Journal of the
Royal Statistical Society
SERIES A
Statistics
in Society

J. R. Statist. Soc. A (2010)
173, *Part* 1, *pp.* 165–184

Exploring sequences: a graphical tool based on multi-dimensional scaling

Raffaella Piccarreta and Orna Lior

Bocconi University, Milan, Italy

[Received March 2008. Final revision March 2009]

Summary. Sequence analysis has become one of the most used and discussed tools to describe life course trajectories. We introduce a new tool for the graphical exploratory analysis of sequences. Our plots combine standard sequence plots with the results that are provided by multi-dimensional scaling. We apply our procedure to describe work and family careers of Israeli women by using data from the Israel Social Mobility Survey. We first focus on some preliminary choices relative to the definition of the sequences: the age span, the length of the sequences and the set of states registered in each time period. We then describe how our plots can be used to gain insights about the main features of sequences and about the relationships between sequences and external information.

Keywords: Coding of states; Dissimilarity; Length of sequences; Multi-dimensional scaling; Optimal matching; Sequence analysis; Sequence plots

1. Introduction and motivation

In recent years sequence analysis (SA) has become one of the most used and discussed techniques to describe life course trajectories (see Halpin (2003) for a comprehensive illustration of the various tools to analyse life courses). In SA the focus is on the whole life course, rather than on the timing of specific events. More precisely, the *activities or states* that are experienced by a given individual during a *specified period* are tracked. Each individual's life course is thus represented by a sequence (ordered collection) of states.

SA is a technique aiming at *describing* life courses. The low frequency generally characterizing each specific sequence (typically each individual will exhibit her peculiar trajectory) makes the description of sequences complicated. Consequently SA usually proceeds by a preliminary *simplification of sequences* via *cluster analysis*. Clusters rather than the singular sequences are described; *typologies* of sequences, or *typical life courses*, are thus deduced by synthesizing clusters. Of course, each cluster should be constituted by (quite) *similar* life courses, which can then be efficiently described by the same typical life course. In addition, different clusters should differ as much as possible from each other: *structurally* different or distinct typical life courses should be distinguished from one another.

To apply cluster analysis the dissimilarity between sequences must be measured appropriately. The centrality of cluster analysis in SA has led to many contributions in this field. The most popular approach remains *optimal matching analysis* (OMA) (Abbott, 1995), modified as necessary to overcome some of its limits (see Elzinga (2003) for a review). Recently, alternative approaches have been proposed, e.g. by Elzinga (2003, 2006).

Address for correspondence: Raffaella Piccarreta, Department of Decision Sciences and "Carlo F. Dondena" Centre for Research on Social Dynamics, Bocconi University, via Guglielmo Röntgen 1, 20136 Milano, Italy.
E-mail: raffaella.piccarreta@unibocconi.it

© 2009 Royal Statistical Society 0964–1998/10/173165

One possible problem of cluster analysis is that different clustering algorithms may lead to different solutions, and, also, the choice of the number of clusters is a rather subjective matter. Hence a detailed inspection of sequences in clusters is fundamental to evaluate the quality of a partition. This can be done by referring to so-called *sequence plots* (Kohler and Brzinsky-Fay, 2005; Scherer, 2001; Halpin, 2003; Lesnard, 2006; Müller *et al.*, 2008). In these plots cases are placed on the horizontal axis and time on the vertical axis. The sequence of each individual is represented by a set of stacked bars with colours and lengths depending on the states visited and their duration.

As emphasized for example by Halpin and Chan (1998) 'cluster analysis will always give a solution even if there is no meaningful structure in the data'. They also suggested analysing dissimilarities by using multi-dimensional scaling (MDS), which 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. MDS maps permit visualization of the dispersion of sequences and 'analysing' the dissimilarity matrix without necessarily grouping cases. If cluster analysis is used, cases in the MDS map may be flagged according to cluster membership, and the quality of a partition, mostly in terms of separation of clusters and of cohesion within clusters, can be graphically inspected.

Although they are useful to analyse the dispersion, MDS maps do not help much in the description of the main features of sequences. For each individual only the position in the map is known, and not the characteristics of the life course.

In this work we introduce a new graphical representation of sequences, permitting exploration of sequences without necessarily referring to the results of cluster analysis. The advantages of graphical representations of life courses were convincingly revealed in Francis and Fuller (1996), where they referred to graphical excellence as stated by Tufte (1961), i.e. giving

> 'the viewer the greatest number of ideas, in the shortest time, with the least ink, in the smallest space, and which tells the truth about data'.

Being inspired by this philosophy, we here propose an improvement of the already mentioned sequence plots in this direction. Our proposal is motivated by the consideration (see for example Halpin (2003)) that sequence plots are useful to describe clusters, constituted by *similar* sequences, but not to visualize heterogeneous sequences. This is quite obvious, since, if cases are randomly ordered on the horizontal axis, it will not be easy or even possible to distinguish any tendency in the sequences displayed. Since our aim is to visualize *all* the sequences in a data set, without the necessity of preliminary grouping into clusters, we here combine *sequence plots* with information arising from MDS, obtaining what we call *MDS sequence plots*, which are introduced in Section 4.

To illustrate our procedure we refer to data from the Social Mobility Survey that was conducted in 2001 on Israeli adults by the Israel Social Science Data Center at the Hebrew University (Web site: `http://isdc.huji.ac.il`). The occupational histories of female respondents and their detailed marital, childbirth and cohabitation history are taken into account. In Section 2 data and sequences are described, and in Section 3 we introduce the dissimilarity matrices that are used in the paper.

In Section 5 we describe how to use MDS sequence plots in the exploratory and preliminary analysis of sequences. First, our graphical displays can support the analyst's choices in the *definition* of the sequence type representation of life courses, with particular attention to the *length* and the *calendar* (age of the respondents) of sequences and to the *states* (*coding*) to take into

account. These preliminary aspects are fundamental (Lesnard, 2006) since they may strongly influence the results of SA.

Second, MDS sequence plots can be used to explore the main characteristics of the sequences that are defined on a given domain. Finally, our proposal can be extended to describe the relationships between sequences and explanatory variables or between sequences in different domains.

2. The data

The Social Mobility Survey was conducted in 2001 on a set of Israeli adults, who were selected by using multistage sampling. In the first stage a set of statistical areas (which are layers defined by geographic, demographic and social characteristics) was sampled. In the second stage households were randomly selected from the statistical areas sampled. In each sampled household only one adult was investigated by using an extended questionnaire. Respondents provided retrospective information on their entire occupational, marital, childbirth and cohabitation history. In our analysis, attention is focused on women.

For each woman we build, on a yearly timescale, a sequence-type representation of two life course domains or 'careers': school or work and family formation. For exposition purposes we initially focus on a wide age span, 15–50 years, and select women with age between 25 and 50 years in 2001, at the moment of the interview ($n = 313$).

Hence to the ith woman, aged a_i in 2001, two sequences are associated: $w_i = \{w_{i;15}, \ldots, w_{i;T_i}\}$ and $f_i = \{f_{i;15}, \ldots, f_{i;T_i}\}$ with $T_i = \min(a_i, 50)$. Clearly for the youngest women only partial information is available (see below for discussion).

For the family formation career we consider whether a woman is cohabitating with a partner (being married, m, or not, u) or not (n), and whether she has children (from zero to three or more children). The combination of these states leads to the alphabet $A_f = \{$n0, n1, n2, n3, u0, u1, u2, u3, m0, m1, m2, m3$\}$.

For the occupational histories of respondents, we 'start' with a simplified school or work coding (see below for discussion): school (s), work part time or full time (wp, wf), a combination of school and part- or full-time work (swp, swf) and unemployment (u). We also consider military service (z), which is compulsory in Israel for women also, to distinguish this activity from unemployment. The resulting *alphabet* of states is $A_w = \{$s, wp, wf, swp, swf, u, z$\}$.

Below are examples of sequences in the work domain (to simplify the notation, we use a *state–duration* representation of sequences: visited states are listed in the order that they were experienced, together with their duration; see Aassve *et al.* (2004)):

(a) s/6 wf/12 wp/4;
(b) s/7 wf/14 wp/5;
(c) s/2 u/34;
(d) s/7 wf/4.

The first woman remains in school until she is 21 years old, then she works full time for 12 years and finally she works part time for 4 years, until the last year of observation. The trajectories of the first two women are similar and could be 'associated' with the same typical career: the decisions to exit school quite late, to enter the labour market after school working full time, and then to work part time. Instead, the third woman exited school quite early, did not enter the labour market and remained unemployed until she was 50 years old. This pattern is really different from the first two. Even if the lengths of the three sequences (22, 26 and 36 years) are different, conclusions can be drawn about their (dis)similarity. Of course, this evaluation is

based on the assumption that a period of 22 years of observation is sufficiently long to make considerations about the (dis)similarity between life courses.

Conversely, comparing the first woman with the fourth, despite observing very similar life courses over the first 11 years of follow-up, we cannot evaluate properly their (dis)similarity, because the fourth sequence is censored. First, given the lengths of the other sequences, 11 years of observations appear not enough to compare properly the two life courses. Second, the dissimilarities between the fourth sequence and the other sequences will be possibly related more to the differences between their lengths than to the difference between the states experienced.

The selection of the age span, and the consequent censoring for some sequences, is clearly a relevant issue. Analysts usually deal with censoring on the basis of *a priori* considerations. Here (see Section 5) we prefer a data-driven approach, based on the inspection of the sequences' features as emphasized by MDS sequence plots.

Similar considerations hold for the choice of the alphabets. For example, for the school or work career, in our data set information is available also on the mode of getting each job, its economic branch, the status at work and managerial status. One could be tempted to define more detailed work sequences, e.g. distinguishing between subcategories of full-time or part-time work. Of course, a more detailed coding is convenient only if there are not rare subcategories, and if careers that are characterized by different subcategories are recognized as different. Otherwise, too complex a set of states may simply increase noise and inflate sequence heterogeneity, without adding relevant information, thus making it difficult to describe sequences and/or to unveil their most important features. Again, in Section 5 we shall base our final choice of the alphabets on the information arising from MDS sequence plots.

Before proceeding we briefly describe how we measure the dissimilarities between sequences.

3. Sequence data and dissimilarity measures

OMA, which was introducsed by Abbott (1995), is so popular in SA that the two techniques are almost regarded as synonymous (Elzinga, 2003). OMA has been applied in different contexts to analyse career paths (Abbott and Hrychak, 1990; Scherer, 2001; Schoon *et al.*, 2001; Stovel *et al.*, 1996, Halpin and Chan, 1998; Chan, 1994, 1995; McVicar and Anyadike-Danes, 2002; Schlich, 2003; Malo and Munoz-Bùllon, 2003; Halpin, 2003; Pollock, 2007; Piccarreta and Billari, 2007).

In OMA the dissimilarity between two sequences is measured as the effort that is needed to transform one sequence into the other. Three operations are considered: insertion of a state, deletion of a state and substitution of a state with another state. To each operation a *cost* is assigned, reflecting how difficult it is to modify a life course according to the operation itself. The total cost (sum of the costs) of the operations that is needed to transform one sequence into another is calculated. The OMA dissimilarity between two sequences is defined as the minimum transformation cost (hence dissimilarity is unequivocally defined, independently of the number of possible transformations). By definition, the dissimilarity is symmetric.

The choice of costs is arbitrary, and this is considered one major weakness of OMA. We shall not enter into the details of this debate (see Elzinga (2003) and Halpin (2003) for a critical review of the main objections to the application of OMA to social sciences). Following what is quite a standard approach, we set the insertion and deletion costs equal to 1. As for the substitution operation, it ideally represents the effort that is needed to move from one state to another. In some references substitution costs are assigned subjectively on the basis of *a priori* knowledge or considerations (e.g. McVicar and Anyadike-Danes (2002)). Here we prefer a data-driven approach (Rohwer and Pötter, 2004; Stovel *et al.*, 1996) and relate substitution costs to

transition frequencies. Frequent transitions are considered less costly than rare transitions. Note that a substitution operation is equivalent to an insertion operation followed by a deletion operation. Thus, for this operation to be effective, we set its cost lower than the sum of insertion and deletion costs, 2.

The MDS sequence plots that we define below can be used with any dissimilarity matrix. However, although other dissimilarity measures have been proposed (e.g. Elzinga (2003, 2006) and Lesnard (2006)), we use OMA because it is currently the most widely used measure. OMA dissimilarity matrices (based on the costs described above) that are analysed throughout the paper were obtained with the CHESA package (which was developed by C. Elzinga and is downloadable from `http://home.fsw.vu.nl/ch.elzinga/`).

4. Visualization of sequences: multi-dimensional scaling sequence plots

We introduce a tool to visualize sequences in a holistic fashion. We start by considering *sequence plots* (Kohler and Brzinsky-Fay, 2005; Halpin, 2003; Müller *et al.*, 2008). In these plots individuals are placed on the horizontal axis and time on the vertical axis. To each individual a set of stacked bars is associated, with colours and lengths depending on the states that are visited and on their duration.

Unfortunately, sequence plots do not help much in describing *heterogeneous* sequences (e.g. all the sequences within a data set). Indeed they are generally dominated by most frequent states in subperiods whereas the other characteristics of sequences are hidden. As Halpin (2003) pointed out, 'browsing tens of hundreds of sequences is not terribly easy, even when colour coded'. This is why Halpin concluded 'I found it extremely useful to view the sequences once they had been clustered'. It is easy to understand that in this last case the visualization is easier since clusters are relatively within homogeneous. Similar sequences are grouped together and consequently their main tendencies are more identifiable.

Intuitively, the visualization of heterogeneous sequences can be improved consistently if they are ordered so that similar trajectories are close on the horizontal axis. Hence a suitable *order* of cases should be related to their (dis)similarities. Our solution to this problem is based on MDS (Borg and Groenen, 1997), which is a factorial technique that represents measurements of dissimilarity between pairs of cases as distances between points of a low dimensional space. One of the most common methods is *metric MDS* (Torgerson, 1958). In this approach the squared dissimilarity matrix, whose elements δ_{ij}^2 are the dissimilarities between all the possible pairs of cases, is converted to the *double-centred* matrix $\mathbf{A}$, with elements $a_{ij} = (\delta_{ij}^2 - \delta_{+j}^2 - \delta_{i+}^2 + \delta_{++}^2)$. δ_{i+}^2, δ_{+j}^2 and δ_{++}^2 are respectively the rows, the columns and the overall means of the squared dissimilarities. $\mathbf{A}$ can be rewritten as $\mathbf{A} = \mathbf{Y}\mathbf{Y}^{\mathrm{T}}$, where $\mathbf{Y}$ is the matrix of (unobserved) co-ordinates. For a given number of dimensions, $\mathbf{Y}$ can be extracted by applying an eigenanalysis (spectral decomposition) to $\mathbf{A}$. The number of latent dimensions is not known *a priori* and must be specified (see Section 5.3). In this approach dimensions are extracted *in decreasing order of explanatory importance* (Mardia *et al.*, 1979).

MDS sequence plots are obtained by combining sequence plots with information that is provided by the MDS dimensions: cases are ordered on the horizontal axis according to their score on each selected dimension. Because dimensions are ordered, the first MDS sequence plot (in which the sequences are plotted against the dimension that explains most variance in the dissimilarity matrix) will describe the most important characteristic of the sequences, the second MDS sequence plot describes the next most important characteristic, and so on.

Before proceeding it is important to point out that dissimilarity matrices that are obtained with different criteria will be characterized by different MDS dimensions, and by possibly different

orderings of sequences. This is not peculiar to our method: in SA results are always conditioned on the dissimilarity criterion that one decides to adopt.

5. Multi-dimensional scaling sequence plots for the exploratory analysis of sequences

In Fig. 1 the first MDS sequence plots are reported for the family and work sequences of Israeli women (ages between 25 and 50 years; age span 15–50 years). Compared with the original sequence plots (which are not reported here) our plots permit an easier exploration of heterogeneous sequences and provide a much clearer visualization of their main features. In both plots we observe clear clustering (and clear oppositions) of sequences dominated by the most relevant states. Also, rare and less relevant states can be easily individuated (n1, n2 and n3, and u1, u2 and u3 for the family domain; swf and swp for the work domain): few sequences are characterized by these states and they are scattered and not clustered along the horizontal axis.

In what follows we illustrate the potential of our graphical displays in SA.

In Sections 5.1 and 5.2 we focus on the definition of sequences (the choice of the age span, of the minimum length required for sequences, and of the coding). In this exploratory analysis, the plots are used only for visualization purposes. Therefore we shall refer for each domain only to the plot that is based on the *first* MDS dimension, which is that explaining at best dissimilarities and hence providing the *best ordering* of cases and the best visualization of all sequences in a single plot.

Next we move to a more detailed analysis of sequences. In Section 5.3 we illustrate how to analyse sequences' features by using MDS sequence plots. In Section 5.4, we show how the 'ordering' idea underlying the definition of MDS sequence plots can be used to analyse the relationships between two domains and between one domain and external information.

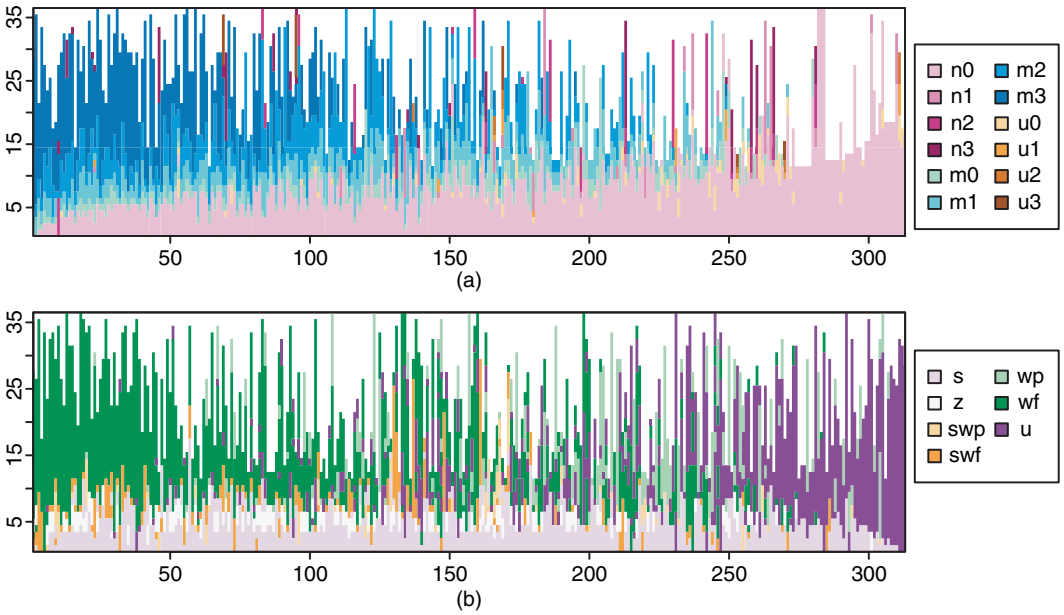

Fig. 1. MDS sequence plots for (a) family and (b) work trajectories

5.1. Length of the sequences

In SA the (chosen) length of the sequences is often considered 'as given' and no detailed explanation is provided about the reasons underlying a particular choice (even if usually suitable). In many references (Piccarreta and Billari, 2007; McVicar and Anyadike-Danes, 2002; Pollock, 2007; Aassve *et al.*, 2007) attention is limited to a given age span, and cases are taken into account only if information is available on the entire period considered. In other references sequences with different lengths are admitted. Each individual may be 'followed' until a given event occurs (e.g. retirement, as in Han and Moen (1999)). Alternatively, sequences may be censored owing to the age of the individuals at the time of the survey (Elzinga, 2003).

In our preliminary definition of work and family careers for Israeli women, we considered a very long age span: 15–50 years. Our data set includes women with different age (at the time of interview): the sequences of the youngest women are therefore censored. Clearly, it is important to choose a cut-off to reduce the disparity between sequence lengths and to avoid dissimilarities to be mostly related to the length of the sequences (the age of women) rather than to their features.

Of course, censoring is not a problem if (and only if) the dissimilarities between sequences are *not related* to their length (see Section 2). Here we show how MDS sequence plots can support the choice of a suitable age span and/or the minimum sequence length to take into account. As was mentioned before, here we refer to the first MDS sequence plot, providing the best visualization of sequences. If the ordering and clustering of sequences along the horizontal axis is related (also) to their length, we can conclude that the length has an effect on the first MDS dimension (and, also, on the dissimilarity). Next, by analysing the main features of the sequences as described by the plot, some considerations can be made about a suitable age span and about the minimum reasonable length of sequences.

Consider first the family domain. Observe from Fig. 1(a) that the ordering of sequences appears to be related to their length. Also, note that women with the longest spell in the n0 state are the youngest, as we may deduce by the short height of their bars. This suggests a postponement of the moment of family formation for the youngest women. Nevertheless, owing to censoring, it is not possible to determine whether the careers of the youngest women will differ from those of the oldest with respect to the subsequent visited states. Some conclusions may also be drawn from the plots about the 'memory' of the sequences. After 20 years of observation (after the 35th year of age) the family careers of women appear quite stable. m3 and u3 are in a sense 'absorbing' states: passages from m3 to u3 or n3 or from u3 to m3 or n3 are not frequent. Dissolution of marriages or life with children but without a partner are not common in this data set. A period of 20 years of observation appears adequate to describe the main differences between women in this domain (and the reduction in the age span is convenient since it leads to a lower number of censored sequences).

For the work domain, instead, the effect of age or censoring is not so evident. The first MDS dimension is seemingly not related to the length of the sequences (Fig. 1(b)).

We decided to refer to the same age span, 15–35 years, for both domains, considering only women for which at least 15 years of observation are available (at least 30 years old in 2001), to avoid an excessive influence of sequence lengths on results. The sample size is now reduced to $n = 241$.

5.2. Coding of states

The choice of the states to take into account when defining sequences is related to the aim of the analysis and to the aspects of life courses that are considered relevant. As already mentioned, from a practical point of view a very detailed coding may result in excessive

heterogeneity, making it difficult to identify common patterns in data. Preliminary analyses are often conducted to verify whether some states are characterized by very low frequencies and progressive simplification of the coding takes place until a satisfactory degree of homogeneity is reached. Nonetheless, it may happen that one state is experienced by many individuals but is so dispersed among the sequences (e.g. small repeated spells) that it is irrelevant or confusing. Hence, determination of a suitable coding of states should preferably be based on all the sequences.

In Fig. 2 MDS sequence plots are reported for the family and work careers defined and selected as described in the previous section. Our attention is focused on states which are rare, i.e. experienced by few women and/or characterized by low durations, and not relevant, i.e. scattered along the horizontal axis.

Consider first the family trajectories (Fig. 2(a)). Most women are married and have children within a marriage. Most cohabitations have short duration and end in marriage. Few women have children without a partner (cohabitation or marriage) or have children during cohabitation before marriage. Also, states u (children within cohabitation) are usually experienced after the corresponding m states (women are cohabiting with their children and the new partner after the breakdown of a marriage).

States u and m were initially distinguished to understand whether women have children outside marriage and/or prefer cohabitation to marriage. The plot evidences that this distinction is not relevant instead. It seems worthwhile to distinguish only between women cohabiting with a partner (independently of marriage) or not.

As for the single motherhood states, they are experienced by few women who are close to each other in the plot *only when they are also similar with respect to other states*. For example, on the right-hand side of the plot (Fig. 2(a)) the single mothers are close together owing to a long spell in the n0 state; nevertheless they are also placed close to sequences that are otherwise completely different except for the duration in n0. We may deduce that the dissimilarity is not

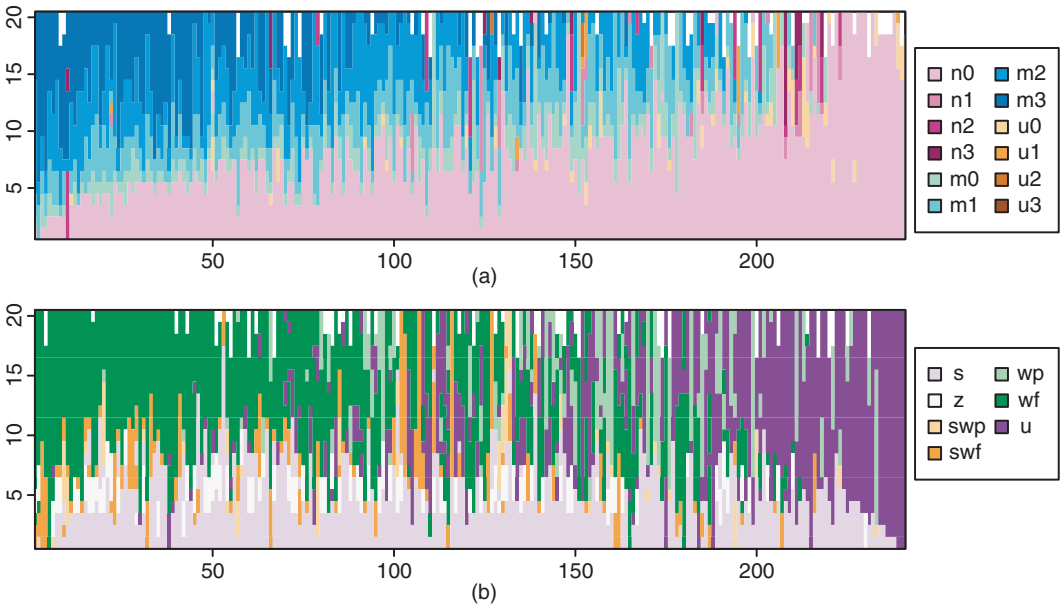

Fig. 2. MDS sequence plots for (a) family and (b) work trajectories (age span, 15–35 years; women at least 30 years old at the time of the interview)

substantially influenced by this state. Hence, these women *could* be grouped in a low frequency cluster and be described by a 'rare' typical trajectory.

Moreover, the results that are provided by MDS and cluster analysis are *related* since both techniques aim at explaining the entries in the dissimilarity matrix. Consequently, we may expect cluster analysis to attribute single mothers to clusters according to the duration or presence of other states in their trajectories. Thus, we might exclude these women from the data set *a priori*. Alternatively, if we want to analyse their similarity with other women, it would be necessary to simplify the coding of sequences further, to focus only on the most relevant states. We decided to follow the latter approach and recoded sequences by using the new alphabet $A_f = \{\text{n0, p0, c1, c2, c3}\}$, where p0 indicates partnership (cohabitation with a partner, being married or not) without children, and c refers to fertility, independently of the union history of the women.

Turning attention to work trajectories (Fig. 2(b)), we observe that state swp is neither frequent nor well represented in the MDS sequence plot. Women who experienced this state are not close together in the plot and their position is related to the presence and/or duration of other states in their career. We therefore simplified the alphabet, collapsing states swp and swf into a single state, sw.

The main advantage of MDS sequence plots over simple frequency tables of transitions or summaries of durations is the possibility of evaluating whether rare states are relevant, i.e. whether trajectories that are characterized by peculiar features are 'recognized' by the dissimilarity criterion (and are thus clustered along the horizontal axis) or whether they are disregarded in the evaluation of the dissimilarities.

This aspect makes MDS sequence plots particularly useful when multiple domains are considered. In some references (Abbott and Hrycak, 1990; Stovel *et al.*, 1996) the joint analysis of domains is based on *sequences* that are built by *combining* in each period the states that are experienced in the single domains. This approach can be followed when the number of combinations is not very high or when the original alphabets are simplified before combination by collapsing some states. There is a clear trade-off between the increase of heterogeneity for combined sequences and the excessive simplification of alphabets. MDS sequence plots can help to compare different simplified coding and to evaluate which states should not be simplified.

5.3. Exploring sequences by using multi-dimensional scaling sequence plots

In this section we illustrate how to use MDS sequence plots to analyse sequences in more detail. In this case we are interested in taking into account the MDS sequence plots that are relative to the relevant dimensions, describing essential structure, and unveiling the most important features of sequences. Thus, an important issue is choosing properly the number of dimensions to consider. This is a general problem in MDS, and various criteria have been introduced in the literature with this aim.

A first selection criterion, which is followed when metric MDS is used, is based on the eigenvalues (the kth eigenvalue measures the contribution of the kth dimension to the explanation of dissimilarities). Eigenvalues are plotted against dimension: usually a distinct break can be observed between the steep slope of the most important dimensions and the gradual trailing off of the others. Other criteria are based on the so-called *stress*, which is a synthesis of the (squared) differences between the observed dissimilarities and the distances that are calculated on the basis of the MDS dimensions. A small stress indicates that points close in the space spanned by the MDS dimensions are close in the original data space also. Some researchers suggest selecting dimensions to minimize the stress (Mardia, 1979). Also, stress can be plotted

against the dimensionality to determine after which number of dimensions it is not reduced substantially. Nevertheless, in many data sets stress decreases with increasing dimensionality and the optimal solution cannot be found or is not convenient (too many dimensions are selected). Some researchers have suggested interpreting stress informally and have indicated 0.1 as the maximum value which can be considered as acceptable (see Kruskal and Wish (1984) for a detailed discussion).

Eigenvalues and stress are reported in Table 1 for the first 10 dimensions in our data. For both domains the first two dimensions are the most important if we consider eigenvalues. The stress function suggests a higher number of dimensions, 3 for the family domain and 3 or 4 for the work domain, if a stress around 0.1 is considered as acceptable (five dimensions are instead necessary to minimize the stress).

Before proceeding, it is important to point out that overinterpretation of MDS sequence plots may be an issue (as for all the other graphical displays) and one should avoid interpreting noise as structure. Therefore, when analysing the plots our attention will generally be focused on coarse patterns; we shall consider as 'structure' a clear clustering on the horizontal axis of sequences that are dominated by the same state or colour and/or characterized by the same sequence of states or colours. Thus, in our context, plots will be considered only relative to MDS dimension providing this kind of 'structured' ordering; actually, this will be our final criterion to evaluate the dimensions that are suggested by the standard criteria mentioned above.

This is why in Fig. 3 only the MDS sequence plots for the first three dimensions are presented for both domains: the subsequent dimensions do not provide interesting or interpretable orderings of sequences.

Analysing the plots for family domains (Figs 3(a)–3(c)), note that the most important contrast (Fig. 3(a)) is between women with three or more children (left-hand side) and women who did not even enter cohabitation (right-hand side). Also observe that the order is related to a combination of age at cohabitation and age at the birth of the first, second and third child. The order of the sequences is *also* related to the number of children but it is not the feature of the life courses that is mostly emphasized by the plot, and thus by the (OMA) dissimilarity considered.

Figs 3(b) and 3(c) provide more detailed description of sequences which are not *opposed* along Fig. 3(a), with particular emphasis on the sequences that are in the central part of Fig. 3(a). In Fig. 3(b), women with fewer than three children (left-hand side) are contrasted with women

Table 1. Eigenvalues and stress functions for family and work MDS dimensions

Dimension	*Results for family domain*		*Results for work domain*	
	Eigenvalue	*Stress*	*Eigenvalue*	*Stress*
1	12.508	0.442	18.075	0.478
2	6.237	0.234	9.081	0.246
3	3.708	0.119	4.771	0.169
4	2.187	0.071	3.658	0.117
5	1.598	0.069	3.113	0.093
6	0.739	0.079	1.790	0.094
7	0.529	0.089	1.147	0.101
8	0.498	0.098	1.075	0.109
9	0.445	0.107	0.948	0.117
10	0.440	0.116	0.841	0.125

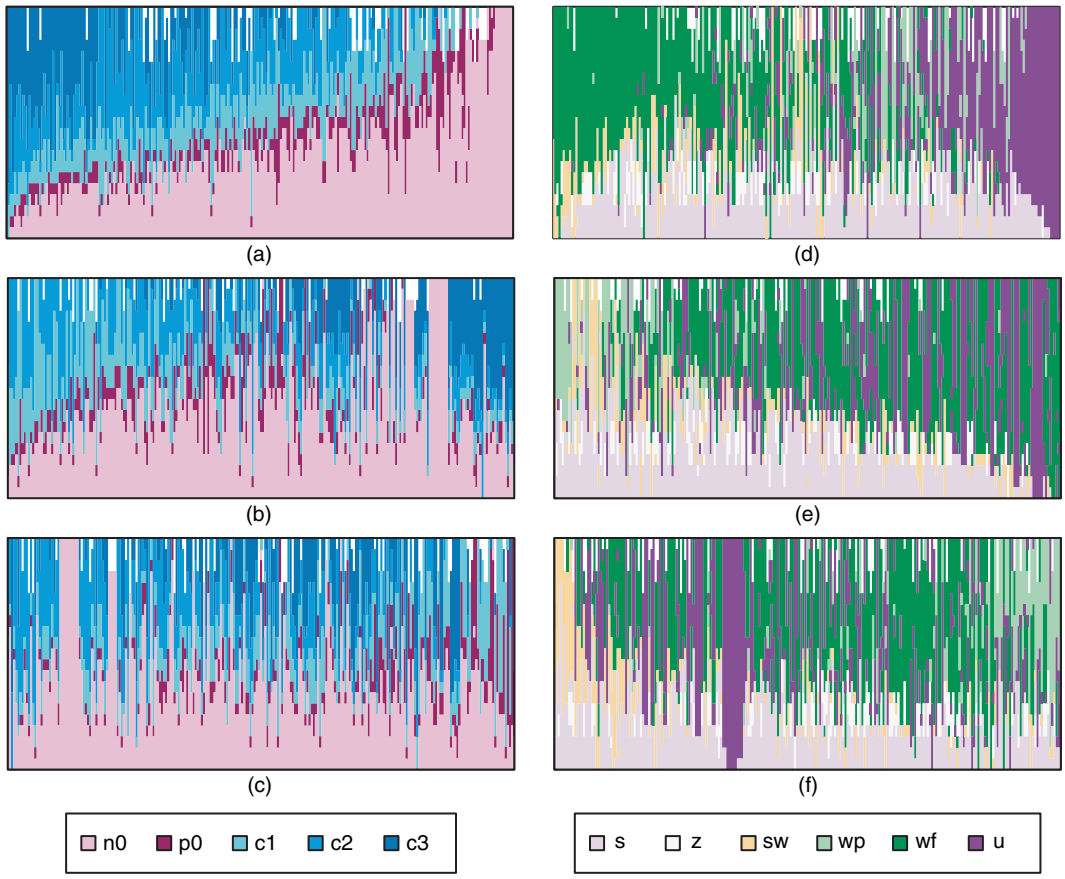

Fig. 3. MDS sequence plots for (a)–(c) family and (d)–(f) work trajectories (age span, 15–35 years; age and alphabet selected as described in Sections 5.1 and 5.2): (a), (d) dimension 1; (b), (e) dimension 2; (c), (f) dimension 3

with three or more children (right-hand side). In the central part of the plot we find women who entered cohabitation and had children quite late. Finally, in Fig. 3(c) women with one child and with early partnership and motherhood (on the right-hand side) are opposed with all the other women, in particular with a small group of women with two children who also started cohabitation and motherhood early.

Hence, we may conclude that the OMA dissimilarity for these women is mostly related to the moment when partnership and motherhood are experienced. The number of children is a relevant but 'secondary' aspect and is described and detailed by the second and third dimensions.

Turning attention to work trajectories (Figs 3(d)–3(f)), we observe a similar hierarchically ordered set of contrasts. The most important, which is described by the first MDS dimension (Fig. 3(d)), is that between women working full time (more specifically, with life courses that are strongly characterized by the wf state) and women unemployed or working part time (with long spells in these states).

Similarity between women (determining their closeness on the horizontal axis) appears related to the states dominating the sequences; small spells in irrelevant states do not have a strong effect on the dissimilarity. For example, the period that is dedicated to military service is not particularly relevant in this ordering. In Fig. 3(e) women working part time and combining school

and work (on the left) are contrasted with all the others, the latter being ordered according to the length of schooling (sequences on the right are characterized by short duration in this state). Finally, Fig. 3(f) contrasts women combining school and work (on the left) with women working part time.

As is evident from this analysis, inspection of the MDS sequence plots in a hierarchical manner allows us to understand which states or durations are most 'responsible' for (and emphasized by) the observed dissimilarities.

MDS sequence plots relative to different dimensions can be analysed jointly combining them with MDS maps. MDS maps provide useful information about the dispersion of sequences, but they may be difficult to interpret since we do not know the features of sequences occupying particular positions in the map; MDS sequence plots can help in the analysis of MDS maps. Fig. 4 presents MDS maps for the first three MDS dimensions for the family domain. The first MDS dimension is on the horizontal axis. The vertical axis represents the second MDS dimension in Fig. 4(a) and the third MDS dimension in Fig. 4(b). The MDS sequence plots for each dimension are reproduced on the appropriate axis.

In the maps, cases are coloured according to cluster membership. (As suggested by Halpin and Chan (1998), we selected a rather high number of clusters, 8, for exploratory convenience; clusters were obtained by using Ward's (1963) algorithm.) The MDS sequence plots along the axes show the features of sequences occupying particular positions in the map. As already noted the first dimension, which is related to a combination between fertility and age at cohabitation, distinguishes between women with many children and women with few children and with no partners. Along the second dimension (vertical axis, Fig. 4(a)) we see the further subdivision between women with high fertility: women with fewer than three children (bottom left-hand side) and women with three or more children (upper left-hand side). In an intermediate position along this axis (the group within the box in the maps) we find the group of women who

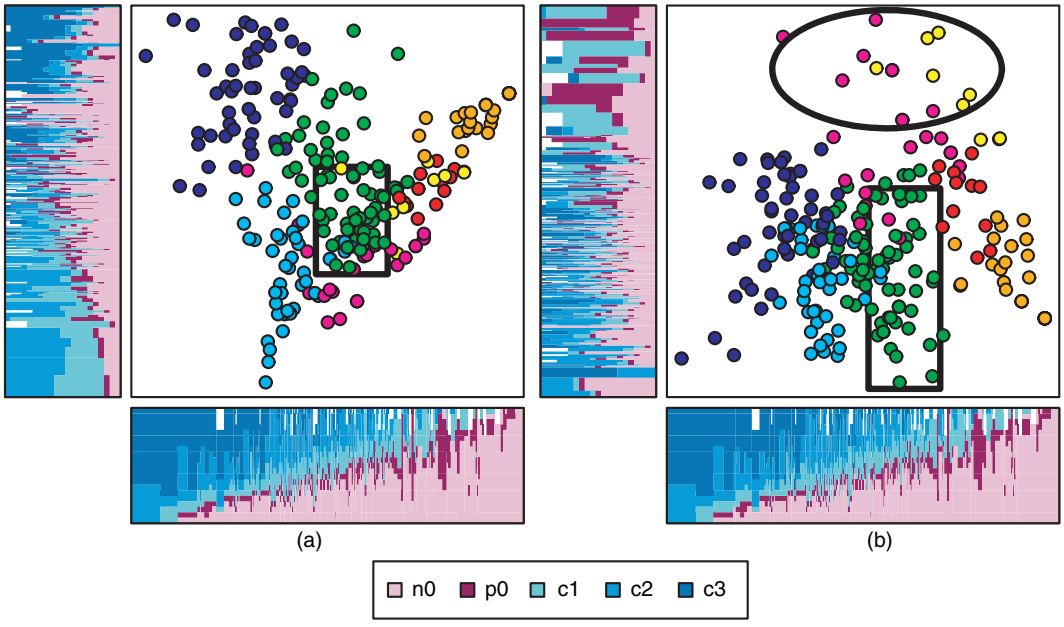

Fig. 4. MDS maps and sequence plots for the family domain (colours in the maps reflect cluster membership): (a) first and second MDS dimensions (horizontal and vertical axis); (b) first and third MDS dimensions (horizontal and vertical axis)

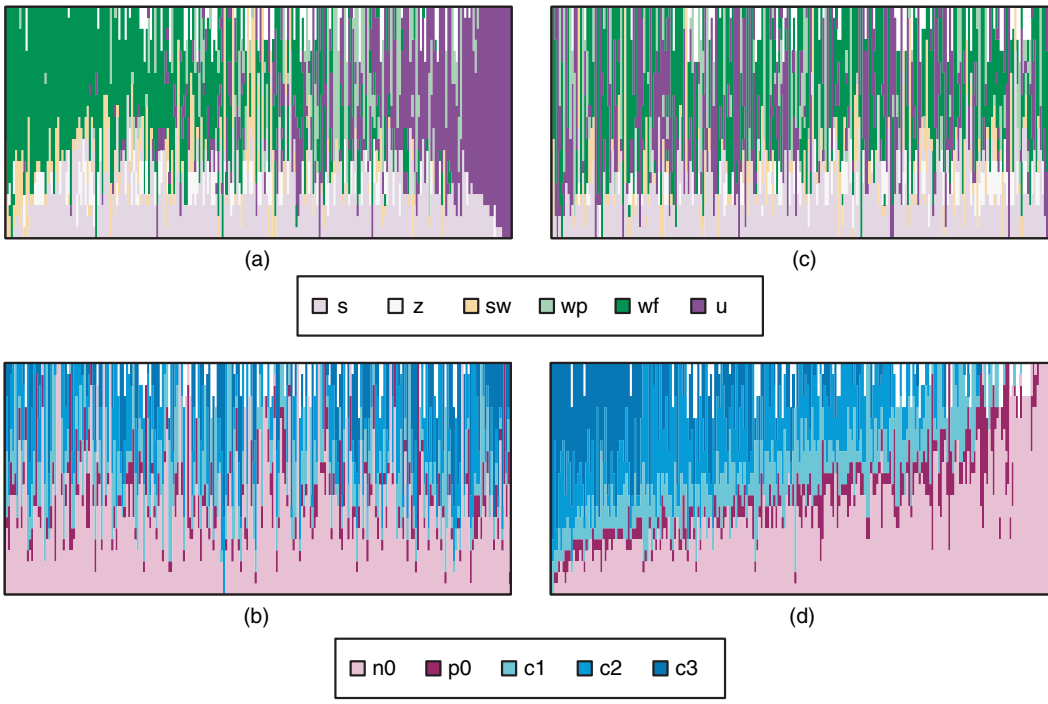

Fig. 5. Sequence plots of (a), (c) work and (b), (d) family domains ordered according to (a), (b) work and (c), (d) family

experienced cohabitation and motherhood quite late. Along the third dimension (vertical axis, Fig. 4(b)) women with one child (the circled group in the map) who experienced cohabitation and motherhood relatively early are distinguished from the others.

These enriched MDS maps permit quite detailed analysis of sequences: the characteristics of the sequences occupying particular positions in the map are evidenced by the ordered sequence plots; the main differences between clusters are highlighted; potentially highly diverse groups can be easily identified. This representation also assists with the interpretation of MDS dimensions. It can also help to evaluate and/or compare the quality of different clustering partitions (different algorithms or different numbers of clusters).

5.4. Sequence plots based on external information

In MDS sequence plots, cases are 'internally' ordered according to their (dis)similarity evaluated with respect to the domain considered. This ordering generally unveils the 'structure' of life courses (if present). The accuracy of the description of the sequences depends of course on their structure and on the significance of the MDS dimensions that are taken into account.

We now illustrate how the idea underlying MDS sequence plots can be extended to analyse the relationship between sequences and *external* information. The idea is to order cases according to the external information (explanatory variables or sequences that are defined on other domains). If there is a relationship between sequences and the external criterion, these *external* sequence plots should exhibit a structure.

5.4.1. Relationships between life courses

We first consider the relationship between two domains. In Fig. 5 the plots for the family

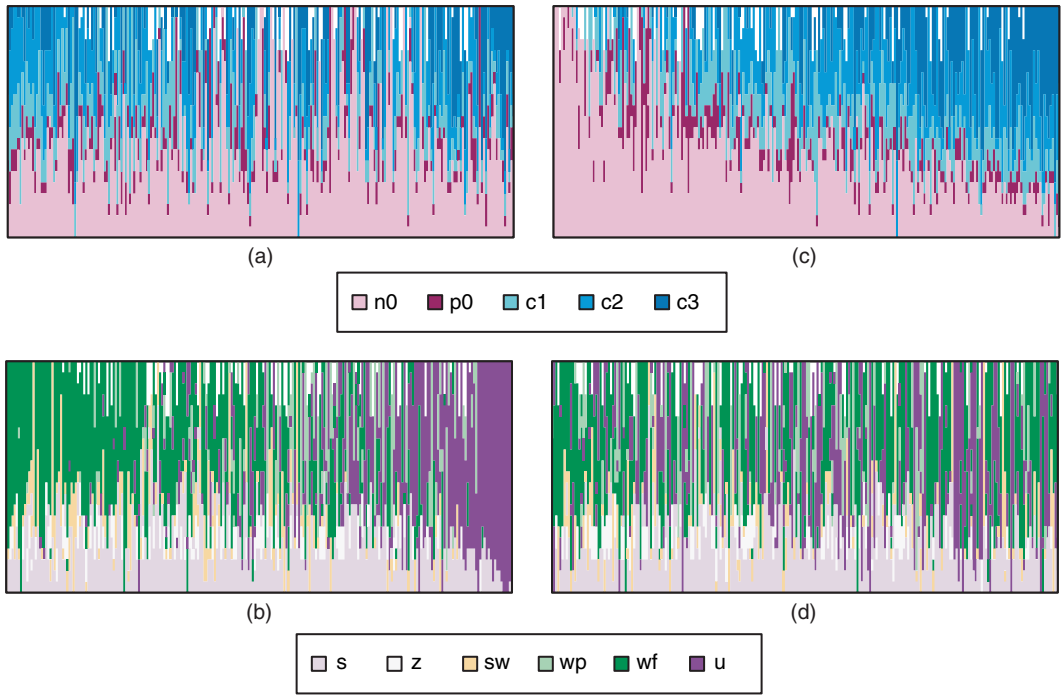

Fig. 6. Sequence plots of (a), (c) the family and (b), (d) the work domains ordered according to (a), (b) the first and (c), (d) the second MDS dimension extracted from the joint dissimilarity matrix

and work domains are reported. In Figs 5(a) and 5(b) (work ordered), sequences are ordered according to the first MDS dimension extracted from the dissimilarity matrix relative to the *work* domain. Instead, in Figs 5(c) and 5(d) (family ordered), sequences are ordered according to the first MDS dimension relative to the *family* domain. Figs 5(a) and 5(c) contain the work sequences whereas Figs 5(b) and 5(d) contain the family sequences.

Some interesting patterns may be observed. Looking at the work-ordered sequences, we note that unemployed women (Fig. 5(a), right-hand side) generally experienced cohabitation and motherhood quite early and had many children (Fig. 5(b), right-hand side). In contrast, women whose work trajectory is dominated by full-time work (Fig. 5(a), left-hand side) generally had fewer children and became mothers later (compared with women on the right-hand side of the plots). Turning to the family-ordered sequences, we observe that single (not cohabiting) women (Fig. 5(d), right-hand side) have work careers that are dominated by full-time work. On the opposite side of the family-ordered plots, we see that women with high fertility are more associated with part-time work (Fig. 5(c), left-hand side).

Despite these marginal patterns, ordering based on one domain is not associated with particularly strong patterns in the other. Hence we can conclude that the association between the two domains is not very strong and that different women make different choices about how to combine work and family (as expected).

Conclusions about association can also be drawn by considering sequence plots that are based on dimensions extracted from the joint dissimilarity matrices. As mentioned in Section 5.2, joint analysis may be conducted (Abbott and Hrycak, 1990; Stovel *et al.*, 1996; Piccarreta and Billari, 2007) by using sequences obtained by *combining* states experienced in the domains considered. In this case we are usually interested in understanding whether the joint dissimilarity

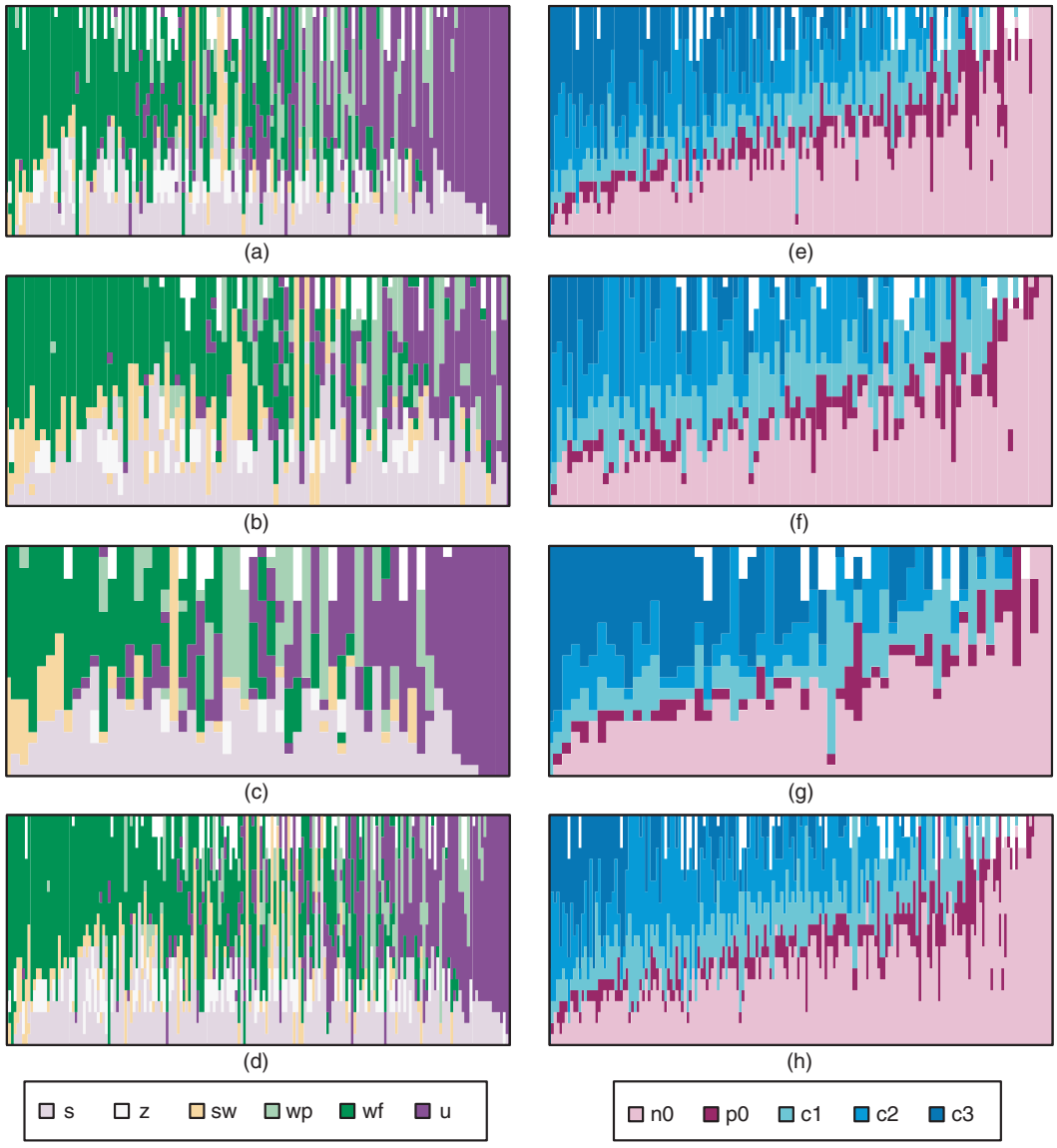

Fig. 7. MDS sequence plots for (a)–(d) the work and the (e)–(h) family domains split according to covariates (ethnicity and religiousness, both with two levels): (a), (e) ethnicity, Asian–African; (b), (f) ethnicity, European–American; (c), (g) religious; (d), (h) not religious

matrix reflects both domains (association) or is instead dominated by one domain (weak or no association). For this we can examine the sequence plots for the two domains, with cases ordered according to the first MDS dimension that is extracted from the joint matrix. For our data, we built sequences by combining activities that are experienced in the family and in the work domain, applied OMA to the combined sequences and extracted MDS dimensions from the resulting joint dissimilarity matrix. In Fig. 6 the sequence plots for family and work are displayed, with sequences ordered according to the first and the second MDS dimensions that are extracted from the joint dissimilarity matrix.

Figs 6(a) and 6(c) are relative to the first joint MDS dimension. Work activities are 'dominating' here: work sequences have quite a clear structure along this dimension, whereas the sequence plot for the family domain appears much more confused. The reverse is true when sequences are ordered according to the second MDS dimension (Figs 6(b) and 6(d)), which is related to the family domain. Since the MDS dimensions are not correlated, the clear separation of the explanation of the two domains along the two dimensions provides further evidence that association is weak and that a joint analysis is not worthwhile. Moreover, reasoning in a 'hierarchical' fashion, we can conclude that the most important distinctions between sequences are those related to work. If cluster analysis is applied to the joint dissimilarity matrix, clusters will be mainly related to the work domain, and the family domain will provide further detail.

5.4.2. Relationships between sequences and external variables

If the relationship with explanatory variables is of interest, different ordered sequence plots may be considered according to the nature of the explanatory variable. If it is quantitative, sequences may be simply ordered according to its values. In the case of categorical explanatory variables, one may instead consider one MDS sequence plot for each level of the variable itself.

For example, in Fig. 7 we split MDS sequence plots according to women's religiousness and ethnicity. In Figs 7(a), 7(b), 7(e) and 7(f), women who were born in or originating from Asian or African countries are distinguished from those from European or American countries or second-generation Israeli Jew. Women who describe themselves as religious or not religious are distinguished in Figs 7(c), 7(d), 7(g) and 7(h). In Figs 7(a)–7(d) are the work sequences plots, ordered according to the first work MDS dimension, whereas in Figs 7(e)–7(h) family sequences are reported, ordered according to the first family MDS dimension.

The sequence plots show that women of African or Asian origin and religious women are more traditional, characterized by higher fertility and smaller delays between subsequent children and by a relatively lower work market activity (note that all the women who never worked are from Africa or Asia). The interesting aspect of these plots is that they permit comparison of the *complete* trajectories relative to the variables considered and not (only) the timing of specific events.

It is interesting to compare what emerges from our plots with the results that were reported by Ekert-Jaffe and Stier (2006), who used bivariate probit models to study the influence of education, religiousness and ethnicity on the timing of the birth of second and third children and on work. They found that ethnicity and religiousness have an influence on the number of children and on the spacing between births, and that work activity does not influence the family variables. All these aspects clearly emerge from the plots above and from those used before to study association between work and family domains.

Interestingly, Ekert-Jaffe and Stier (2006) also found that, contrary to their expectations, religiosity has a positive influence on labour force participation:

> '. . . the necessity of co-providing income in order to insure the minimum standard of living for the family is overcoming the traditional reluctance of observant women to participate in the labour market'.

Fig. 7(c) makes it clear that religious women are relatively more associated with part-time work. Needless to say, it would be interesting to investigate the work career of these women in more detail. This could be done, for example, by analysing more detailed coding of sequences, adding information about the job (status at work and managerial status), to evaluate the possible association with particular kinds of work careers. It would instead be difficult to model highly detailed response variables by using probit models.

Of course exploratory plots provide only descriptive results. Also, contrary to models, they do not give information about the marginal contribution of one explanatory variable given the

others, and they cannot be used for inferential purposes. Still, graphical tools can suggest which events might be modelled and can be used to explore results and to visualize the relationships between trajectories and external information, without necessarily considering syntheses of the life courses themselves (e.g. age when one state was first experienced, average duration of a state and number of times that a state was visited).

6. Conclusions and direction of future research

In this paper we introduced a graphical tool for the exploratory analysis of sequences. Our proposal combines sequence plots (Kohler and Brzinsky-Fay, 2005; Scherer, 2001; Halpin, 2003; Lesnard, 2006; Müller *et al.*, 2008) and MDS dimensions.

As underlined by Francis and Fuller (1996) scientific displays are designed to be viewed on a high resolution computer screen rather than on paper. Their advantages become even clearer when operations such as zooming, rotation and manipulation of colours can be fully exploited. Probably this 'paper version' makes it difficult to appreciate fully the advantages of this graphical visualization of sequences.

The aim of this paper was to point out that sequence visualization is helpful and important in SA. We illustrated how plots can be used to make data-driven decisions about the analysis, supporting the analysts in their choice of the length of sequences and of the coding. We also showed how these plots may be used to visualize the main patterns of sequences in a simple way.

We then extended the technique to obtain graphical displays of the relationships between sequences and external information. This last representation in our opinion is particularly useful since at the moment there is no established technique for exploration of these relationships from a descriptive point of view.

Moreover, as mentioned throughout the paper, inspection of the MDS sequence plots in a hierarchical manner permits understanding which states or durations are most 'responsible' for (and emphasized by) a given dissimilarity measure. Hence, MDS sequence plots can also graphically describe the possibly different features of sequences that are mostly focused by alternative dissimilarity criteria (e.g. OMA with different choices of costs). In this sense, the plots can support sensitivity analyses (see for example McVicar and Anyadike-Danes (2002) and Pollock (2007)) that are conducted to evaluate whether and to what extent SA results depend on the chosen dissimilarity measure. This is particularly interesting since the choice of the dissimilarity criterion in SA is a still open problem.

This exploratory approach may also be useful for comparing methods to obtain the joint dissimilarity matrix when dealing with multiple domains. In Section 5.4 we referred to the approach that is based on the combination of domains (Abbott and Hrycak, 1990; Stovel *et al.*, 1996; Piccarreta and Billari, 2007). Nevertheless other criteria may be followed. For example, when the number of combined states is too large, some researchers (Han and Moen, 1999; Pollock, 2007) proposed to *combine the dissimilarity matrices* that are obtained for different domains into a 'compromise' dissimilarity matrix (using the sum, for example). Our plots may be used to analyse the main differences between different criteria. Also, they can be used to individuate one or more domains which are not explained in a satisfactory way by the joint dissimilarity matrix, whatever the criterion that is used to obtain it.

Before concluding, it is important to point out some possible limits of our proposal. A first consideration concerns the case of large data sets. The CHESA program that we used to calculate dissimilarities imposes no limitations on the size of the data set. Nevertheless, the calculation of an OMA dissimilarity matrix when the sample size and/or the length of sequences is high can be very slow, as underlined by Elzinga (2007) himself (`http://home.fsw.vu.nl/ch.elzinga/`).

Recently, OMA has been implemented in STATA by Brzinsky-Fay *et al.* (2006). As they commented, the program

> 'seems capable of working with a moderate number of relatively short sequences. It has been tested for around 2000 sequences of length up to 100 positions.'

Our experience suggests that the same limits also apply to CHESA. Hence we would conclude that, at least at the moment, OMA can be applied to data sets with up to around 2000 cases (depending on the length of the sequences). With these sizes, metric MDS can be applied with no problems, e.g. with R (R Development Core Team, 2007).

Nevertheless our procedure is not necessarily related to OMA. Alternative dissimilarity measures are available for very large databases (Lesnard, 2006) but, although it might be possible to calculate dissimilarity, the size limitations of MDS become a limiting factor. Recently Tzeng *et al.* (2008) developed a rapid metric MDS for large data sets, but with standard statistical software MDS cannot be applied to very large data sets (say with more than 3000 cases, at least in R).

When many sequences are available and MDS can be applied, standard sequence plots are not very useful so MDS sequence plots constitute an important exploratory tool. Sequences in large data sets are not necessarily more heterogeneous than those in small data sets. Still, in MDS plots describing many sequences, the less important states will probably be masked unless sequences presenting these states are close on the horizontal axes. To have a very detailed visualization of rare states one might consider MDS plots for a subset of cases, or interactively use the plots, e.g. zooming into the careers of some individuals. In any case, if rare states are not visible in the MDS sequence plots they will probably not emerge with cluster analysis either. Also, if a large database contains highly heterogeneous sequences, MDS sequence plots can be useful in cluster analysis: they can be used to explore partitions with a relatively low number of clusters, i.e. MDS sequence plots may be useful to identify the features that are common to sequences that are clustered together.

Finally, when the sample size is so large that MDS cannot be applied, it is possible to gain some insights about the main characteristics of the sequences by considering MDS sequence plots for a subset of observations.

Another important point concerns the possible overinterpretation of the plots and/or the 'overfitting' of the plots to the cases that they are based on. At least with reference to the plots that were considered in Section 5.3 (used to describe the main patterns of sequences), it could be worthwhile to evaluate the 'stability' of the (appearance of the) plots. Validation of the MDS solutions is a still open problem and there is not a well-defined and common approach to this aim (see Hair *et al.* (2007)). In our context we suggest a validation procedure based on the split-sample approach. A random subset of cases is removed from the data set and the MDS solutions (and thus sequence plots) are extracted from the remaining observations. Alternatively, the data set can be randomly divided into K groups and different solutions are obtained by removing one subsample at a time. Under model stability, the MDS sequence plots that are obtained on the basis of the reduced data sets should exhibit the same structure, i.e. the same clustering of states and the same main oppositions.

As a final remark, we point out that an interesting improvement on our plots would be simultaneous representation of multiple domains on the same plot. A promising idea would be to combine our ordering idea with the plots that were introduced by Francis and Fuller (1996).

Acknowledgements

The authors are grateful to two referees, the Joint Editor and the Associate Editor of the journal for helpful and important comments on a previous version of the paper. R. Piccarreta is grateful

to C. Elzinga, for his openness to discussion and his helpful comments, and for making his software available, to A. C. Liefbroer for interesting discussions about sequence analysis and to Jane Klobas for the valuable and detailed comments on the paper. She is also grateful to colleagues at the Department of Social Science Research Methods, Vrije Universiteit Amsterdam (The Netherlands) and to the researchers at the Netherlands Interdisciplinary Demographical Institute, The Hague (The Netherlands) for their comments and suggestions on a previous version of this paper.

The R code to obtain MDS sequence plots is available on request from R. Piccarreta.

The ideas that are expounded in the paper are the result of the joint work of the authors and were first investigated in the graduation thesis of O. Lior (supervised by R. Piccarreta). With reference to the present contribution, R. Piccarreta drafted Sections 1, 4 and 5, and O. Lior drafted Section 6. Sections 2 and 3 were jointly drafted.

References

Aassve, A., Billari, F. C. and Piccarreta, R. (2004) Sequence analysis of BHPS life course data. In *New Developments in Classification and Data Analysis* (eds M. Vichi, P. Monari, S. Mignani and A. Montanari), pp. 275–284. Heidelberg: Springer.

Aassve, A., Billari, F. C. and Piccarreta, R. (2007) Strings of adulthood: a sequence analysis of young British women's work-family trajectories. *Eur. J. Popln*, **23**, 369–388.

Abbott, A. (1995) Sequence analysis: new methods for old ideas. *A. Rev. Sociol.*, **21**, 93–113.

Abbott, A. and Hrychak, A. (1990) Measuring resemblance in sequence data: an optimal matching analysis of musicians' careers. *Am. J. Sociol.*, **96**, 144–185.

Borg, I. and Groenen, P. (1997) *Modern Multidimensional Scaling: Theory and Applications*. Berlin: Springer.

Brzinsky-Fay, C., Kohler, U. and Luniak, M. (2006) Sequence analysis with STATA. *Stata J.*, **6**, 435–460.

Chan, T. W. (1994) Tracing typical mobility paths. *Manuscript*. Nuffield College, Oxford.

Chan, T. W. (1995) Optimal matching analysis: a methodological note on studying career mobility. *Wrk Occupn*, **22**, 467–490.

Ekert-Jaffe, O. and Stier, H. (2006) Normative or economic behavior?: fertility and women's employment in Israel. *Population Association of America Meet., Los Angeles, March 30th–April 1st.*

Elzinga, C. H. (2003) Sequence similarity: a nonaligning technique. *Sociol. Meth. Res.*, **32**, 3–29.

Elzinga, C. H. (2006) Sequence analysis: metric representation of categorical time series. To be published.

Elzinga, C. H. (2007) *CHESA 2.1 User Manual*. Amsterdam: Vrije Universiteit Amsterdam. (Available from `home.fsw.vu.nl/ch.elzinga/`.)

Francis, B. and Fuller, M. (1996) Visualization of event histories. *J. R. Statist. Soc.* A, **159**, 301–308.

Hair, J. F., Black, W., Babin, B., Anderson, R. E. and Tatham, R. (2007) *Multivariate Data analysis*, 6th edn. Englewood Cliffs: Prentice Hall.

Halpin, B. (2003) Tracks through time and continuous processes: transitions, sequences and social structure. *Conf. Frontiers in Social and Economic Mobility, Ithaca, Mar*.

Halpin, B. and Chan, T. W. (1998) Class careers as sequences: an optimal matching analysis of work-life histories. *Eur. Sociol. Rev.*, **14**, 111–130.

Han, S.-K. and Moen, P. (1999) Clocking out: temporal patterning of retirement. *Am. J. Sociol.*, **105**, 191–236.

Kohler, U. and Brzinsky-Fay, C. (2005) Stata tip 25: sequence index plots. *Stata J.*, **5**, 601–602.

Kruskal, J. B. and Wish, M. (1984) *Multidimensional Scaling*. Beverly Hills: Sage.

Lesnard, L. (2006) Optimal matching and social sciences. Institut National de la Statistique et des Etudes Economiques, Paris.

Malo, M. A. and Munoz-Bùllon, F. (2003) Employment status mobility from a life–cycle perspective: a sequence analysis of work-histories in the BHPS. *Demogr. Res.*, **9**, 119–161.

Mardia, K. V., Kent, J. T. and Bibby, J. M. (1979) *Multivariate Analysis*. London: Academic Press.

McVicar, D. and Anyadike-Danes, M. (2002) Predicting successful and unsuccessful transitions from school to work by using sequence methods. *J. R. Statist. Soc.* A, **165**, 317–334.

Müller, N. S., Lespinats, S., Ritschard, G., Studer, M. and Gabadinho, A. (2008) Visualisation et classification des parcours de vie. *Rev. Nouv. Technol. Inform.*, **11**, 499–510.

Piccarreta, R. and Billari, F. C. (2007) Clustering work and family trajectories by using a divisive algorithm. *J. R. Statist. Soc.* A, **170**, 1061–1078.

Pollock, G. (2007) Holistic trajectories: a study of combined employment, housing and family careers by using multiple-sequence analysis. *J. R. Statist. Soc.* A, **170**, 167–183.

R Development Core Team (2007) *R: a Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. (Available from `http://www.R-project.org`.)

Rohwer, G. and Pötter, U. (2004) *TDA User's Manual*. Bochum: Ruhr-Universität Bochum.
Scherer, S. (2001) Early career patterns: a comparison of Great Britain and Germany. *Eur. Sociol. Rev.*, **17**, 119–144.
Schlich, R. (2003) Homogeneous groups of travellers. *10th Int. Conf. Travel Behaviour Research, Lucern, Aug.*
Schoon, I., McCullough, A., Joshi, H., Wiggins, R. and Bynner, J. (2001) Transitions from school to work in a changing social context. *Young*, **9**, 4–22.
Stovel, K., Savage, M. and Bearman, P. (1996) Ascription into achievement. *Am. J. Sociol.*, **102**, 358–399.
Torgerson, W. S. (1958) *Theory and Methods of Scaling*. New York: Wiley.
Tufte, E. R. (1961) *Visual Display of Quantitative Information*. Cheshire: Graphics Press.
Tzeng, J., Lu, H. H.-S. and Li, W.-H. (2008) Multidimensional scaling for large genomic data sets. *BMC Bioinformatics*, **9**, article 179.
Ward, J. H. (1963) Hierarchical grouping to optimize an objective function. *J. Am. Statist. Ass.*, **58**, 236–244.