Predicting Work and Family Trajectories

Raffaella Piccarreta[1] and Francesco C. Billari[2]

[1] Istituto di Metodi Quantitativi, Università Bocconi, Milano, Italy
raffaella.piccarreta@uni-bocconi.it
[2] Istituto di Metodi Quantitativi and IGIER, Università Bocconi, Milano, Italy
francesco.billari@uni-bocconi.it

Preliminary version

Abstract. In the recent literature the study of life courses (or trajectories) as conceptual units has been approached from the statistical point of view. Techniques based on the analysis of sequences of states, and in particular Optimal Matching Analysis (OMA), have been used to build clusters of life courses. Ideal-typical sequences in groups can then be extracted. Proposals on the prediction of life course sequences have been made by McVicar and Anyandike-Danes (2001), who use multinomial logit models to study the determinants of cluster membership, adopting thus a two–step approach. The main problem with this approach is that cluster found not considering the prediction purpose may be hardly predictable. We here propose a new algorithm, modifying the first step of this procedure. Clusters are still obtained considering OMA as a basis for the definition of distance between individual trajectories. Nevertheless, the predictive problem is taken into account also in the first step, when clusters are formed. The aim is to define clusters that are better predictable given a set of covariates. We apply this algorithm to data from the British Household Panel Survey (BHPS) on the employment and family trajectories of women, and we show the advantage of the proposed algorithm.

Keywords: life course analysis, cluster analysis, multinomial logit model, employment and family trajectories.

1 Introduction

In this paper we consider how to analyze individual trajectories over the life course, as collected using retrospective or panel surveys. We start from a representation of life course as sequences of states and we analyze such sequences as a whole conceptual unit, adopting the approach known as "sequence analysis". In particular, our attention is devoted to the problem of *predicting* life courses. More specifically, if S denotes the trajectory followed by a given individual (say, from age a to age b), our aim is to predict S on the basis of a set of explanatory variables that are available before age a.

When analyzing sequences of states, the frequency of specific trajectories will in general be very low, so that it is not possible to directly predict S

using standard techniques such as multinomial logit models, or classification trees (Breiman et al., 1984). This is probably the reason why the prediction problem has so far received less attention in the literature on the sequence analysis of life courses. Most attention has been devoted to the simplification of the data structure; this aim has been often achieved by applying cluster analysis to individuate ideal-types of trajectories.

The key methodological problem of clustering sequences is usually solved in the literature by applying standard clustering methods to a properly defined distance (or, better, dissimilarity) matrix. The main methodological challenge in this case is how to suitably measure dissimilarity between two sequences. *Optimal Matching Analysis* (OMA), introduced to the social sciences by Abbott (see Abbott, 1995) is the most popular approach.

A first proposal in the direction of predicting sequences is the two–step procedure proposed by McVicar and Anyandike-Danes (2001). In this papers, the response variable in the prediction step is not the whole sequence, S, but, rather, the output of the clustering procedure based on OMA, C. Cluster membership is hence explained on the basis of the available explanatory structure, via multinomial logit models.

In the first step of this procedure, we have a first simplification, from the sequence, S, to the cluster, C. In this simplification the prediction problem is not taken into account.

In the second step of the procedure, we have a prediction for C obtained via the multinomial logit model, $\widehat{C}$.

In a sense, the described approach consists of two steps which are not "coherent", in the sense that two different criteria and scopes are pursued.

The main problem with this kind of approach is that a great effort is spent in the description and the characterization of the obtained clusters, C. Nevertheless, as we will show in Section 4, the clusters $\widehat{C}$ obtained after multinomial logit models are likely to have different characteristics when compared to the original clusters. This is due to the fact that the aim of standard clustering techniques is to obtain homogeneous clusters, so that the prediction problem, which is the main object of analysis, is not taken into account when forming clusters.

In this paper a new criterion is presented to implement this two–step procedure. In particular, the first step of the procedure is modified, and clusters are obtained by explicitly taking into account the prediction problem.

We here innovate on the existing literature on sequence analysis by introducing a new technique that allows to obtain clusters of life course sequences that are predictable, given a set of observed covariates. This problem is particularly relevant if one would like to build ideal-typical trajectories for which it is important to assess the determinants. As in the papers by McVicar and Anyandike-Danes (2001), this is clearly motivated by policy purposes.

The paper is structured as follows. In Section 2 we present the data we focus on, and the coding of states for our sequences. In Section 3, we shortly

review the methodological approaches that have been followed in the literature on sequence analysis to obtain clusters. Section 4 describes the two–step approach proposed by McVicar and Anyandike-Danes (2001) and illustrates results obtained for our data. Section 5 is devoted to the description of our prediction–oriented clustering algorithm and to the results of the application to our data.

2 Family-work sequence data and the coding of states

For the empirical analysis of this paper, we focus on family-work trajectories during early adulthood of British women, as surveyed in the British Household Panel Study (BHPS from now onwards). We select women born from 1960 to 1968, and we focus on the age span 13-30. For each woman we build, on a monthly time scale, a sequence-type representation of three life course domains: employment, co-resident partnership, and childbearing. After excluding cases with missing information, we could analyze data on 578 women, each of them with 204 time points.

For the coding of states, we proceed as follows. Employment (**W**) and partnership status (**U**) for each month are represented in a dichotomous manner. For what concerns fertility, each woman has 4 possible states according to her number of children (from 0 to 3 children and over). The (joint) states in the sequences are obtained by combining the categories of the involved domains. For example, **U** means in a certain period a woman has a partner (does not work and has no children), **WU** means a woman is employed and has a partner and **1WU** means a woman is employed, has a partner and has one child. The employment and partnership status can change in either direction at any time. As concerns fertility, once a number of children is reached, women cannot reverse to a lower number of children. All possible life course combinations in a given month yield a total number of 16 states. Table 1 contains a representation of such possibilities.

Table 1. *Life course states possible during each month.*

State	0	U	W	WU	1	1U	1W	1WU	2	2U	2W	2WU	3	3U	3W	3WU
No. children	0	0	0	0	1	1	1	1	2	2	2	2	3+	3+	3+	3+
Employment	N	N	Y	Y	N	N	Y	Y	N	N	Y	Y	N	N	Y	Y
Union	N	Y	N	Y	N	Y	N	Y	N	Y	N	Y	N	Y	N	Y

As concerns the explanatory variables, we have information about the following indicators: Lclassmu and Lclassda, indicating whether or not the mother and the father of the individual has a low social class; Bothpar,indicating whether or not the individual lived with both parents at age 16; Ethnic indicating whether or not the individual has Non-white origin. We also consider the Region and the year of birth (Doby).

3 Clustering sequences: methodological issues

The standard approach to the analysis of life course sequences (or trajectory) is to first define a dissimilarity matrix between trajectories, and then to apply cluster analysis. The definition of this dissimilarity is done using OMA, a method measuring the distance between two individuals by taking into account their full trajectories. The method, originally introduced in molecular biology to study protein or DNA sequences (Sankoff and Kruskal, 1983), was extended to sociology by Andrew Abbott. It was then applied in a number of papers, in most cases to analyze career paths (see among others Abbott and Hrychak, 1990; Scherer, 1999; Schoon et al., 2001; Stovel et al., 1996, Halpin and Chan, 1998; Chan, 1994, 1995; McVicar and Anaydike-Danes, 2001; Schlich, 2003; Malo, Munoz-Bùllon, 2003; Stovel and Bolan, 2004). Abbot and Tsay (2000) provide an overview of the applications in sociology.

The basic idea of OMA is to measure the dissimilarity by properly quantifying the effort needed to transform one sequence into another. In the most elementary approach, a set composed of three basic operations to transform sequences is used, $\Omega = \{\iota, \delta, \sigma\}$, where:

insertion (ι): one state is inserted into the sequence;
deletion (δ): one state is deleted from the sequence;
substitution (σ): one state is replaced by another one.

To each elementary operation ω_i, $\omega_i \in \Omega$, a specific cost can be assigned, $c(\omega_i)$. Suppose that k basic operations have to be performed to transform one sequence into another one. Then, the cost of applying a series of k elementary operations can be computed as:

$$c\left(\omega_1, \omega_2, ..., \omega_k\right) = \sum_{i=1}^{k} c\left(\omega_i\right)$$

The distance between two sequences can thus be defined as the minimum cost of transforming one sequence into the other one. Thus the OMA-distance takes into account the entire sequences.

A main problem in the application of OMA concerns the choice of the costs, which is arbitrary. It is common practice to set $c(\iota) = c(\delta)$ and $c(\sigma) = 2c(\iota)$. Moreover, $c(\iota)$ is usually set equal to 1. There is not general agreement about this choice, and this is maybe one of the major weakness evidenced for OMA by some authors (see Schlich, 2003, for a discussion; see Wu, 2000, and Levine, 2000, for criticisms about OMA, and Abbott, 2000, for a reply to criticisms). As concerns the deletion and insertion operation, their cost is usually set equal to 1.

In the following analyses we adopt a data-driven approach to define substitution costs. We define substitution costs as inversely proportional to observed transition frequencies; this suggestion by Rohwer and Pötter (2002) is implemented in the TDA package. More specifically, consider two states,

a and b. Let $N_t(a)$ and $N_t(b)$ be the number of individuals experiencing respectively state a and state b at time t, and $N_{t,t+1}(a,b)$ be the number of individuals experiencing state a at time t and state b at time $(t+1)$. The transition frequency from state a to state b is:

$$p_{t,t+1}(a,b) = \frac{\sum_{t=1}^{T-1} N_{t,t+1}(a,b)}{\sum_{t=1}^{T-1} N_t(a)}.$$

The substitution cost between state a and b is:

$$c(\sigma; a, b) = 2 - p_{t,t+1}(a,b) - p_{t,t+1}(b,a) \quad \text{if } a \neq b$$

Once the dissimilarity matrix has been defined, standard techniques can be applied to obtain clusters of individuals (McVicar and Anaydike-Danes, 2001, discuss criteria not involving the OMA–step to obtain clusters of sequences). In Aassve et al. (2003), different clustering algorithms were tried (single linkage, complete linkage, centroid, ward, median) to cluster the considered data. The solution provided by Ward's minimum variance algorithm was chosen (Ward, 1963). Actually, the other algorithms tend to define some clusters having very high sizes from small, residual clusters. In what follows, given our emphasis on prediction, we will need to define a sufficiently small number of clusters that allow to estimate a multinomial logit model of cluster membership.

3.1 Ward's clustering algorithm

Before proceeding further, we briefly describe the main characteristics of Ward's clustering algorithm, which is strictly connected to the method we are going to propose.

Consider N individuals to be clustered according to their life sequences. Let $d(i,j)$ denote the distance, or dissimilarity, between the i–th and the j–th individual (in our setting $d(i,j)$ is the OMA-distance between the two individuals).

A measure of the total variability or, better, heterogeneity, characterizing the whole data set is:

$$\mathbf{T} = \sum_{i,j} d(i,j). \tag{1}$$

Suppose now the whole sample is partitioned into G clusters by the partition $\mathcal{C}_G$; a measure of the dispersion within the g–th cluster, C_g, is given by:

$$W(C_g) = \sum_{(i,j)\in C_g} d(i,j), \tag{2}$$

and a measure of within-groups dispersion is given by:

$$\mathbf{W}(\mathcal{C}_G) = \sum_{C_g \in \mathcal{C}_G} W(C_g). \tag{3}$$

The quantity $\mathbf{B}(\mathcal{C}_G) = \mathbf{T} - \mathbf{W}(\mathcal{C}_G)$ can be regarded as a synthesis of the distances *among* or *between* the G groups.

A commonly used criterion to evaluate the adequacy of a clustering solution is to compute $R^2(\mathcal{C}_G) = \mathbf{B}(\mathcal{C}_G)/\mathbf{T} = 1 - [\mathbf{W}(\mathcal{C}_G)/\mathbf{T}]$, i.e., the proportion of the total dispersion accounted for by the G clusters constituting the partition $\mathcal{C}_G$.

In a hierarchical agglomerative clustering algorithm, at each step two clusters have to be joined to form a single cluster. Consider now a G–cluster partition, $(\mathcal{C}_G)$, and suppose that $(G-1)$ clusters have to be obtained by joining two clusters into a single one. Suppose that two clusters, say C_w and C_z are joined to form cluster C_{wz}, and let $\mathcal{C}_{G-1}$ be the resulting partition. By definition, it will be $\mathbf{W}(\mathcal{C}_G) < \mathbf{W}(\mathcal{C}_{G-1})$, and $R^2(\mathcal{C}_G) > R^2(\mathcal{C}_{G-1})$. The increase in the within-groups heterogeneity, will be:

$$\begin{aligned} \Delta_R(\mathcal{C}_G, \mathcal{C}_{G-1}) &= W(C_{wz}) - W(C_z) - W(C_w) \\ &= \mathbf{W}(\mathcal{C}_{G-1}) - \mathbf{W}(\mathcal{C}_G). \end{aligned} \tag{4}$$

In Ward's algorithm the two clusters to be joined are selected by *minimizing* the quantity $\Delta_R(\mathcal{C}_G, \mathcal{C}_{G-1})$ i.e., by minimizing $\Delta R^2_{G-1} = R^2(\mathcal{C}_G) - R^2(\mathcal{C}_{G-1})$. Hence at each step the clusters to be joined are selected so as to minimize the decrease in R^2. Of course, due to the hierarchy of the procedure, the maximization of the criterion is conditioned to the solution (partition) at the previous step.

In the following sections, we will describe the results obtained by applying Ward's algorithm to BHPS data, and the results obtained by applying a multinomial logit model to predict cluster–membership.

4 Clusters and their prediction: a "traditional" approach

In Table 2 we describe the main characteristics of the clusters we obtained using OMA and Ward's algorithm, as concerns the degree of within–heterogeneity. For each cluster we report the sum of dissimilarities between cases in the cluster, $W(C_g)$ (cfr. equation (2)), the number of cases in the cluster, n_g, and the average $\overline{W}(C_g) = W(C_g)/n_g$.

Table 2 Distance Within Clusters

Cluster: g	$W(C_g)$	n_g	$\overline{W}(C_g)$
1	2705244.7	157	17230.858
2	127893.3	100	12789.133
3	241748.9	56	4316.9446
4	3905192.0	150	26034.613
5	2988869.1	115	25990.166
$R^2 = 0.270$			

To better characterize and describe the obtained clusters, we need now a synthesis of cases within the cluster itself. In Aassve et al. (2003) criteria to synthesize clusters of sequences are reviewed and discussed. Following the therein indications, we consider the *medoid* (see Kauffman and Rousseeuw, 1990) of each cluster, i.e., the individual (sequence) which is less distant from all the other individuals in the cluster. The synthesis of the observations in the g–th cluster is then:

$$\hat{g} = \arg\min_{j} \sum_{i \in C_g} d(i,j).$$

In Table 3 we report the medoid of each cluster, together with some summary statistics: the total distance between sequences in the cluster and the medoid, W_M, the averaged distance, $\overline{W}_M$, and the maximum distance W_M^{Max}. Since the number of states is very high, we use a s/p (*state-permanence*)-representation of the medoid–sequence (Aassve et al., 2003). The *state*-sequence is the sequence of the *states* "visited" by an individual (note that the ordering of the visits is important since some states can be visited more than once). The *permanence*-sequence is the sequence of the length of the periods an individual "remained" in each of the visited states. The $s/p-$sequence is obtained by combining state and permanence sequence; for example, the s/p-sequence of the sequence **W-W-WU-WU** is $\mathbf{W_2\text{-}WU_2}$.

Table 3 Medoids and summary statistics

Cluster: 1 W_M =12863.97 $\overline{W}_M$ =81.94 W_M^{Max} =221.12
Medoid: N_{57}-W_{117}-WU_{30}
Cluster: 2 W_M =9302.44 $\overline{W}_M$ =93.02 W_M^{Max} =234.57
Medoid: N_{116}-W_{39}-WU_{49}
Cluster: 3 W_M =3068.79 $\overline{W}_M$ =54.80 W_M^{Max} =119.52
Medoid: N_{63}-W_{40}-WU_{101}
Cluster: 4 W_M =20221.03 $\overline{W}_M$ =134.81 W_M^{Max} =241.47
Medoid: N_{44}-W_{48}-WU_{29}-WUC_7-UC_{25}-WUC_{15}-$WUCC_{16}$-UCC_9-$WUCC_{11}$
Cluster: 5 W_M =21112.35 $\overline{W}_M$ =183.58 W_M^{Max} =338.76
Medoid: N_{65}-W_{15}-WU_4-WUC_2-UC_9-UCC_{65}-$UCCC_{44}$

Our interpretation of the clusters is based on Hakim (2002; 2003), who studies work-family interrelationship by arguing that there is a subset of family-oriented women, who are willing to give priority to family formation over work; there is a subset of work-oriented women, who are on the contrary focused on work and careers; there is a majority of women who try hard to "have the best of both worlds", combining work and family. This categorization will help us in characterizing the five clusters obtained from now onwards.

In Table 4 we report the average number of months spent in each state for each cluster. Notice that the medoids give information about the followed trajectories, whereas the average months give information about the relative importance of each state within the cluster.

Table 4 Months (mean) spent in each state

	Cluster				
Variable	1	2	3	4	5
N	55.23	112.20	57.02	52.15	50.17
U	1.51	5.81	3.82	4.59	1.72
W	109.35	39.95	41.46	44.20	23.96
WU	29.54	36.01	98.53	28.17	6.35
C	0.41	0.25		0.81	12.70
UC	1.90	2.02	1.19	15.53	13.25
WC	0.97	0.76	0.00	1.84	5.83
WUC	2.74	3.62	1.64	19.43	3.39
CC	0.04	0.05		0.48	3.83
UCC	1.77	2.35	0.03	9.06	31.89
WCC	0.08			1.43	1.23
WUCC	0.43	0.88	0.28	20.91	6.95
CCC				0.01	4.81
UCCC				3.16	27.75
WCCC					0.85
WUCCC		0.10		2.23	9.33

Table 3 and Table 4 help us in understanding the main features of each cluster. In *Cluster 1*, the medoid woman studies for 57 months (that is, up to almost 18 years). For 117 months (almost 10 years) she then works without forming a family. At the age of about 28 years she enters a union. Looking at time spent in each state we notice that the **W** (more than 9 years on average) and **WU** states are prevalent, with the presence of children being rare. Women belonging to this cluster thus have work-oriented trajectories, with average education. A similar type of trajectory, with one important difference, is found for women belonging to *Cluster 2*. The medoid women spends about 10 years in the initial state, which indicates a prolongation of education. Only then she starts working, and later starts a union. The prevalence of a long period spent in the initial state is confirmed by the analysis of mean state presence from Table 4. Women belonging to Cluster 2 have work-oriented trajectories, and (probably) higher education. In *Cluster 3*, the difference is that the degree of combination between work and family life is higher with respect to Clusters 1 and 2. In fact, women in this cluster live on average more than 8 years in the **WU**, and the medoid gives a picture consistent with this. Cluster 3 contains women who combine work and partnership, without having children in their early adulthood. A different combination strategy seems characteristic of *Cluster 4*: the medoid woman has a child, then interrupts work for about two years, goes back to work, has a second child, leaves work for less than a year and goes back to work. Women who combine working and having children by going in and out of the labor market belong to this cluster. In fact, Table 4 shows that there is a greater variability in the distribution of time spent in each state. Interestingly, almost 40 months are spent either in **WUC** or in **WUCC**, while almost two years are spent either

in **UC** or in **UCC**. *Cluster 5* contains the more family-oriented trajectories. The medoid woman is an example: she leaves work after the birth of her first child and then goes on up to three children without re-entering the labor market. Women in this cluster spend on average five years either in **UCC** or in **UCCC**, as mothers of two or more children who are not working.

4.1 Application of multinomial models

Following the strategy proposed by McVicar and Anyandike-Danes (2001), we applied multinomial logit models to explain cluster–membership on the basis of the considered explanatory structure. In Table 5 we report the results relative to global effects. In particular, for each effect, we computed the LR statistics as the difference between the log-likelihood of the global model (containing all the variables) and the nested sub-model obtained by excluding the variable itself (the results were obtained using STATA). Hence the null hypothesis states that the reduced model is as significant as the complete one (low p–values "flag" covariates significantly contributing to the prediction of cluster membership).

Table 5 LR Results

Source	DF	LR (chi2)	Pr >chi2
Intercept	4	11.54	0.0211
LCLASSMU	4	13.89	0.0076
LCLASSDA	4	28.71	$< .0001$
ETHNIC	4	5.75	0.2184
BOTHPAR1	4	19.11	0.0007
REGION	68	93.71	0.0211
DOBY	32	38.05	0.2132
LR(full)	116	211.6	0.0000
Pseudo R^2		0.1177	

We notice that the effects are significant, with an exception for ETHNIC and DOBY. Since we are interested in the prediction of sequences, we now analyze the predicted clusters obtained with the multinomial model (hence, for each case we predict the predicted cluster, $\widehat{C}$). In Table 6a and 6b we analyze the cross-tabulation between the clusters obtained with Ward's algorithm and their prediction, and report some association measures.

Table 6(a) Cross-Tabulation of $(C, \widehat{C})$

	Cluster Prediction					
Cluster	1	2	3	4	5	Total
1	63	27	0	46	21	157
2	32	41	1	18	8	100
3	26	5	1	16	8	56
4	37	15	0	69	29	150
5	18	8	0	34	55	115
Total	176	96	2	183	121	578

Table 6(b) Measures of association between C and $\widehat{C}$

Statistic	DF	Value	Prob
Chi-Square	16	143.4173	$< .0001$
Likelihood Ratio Chi-Square	16	128.4685	$< .0001$
Mantel-Haenszel Chi-Square	1	56.4233	$< .0001$
Phi Coefficient		0.4981	
Contingency Coefficient		0.4459	
Cramer's V		0.2491	

Notice that one of the five cluster (Cluster 3) is almost never actually predicted with the available sample.

We are now interested in analysing the characteristics of the cluster predicted using multinomial model, to evaluate if the main characteristics of the original clusters are reproduced or not. At this aim, in Tables 7–9, we analyze predicted clusters using the same criteria referred to in the description of the original clusters.

We observe from Table 7 that there is a consistent decrease in the R^2. By analyzing medoids and the time spent in each state, we observe that for Cluster 1 and Cluster 2 the more relevant states coincide with those characterizing the original clusters, even if we lost the information about the fact that women in Cluster 1 enter a union later as compared to other clusters. As concerns Cluster 4 and 5 we notice that they are more confused as compared to the original ones.

Table 7 Distance Within Predicted Clusters

Pred. Cluster: g	W_g	n_g	$\overline{W}_g$
1	5716297.4	176	32478.96
2	1634689.4	96	17028.01
3	279.4	2	139.73
4	6751448.0	183	36893.16
5	3374424.2	121	27887.80
$R^2 = 0.033$			

Table 8 Medoids and summary statistics (Pred. clusters)

Pred. Cluster: 1 W_M =23863.98 $\overline{W}_M$ =135.5908 W_M^{Max} =309.75
Medoid: N_{64}-W_{77}-WU_{63}
Pred. Cluster: 2 W_M =12463.13 $\overline{W}_M$ =129.82427 W_M^{Max} =309.99
Medoid: N_{106}-W_{58}-WU_{40}
Pred. Cluster: 3 W_M =139.73 $\overline{W}_M$ =69.864998 W_M^{Max} =139.73
Medoid: N_{35}-W_{43}-WU_{85}-WUC_7-UC_9-WUC_7-$WUCC_{16}$-UCC_2
Pred. Cluster: 4 W_M =27807.21 $\overline{W}_M$ =151.95197 W_M^{Max} =339.39
Medoid: N_{65}-W_{73}-WU_{42}-WUC_6-UC_{10}-WUC_1-UC_7
Pred. Cluster: 5 W_M =21848.56 $\overline{W}_M$ =180.56661 W_M^{Max} =291.69
Medoid: N_{66}-W_{61}-WU_{31}-WUC_6-UC_{12}-UCC_5-$WUCC_{10}$-UCC_6-$WUCC_7$

Table 9 Months (mean) spent in each state

	Predicted Cluster				
Variable	1	2	3	4	5
N	63.89	84.49	70.00	57.98	54.30
U	3.25	4.33		2.87	3.34
W	63.96	61.51	37.00	55.82	44.78
WU	40.10	27.10	73.00	32.25	24.83
C	1.95	2.00		2.78	5.18
UC	5.42	4.29	4.50	9.60	10.65
WC	2.30	2.20		2.09	1.45
WUC	7.54	5.04	10.50	8.47	6.66
CC	0.61	0.08		0.74	2.26
UCC	5.81	4.945	1.00	9.75	18.69
WCC	0.26	0.49		0.70	1.20
WUCC	5.41	4.06	8.00	9.12	8.93
CCC		0.50		1.11	2.49
UCCC	2.04	2.1563		8.05	13.43
WCCC				0.29	0.36
WUCCC	1.43	0.80		2.35	5.44

The main results of this exercise show that the predicted clusters may differ from the original ones, and that when adopting this two–step procedure attention must be paid not only to the significance of parameters and of the model but also to the description of the predicted clusters. In general, the use of the multinomial logit two-step approach may lead to predicted clusters which do not "reproduce" the characteristics of the original ones. We think that this may be depend upon the fact that a "good" cluster solution (homogeneous clusters) is not necessarily a response which can be satisfactory predicted.

Hence, a first conclusion of this paper is that a great caution has to be used in the analysis of what we are going to predict using multinomial logit models.

5 A new algorithm for predicting sequences

In this section we introduce a new agglomerative hierarchical clustering algorithm, which takes explicitly into account the fact that cluster analysis is not the primary aim of analysis but, rather, constitutes an intermediate step. This algorithm is tailored for analyses in which the real aim is to obtain a prediction of the sequences or, as an alternative, of a proper simplification of them. This approach can be applied when the explanatory variables are qualitative.

To do this, we start considering what we call "predictable clusters". Please recall that our aim is to predict the response S on the basis of a set of covariates, $X_1, \ldots, X_Q$. Consider now the vector of covariates, $\boldsymbol{X}$.

In the particular case when all the covariates are qualitative, the number of realization of $\boldsymbol{X}$ will possibly be lower than the number of sequences, N. Let now $S_{\boldsymbol{x}}$ be the set of sequences relative to individuals characterized by a given vector of covariates, $\boldsymbol{x}$. Clusters of sequences obtained in this way are the clusters which can be predicted at best on the basis of the available explanatory structure. Let now $\mathcal{C}_{K^*}$ be the qualitative variable indicating cluster–membership corresponding to this partition, and let X^* indicate the univariate qualitative variable obtained by compounding the categories of the covariates (i.e., X^* is the qualitative variable corresponding to the realizations of vector of covariates $\boldsymbol{X}$). K^* indicates the number of clusters in the initial partition and, by definition, it coincides with the number of categories of X^*.

The initial partition $\mathcal{C}_{K^*}$ can now be evaluated from two different points of view.

Remaining in logic described in the previous section, we can evaluate the within–groups heterogeneity characterizing the partition, $\mathbf{W}(\mathcal{C}_{K^*})$ (see equation (3)). The quality of the initial partition can be evaluated by referring to the R^2–type measure:

$$R^2(\mathcal{C}_{K^*}) = 1 - \frac{\mathbf{W}(\mathcal{C}_{K^*})}{\mathbf{T}},$$

$\mathbf{T}$ indicating the heterogeneity within the whole sample as in (1).

Remember from the Section 3 that, following this logic, the criterion to evaluate the passage from a K–clusters partition $\mathcal{C}_K$, to a $(K-1)$–clusters partition, $\mathcal{C}_{K-1}$ is:

$$\Delta_R(\mathcal{C}_G, \mathcal{C}_{G-1}) = \mathbf{W}(\mathcal{C}_{G-1}) - \mathbf{W}(\mathcal{C}_G). \tag{5}$$

Hence the clusters to be joined at each step are selected so as to minimize $\Delta_R(\mathcal{C}_G, \mathcal{C}_{G-1})$, i.e., by minimizing $\Delta R^2 = R^2(\mathcal{C}_K) - R^2(\mathcal{C}_{K-1})$. In particular, Δ_R can be considered as the distance between two clusters (the less distant clusters are joined).

Notice that, should we follow this agglomerative procedure, the explanatory structure would not be taken into account in building clusters, except at the initial step (the initial partition).

As it was mentioned before, the initial partition can also be evaluated from a second point of view. In particular, the association between the initial partition and the compounded explanatory variable X^* can be measured by means of the likelihood-ratio statistic G^2:

$$G^2(\mathcal{C}_{K^*}, X^*) = 2 \sum_{k=1}^{K^*} \sum_{j=1}^{K^*} m_{kj} \log \frac{m_{kj} \cdot N}{n_k n_j},$$

where m_{kj} is the number of cases in the k–th cluster characterized by the j–th category of X^* and n_k and n_j are the marginal absolute frequencies.

Consider now a K–clusters partition $\mathcal{C}_K$, and suppose that $(K-1)$ clusters have to be obtained by joining two clusters into a single one. Suppose that two

clusters, say C_w and C_z are joined to form cluster C_{wz}. In the association–logic, we are substantially collapsing two rows of the two-way contingency table displaying the distribution of the N cases according to the categorical variables, X^* and $\mathcal{C}_K$. Let $G^2(\mathcal{C}_{K-1}, X^*)$ denote the G^2 of the new table, and let $G^2(C_{wz}, X^*)$ denote the G^2 characterizing the sub-table constituted by the two joined rows. It can be easily shown that:

$$G^2(\mathcal{C}_K, X^*) = G^2(\mathcal{C}_{K-1}, X^*) + G^2(C_{wz}, X^*)$$

Hence, as the number of clusters decreases, the association between the partition and the explanatory structure decreases. In particular, it is:

$$\Delta_G(\mathcal{C}_G, \mathcal{C}_{G-1}) = G^2(\mathcal{C}_K, X^*) - G^2(\mathcal{C}_{K-1}, X^*) = G^2(C_{wz}, X^*) \qquad (6)$$

By referring to the G^2–logic, the two clusters to be joined should be selected by *minimizing* Δ_G. Notice that, should we follow *only* this agglomerative procedure, the heterogeneity within clusters, at the basis of the R^2 approach, would not be taken into account in building clusters.

Now notice that in the two–step approach followed by McVicar and Anyandike-Danes (2001), clusters are firstly determined under a purely R^2–logic. Subsequently we refer to the explanatory variables to predict cluster membership via multinomial logit models, following, in a sense, a G^2–logic.

Our idea is to obtain clusters by explicitly taking into account that in the second step of analysis they have to be predicted on the basis of a multinomial logit model. Hence also the G^2 logic should be taken into account in the formation of clusters. To do this, we consider as a starting point the partition $\mathcal{C}_{K^*}$, induced by the (compounded) categories of the explanatory variables. Remember that this partition is the one which can be predicted at best on the basis of the available explanatory structure.

At each step of the procedure, the clusters to be joined are selected by referring either to the R^2 and to the G^2 logic. This is done by considering the average of Δ_R and Δ_G (more precisely both the Δ's are adjusted so as to have the same relative importance, since they have different range). The average is taken as a measure of the distance between two clusters. Of course one can decide to give more or less weigth to one of the component[1].

In Table 10–12 we describe the clusters obtained using our algorithm. Notice from Table 10 that the R^2 of the obtained partition is lower than that characterising Ward's solution (which is only R^2–oriented).

[1] It is maybe worthwhile to point out that it is very simple to update the dissimilarity matrix once two clusters are joined so the algorithm is not more computational expensive than Ward's algorithm.

Table 10 Distance Within Clusters

Cluster: g	W_g	n_g	$\overline{W}_g$
1	784203.26	77	10184.458
2	8188211.00	199	41146.789
3	960138.12	81	11853.557
4	1195221.80	100	11952.218
5	3115593.60	121	25748.707
$R^2 = 0.15$			

Table 11 Medoids and summary statistics

Cluster: 1 W_M =7336.92 $\overline{W}_M$ =95.28 W_M^{Max} =255.55
Medoid: N_{77}-W_{40}-WU_{87}
Cluster: 2 W_M =32306.32 $\overline{W}_M$ =162.34 W_M^{Max} =339.21
Medoid: N_{66}-W_{61}-WU_{31}-WUC_6-UC_{12}-UCC_5-$WUCC_{10}$-UCC_6-$WUCC_7$
Cluster: 3 W_M =8468.25 $\overline{W}_M$ =104.55 W_M^{Max} =295.95
Medoid: N_{107}-W_{58}-WU_{39}
Cluster: 4 W_M =8699.74 $\overline{W}_M$ =87.00 W_M^{Max} =211.79
Medoid: N_{61}-W_{107}-WU_{36}
Cluster: 5 W_M =20782.38 $\overline{W}_M$ =171.75 W_M^{Max} =309.47
Medoid: N_{40}-W_{42}-WU_{27}-WUC_6-UC_{22}-UCC_5-$WUCC_6$-UCC_{21}-$WUCC_{18}$-UCC_2-$WUCC_{15}$

Table 12 Months (mean) spent in each state

	Cluster				
Variable	1	2	3	4	5
N	68.82	55.32	102.39	60.52	49.78
U	9.16	2.42	2.27	1.80	3.03
W	38.57	60.85	49.77	101.30	29.99
WU	76.03	26.33	27.36	34.79	15.89
C		7.13	0.38	0.02	1.81
UC	4.25	10.78	3.47	1.48	12.55
WC		4.33	0.93		1.97
WUC	6.12	11.89	3.91	1.67	7.16
CC		1.29	0.10	0.06	2.10
UCC	0.54	6.79	7.80	1.99	27.44
WCC		0.45	0.16	0.12	2.09
WUCC	0.52	6.31	3.89	0.25	20.44
CCC		0.90			3.10
UCCC		4.55	0.52		22.45
WCCC		0.43	0.02		0.08
WUCCC		4.22	1.02		4.09

For the sake of simplicity, we interpret these clusters in reference to the ones found using Ward's algorithm. *Cluster 1* is a cluster that resembles Cluster 3 in Table 4, with a combination of work and family without a key role of childbearing. The **WU** state is the most frequent on average and also for the medoid. *Cluster 2* and *Cluster 3* present combination trajectories, with lower education for Cluster 2, and the presence of children. The medoid

of Cluster 2 has family-related interruptions of work (similarly to Cluster 4 in Table 4). For what concerns *Cluster 4*, this is clearly a cluster of work-oriented trajectories, with a clear prevalence of the **W** state, similar to Cluster 1 in Table 4. *Cluster 5* includes the most family-oriented trajectories, with around 50 months spent in either **UCC** or **UCCC** – this is similar to Cluster 5 obtained using Ward's algorithm, although a higher level of labor force participation is visible from Table 13 or from the medoid woman in Table 12.

A multinomial logit model was applied to explain cluster membership. Notice that this model can not be directly compared to the one relative to Ward's solution, since the response variable is different. Nevertheless, from Table 13 it is possible to notice an improvement in the significance of the explanatory variables (they are now all significant and the significance level is increased) as well as in the global measures (Likelihood Ratio and Pseudo R^2). Results in Tables 14a–14b emphasize the higher association between clusters and their predictors.

Table 13 LR Results

Source	DF	LR (chi2)	Pr >chi2
Intercept	4	11.54	0.0211
LCLASSMU	4	16.36	0.0026
LCLASSDA	4	50.40	0.0000
ETHNIC	4	14.79	0.0052
BOTHPAR1	4	39.72	0.0000
REGION	68	186.44	0.0000
DOBY	32	92.85	0.0000
LR(full)	116	411.85	0.0000
Pseudo R^2		0.2311	

Table 14(a) Cross-Tabulation of $(C, \widehat{C})$

	Cluster Prediction					
Cluster	1	2	3	4	5	Total
1	19	19	11	13	15	77
2	6	153	10	9	21	199
3	13	20	32	10	6	81
4	11	30	14	32	13	100
5	8	36	6	3	68	121
Total	57	258	73	67	123	578

Table 14(b) Measures of association between C and $\widehat{C}$

Statistic	DF	Value	Prob
Chi-Square	16	309.6489	$< .0001$
Likelihood Ratio Chi-Square	16	269.6728	$< .0001$
Mantel-Haenszel Chi-Square	1	107.3682	$< .0001$
Phi Coefficient		0.7319	
Contingency Coefficient		0.5906	
Cramer's V		0.3660	

Turning now attention to the predicted clusters, we observe from Table 15 that the increase in the R^2 is not so significant (it was 0.033 for the predictors of Ward's clusters).

Table 15 Distance Within Predicted Clusters

Pred. Cluster: g	W_g	n_g	$\overline{W}_g$
1	566635.62	57	9940.97
2	13655021.00	258	52926.44
3	977891.22	73	13395.77
4	751760.94	67	11220.31
5	3353344.3	123	27262.96
$R^2 = 0.039$			

Table 16 Medoids and summary statistics (Pred. clusters)

Pred. Cluster: 1 W_M =7655.07 $\overline{W}_M$ =134.30 W_M^{Max} =267.96
Medoid: N_{64}-W_{77}-WU_{63}
Pred. Cluster: 2 W_M =40658.07 $\overline{W}_M$ =157.59 W_M^{Max} =339.96
Medoid: N_{63}-W_{84}-WU_{57}
Pred. Cluster: 3 W_M =9872.44 $\overline{W}_M$ =135.24 W_M^{Max} =309.99
Medoid: N_{106}-W_{58}-WU_{40}
Pred. Cluster: 4 W_M =8122.01 $\overline{W}_M$ =121.22 W_M^{Max} =305.79
Medoid: N_{69}-W_{83}-WU_{52}
Pred. Cluster: 5 W_M =21237.22 $\overline{W}_M$ =172.66 W_M^{Max} =291.69
Medoid: N_{66}-W_{61}-WU_{31}-WUC_6-UC_{12}-UCC_5-$WUCC_{10}$-UCC_6-$WUCC_7$

Table 17 Months (mean) spent in each state

	Predicted Cluster				
Variable	1	2	3	4	5
N	77.05	58.69	81.75	65.57	55.14
U	4.39	2.93	4.96	2.03	3.36
W	55.56	60.65	55.51	68.10	44.23
WU	38.58	31.25	27.82	42.82	28.86
C	2.39	3.65	3.14	0.12	2.89
UC	5.65	8.77	4.85	4.48	9.61
WC	0.96	2.67	2.51		2.02
WUC	6.17	8.94	6.68	4.42	6.07
CC	0.23	0.67	0.40	0.13	2.45
UCC	5.95	8.81	6.89	5.92	16.54
WCC	0.70	0.80	0.16	0.30	0.72
WUCC	6.37	8.03	3.97	2.66	9.79
CCC		0.68	1.26		2.33
UCCC		5.05	3.75	3.16	15.26
WCCC		0.21	0.20	0.03	0.22
WUCCC		2.21	0.14	4.25	4.49

The analysis of the medoids and of the relative importance of each state, in terms of months spent in each of them, evidence a higher capability of the

predicted cluster to reproduce the original clusters. The only exception is for Cluster 2, which is characterized by a medoid quite different from the one characterizing the original cluster. Nevertheless, by analyzing Table 17 we can observe that the relevant states are the same, even if their importance is lower (and this explains the difference in the medoids).

The obtained results evidence that, differently from what happens with Ward's clusters, "our" clusters are all predicted by the multinomial logit model (observe also that the misclassification rate is lower for our clusters). Moreover, the explanatory variables are more significant. In a sense, we are "trying to exploit at best" the predictive capability of the explanatory variables.

We can now draw some general conclusions. As a first consideration, we can say that the the cluster solution provided by our combined agglomerative criterion is worse than the one obtained by applying Ward's algorithm in terms of R^2. This means that the latter algorithm leads to clusters which are more homogeneous. Nevertheless, the main point here is that if the aim of analysis is to predict sequences, the R^2 criterion can not be the only one referred to. Actually, we have to evaluate if the best R^2 partition is also a predictable one.

In this sense, our algorithm favours a partition which is maybe less homogeneous but which can be predicted better. This is particularly important in cases when the multinomial model leads to predictors which do not reproduce in a satisfactory way the R^2–best clusters.

This is particularly relevant when conclusions drawn on the basis of the results of multinomial logit models have to be referred to by policy makers.

References

Aassve, A., Billari, F. C., Piccarreta, R. (2003) Sequence Analysis of BHPS Life Course Data, *Book of Short Papers, CLADAG 2003, Classification and Data Analysis Group Italian Statistical Society*, Bologna, September 2003, 13–16.

Abbott, A. (1995) Sequence Analysis: New Methods for Old Ideas, *Annual Review of Sociology*, **21**, 93–113.

Abbott, A. (2000) Reply to Levine and Wu, *Social Methods and Research*, **29**, 65–76.

Abbott, A. and Hrychak, A. (1990) Measuring Resemblance in Sequence Data: An Optimal Matching Analysis of Musicians' Careers, *American Journal of Sociology*, **96**, 144–185.

Abbott, A. and Tsay, A. (2000), Sequence Analysis and Optimal Matching Methods in Sociology: Review and Prospect, *Social Methods and Research*, **29**, 3–33.

Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. (1984) *Classification and Regression Trees*, Belmont CA: Wadsworth.

Chan, T.W. (1994) *Tracing Typical Mobility Paths*, Ms Nuffield Coll. Oxford.

Chan, T.W. (1995) Optimal Matching Analysis: A Methodological Note on Studying Career Mobility, *Work and Occupations*, **22**, 467–490.

Crawford, R. M. M. and Wishart, D. (1967) A Rapid Multivariate Method for the Detection and Classification of Groups of Ecologically Related Species, *J. Ecology*, **55**, 505–524.

Dex, S. (Ed.) (1991) *Life and Employment History Analyses: Qualitative and Quantitative Developments*, London: Routledge.

Edwards, A. W. F. and Cavalli-Sforza, L. L. (1965) A Method for Cluster Analysis, *Biometrics*, **21**, 363–375.

Everitt, B. S. (1993) *Cluster Analysis*, London: Arnold.

Gini, C. (1954) *Variabilità e concentrazione*, Roma: Veschi.

Hakim, C. (2002): Lifestyle Preferences as Determinants of Women's Differentiated Labor Market Careers, *Employment and Occupations*, **29**, 428–459.

Hakim, C. (2003) A New Approach to Explaining Fertility Patterns: Preference Theory, *Population and Development Review*, **29**, 349–374.

Halpin, B. and Chan, T.W. (1998) Class Careers as Sequences: An Optimal Matching Analysis of Work–Life Histories, *European Sociological Review*, **14**, 111–130.

Hubert, L. (1973) Monotone Invariant Clustering Procedure, *Psychometrika*, **38**, 47–62.

Jobson, J. D. (1992) *Applied Multivariate Data Analysis*, Vol. II: *Categorical and Multivariate Methods*, New York: Springer-Verlag.

Kauffman, L. and Rousseeuw, P. J. (1990) *Finding Groups in Data*, New York: John Wiley and Sons.

Lance, G. N. and Williams, W. T. (1965) Computer Programs for Monothetic Classification (Association Analysis), *Computer J.*, **8**, 246–249.

Lance, G. N. and Williams, W. T. (1968), Note on a New Information–Statistic Classificatory Program, *Computer J.*, **11**, 195.

Levine, J.H. (2000) But What Have You Done for Us Lately?, Commentary on Abbott and Tsay, *Social Methods and Research*, **29**, 34–40.

Macnaughton–Smith, P., Williams, W. T., Dale, M. B., and Mockett, L. G. (1964) Dissimilarity Analysis a New Technique of Hierarchical Sub–division, *Nature*, **202**, 1034–1035.

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, *Demographic Research*, **9**, 119–161.

McVicar, D. and Anyadike-Danes, M. (2001), Predicting Successful and Unsuccessful Transitions from School to Work by Using Sequence Methods, *Journal of the Royal Statistical Association, Series A*, **165**, 317–334.

Piccarreta R. (2003) CART for Distance or Dissimilarity Matrices, *Studi Statistici*, **80**, Istituto di Metodi Quantitativi, Bocconi University , Milan, December 2003.

Rohwer, G., Pötter, U. (2000) *TDA User's manual*, Ruhr-Universitat Bochum, Bochum.

Sankoff, D. and Kruskal, J.B. (1983) *Time Warps, String Edits and Macromolecules*, Reading, MA: Addison-Wesley.

Scherer, S. (1999) Early Career Patterns: A Comparison of Great Britain and Germany, *European Sociological Review*, **17**, 119-144.

Schlich, R. (2003), Homogeneous Groups of Travellers, Paper presented at the 10th International Conference on Travel Behaviour Research, Lucern, August 2003.

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.

Scott, A. J. and Symons M. J. (1971) On the Edwards–Cavalli Sforza Method of Cluster Analysis, *Biometrics*, **27**, 217–219.

Siegers, J. J., De Jong-Gierveld, J., and Van Imhoff, E. (Eds.) (1991) *Female Labour Market Behaviour and Fertility: A Rational-Choice Approach*, Berlin/Heidelberg/New York: Springer-Verlag.

Stovel, K., and Bolan, M. (2004) Residential Trajectories. Using Optimal Alignment to Reveal the Structure of Residential Mobility, *Sociological Methods & and Research*, **32**, 559–598.

Stovel, K., Savage, M. and Bearman, P. (1996) Ascription into Achievement, *American Journal of Sociology*, **102**, 358–399.

Ward, J.H. (1963) Hierarchical Grouping to Optimize an Objective Function, *Journal of the American Statistical Association*, **58**, 236–244.

Williams, W. T. and Lambert, J. M. (1959) Multivariate Methods in Plant Ecology I. Association Analysis in plant communities, *J. Ecol.* **47**, 83–101.

Wu, L. L. (2000) Some Comments on Sequence Analysis and Optimal Methods in Sociology: Review and Prospect, *Sociological Methods and Research*, **29**, 41–64.