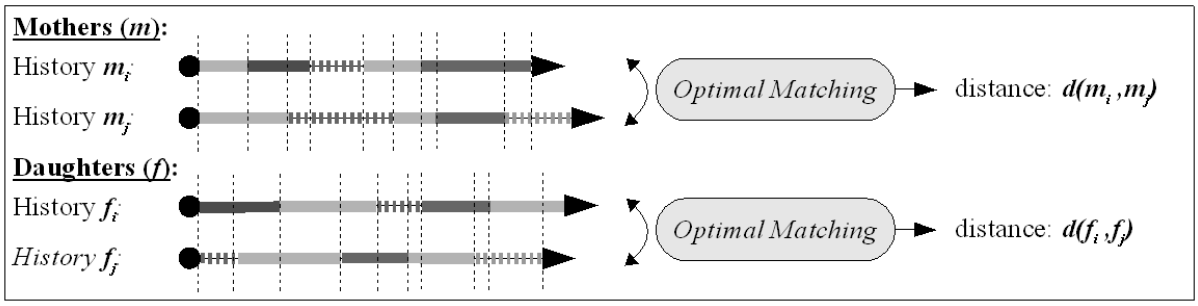
Our objective here is to propose an approach which is parsimonious and takes into account multidimensionality as well as global interdependence between parents' and children's sequences. In effect, parents' sequences and those of their children can be of very different nature, reflecting the norms of their time, and our focus is on the relationship between sequences as wholes rather than on their synchronization in terms of age. Thus our approach presents a unique combination of the criteria defined above, well-suited for the study of intergenerational transmission of life course patterns, and it provides a useful complement to the existing strategies.

To facilitate the statistical presentation of our approach, we build on the study of the transmission of patterns of employment history between mothers and daughters, which we then empirically explore.

3. A "GLOBALLY INTERDEPENDENT MULTIPLE SEQUENCE ANALYSIS" APPROACH

This approach, which we called "Globally Interdependent Multiple Sequence Analysis" (GIMSA), breaks down into several steps. Let's consider that each unit of analysis is composed of one sequence for a mother and one for her daughter, i.e. a set of paired trajectories.

FIGURE 1. First Step Of GIMSA



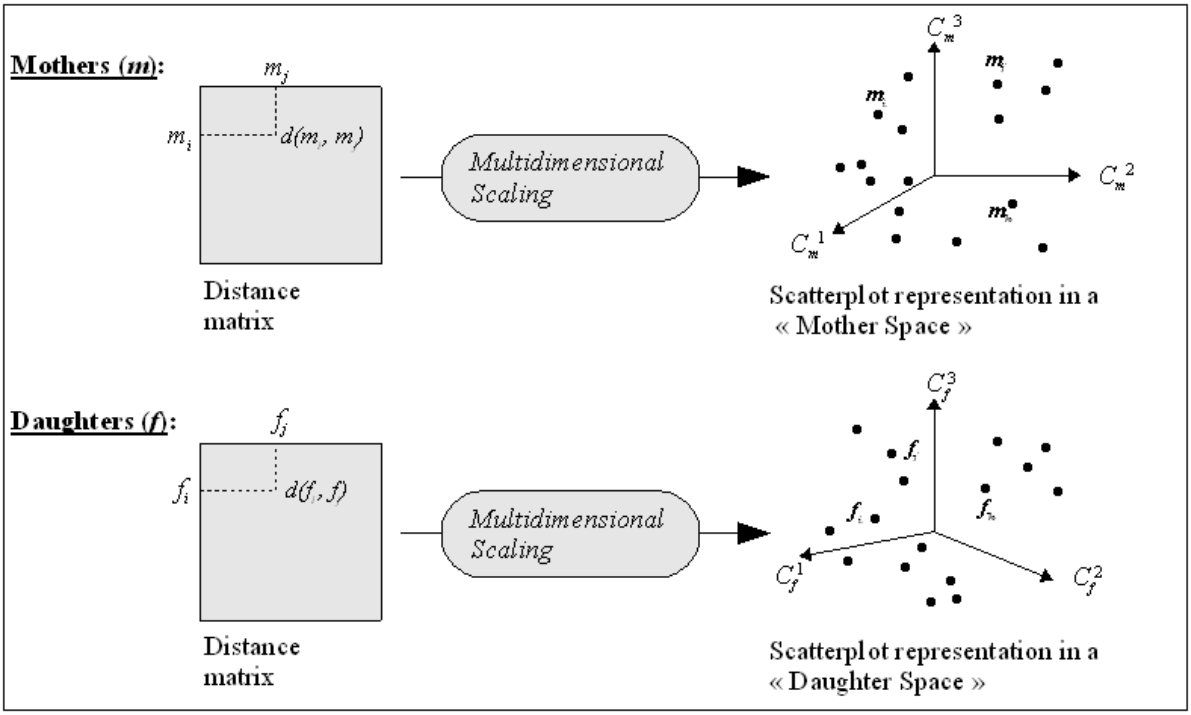
At the first step (Figure 1), Optimal Matching (OM), or any other standard sequence dissimilarity measure, is used to calculate distances associated with each pair of mother's and daughter's histories. This stage gives us two symmetric distance matrices $\boldsymbol{M} = (m_{ij})_{i,j}$ where $m_{ij} = d(m_i, m_j) = d(m_j, m_i)$ (respectively $\boldsymbol{F} = (f_{ij})_{i,j}$ where $f_{ij} = d(f_i, f_j) = d(f_j, f_i)$) in which the diagonal is zero.

Whether or not GIMSA is highly sensitive to a given kind of sequence pattern depends very much on this first stage. Indeed, some sequence metrics may be more suited to capturing patterns in terms of durations or timing, while others may rely more on the order, on reversals or repetitions. The specificities of the major sequence dissimilarity measures used in the social sciences are discussed and tested with empirical and simulated data in Aisenbrey & Fasang 2010, Robette & Bry 2012 or Studer 2012. It should be kept in mind that in the end, most of these measures, when applied to empirical data, lead to relatively similar results. The

OM algorithm is presented briefly in the introduction of this paper (for a more detailed presentation, see for instance MacIndoe and Abbott 2004).

Moreover, different dissimilarity measures may be chosen for the various dimensions, emphasizing global interdependence.

FIGURE 2. Second Step Of GIMSA

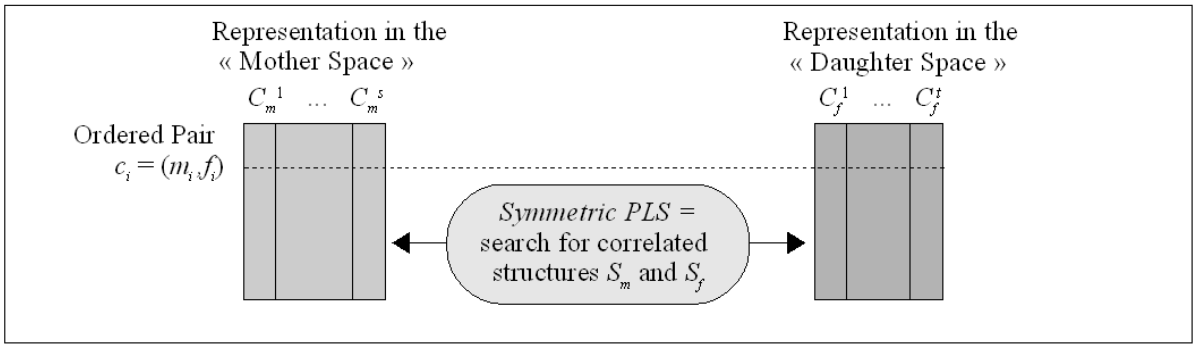


At the second step (Figure 2), we use multidimensional scaling (MDS) (Kruskal and Wish 1984). MDS is “a factorial technique that provides a visual representation of a dissimilarity matrix. Sequences are projected in a low dimension factorial space in such a way that the distance between cases in this space resembles as much as possible the original dissimilarity between them” (Piccarreta and Lior 2010). Here, the matrix M (respectively F) of distances between the mothers’ careers (respectively their daughters’) is converted into a spatial representation by MDS, i.e. represented as a scatterplot of points in a multidimensional “mothers' space” (resp. “daughters' space”) with respect to a system of principle components C_m^1, C_m^2, etc. (resp. C_f^1, C_f^2, etc.). For a brief mathematical presentation of the metric MDS we

use, see Appendix 1[9]. The sequence of components C_m^1, C_m^2, etc. (resp. C_f^1, C_f^2, etc.) provides a hierarchical breakdown of the heterogeneity of the mothers' (resp. the daughters') histories, in the sense that each component complements the preceding ones in an optimal manner. At the end of this stage, each mother-daughter pair is described both by the coordinates of the mother's history in the mothers' space and those of the daughter's history in the daughters' space.

The choice of the numbers of components[10] to be kept for the next stage may follow several criteria. The quality of MDS results against their dimensionality can be assessed with indicators such as *eigenvalues* or with a stress function. One may also opt for a close examination of whether a given dimension reveals a clear structure in sequence data, i.e. whether the order of sequence dyads in this dimension can be easily interpreted in a sociologically relevant way. More generally, the selection of the relevant number of components is a trade-off between two extremes: on the one hand, it may be desirable to keep a maximum amount of sequence data information (i.e. to keep many components); on the other, noise reduction is important to prevent overinterpretation. Here again, Piccarreta and Lior (2010) provide a helpful guideline for this step of GIMSA.

FIGURE 3. Third Step Of GIMSA

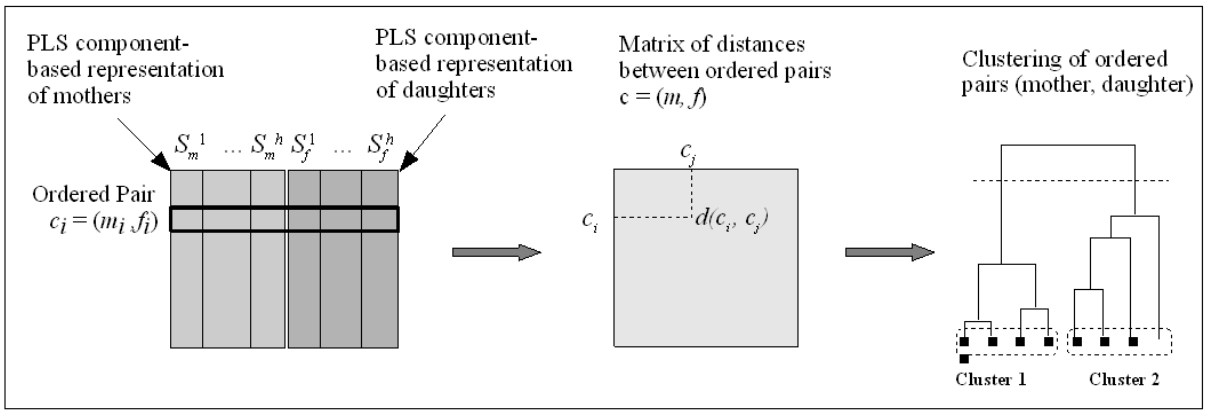


We next look for structural links between the mothers' and daughters' histories (Figure 3). While standard factor analysis techniques deal with a single set of variables, we need to study

simultaneously several sets of variables, i.e. to take into account the relationship between two groups of variables. A few methods have been specifically designed to suit this partitioning of the data. Among them, we use the symmetric (or canonical) PLS method (Partial Least Square) (Bry 1996; de Jong et al. 2001), which seeks structures *common to* the variability of the mothers' histories and that of the daughters' histories[11]. These structures are extracted under the form of components denoted S_m^k (respectively S_f^k) for the mothers (resp. daughters). In short, symmetric PLS seeks, amongst the mothers' and daughters' MDS data, pairs of components having maximum covariance. Indeed, the covariance between a mother's component and a daughter's component takes into account both their correlation, i.e. their linear link, and their variances, interpreted as their principal-component-type structural strength. For a quick mathematical presentation of symmetric PLS, see Appendix 2.

Since a noise reduction step has been performed at the previous stage of GIMSA, here we keep all the components produced by PLS computation.

FIGURE 4. Fourth Step Of GIMSA



Finally, the Euclidean coding of the mothers' and daughters' histories, restricted to "common" components S_m^k and S_f^k, provides a base for clustering the mother-daughter pairs (Figure 4). A Euclidean distance matrix between these pairs is calculated from the PLS components and will be used as an input for clustering.

As one dimension of the dyads - e.g. mothers' sequences - may be more diverse than the other, the results may become excessively driven by this dimension, thus masking the daughters' heterogeneity. That is why the weighting issue should be considered at this stage. Indeed, it is essential to give the same importance to the encodings of mothers' and daughters' careers in the calculation of the distance. Two options may be considered. In the first option, PLS components are merely standardized, which is equivalent to weighting each with the inverse of its variance. Doing so makes them all equally important, whatever their original variance. By contrast, the second option consists in giving an equivalent weighting to all the original PLS-components of mothers (respectively daughters), chosen so as to make the variances of mothers' components comparable to those of daughters. In this way, the PLS-components of mothers (resp. daughters) keep their relative importance. For such weights, the inverse of the first *eigenvalue* of mothers' (resp. daughters') MDS may be used, or the inverse of the number of distinct sequences among the whole set of mothers' (resp. daughters') sequences.

The distance matrix is submitted to a Hierarchical Clustering Analysis with Ward criterion[12] (although other clustering techniques can potentially be used). This gives a typology of pairs based on structures common to the discrepancies of the mothers' histories and to those of their daughters.

Once again, it should be noted that at each step of GIMSA, several technical alternatives may be considered, e.g. k-means instead of hierarchical clustering. We are focusing on the combination of sequence analysis, data reduction (MDS and PLS) and clustering, and it is assumed that at each step, each of them is applied coherently and in a substantive way (or in the way that the researcher considers as 'satisfactory' when not optimal).

4. AN EMPIRICAL ILLUSTRATION OF GIMSA: EMPLOYMENT HISTORIES OF MOTHERS AND DAUGHTERS

4.1. *Data*

To analyze sequences related to linked individuals, appropriate data must be available. INED's *Biographies et entourage* survey is a retrospective life-event history survey of 2,830 residents of the Île-de-France[13] area aged between 50 and 70, and those of their contact circles. The sample interviewed was representative of the Île-de-France region population in 2000, the year the survey data were collected. The 'contact circle' (*entourage*) includes family members (lineal kin and relations by marriage) across four generations, plus all those with whom the respondents have cohabited and any others, related or not, who have played a major role in their life (Lelièvre and Vivier 2001; GRAB 2009). The *Biographies et entourage* questionnaire recorded all stages in the residential, occupational and family trajectories of the respondents and the members of their contact circle, year by year, using an extended Life History Calendar. It is thus possible to reconstitute detailed individual trajectories and to consider the employment histories of more than one generation, notably the female respondents and their mothers[14].

We thus have a respondent's entire occupational history, including periods of economic inactivity (all periods of more than a year were included). Each stage is characterized by the occupation declared by the respondent, their employment status, a description of the employer (public or private sector, economic branch, location, firm size) and the proportion of time spent at work. The employment histories of the female respondents in the *Biographies et entourage* survey can be summarized as sequences of employment statuses, year by year, between the ages of 14 (end of mandatory schooling for the cohorts studied) and 50 (the age of youngest respondents at the time of the survey). In order to demonstrate GIMSA's capacity to handle different types of paired sequences, in the application presented here we will restrict

our attention to the way in which the transition from school to work unfolds and how employment histories begin. We therefore gather the successive employment statuses of each female respondent from the completion of education to 15 years later. This gives a set of 1,413 sequences, all lasting 15 years, and each state has four categories representing the four employment statuses, i.e. education, inactivity, part-time employment, full-time employment. The time window is defined from an initial event[15], and not in terms of age or calendar year as is usually the case.

Less precision was required for the employment histories of the respondent's parents, but it is possible to reconstruct their employment statuses and job types, as well as any career interruptions and their nature. Again, for the purpose of our application, we here use the successive occupational statuses of the respondents' mothers from ages 14 to 60. So we have a set of 1,413 sequences, all lasting 47 years, and an alphabet comprising the five following states: education, inactivity, self-employment, lower-level occupation, higher-level or intermediate occupation. We therefore choose a 15-year life span for the respondents with four different statuses and a 47-year occupational career for their mothers with five distinct statuses.

4.2. *Applying standard sequence analysis and GIMSA*

As a preliminary exploration, we used standard sequence analysis (and not GIMSA yet) to build separate typologies of employment histories for the female respondents and their mothers. We wanted to distinguish female respondents finely according to the timing of the transitions in their early employment histories, when family formation competes strongly with their working career, and distinguish between women who leave school for the job market but shift to inactivity after a very short time, and those who establish a career before they stop working. As a result, we used the Hamming distance for dissimilarity measure as it emphasizes timing. The choice of costs is an important aspect for sequence analysis

techniques related to OM (Lesnard 2010), as is the case for Hamming distance. For our study, substitution costs were set at the same value regardless of the elements replaced. Using data-driven costs based on transition likelihoods was considered but this leads to almost identical results (see Robette and Bry 2012 for a systematic comparison). Deriving costs from a theoretical hierarchy of statuses would be possible as well, e.g. replacing education with part-time employment would cost 1 while replacing education with full-time employment would cost 2. But we believe that these situations are more different in nature than in intensity and that there is no perfectly satisfying way of quantifying the difference between them[16]. Besides, for mothers the focus is on the shapes their occupational careers have taken (always working *vs* inactive, etc.), on the transitions that compose them, on whether they are characterized by stability or mobility between labor force and inactivity, etc. For this reason, we chose a dissimilarity measure which favors order over timing, i.e. the Longest Common Subsequence metric (LCS, Elzinga 2008).

Distance matrices were computed[17] and used as input for clustering (here Hierarchical Clustering Analysis with Ward criterion). In order to choose appropriate numbers of clusters, we computed several validation indicators (Milligan and Cooper 1985): Hubert's Gamma (Hubert and Arabie 1985) and Hubert's C[18]. Both indicated 6 clusters for female respondents and 4 for mothers (see state distribution plots in Appendices 3 and 4[19]). The clusters of female respondents employment histories can be summarized as follows: mostly in full-time jobs; mostly in part-time jobs; full-time jobs and then shift to inactivity about ten years after completing education; shift from part-time to full-time jobs between three and eight years after completing education; mostly inactive; full-time jobs interrupted by an inactivity spell around five to ten years after completing education. The four patterns of mothers occupational histories which are clustered can be crudely defined by the status held during most of the

career[20]: inactivity; lower-level occupations; self-employment; higher-level or intermediate occupations.

The aim now is not to identify the influence of the mother's employment history on that of her daughter, but to see whether we can identify preferential dyads among the lineages, i.e. recurrent sequences conditioned by the social cohesion between mother and daughter: an internal *structure* of transmission in the lineage concerning employment patterns.

We could stop here and simply cross-tabulate the typology of daughters' sequences with that of their mothers as determined above, producing a contingency table of 6x4=24 cells. Still a 24-cluster distribution is hard to describe[21] and some groupings would be needed: this result clearly lacks parsimony. While some cells have low numbers, not one is empty (Table 2), and there is no straightforward rule to allocate less frequent combinations of mother and daughter clusters to fuller cells. This is particularly problematic the aim is to give an exhaustive description of the population under study. Moreover, such a grouping would be based on the specific characteristics of the daughter's (respectively the mother's) individual employment histories, whose correlations would then be identified *ex-post*: the characteristics underlying mother-and-daughter dyads would be hidden. At least, this strategy shows that mothers and daughters' sequences are significantly associated[22]. But we need to use a method that specifically and parsimoniously focuses on the linked characteristics of the two histories.

TABLE 2

Cross-tabulation Of Mothers' Typology And Daughters' Typology

Number of dyads		Mothers				*Total*
		Mostly inactive	Mostly low	Mostly self	Mostly high/interm	
Daughters	Mostly FT	437	223	162	76	*898*
	Mostly PT	24	8	10	8	*50*
	From FT to inactivity	93	38	33	11	*175*
	From PT to FT	20	10	13	3	*46*
	Mostly inactive	129	24	23	10	*186*
	Interruption	30	15	10	3	*58*
Total		*733*	*318*	*251*	*111*	*1413*

Source: *Biographies et entourage* (2000)
Reference population: the 1413 female respondents and their mothers
Reading: low = lower-level occupations; high/interm. = higher-level or intermediate occupations; FT = full-time employment; PT = part-time employment

In our approach, we are not interested in synchronizing mothers and daughters, i.e. locally aligning mothers' and daughters' sequences within dyads, and the interdependence between mothers and daughters' histories is not focused on employment statuses at a given point in time. Indeed the daughter's situation at a given age is unlikely to be linked to her mother's situation at the same age, but rather it is the mother's whole employment history which is considered as part of the daughter's social background. The question of interest is more to identify underlying intergenerational transmission at work at individual lineage level: the individual translation of the global trend (the massive entry into the labor force) concomitant to the baby-boom. Besides, given the very different time windows (one is defined according to an initial event and lasts 15 years while the other is defined according to age and lasts 47 years), synchronizing mothers and daughters' sequences would be technically impractical[23]. For these reasons, the second strategy identified earlier, i.e. "Multi-channel sequence analysis" (Gauthier et al. 2010), would not be well-suited to handle this application. Rather, our objective is better approached by building a typology of pairs of mother-daughter employment histories that captures in detail the correspondence between some patterns of the

daughter's history and some of her mother's *taken as wholes*, i.e. emphasizing *global interdependence* and illustrating intergenerational transmission. Thus, we now apply "Globally Interdependent Multiple Sequence Analysis" (GIMSA) to the mothers' and daughters' employment histories from the *Biographies and entourage* survey[24].

For the first step of GIMSA, we use the dissimilarity matrices that were computed earlier for the separate typologies. We recall that two different dissimilarity measures are at work: Hamming distance for daughters and the Longest Common Subsequence metric for mothers[25].

Then mothers' and daughters' dissimilarity matrices are both submitted to MDS. In order to choose the number of MDS components to retain for the next stage of GIMSA, we have a look at *eigenvalues* and a stress function (Table 3).

TABLE 3

***Eigenvalues* And Stress Functions For Mothers And Daughters MDS Dimensions**

Dimension	Results for mothers		Results for daughters	
	Eigenvalue	Stress	Eigenvalue	Stress
1	1,000	0,467	1,000	0,373
2	0,600	0,253	0,282	0,203
3	0,322	0,119	0,199	0,136
4	**0,127**	0,094	**0,132**	**0,134**
5	**0,087**	**0,080**	**0,076**	0,151
6	0,068	0,081	0,057	0,166
7	0,055	0,087	0,042	0,177
8	0,051	0,095	0,033	0,187
9	0,039	0,103	0,030	0,197
10	0,030	0,109	0,025	0,204

Source: *Biographies et entourage* (2000)
Reference population: the 1,413 female respondents and their mothers

The stress function indicates that the first five dimensions are the most important for mothers, and the first four dimensions for daughters. The situation is less clear-cut for *eigenvalues*, but they seem to point to four or five dimensions for mothers and four or five for daughters. Using MDS sequence plots (Piccarreta and Lior 2010), the first four mothers' dimensions are easily to interpret as follows[26]: the first dimension contrasts inactivity with full-time employment; the second one contrasts low level occupations with other occupations; the third

one contrasts high and intermediate occupations with self-employment; the fourth one ranks mothers according to the length of their studies. Interpretation is less straightforward for daughters: while the first dimension clearly contrasts full-time jobs with inactivity and the second and third ones contrasts part-time jobs with the other statuses, the meaning of the following dimensions remains unclear.

Then comes the third stage of GIMSA: canonical PLS is computed on the MDS components (five for mothers and four for daughters), which leads to two sets of four components[27]. For the last stage of GIMSA, all these components are weighted according to the first option mentioned above – i.e. PLS components are weighted with the inverse of their variance (see section 3, step 4) - , a Euclidean distance matrix is computed and used as input for clustering (here Hierarchical Clustering Analysis with Ward criterion). Hubert's Gamma and Hubert's C both reach a local optimum for 10 clusters. Furthermore, these criteria are only guidelines, as the creation of a taxonomy in social sciences should be guided above all by background theories, heuristic views and a balance between parsimony and cluster homogeneity: "Classifications so produced can never be true or false, or even probable or improbable; they can only be profitable or unprofitable" (Williams and Lance 1965). Eventually, we opt for a 10-clusters solution.

4.3. *Results*

The typology of mother-daughter dyads resulting from this approach comprises 10 clusters, as shown in Table 4. This typology reveals a large diversity of employment and occupational histories, for mothers as well as for daughters. This diversity seems well-balanced between the two.

As compared to the fourth strategy, the GIMSA-typology leads to a lower number of clusters (10 *vs* 24 clusters). Still, as the fourth strategy builds clusters of mothers and daughters independently, its final typology comprises more homogeneous clusters. Indeed, this typology

explains 59.6% of the discrepancy in daughters' sequences (respectively 50.3% of mothers'), while the GIMSA-typology explains 45.6% (resp. 39.5%). However, the difference is not massive. Besides, the 6 patterns of "unidimensional" daughters' sequences can be found in the GIMSA-typology (see Table 2 for the pattern labels of "unidimensional" sequence analysis, i.e. the fourth strategy), as well as the 4 patterns of "unidimensional" mothers' sequences, and some new patterns appear with GIMSA (e.g. "from inactivity to low" among mothers).

In addition, GIMSA seems more efficient than the fourth strategy to describe the matrix of distances between mother-daughter dyads. Indeed, the GIMSA typology in 10 clusters explains a slightly smaller share of the discrepancy of this distance matrix than the typology in 24 clusters of the fourth strategy (37.0% *vs* 41.5%), but with a much lower number of clusters. Moreover, a GIMSA typology in 24 clusters would explain 50.9% of the discrepancy, i.e. significantly more than the fourth strategy. This is explained by the fact that with the fourth strategy, the linking of mothers and daughters is made from severely reduced information, i.e. in our example a typology in 4 (resp. 6) clusters, which explains only 50.3% (resp. 59.6%) of the discrepancy of mothers' (resp. daughters') sequences, as we have just seen. On the other side, with GIMSA, the linking of mothers and daughters is made from the MDS components, which retain a larger share of information: the 5 (resp. 4) MDS components of mothers (resp. daughters) explain 69.5% (resp. 73.4%) of the discrepancy of mothers' (resp. daughters') sequences. And this share could be further increased keeping more MDS components without a loss of parsimony in the final typology.

Finally, the balance between the number of clusters and the homogeneity of these clusters (i.e. the parsimony issue) seems favorable to GIMSA[28], as can be shown by close examination of the 10 clusters[29] (see state distribution plots in Appendix 6[30]; a cross-tabulation of the GIMSA typology and fourth strategy's typology are given in Appendix 7).

TABLE 4

Typology Of Mother-Daughter Employment Histories

Cluster	Main features of the dyads		N	%
	Mothers	**daughters**		
1	inactivity (or early shift from low to inactivity)	FT	276	19.5
2	self-employment	FT	224	15.9
3	Inactivity	from FT to inactivity (shift after 5 to 10 years)	211	14.9
4	Low	FT	173	12.2
5	from inactivity to low	FT	157	11.1
6	Inactivity	inactivity	100	7.1
7	high/interm.	mostly FT	95	6.7
8	Inactivity	interruption (back to work about 10 years after completing education)	81	5.7
9	Diverse	mostly PT employment	55	3.9
10	Diverse	shift from PT to FT (after around 5 to 10 years)	41	2.9
	Total		*1413*	*100.0*

Source: *Biographies et entourage* (2000)
Reference population: the 1,413 female respondents and their mothers
Reading: low = lower-level occupations; high/interm. = higher-level or intermediate occupations; FT = full-time employment; PT = part-time employment

Continuous inactivity is the most common pattern among mothers (it characterizes 4 out of 10 clusters, which together represent 47% of the sample), while a transition from school to continuous full-time employment is the most common among daughters (also 4 clusters, 59% of the sample). These clusters of female respondents' transition to continuous full-time employment account for 4 of the 5 largest clusters. In these, mothers are continuously inactive or shift to inactivity early in their life course (cluster 1), they are self-employed (cluster 2), they hold lower-level occupations (cluster 3) or they stay inactive for a while and then shift to lower-level occupations (cluster 4). Clusters 2 and 4 on one side, and cluster 5 on the other, contrast sharply in terms of the daughters' socio-demographic profiles (see Appendix 8). Indeed, while female respondents in clusters 2 and 4 tend to belong to older cohorts, have few qualifications and lower-level occupations at the time of the survey, and are often an only child, daughters in cluster 5 belong to younger cohorts, are often the eldest child and have the

highest level occupations. They have a relatively low level of qualification, however, which suggests upwards career mobility career[31].

The transmission of continuous inactivity is unusual, as it is found in cluster 6 which represents only 7% of the sample. Continuous inactivity is here associated with high fertility, as these respondents have the highest number of children at the time of the survey. Still, inactivity appears in other forms among daughters' early careers: in cluster 3, they shift from full-time employment to inactivity five to ten years after completing education, while in cluster 8 they interrupt their career for a given period, which may vary in duration but almost never ends later than 10 years after completing education. Like cluster 6, cluster 3 is characterized by the high fertility rates of the female respondents; these women also often belong to older cohorts and have low levels of qualification.

There is one cluster of upper-class mothers: they hold higher-level or intermediate occupations, and their daughters - whose transition is to continuous full-time - work during most of their first 15 years after completing education (cluster 7). Here, female respondents are relatively young, often hold higher-level or intermediate occupations at the time of the survey and have a high level of qualification, which emphasizes the fact that intergenerational transmission of high social status is observed on the side of female lineages as well.

Lastly, the two smallest clusters contrast female respondents whose early career comprises part-time employment: in cluster 9, they work part-time almost continuously, but shift from part-time to full-time after five to ten years in cluster 10. Their mothers have heterogeneous occupational careers. Female respondents in cluster 9 are the youngest of the sample, and they are relatively young in cluster 10 as well, which reflects the historically late appearance of part-time employment in France in the 1980s (Maruani 2000). These women are also highly qualified, which reflects the fact that part-time employment for these cohorts initially developed among well-educated women, whatever their mothers' careers.

To highlight an interesting point in these results, we can see that, while mothers' inactivity is often linked to daughters' inactivity (and mothers' activity to daughters' full-time employment), there are also significant patterns where the daughters of inactive mothers have full-time early employment careers: transmission does not automatically mean that mother and daughter follow identical paths.

4.4. *Robustness checks*

Applying GIMSA implies of series of methodological choices and it is important to assess the influence of these choices on the results. At the first step of the process, dissimilarity measures must be chosen. As stated in section 3 step 1, these are numerous and they have already been widely discussed in the social science literature: it is beyond our scope to investigate the characteristics and impact of one metric or another. More importantly, one of the advantages of GIMSA is that it allows a choice of metrics that specifically fit the theoretical questions and data limitations under study, in a separate way for each dimension.

The second step of GIMSA is an operation to reduce noise by retaining only a small number of MDS components. To assess the impact of this noise reduction step, we replicated the previous analysis retaining 20 MDS components for mothers and daughters (instead of five and four, respectively): noise reduction should be much weaker in this case. So we built a new ten-cluster typology, which we then compared to the previous one. The Rand index (Saporta and Youness 2002) is relatively high (0.649), which means the typologies are rather similar. However, looking more closely at the new typology, we can observe some differences. First, a very large group emerges, made of daughters with transition to full-time employment and heterogeneous mothers (they are inactive and/or in lower-level occupations). On the other hand, several very small clusters emerge, highlighting very marginal patterns: mothers who studied for many years, mothers who shift from inactivity to higher-level occupations, and two groups with daughters who return to education a few years after completing initial

education. These rare patterns are of limited interest compared to more regular profiles: the noise reduction step – via the selection of a smaller number of MDS components - seems to improve significantly the substantive quality of the results.

The third step of GIMSA does not imply any choices, as all PLS components are retained. The fourth step includes an important operation: the weighting of the two dimensions of dyads, so that their influence may be balanced in the building of the typology. As explained in section 3, several weighting schemes are possible. In our empirical application, we chose to weight PLS components by the inverse of their variance (w1). We replicated the analysis applying three alternative weighting schemes: no weighting at all (w0); mothers' (resp. daughters') PLS components weighted by the inverse of the number of distinct mothers' (resp. daughters') sequences in the sample (w2); or by the inverse of the first mothers' (resp. daughters') MDS *eigenvalue* (w3). The 10-cluster typologies are compared with the Rand index:

TABLE 5

Rand Indices For Various Weighting Schemes

	w0	w1	w2	w3
w0	0.000			
w1	0.850	0.000		
w2	0.869	0.860	0.000	
w3	0.775	0.804	0.821	0.000

All the Rand indices are high, ranging from 0.775 to 0.869, suggesting rather similar typologies: the results are robust to different schemes of weighting. When we examine more carefully the various typologies, a few comments may be added. First, the typology with no weighting comprises many clusters where daughters have comparable patterns: variety is significantly higher on the mother's side. Thus the balance between mothers' and daughters' dimensions is not satisfying, which emphasizes the added-value of using a weighting strategy. Moreover, every major pattern comes to light whether w1, w2 or w3 is employed. The

differences concern small clusters and minority patterns, which are more or less aggregated according to the weighting scheme. While w1 distinguishes two clusters with part-time employed daughters, w2 only gives one but contrasts three patterns with daughters who shift from full-time jobs to inactivity according to the age of the shift. w3 identifies three groups with daughters in part-time work: one where women shifts from full-time to part-time, another with the opposite shift, and the last one with continuous part-time employment.

5. CONCLUSION

In this paper, we examined the intergenerational pattern of women's careers within lineages, pairing the employment histories of mothers and daughters. Using the rich data from the *Biographies et entourage* survey (INED, 2000), from which we can track respondents' employment and occupational careers and also those of their mothers, we applied an approach we called "Globally Interdependent Multiple Sequence Analysis" (GIMSA) to make a typological analysis of mother-daughter employment histories: this method combines standard sequence analysis (e.g. OMA), Multidimensional Scaling, Canonical PLS and clustering techniques.

Intentionally we devised an application example more suited for discussion of the methodology than its substantive sociological output. Nevertheless the first results presented here are promising. They open perspectives for studying long term trends and understanding specific intra-family features and continuities that contribute to the overall macro changes in women's involvement in paid activity. The typologies obtained shed light on intergenerational transmission, leaving aside a mechanical determinism and showing the relative multiplicity of career pathways open to children starting from similar parental background in terms of their mother's labor market participation: some never employed mothers have daughters with incomplete careers, but others' daughters always work full-time, and so on. The differences

depend on characteristics such as educational background, cohort or birth order among siblings.

From a methodological point of view, GIMSA provides a flexible way to uncover patterns of dyads of sequences. It is parsimonious and takes multidimensionality into account, i.e. each dimension's contribution remains explicit and can be specifically parameterized. Moreover, GIMSA emphasizes *global interdependence* between sequences within dyads, i.e. the relationship between *sequences taken as wholes*. It presents very few constraints about the data: sequences within the dyads do not have to be synchronized or to have the same nature: they may have completely different alphabets, time windows and time units, and distinct aspects of temporality (e.g. order *vs* timing) can be emphasized separately for each dimension. Besides the theoretical advantages that this flexibility provides, this also means that the quality of the data does not have to be comparable between the dimensions: detailed information about respondents' histories and cruder material about their parents can be used together. Combining OMA with Euclidean tools, GIMSA offers a straightforward and computationally efficient multidimensional approach to sequence pattern mining, complementary to existing strategies like "Multi-channel sequence analysis" (Gauthier et al. 2010): the choice of one method or the other will depend on theoretical and data issues. While MCSA is best suited to analyze the interdependence between various dimensions at each point in time (i.e. *local interdependence*), GIMSA favors *global interdependence*.

GIMSA may be used for distinct purposes, as may clustering approaches in general. Indeed, when performing an in-depth exploration of the data, one tends to split the data into a large number of clusters in order to identify homogeneous patterns and reflect the diversity embedded in the sequences. On the other hand, to summarize the heterogeneity of the sequences, it is preferable to use a smaller number of clusters, which renders further analysis

manageable, as we did in this paper. This simplification of the data is not specific to GIMSA; it is a procedure the users undertake in relevance with their research question.

Still, GIMSA focuses on a specific kind of data, e.g. linked life courses. This method allows to identify patterns of sequences, but does not provide the degree of association between the various dimensions of the sequence data. It will find clusters whatever the degree of association between dimensions. To avoid drawing conclusions from weakly associated dimensions, it may be useful to use strategy 4 and an association index (see note xx) prior to GIMSA, to assess whether pattern searching is relevant or not. Moreover, once GIMSA has been performed, one may complement it with intra-cluster homogeneity measures (e.g. average intra-cluster sequence dissimilarity). Indeed, in a given typology, a cluster may result from a strong association between dimensions, and then sequences will then be homogeneous within each dimension, while another cluster may present only a weak association, and then sequences will probably be considerably more homogeneous in a dimension than in the other. In this view, sequence homogeneity measures, as well as sequence plots, are useful guidelines to identify variations in the strength of the link between dimensions.

One could argue that sequence analysis has not produced any blockbuster applications yet, as Abbott (2000) stated some years ago. Still, blockbuster applications are rare whatever the methodological approach, and the relevant issue is whether sequence analysis helps to better understand some parts of the social world. For more than 25 years, sequence analysis has played a very positive role in helping to understand the complexity of life courses and careers. From this angle, GIMSA provides an additional element in the inherited toolbox. As illustrated by our application, GIMSA may offer a different viewpoint on social mobility. Indeed, mobility is often analyzed by comparing social positions at a given point in time. But this gives only a partial view of an individual’s position, which could be better captured by examining his or her trajectory (or a part of it). That is why comparing multiple sequences

instead of single states or events may be a fruitful avenue for research on social mobility and intergenerational transmission. For instance, it is conceivable to explore the global interdependence between siblings' and parents' life courses or careers all together, whether trajectories have different time windows or not (for instance with children's sequences focusing on school-to-work transition and parents' on the whole career).

The range of potential applications for GIMSA thus extends well beyond intergenerational social mobility studies and intergenerational transmission in general. They can be classified on the basis of several dichotomies. First, the entities associated with the sequences under study can be human (individuals) or non-human (e.g. nations, firms, etc.). Second, the various dimensions that make up the multi-dimensional sequences can characterize a single entity (e.g. an individual) or more entities, as in this article where one dimension characterizes the mother and another dimension characterizes her daughter. The latter case can be extended to the relationship between parents and children in general, but also to siblings or peer groups (friends, colleagues, etc.), for instance. In addition, the various dimensions of the sequences may correspond to trajectories whose nature is either similar (as is the case here, where the careers of mothers and their daughters are analyzed) or different (for example, when comparing family and employment histories). From the perspective of temporality, the various dimensions of the sequences can describe contemporaneous trajectories, in terms of age or historical period, or asynchronous ones. In our example, the dimensions are asynchronous, insofar as they describe the school-to-work transition for daughters and the entire career for mothers. Again, the dimensions may or may not have the same time unit (year, month, etc.) and/or the same length. Note that in the case of contemporaneous dimensions, MCSA is probably more appropriate than GIMSA. Finally, the sequences may have two dimensions, as in this paper, or a larger number. In the latter case, canonical PLS should be replaced by an

alternative factor analysis technique, such as Multiple Factor Analysis (Escofier and Pagès 1994).

REFERENCES

Aassve, Arnstein, Francesco C. Billari, and Raffaella Piccarreta. 2007. “Strings of adulthood: a sequence analysis of young British women's work-family trajectories.” *European Journal of Population* 23:369-388.

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

Abbott, Andrew, and Emily Barman. 1997. “Sequence comparison via alignment and Gibbs sampling.” *Sociological Methodology* 27:47-87.

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 ressemblance in sequence data: 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 & Research* 29:3-33.

Aisenbrey, Silke, and Anette E. Fasang. 2010. “New Life for Old Ideas: The ‘‘Second Wave’’ of Sequence Analysis Bringing the ‘‘Course’’ Back Into the Life Course.” *Sociological Methods & Research* 38:420-462.

Barrère-Maurisson, Marie-Agnès. 1992. *La division familiale du travail. La vie en double*. Paris : PUF.

Beller, Emily. 2009. “Bringing Intergenerational Social Mobility Research into the Twenty-first Century: Why Mothers Matter.” *American sociological review* 74:507-528.

Biemann, Torsten. 2011. “A transition-oriented approach to optimal matching.” *Sociological Methodology* 41:195-221.

Blair-Loy, Mary. 1999. “Career patterns of executive women in finance: an optimal matching analysis.” *American Journal of Sociology* 104:1346-1397.

Blanchard, Philippe. 2005. “Multi-dimensional biographies. Explaining disengagement through sequence analysis.” Presented at the 3rd ECPR Conference, Budapest, Hungary.

Blanchard, Philippe. 2010. *Analyse séquentielle et carrières militantes*. (http://halshs.archives-ouvertes.fr/hal-00476193/)

Bry, Xavier. 1996. *Analyses factorielles multiples*. Paris : Economica Poche.

Brzinsky-Fay, Christian, and Ulrich Kohler. 2010. “New Developments in Sequence Analysis.” *Sociological Methods & Research* 38:359–364

Chaloupkova, Jana. 2010. “The De-standardisation of Early Family Trajectories in the Czech Republic: A Cross-cohort Comparison.” *Czech Sociological Review* 46:427-451.

Courgeau, Daniel, and Eva Lelièvre. 1992. Event History Analysis in Demography. Oxford: Clarendon.

De Jong, Sijmen, Barry M. Wise, and N. Lawrence Ricker. 2001. “Canonical partial least squares and continuum power regression.” *Journal of Chemometrics* 15:85-100.

Dijkstra, Wil, and Toon Taris. 1995. “Measuring the Agreement between Sequences.” *Sociological Methods & Research* 24:214-231.

Elzinga, Cees H. 2003. “Sequence similarity: a nonaligning technique.” *Sociological Methods & Research* 32:3-29.

Elzinga, Cees H. 2008. “Sequence analysis: Metric representations of categorical time series.” *Technical report*. Department of Social Science Research Methods, Vrije Universiteit, Amsterdam.

Elzinga, Cees H., and Aart C. Liefbroer 2007. “De-standardization of Family-Life Trajectories of Young Adults: A Cross-National Comparison Using Sequence Analysis.” *European Journal of Population* 23:225-250.

Escofier, Brigitte, and Jérôme Pagès. 1994. “Multiple factor analysis (AFMULT package).” *Computational statistics & data analysis* 18:121-140.

Falcon, Julie. 2012. “Behind the black box: the effect of intragenerational social mobility on intergenerational social mobility.” Presented at the Lausanne Conference On Sequence Analysis, Lausanne, Switzerland.

Fasang, Anette E., and Marcel Raab. 2014. “Beyond Transmission: Intergenerational Patterns of Family Formation Among Middle-Class American Families.” *Demography* published online.

Gabadinho, Alexis, Gilbert Ritschard, Nicolas S. Müller, and Matthias Studer. 2011. “Analyzing and visualizing state sequences in R with TraMineR.” *Journal of Statistical Software* 40:1-37.

Gauthier, Jacques-Antoine, Éric D. Widmer, Philipp Bucher, and Cédric Notredame. 2010. “Multichannel sequence analysis applied to social science data.” *Sociological Methodology* 40:1-38.

GRAB. 2009. *Biographies d'enquêtes : bilan de 14 collectes biographiques*. Paris: INED. (http://grab.site.ined.fr/fr/editions_en_ligne/biographies_enquetes/)

Grelet, Yvette. 2002. “Des typologies de parcours. Méthodes et usages. ” *Document Génération 92* 20:1-47.

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.

Han, Shin-Kap, and Phyllis Moen. 1999. “Clocking out: temporal patterning of retirement.” *American Journal of Sociology* 105:191-236.

Hubert, Lawrence, and Phipps Arabie. 1985. “Comparing Partitions.” *Journal of Classification* 2:193–218.

Kruskal, Joseph B., and Myron Wish. 1984. *Multidimensional Scaling*. Beverly Hills: Sage.

Lelièvre, Eva, and Géraldine Vivier. 2001. “Evaluation d'une collecte à la croisée du quantitatif et du qualitatif : l'enquête Biographies et entourage.” *Population* 56:1043-1073.

Lelièvre, Eva, and Nicolas Robette. 2010. “A Life Space Perspective to Approach Individual Demographic Processes.” *Canadian Studies in Population* 37:207-244.

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. 2010. “Setting Cost in Optimal Matching to Uncover Contemporaneous Socio-Temporal Patterns.” *Sociological Methods & Research* 38:389-419.

Lesnard, Laurent, and Man Yee Kan. 2011. “Investigating scheduling of work: a two-stage optimal matching analysis of workdays and workweeks.” *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 174:349-368.

Liefbroer, Aart C., and Cees H. Elzinga. 2012. “Intergenerational Transmission of Behavioural Patterns: How Similar are Parents’ and Children’s Demographic Trajectories?” *Advances in Life Course Research* 17:1-10.

MacIndoe, Heather, and Andrew Abbott. 2004. “Sequence analysis and optimal matching

techniques for social science data." pp. 387-406 in *Handbook of Data Analysis*, edited by M. Hardy and A. Bryman. London: Sage.

Maruani, Margaret. 2000. *Travail et emploi des femmes*. Paris: La Découverte.

Milligan, Glenn W., and Martha C. Cooper. 1985. "An examination of procedures for determining the number of clusters in a data set." *Psychometrika* 50(2):159–179.

Nakache, Jean-Pierre, and Josiane Confais. 2005. *Approche pragmatique de la classification*. Ed. Technip.

Oh, Man-Suk, and Adrian E. Raftery. 2001. "Bayesian Multidimensional Scaling and Choice of Dimension." *Journal of the American Statistical Association* 96:1031-44.

Piccarreta, Raffaella, and Orna Lior. 2010. "Exploring sequences: a graphical tool based on multi-dimensional scaling." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 173:165-184.

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

R Development Core Team. 2013. *R: a Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. (Available from http://www.R-project.org.)

Robette, Nicolas. 2010. "The diversity of pathways to adulthood in France: evidence from a holistic approach." *Advances in Life Course Research* 15 :89-96.

Robette, Nicolas, and Xavier Bry. 2012. "Harpoon or bait? A comparison of various metrics to fish for life course patterns." *Bulletin of Sociological Methodology* 116:5-24.

Robette, Nicolas, and Nicolas Thibault. 2008. "Comparing qualitative harmonic analysis and optimal matching. An exploratory study of occupational trajectories." *Population-E* 64:621-646.

Robette, Nicolas, Anne Solaz, and Ariane Pailhé. 2009. "Work and family over the life-cycle: a typology of couples." Presented at the XXVIth IUSSP International Population Conference, Marrakech, Marocco.

Salmela-Aro, Katariina, Noona Kiuru, Jari-Erik Nurmi, and Mervi Eerola. 2011. "Mapping pathways to adulthood among Finnish university students: Sequences, patterns, variations

in family- and work-related roles." *Advances in Life Course Research* 16:25-41.

Saporta, Gilbert, and Genane Youness. 2002. "Comparing Two Partitions: Some Proposals and Experiments." Pp. 243-248 in *Compstat. Proceedings in Computational Statistics*. Physica-Verlag HD.

Shepard, Roger N. 1962. "The Analysis of Proximities: Multidimensional Scaling with an Unknown Distance Function, II." *Psychometrika* 27:219-46.

Stovel, Katherine, Michael Savage, and Peter Bearman. 1996. "Ascription into achievement: models of career systems at Lloyds Bank, 1890-1970." *American Journal of Sociology* 102:358-399.

Studer, Matthias. 2012. *Étude des inégalités de genre en début de carrière académique à l'aide de méthodes innovatrices d'analyse de données séquentielles*. Thèse SES 777, Faculté des sciences économiques et sociales, Université de Genève.

Vandeschelden, Mélanie. 2006. "Homogamie socioprofessionnelle et ressemblance en termes de niveau d'études : constat et évolution au fil des cohortes d'unions." *Economie et statistique* 398-399:33-58.

Williams, W.T. and G.N. Lance. 1965. "Logic of computer-based intrinsic classifications." *Nature* 207(4993): 159-161.

Wilson, Clarke. 1998. "Activity pattern analysis by means of sequence-alignment methods." *Environment and Planning A* 30:1017-1038.

Wu, Lawrence L. 2000. "Some comments on "Sequence analysis and optimal matching methods in sociology: Review and prospect"." *Sociological Methods & Research* 29:41-64.

APPENDIX 1. Multidimensional Scaling (MDS)

Given the matrix of distances $\|\xi_i - \xi_j\|$ between *n* points $\{\xi_i ; i = 1,n\}$ in a euclidean space, MDS provides a means to rebuild the image of the unit scatterplot in the basis of its principal components:

■ Finding the scalar product matrix of vectors centered on their centroid:

Centering vectors on their centroid: let $\bar{\xi} = \frac{1}{n}\sum_{i=1}^{n} \xi_i$ and $x_i \overset{\text{def.}}{=} \xi_i - \bar{\xi}$ $\forall i = 1,n$.

Then:

$$\forall i, j = 1,n : \|\xi_i - \xi_j\|^2 = \|x_i - x_j\|^2 = \|x_i\|^2 + \|x_j\|^2 - 2\langle x_i \mid x_j \rangle$$

$$\Leftrightarrow \langle x_i \mid x_j \rangle = \frac{1}{2}\left(\|x_i\|^2 + \|x_j\|^2 - \|x_i - x_j\|^2\right) (1)$$

Besides, in view of the Koenig equality applied to $\{x_i ; i = 1,n\}$ with centroid $\bar{x} = 0$:

$$\forall i = 1,n : \sum_{j=1}^{n} \frac{1}{n}\|x_i - x_j\|^2 = \|x_i - 0\|^2 + \sum_{j=1}^{n} \frac{1}{n}\|x_j - 0\|^2 \ (2_i)$$

Let $D_0 = \sum_{j=1}^{n} \frac{1}{n}\|x_j\|^2$ and $D_i = \sum_{j=1}^{n} \frac{1}{n}\|x_i - x_j\|^2$. From ($2_i$), we have:

$$\forall i = 1,n : \|x_i\|^2 = D_i - D_0 \ (3)$$

Summing up equations (2_i) over *i* and dividing by *n*, we get:

$$\frac{1}{n^2}\sum_{j=1}^{n} \|x_i - x_j\|^2 = 2D_0 \Leftrightarrow D_0 = \frac{1}{2n^2}\sum_{j=1}^{n} \|x_i - x_j\|^2 \ (4)$$

So, from (4) and (3), draw every $\|x_i\|^2$, and then, from (1), every $\langle x_i \mid x_j \rangle$.

■ Finding the principal components:

Let *S* be the matrix $(\langle x_i \mid x_j \rangle)_{i,j=1,n}$. Let λ_k denote the k^{th} eigenvalue in decreasing order and v_k be the associated unit norm eigenvector. Then, the principal components f^k of $\{x_i ; i = 1,n\}$ are:

$$f^k = \sqrt{n\lambda_k}\, v_k$$

APPENDIX 2. Symmetric PLS

Given two data matrices $X(n,p)$ and $Y(n,q)$ containing respectively p and q numeric variables describing the same n statistical units, the purpose of symmetric PLS is to extract two sequences of uncorrelated components $\left\{ f^k, k = 1, K \right\}$ and $\left\{ g^k, k = 1, K \right\}$, such that, $\forall k$:

- f^k (respectively g^k) belongs to the space spanned by X's (resp. Y's) columns;
- f^k (respectively g^k) captures as much as possible of X's (resp. Y's) variance unaccounted for by previous components;
- f^k and g^k are as correlated as possible.

Such components are extracted through the following algorithm.

Rank 1 components:

Let $f^1 = Xu_1$ with $\|u_1\| = 1$; $g^1 = Yv_1$ with $\|v_1\| = 1$

Vectors u^1 and v^1 are the solutions of the following program :

$$\mathbf{Q}(X,Y): \max_{\substack{u \in \mathbb{R}^p,\, u'Pu=1 \\ v \in \mathbb{R}^q,\, v'v=1}} cov(f,g) \Leftrightarrow \max_{\substack{u \in \mathbb{R}^p,\, u'Pu=1 \\ v \in \mathbb{R}^q,\, v'v=1}} \langle Xu | Yv \rangle_P \text{ where } P = \frac{1}{n} I$$

$$L = v'Y'PXu - \lambda(u'u - 1) - \mu(v'v - 1)$$

$$\nabla_u L = 0 \Leftrightarrow X'PYv = 2\lambda u \ (1) \ ; \ \nabla_v L = 0 \Leftrightarrow Y'PXu = 2\mu v \ (1')$$

u' (1) and v' (1') give:

$$u'X'PYv = 2\lambda u'u = 2\lambda \ ; \ v'Y'PXu = 2\mu v'v = 2\mu$$

$$\Rightarrow \max_{\{u,v\}} \{f'g : \|u\| = \|v\| = 1\} = \sqrt{\eta}$$

which implies that η be maximum.

Besides:

$$(1,1') \Rightarrow X'PYY'PXu = \eta u \ (2) \ ; \ Y'PXX'PYv = \eta v \ (2')$$

So, the solution vector u (resp. v) is the eigenvector characterized by (2) (resp. (2')) associated with the largest eigenvalue.

Rank k>1 components:

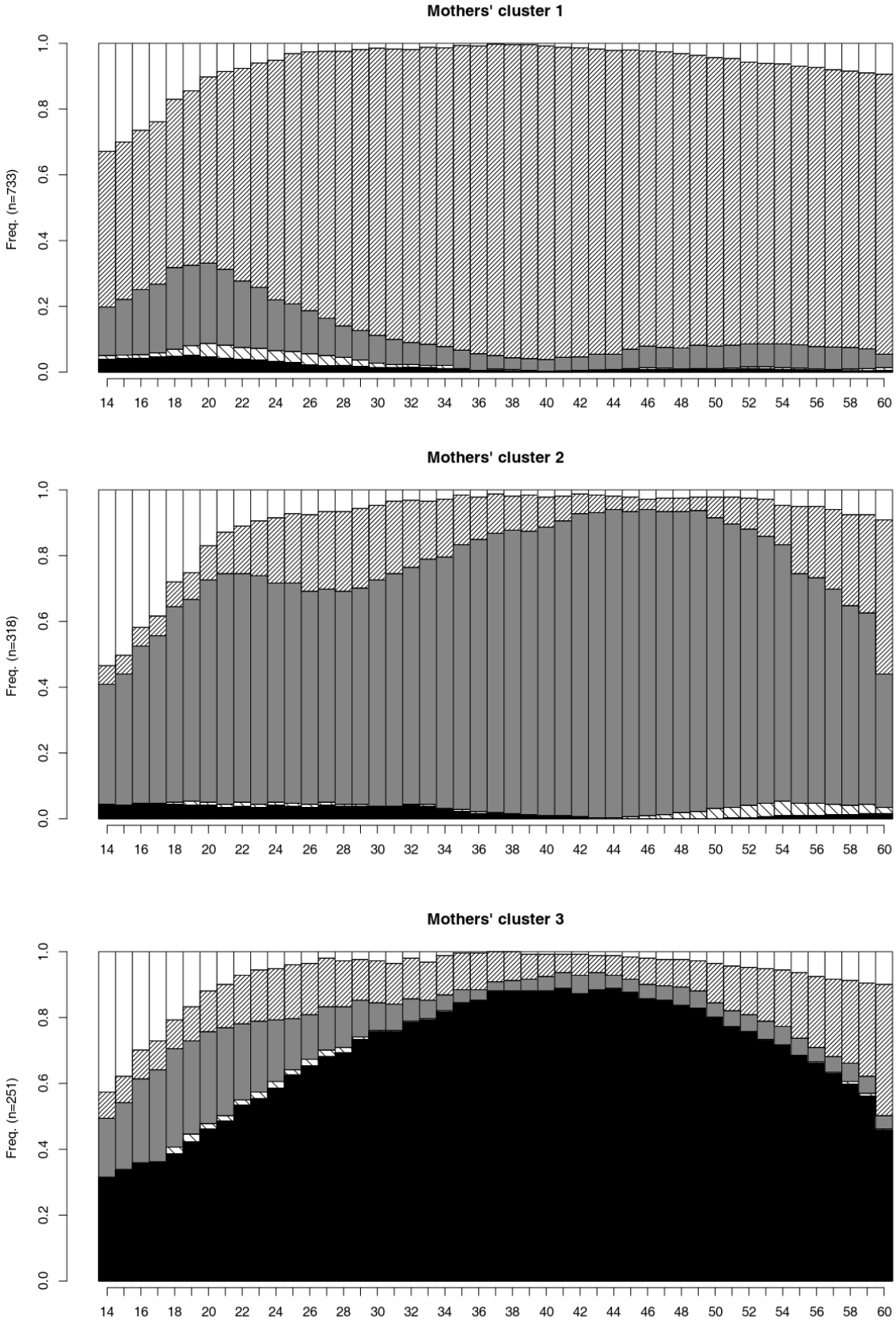
Rank k component f^k (resp. g^k) must be uncorrelated to the former rank ones f^1, … f^{k-1} (resp. g^1, … g^{k-1}). To ensure that, we define:

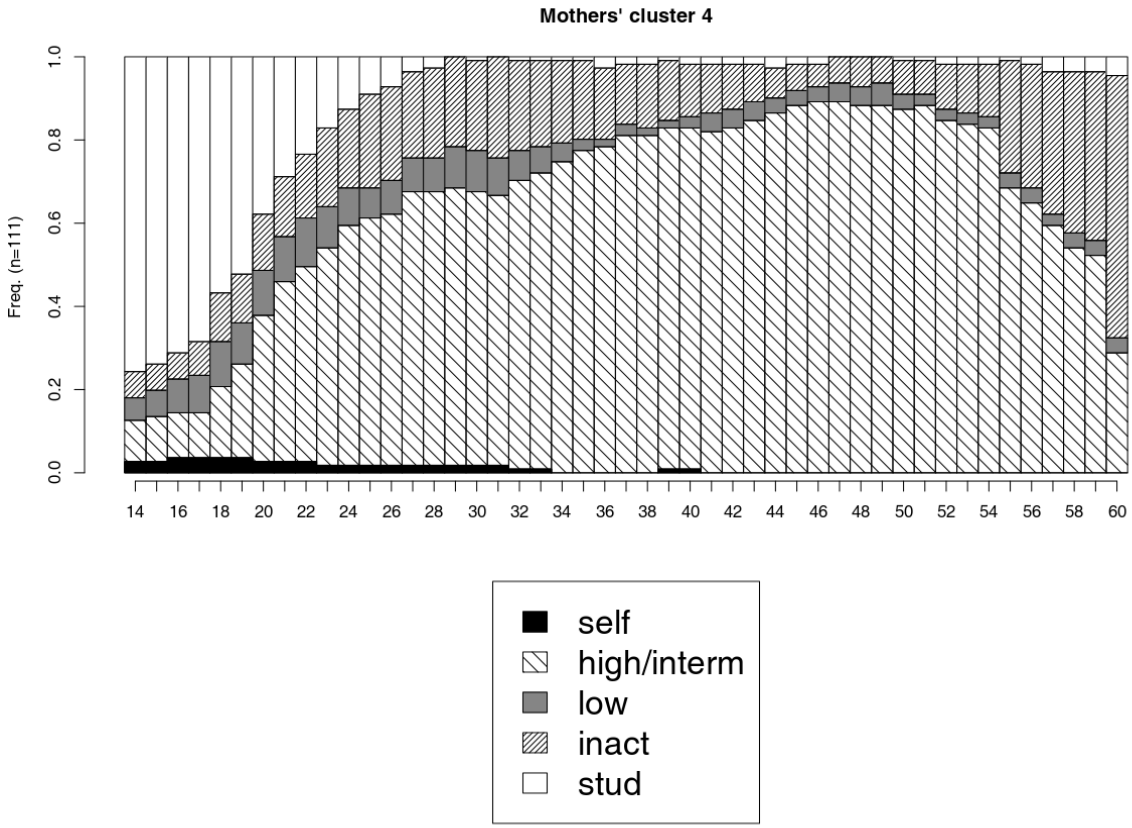
$$X_0 = X; Y_0 = Y \text{ and } \forall k > 1: X_k = \Pi_{\langle f^k \rangle^\perp} X_{k-1}, Y_k = \Pi_{\langle g^k \rangle^\perp} Y_{k-1}$$

To put it more statistically, X_k (resp. Y_k) is made of the residuals of X_{k-1} (resp. Y_{k-1}) regressed on f^k (resp. g^k). Then, we look for:

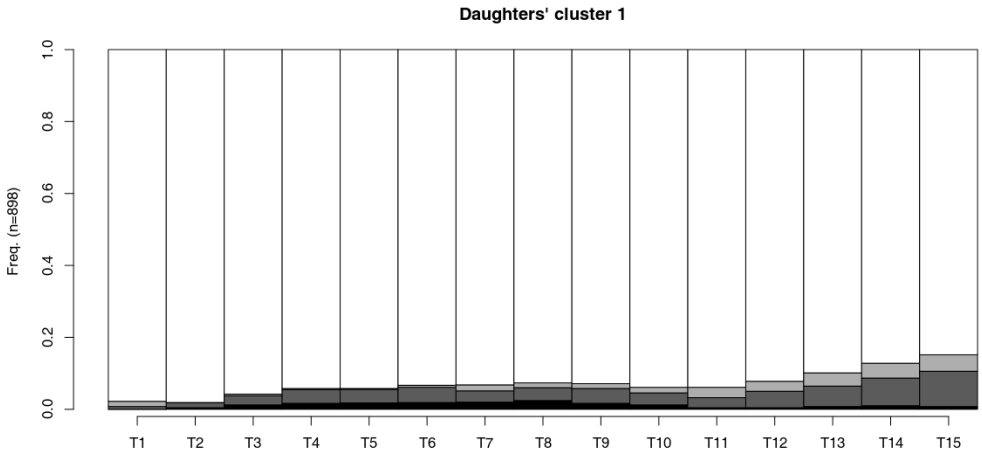
$$(u_k, v_k) = \text{sol. of } \mathbf{Q}(X_{k-1}, Y_{k-1})$$

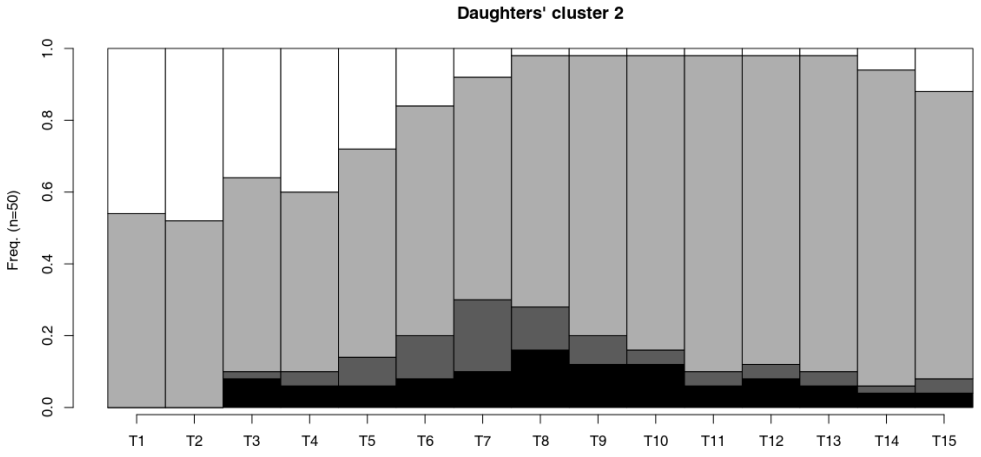
APPENDIX 3. State Distribution Plots For Mothers' Clustering

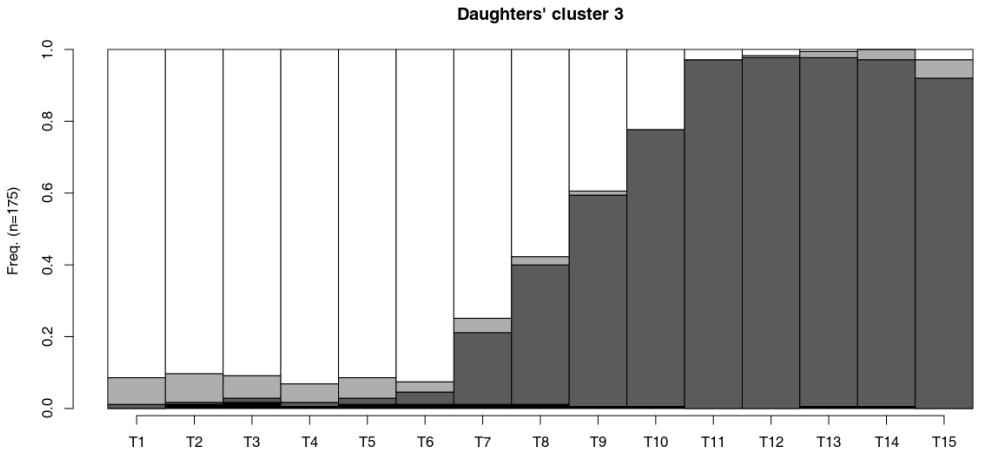


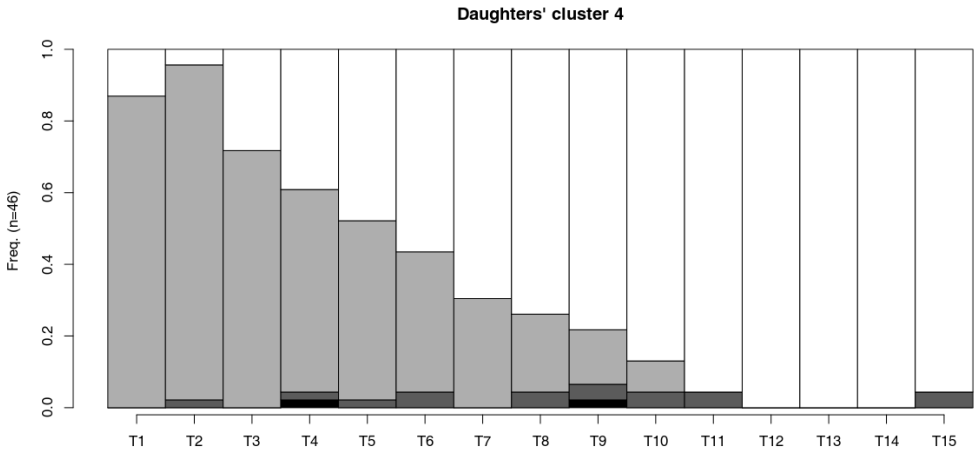

Mothers' cluster 4
Freq. (n=111)
0.0
0.2
0.4
0.6
0.8
1.0
14
16
18
20
22
24
26
28
30
32
34
36
38
40
42
44
46
48
50
52
54
56
58
60
self
high/interm
low
inact
stud

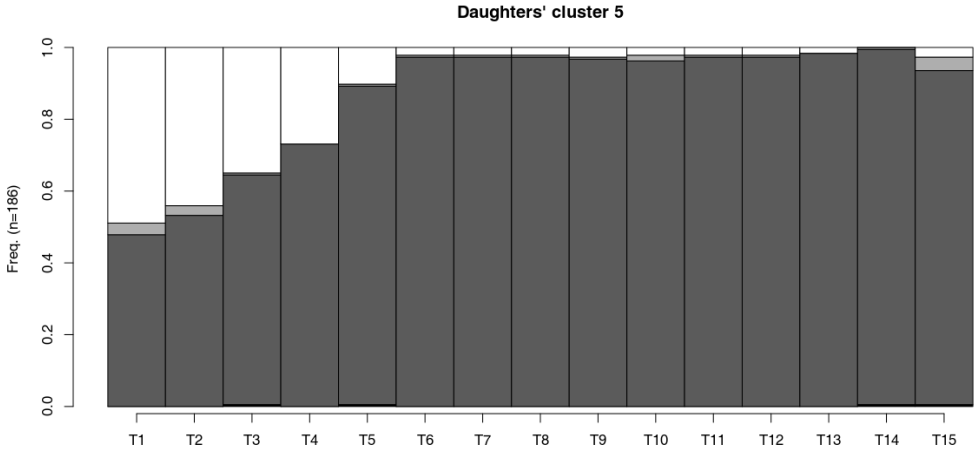
APPENDIX 4. State Distribution Plots For Daughters' Clustering

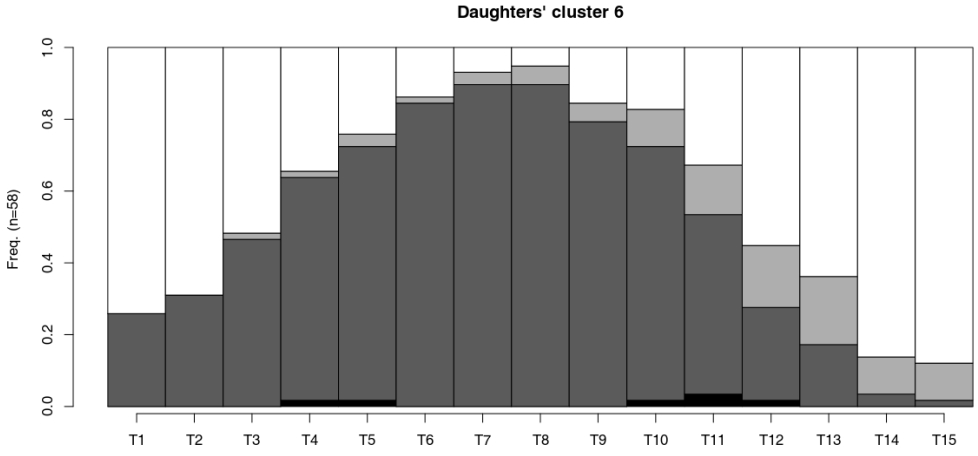


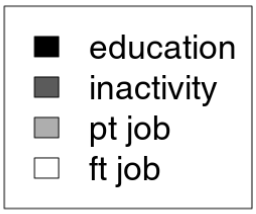





Daughters' cluster 4
Freq. (n=46)
0.0
0.2
0.4
0.6
0.8
1.0
T1
T2
T3
T4
T5
T6
T7
T8
T9
T10
T11
T12
T13
T14
T15


Daughters' cluster 5
Freq. (n=186)
0.0
0.2
0.4
0.6
0.8
1.0
T1
T2
T3
T4
T5
T6
T7
T8
T9
T10
T11
T12
T13
T14
T15


Daughters' cluster 6
Freq. (n=58)
0.0
0.2
0.4
0.6
0.8
1.0
T1
T2
T3
T4
T5
T6
T7
T8
T9
T10
T11
T12
T13
T14
T15

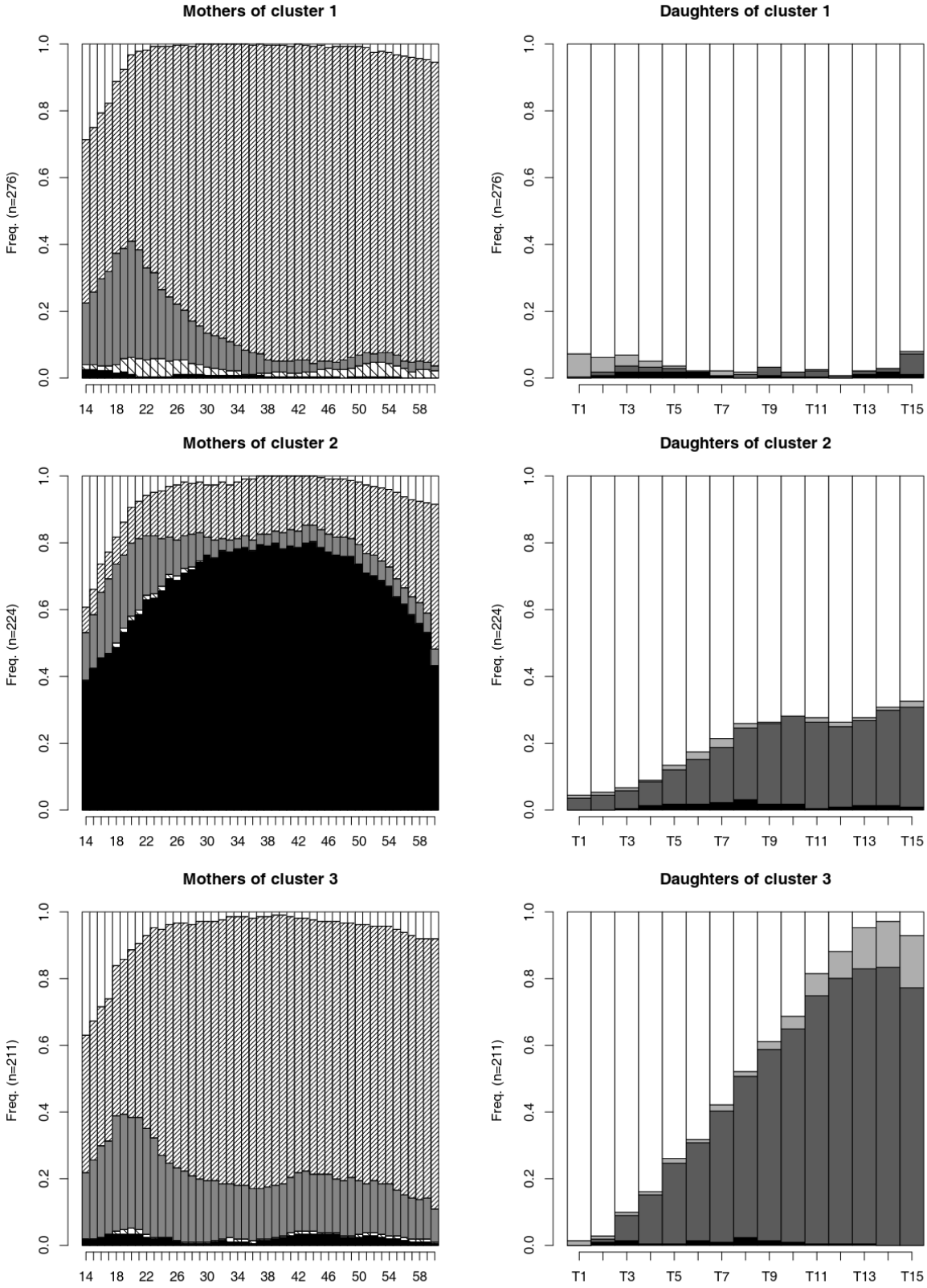

education
inactivity
pt job
ft job

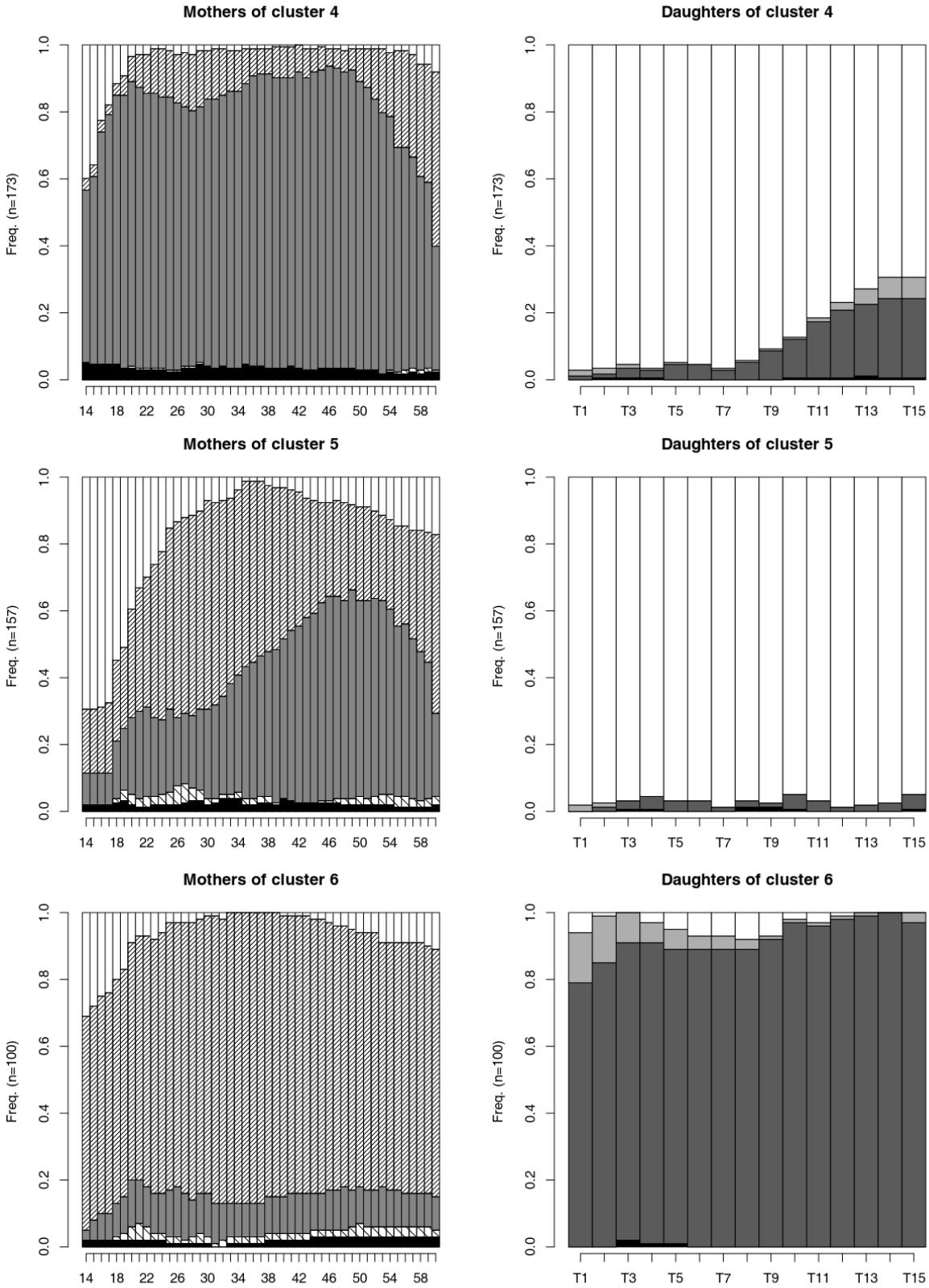
APPENDIX 5. R Program For The GIMSA Application

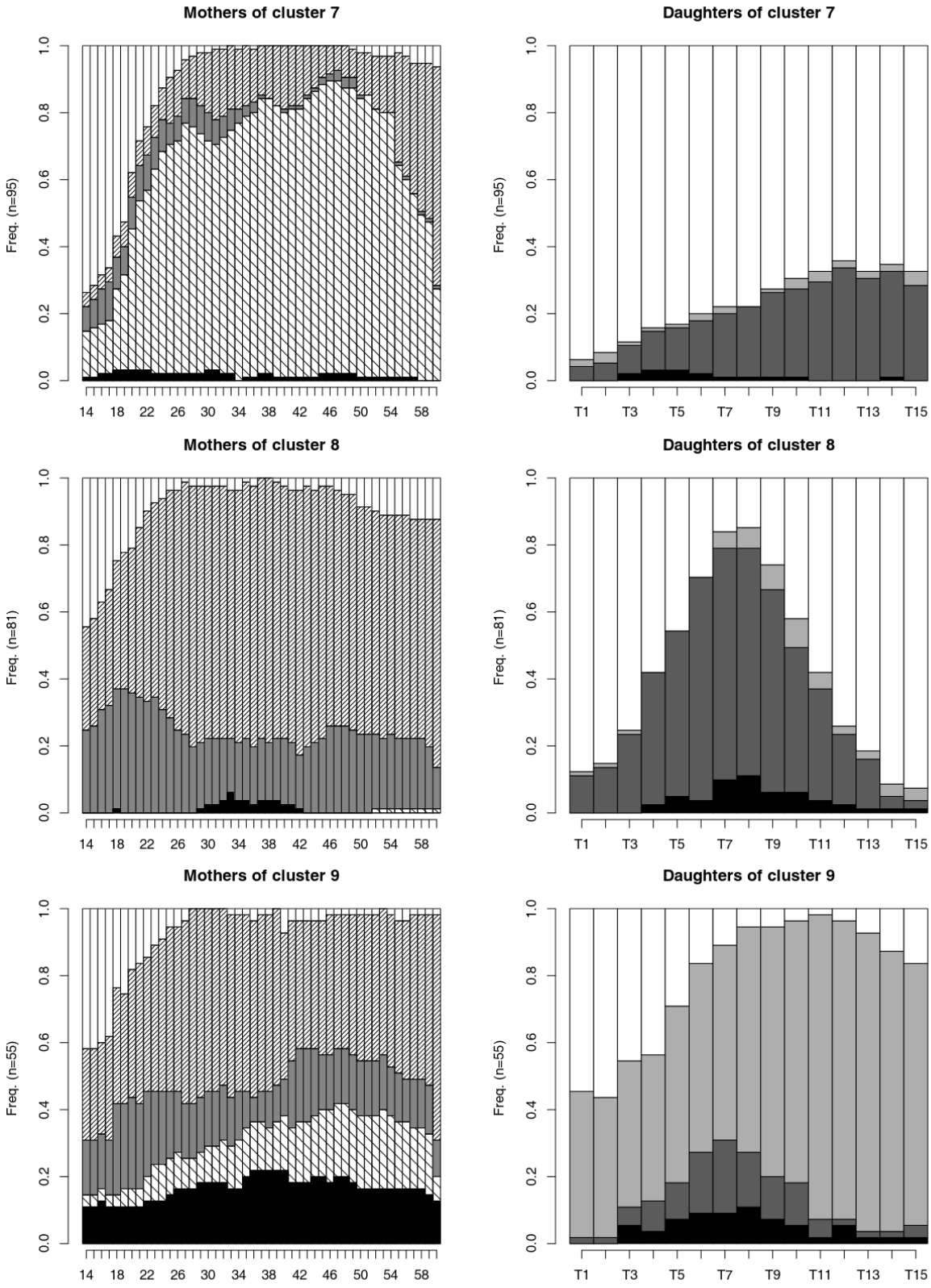
```
## ============
## Data import, etc
## ============

library(TraMineR)
library(cluster)

seqmother <- read.table("mydata_mother.txt")
seqdaughter <- read.table("mydata_daughter.txt")


## ===============
## Sequence definition
## ===============

mother.lab <- c("indep","high/mid lev","low lev","inact","stud")
mother.seq <- seqdef(seqmother,lab=mother.lab)

daughter.lab <- c("education","inactivity","pt job","ft job")
daughter.seq <- seqdef(seqdaughter,lab=daughter.lab)


## ============================
## 1st GIMSA step: dissimilarity measures
## ============================

mother.om <- seqdist(mother.seq,method="LCS")

daughter.cost <- seqsubm(daughter.seq,method="CONSTANT",cval=2)
daughter.om <- seqdist(daughter.seq,method="HAM",sm=daughter.cost)


## ============================
## 2nd GIMSA step: Multidimensional Scaling
## ============================

mother.mds <- cmdscale(mother.om,k=20,eig=TRUE)
daughter.mds <- cmdscale(daughter.om,k=20,eig=TRUE)

# To choose the number of MDS components to retain
stress <- function(omres,mdsres) {
  datadist <- as.dist(omres)
  res <- numeric(length=ncol(mdsres$points))
  for(i in 1:ncol(mdsres$points)) {
    fitteddist <- dist(mdsres$points[,1:i],diag=TRUE,upper=TRUE)
    res[i] <- sqrt(sum((datadist-fitteddist)^2)/sum(datadist^2))
    }
  res
}
stress(mother.om,mother.mds)
mother.mds$eig[1:10]/mother.mds$eig[1]
seqIplot(mother.seq,sort=mother.mds$points[,1])
stress(daughter.om,daughter.mds)
daughter.mds$eig[1:10]/daughter.mds$eig[1]
seqIplot(daughter.seq,sort=daughter.mds$points[,1])

nbmds.mother <- 5
nbmds.daughter <- 4


## ================================
## 3rd GIMSA step: symetrical (ie canonical) PLS
## ================================

a <- mother.mds$points[,1:nbmds.mother]
b <- daughter.mds$points[,1:nbmds.daughter]

symPLS <- function(a,b) {
  k <- min(ncol(a),ncol(b),nrow(a),nrow(b))
  X <- vector("list", k)
  Y <- vector("list", k)
  X[[1]] <- scale(a,scale=FALSE)
  Y[[1]] <- scale(b,scale=FALSE)
```

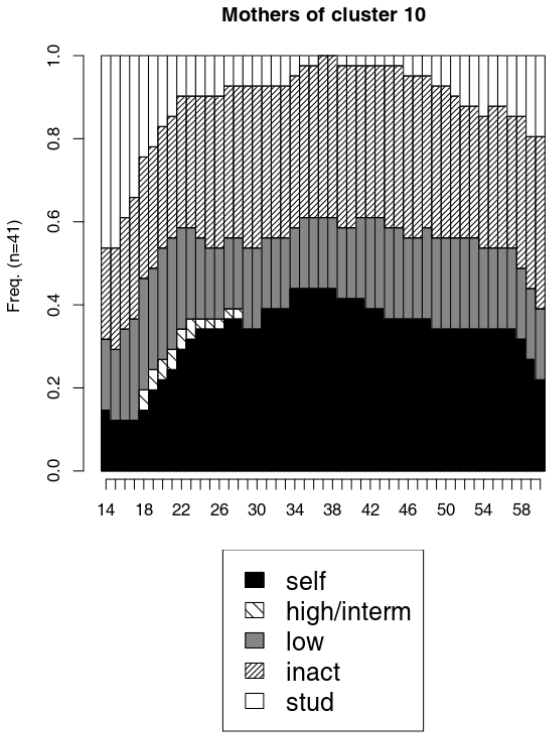
```
  F <- matrix(nrow=nrow(X[[1]]), ncol=k)
  G <- matrix(nrow=nrow(X[[1]]), ncol=k)
  f <- matrix(nrow=nrow(X[[1]]), ncol=k)
  g <- matrix(nrow=nrow(X[[1]]), ncol=k)
  vF <- vector(mode="numeric", length=k)
  vG <- vector(mode="numeric", length=k)
  corr <- vector(mode="numeric", length=k)
  for(i in 1: k) {
              u <- eigen(t(X[[i]])%*%Y[[i]]%*%t(Y[[i]])%*%X[[i]])$vectors[,1]
              F[,i] <- X[[i]]%*%u
              v <- t(Y[[i]])%*%X[[i]]%*%u
              v <- v*as.vector(1/((t(v)%*%v)^0.5))
              G[,i] <- Y[[i]]%*%v
              f[,i] <- F[,i]*as.vector(1/((t(F[,i])%*%F[,i])^0.5))
              g[,i] <- G[,i]*as.vector(1/((t(G[,i])%*%G[,i])^0.5))
              X[[i+1]] <- X[[i]] - f[,i]%*%t(f[,i])%*%X[[i]]
              Y[[i+1]] <- Y[[i]] - g[,i]%*%t(g[,i])%*%Y[[i]]
              vF[i] <- var(F[,i])
              vG[i] <- var(G[,i])
              corr[i] <- cor(x=F[,i], y=G[,i], method="pearson")
              }
  res <- list(F=F,G=G,vF=vF,vG=vG,corr=corr)
  rm(k,X,Y,f,g,u,v)
  return(res)
  }

pls <- symPLS(a,b)


## ================================
## 4th GIMSA step: distance matrix and clustering
## ================================

# no weighting (w0)
F <- pls$F
G <- pls$G

# weighting by variance of PLS components (w1)
F <- apply(pls$F,2,scale,center=FALSE)
G <- apply(pls$G,2,scale,center=FALSE)

# weighting by number of distinct sequences (w2)
F <- pls$F/nrow(seqtab(mother.seq,tlim=0))
G <- pls$G/nrow(seqtab(daughter.seq,tlim=0))

# weighting by MDS 1st eigenvalue (w3)
F <- pls$F/mother.mds$eig[1]
G <- pls$G/mother.mds$eig[1]

# distance computation
diff2 <- function(X) return(as.matrix(dist(X,upper=T,diag=T)^2,nrow=nrow(X)))
D <- (diff2(F)+diff2(G))^0.5

# clustering
seq.dist <- as.dist(D)
seq.agnes <- agnes(seq.dist, method="ward", keep.diss=FALSE)
seq.part <- cutree(seq.agnes, nbcl<-10)
```

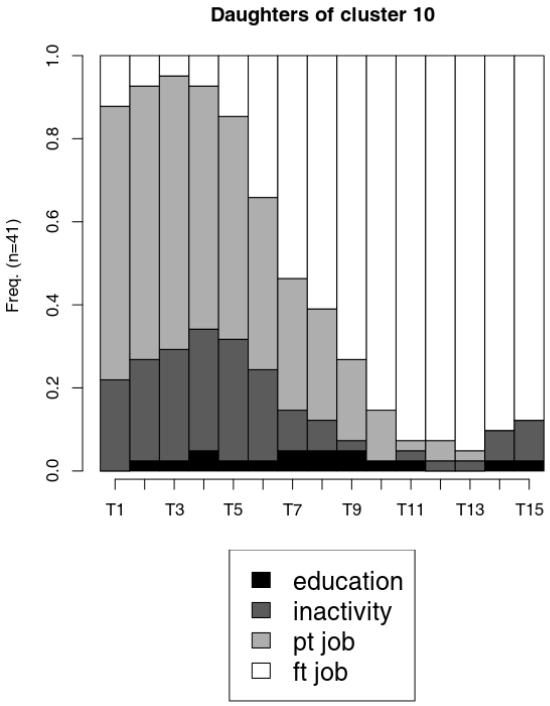
APPENDIX 6. State Distribution Plots Of The Clusters




Mothers of cluster 4
Daughters of cluster 4
Mothers of cluster 5
Daughters of cluster 5
Mothers of cluster 6
Daughters of cluster 6
Freq. (n=173)
Freq. (n=157)
Freq. (n=100)
0.0
0.2
0.4
0.6
0.8
1.0
14 18 22 26 30 34 38 42 46 50 54 58
T1 T3 T5 T7 T9 T11 T13 T15


Mothers of cluster 7
Daughters of cluster 7
Mothers of cluster 8
Daughters of cluster 8
Mothers of cluster 9
Daughters of cluster 9
Freq. (n=95)
Freq. (n=81)
Freq. (n=55)
0.0
0.2
0.4
0.6
0.8
1.0
14 18 22 26 30 34 38 42 46 50 54 58
T1 T3 T5 T7 T9 T11 T13 T15


Mothers of cluster 10
Freq. (n=41)
0.0
0.2
0.4
0.6
0.8
1.0
14 18 22 26 30 34 38 42 46 50 54 58
self
high/interm
low
inact
stud


Daughters of cluster 10
Freq. (n=41)
0.0
0.2
0.4
0.6
0.8
1.0
T1 T3 T5 T7 T9 T11 T13 T15
education
inactivity
pt job
ft job

APPENDIX 7. Cross-tabulation Of GIMSA Typology And Fourth Strategy's Typology

Fourth Strategy's Typology (daughters || mothers)	**GIMSA Typology**										***Total***
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	
Mostly FT || Mostly inactive	248	22	48	3	74	0	0	35	0	7	*437*
Mostly PT || Mostly inactive	0	0	1	0	0	0	0	1	22	0	*24*
From FT to inactivity || Mostly inactive	0	2	78	1	0	6	5	0	0	1	*93*
From PT to FT || Mostly inactive	13	0	0	0	1	0	0	0	0	6	*20*
Mostly inactive || Mostly inactive	0	1	45	0	0	78	2	3	0	0	*129*
Interruption || Mostly inactive	0	1	0	0	0	1	0	26	1	1	*30*
Mostly FT || Mostly low	6	4	3	136	72	0	0	2	0	0	*223*
Mostly PT || Mostly low	0	0	2	0	0	0	0	0	5	1	*8*
From FT to inactivity || Mostly low	0	1	14	23	0	0	0	0	0	0	*38*
From PT to FT || Mostly low	0	0	0	2	1	0	0	0	0	7	*10*
Mostly inactive || Mostly low	0	0	14	0	0	10	0	0	0	0	*24*
Interruption || Mostly low	0	0	0	0	0	0	0	13	2	0	*15*
Mostly FT || Mostly self	3	140	1	7	7	0	0	1	2	1	*162*
Mostly PT || Mostly self	0	0	0	0	0	0	0	0	10	0	*10*
From FT to inactivity || Mostly self	0	26	4	0	0	1	1	0	0	1	*33*
From PT to FT || Mostly self	0	0	0	1	0	0	0	0	0	12	*13*
Mostly inactive || Mostly self	0	21	0	0	0	2	0	0	0	0	*23*
Interruption || Mostly self	0	6	0	0	0	0	0	0	0	4	*10*
Mostly FT || Mostly high/interm	6	0	0	0	2	0	64	0	4	0	*76*
Mostly PT || Mostly high/interm	0	0	1	0	0	0	0	0	7	0	*8*
From FT to inactivity || Mostly high/interm	0	0	0	0	0	2	9	0	0	0	*11*
From PT to FT || Mostly high/interm	0	0	0	0	0	0	3	0	0	0	*3*
Mostly inactive || Mostly high/interm	0	0	0	0	0	0	10	0	0	0	*10*
Interruption || Mostly high/interm	0	0	0	0	0	0	1	0	2	0	*3*
Total	*276*	*224*	*211*	*173*	*157*	*100*	*95*	*81*	*55*	*41*	*1413*

Source: *Biographies et entourage* (2000)
Reference population: the 1413 female respondents and their mothers
Reading: low = lower-level occupations; high/interm. = higher-level or intermediate occupations; FT = full-time employment; PT = part-time employment

Note that GIMSA is more parsimonious than strategy 4 (which is one of its advantages), but it cannot be seen as a means to simplify the combinations of the contingency table produced with strategy 4: this is a completely different method, which leads to a different clustering. Looking at the two typologies, one notices that some of their clusters are very similar, while some patterns are only visible in one method or the other. For instance, GIMSA distinguishes, among daughters whose mothers are mostly inactive, between those who are mostly inactive (cluster 6) and those who shift from full-time employment to inactivity (cluster 3), while strategy 4 lumps them together (row 5). On the contrary, GIMSA groups daughters whose mothers have lower-level occupations (cluster 4), whether they are mostly employed full-time (row 7) or they shift from full-time to inactivity (row 9).

APPENDIX 8. Descriptive Variables Distribution By Cluster

Characteristic		Cluster										total
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	
Daughter's year of birth	1930-1939	32.6	40.6	37.9	37	29.3	37	25.3	29.6	16.4	31.7	33.8
	1940-1945	28.3	23.2	28	24.9	26.1	34	28.4	37	30.9	24.4	27.7
	1946-1950	39.1	36.2	34.1	38.2	44.6	29	46.3	33.3	52.7	43.9	38.5
Daughter's occupation at the time of the survey	inactive	12	16.5	42.7	20.2	10.8	61	18.9	13.6	10.9	19.5	22.4
	self-employed	0.7	2.2	1.4	1.2	2.5	1	3.2	3.7	1.8	2.4	1.8
	intermediate occupation	27.9	12.1	9	8.1	14	4	31.6	25.9	25.5	22	16.8
	higher-level occupation	28.3	24.1	13.7	23.1	36.3	11	28.4	24.7	30.9	17.1	24.1
	clerical and sales	29.3	41.5	31.3	43.4	33.8	21	17.9	27.2	23.6	34.1	32.2
	manual worker	1.8	3.6	1.9	4	2.5	2	0	4.9	7.3	4.9	2.8
Daughter's qualification	None	8	9.4	11.8	7.5	5.1	31	0	4.9	14.5	14.6	9.8
	< baccalauréat	38	56.7	54	66.5	57.3	27	26.3	43.2	18.2	41.5	47.1
	baccalauréat	21.4	16.5	20.9	15.6	19.7	19	28.4	23.5	16.4	9.8	19.5
	> baccalauréat	32.6	17.4	13.3	10.4	17.8	23	45.3	28.4	50.9	34.1	23.6
Daughter's number of children	0	23.6	10.7	1.4	12.7	20.4	1	10.5	4.9	14.5	22	12.6
	1	26.1	14.3	11.4	20.2	30.6	6	23.2	21	14.5	19.5	19.2
	2	31.9	42	42.2	42.8	29.3	20	38.9	50.6	40	29.3	37
	3 or more	18.5	33	45	24.3	19.7	73	27.4	23.5	30.9	29.3	31.1
Daughter's birth order	only child	5.4	12.1	7.6	27.7	8.9	2	9.5	6.2	10.9	17.1	10.5
	eldest	29	26.3	35.1	34.1	35	25	26.3	30.9	30.9	22	30.3
	in between	38.4	36.2	38.4	18.5	28.7	50	37.9	45.7	34.5	39	35.6
	youngest	27.2	25.4	19	19.7	27.4	23	26.3	17.3	23.6	22	23.6
Daughter's father's occupation	farmer	8	12.5	10	9.2	8.9	9	9.5	6.2	9.1	4.9	9.3
	self-employed	14.1	12.9	18.5	16.8	17.8	11	17.9	18.5	9.1	12.2	15.4
	intermediate occupation	20.3	17.4	17.5	24.3	21.7	22	11.6	13.6	21.8	29.3	19.5
	higher-level occupation	12	11.6	12.8	13.3	7	15	13.7	12.3	10.9	7.3	11.8
	clerical and sales	15.2	9.8	12.3	11.6	10.8	12	12.6	13.6	14.5	14.6	12.5
	manual worker	26.1	29.5	21.8	19.7	27.4	23	27.4	27.2	30.9	29.3	25.5
	inactive	4.3	6.2	7.1	5.2	6.4	8	7.4	8.6	3.6	2.4	6

Source: *Biographies et entourage* (2000)
Reference population: the 1,413 female respondents and their mothers
Reading guideline: For example, 32.4% of daughters belonging to cluster 1 are born between 1930 and 1939.

[1] For an extensive review of these two sets of methods, see Robette and Bry 2012.
[2] For a detailed presentation of the different stages of a sequence analysis by Optimal Matching see for example MacIndoe and Abbott 2004.
[3] The same kind of problems have been considered for more atypical sequences, e.g. activity patterns from diary data (Wilson 1998) or activism careers (Blanchard 2005).
[4] i.e. the frequency of transition between the states (or a function of it).
[5] It should be noted that some authors have compared several of these approaches using the same data set (Blanchard 2010). Furthermore, Piccarreta and Lior (2009) use Multidimensional Scaling to sort one-dimensional index plots according to a second dimension; the scope of our review focuses on the dominant view, however, i.e. typological approaches.
[6] These criteria are inspired by Gauthier et al. 2010.
[7] More precisely, a typology may be considered as more parsimonious than another if it keeps the same amount of information with fewer clusters, or if it keeps more information with the same number of clusters. Practically, parsimony is a balance between the amount of information and the number of clusters and the decisions it implies are not always straightforward.
[8] Another example could the dyad of sequences formed by the current workday or workweek, as gathered in time-use surveys (Lesnard and Kan 2011), and past occupational career.
[9] Other MDS techniques exist, such as nonmetric MDS (Shepard 1962) or Bayesian MDS (Oh and Raftery 2001). We follow Piccarreta and Lior's (2010) approach by applying metric MDS to sequence dissimilarity measures. Another application of MDS to sequence analysis may be found in Halpin and Chan 1998.
[10] The number of retained components may be different for mothers and daughters.
[11] Symmetric (or canonical) PLS can only deal with two sets of variables, still in this case it is probably the most powerful method (Bry 1996). In cases of three sets of variables or more, alternative factor analysis techniques may be used, such as Multiple Factor Analysis or STATIS (Escofier and Pagès 1994).
[12] Ward's criterion is known to produce homogeneous and compact clusters (Nakache and Confais 2004). It has been widely and successfully applied with sequence analysis.
[13] The Île-de-France region comprises the Paris metropolitan area plus the outer suburbs, and was home to 19% of the French population in 2000.
[14] In our survey, the interviewees are the children generation and no siblings were interviewed, so each mother is present only once in our sample.
[15] Completing education, at ages which range from 14 to 28 years.
[16] Added to that, in our survey, part-time employment covers jobs at 80% of full-time hours , as well as jobs at 50% or even less.
[17] These analyses were done using R software (R Development Core Team 2013) and the TraMineR package (Gabadinho et al. 2011).
[18] These computations were done using the WeightedCluster package (Studer 2012) in R.
[19] Colored index plots are available at http://nicolas.robette.free.fr/publis_eng.html
[20] Despite the choice of a dissimilarity measure emphasizing order.
[21] The 24 state distribution plots are available from the authors.
[22] To test the association between the typologies, we first perform a Pearson’s Chi-squared test: p-value = 0.0008624. However, as some cell counts are lower than 5, Fisher’s exact test is best-suited: p-value = 0.0004998. Both tests show that the typologies are significantly associated.
[23] From a strictly practical view, using MCSA here would imply adding 32 ‘non-response’ states within every daughter’s sequence. And it is not clear where these states should be added (at the beginning, at the end or somewhere in the middle), due to the difference in the definition of the time windows. While it is not technically impossible, the interpretation of the results would be blurred: to what extent would these results be driven by the reshaping of daughters’ sequences? But our major argument for using GIMSA to analyze the data remains the difference between local and global interdependence.
[24] The R program of the GIMSA application process in given in Appendix 5.
[25] As mentioned before, this characteristic of GIMSA emphasizes *global interdependence*.
[26] The figures are available from the authors.
[27] i.e. Min(5,4,1413).
[28] It would be all the more so in the case of more than two dimensions.
[29] Automatic clustering procedures for empirical data are hierarchical: this means that each cluster could be further divided into several sub-clusters, which are distinct and more homogeneous than the main cluster. Thus a typology is based on a delicate choice between the aggregated profile of the whole sample and the idiosyncrasy of individual cases.
[30] Colored index plots are available at http://nicolas.robette.free.fr/publis_eng.html
[31] Social origin as measured by daughters’ father occupation is not significantly associated with the clusters. This may be explained by the fact that daughters’ mother occupation is taken into account in the dyads which lead to the clusters, combined with the importance of homogamy in France (Vanderschelden 2006).