Class Careers as Sequences: an Optimal Matching Analysis of Work-Life Histories

Brendan Halpin   Tak Wing Chan

We apply optimal matching techniques to class careers from age 15 to age 35 for two moderately large samples, as a means of exploring the utility of this sequence-oriented approach for the analysis of work-life social mobility. We first apply multi-dimensional scaling techniques to the inter-sequence distances generated by the optimal matching algorithm in order to test whether the technique locates sequences in a coherent and interpretable space. We find the space to be highly patterned and reasonably interpretable. Next we run the two moderately large samples (each approximately 1,500 sequences) through the analysis and examine the nature of the set of clusters that emerges. We find the clusters to be distinct and an intuitively attractive grouping of the sequences. Finally, we consider how the clusters are distributed across cohorts: the distributions change markedly, though this is largely due to changes in the distribution of classes over time. We briefly discuss means of separating 'pure sequence' change from change in the gross 'class time-budget' of cohorts, and consider means of applying statistical models to the problem. We conclude by endorsing Optimal Matching Analysis, especially as a means of exploratory analysis of longitudinal data.

Introduction: enthusiasm and scepticism

Data on individuals' life histories can be represented as sequences of states, and the sociological analysis of such data can benefit from the importation from other disciplines of techniques designed to work with sequence data. This paper reports our experiences applying a technique well-known in molecular biology, Optimal Matching Analysis, to longitudinal data on class careers.

Techniques currently popular in sociology for the analysis of life-history and other longitudinal data, such as models of transition matrices or hazard models, are powerful and flexible, but generally do not deal with sequences holistically. The inherent complexity of sequences (for an *n*-category state space, over a period *m* units long, there are $\sum_{i=1}^{m} n^i$ different possible sequences), and the lack of techniques to handle them directly, means that it is difficult to get an overview of the patterns in a longitudinal data set. Sequence analysis techniques such as optimal matching appear to offer a useful addition to the longitudinal analyst's tool-kit: the ability to generate typologies of sequences empirically. They do this by calculating an inter-sequence distance measure which can subsequently be subjected to a cluster analysis.

We came to Optimal Matching Analysis (OMA) with a mixture of enthusiasm and scepticism. We are predisposed to be wary of black boxes, of cluster analysis in general, and of claims that OMA and cognate procedures represent a movement from variable-centred to narrative-centred analyses (Abbott, 1988), but a means of comparing *sequences* is precisely what existing analyses of longitudinal data on class careers is lacking (Chan, 1995). Furthermore, the optimal matching algorithm is, at least *prima facie*, intuitively reasonable. We therefore decided to explore the procedure, to see what sorts of results it generates and to see what sort of sense it makes. We applied it to a problem Halpin had been dealing with using other techniques (Halpin, 1993), namely to test for historical change evidenced in cross-cohort difference in patterns of class mobility during the work life.

The Optimal Matching algorithm

Optimal Matching Analysis is a set of techniques routinely used by molecular biologists in the study of DNA or protein sequences, often as a tool for reconstructing evolutionary trees. It was introduced into sociological analysis by Andrew Abbott who has applied OMA to several substantive issues, such as the development of the welfare state (Abbott and DeViney, 1992), the order of professionalisation (Abbott, 1991), the careers of musicians (Abbott and Hrycak, 1990), and Morris dances (Abbott and Forrest, 1986).[1]

OMA offers a means to analyse data in the form of complete sequences of events, such as career histories. Its basic idea is very simple. Suppose the career history of two people, observed at five-year intervals, can be represented as follows:

- person A
  - unskilled manual worker
  - unskilled manual worker
  - skilled manual worker
  - unskilled manual worker
  - unskilled manual worker
- person B
  - unskilled manual worker
  - unskilled manual worker
  - foreman
  - small employer

What OMA does is to count how many substitutions, insertions or deletions ('elementary operations') are needed in order to turn sequence A into sequence B, or vice versa. In this example, both A and B started their career as unskilled manual workers. A took up a skilled manual job mid-way through his work life, but he returned to an unskilled manual position in the end. This is largely a case of career immobility. In contrast, B became a foreman at about the same time as A took up the skilled manual job. She then moved on to become a small employer. One possible way to turn sequence A into sequence B is to substitute ‘foreman’ for ‘skilled manual worker’ (third observation), ‘small employer’ for ‘unskilled manual worker’ (fourth observation), and then delete ‘unskilled manual worker’ (fifth observation).

Note that each elementary operation deals with a pair of units between two sequences, rather than the position of these units in relation to other units in their own sequences. In this sense, each elementary operation is blind to the temporal order of events. However, in comparing two sequences a string of these elementary operations is carried out, and it is this repeated execution of local elementary operations that carries the sequential and temporal information. Now consider that each substitution, insertion or deletion incurs a ‘cost’ to the pair of sequences under comparison – the more substitutions one makes, the higher the cost, and the greater the distance between the pair (the same for insertion and deletion). Optimal matching algorithms are designed to find (or approximately find, where exact solutions are impractical) the ‘cheapest’ set of transformations between sequence pairs, and thus an overall similarity score (or distance score, depending on how the calculation is programmed) can be given to each pair of sequences. Deciding these costs is a key theoretical issue in OMA, to which we return later. But once such comparison is repeated for all sequence pairs in the sample, we have a measure of how much every sequence resembles every other. These similarity scores can then be used in a clustering procedure which will suggest how many clusters of sequences there are, and what the typical sequence of each cluster looks like.

One major reservation we had to begin with was the relative indeterminacy of cluster analysis. Cluster analysis is not supported by a large body of statistical reasoning. It can be sensitive to the specific linkage mechanism employed – different rules of cluster formation can give different solutions to the same data set. At the same time, cluster analysis will always give a solution even if there is no meaningful structure in the data (Aldenderfer and Blashfield, 1984). Moreover, it can be very sensitive to the actual sample it deals with, so a large sample is necessary in order to expect to get repeatable, stable results. Similarly, since the clusters depend on the sample, it is not possible to directly relate the clusters in one sample to those in another: there is no guarantee that they will correspond.

We dealt with this reservation in two separate ways: first we avoided cluster analysis altogether and examined the inter-sequence distances in terms of multidimensional scaling (MDS), and second we used as large an *N* as possible in cluster analyses of entire samples (we thus reduced sampling variability to the minimum possible, and were able to examine how the distribution across clusters varied by sub-group within the large sample, rather than having to compare separate sets of clusters across subsamples).

However, as Abbott and Forrest remark (1986), there is an inherent problem with analyses which involve procedures of pairwise comparison: the processing time required rises with the square of, rather than in direct proportion to, the number of cases. This results in strong constraints on the size of sample we can deal with. In the case of the MDS analysis we report below we could not exceed 400 cases at a time, but we managed to process entire samples (approximately 1,500 cases) in the cluster analyses.

Before reporting our findings, we briefly review existing tools for analysing longitudinal work-life data.

Longitudinal work-life data and analysis

Analyses of class careers

Conventionally the analysis of social mobility is in terms of two-point inter-generational trajectories, analysed as a table of class of family of origin by respondent's mature class position (see, for example, Goldthorpe, 1987). As fuller data came to be available, three-point trajectories (class of origin, class at entry to the labour market, mature class) were analysed (Goldthorpe, 1987, ch. 5) A large body of research now exists focusing on each of the two legs of this three-point journey. Much research on education is addressed to the issue of how it places individuals in the labour market, and how its effect coexists and interacts with the effect of class of origin (*e.g.*, König and Müller, 1986; Jonnson, 1993). Correspondingly, with the collection of more and more longitudinal data covering the whole work life, more analyses of work-life mobility are being carried out. However, the greatest problem of longitudinal data is often its richness: it contains so much information that it is difficult to use fully. Powerful statistical techniques can be applied to it, but they almost always involve discarding some part of the information. For example, the entry-class by mature-class table can be analysed by powerful loglinear models giving useful insights into the overall structure of the processes governing work-life mobility. But this discards all information on what happens between these two points.

A second approach translates the individual trajectories into class spells and tabulates spells (*i.e.*, one or more per individual) according to the class in which they occur and their outcome (either transition to another class or persistence until the moment of data collection, 'censoring') (Featherman and Selbee, 1988; Halpin, 1993, ch. 5). This has the advantage of focusing explicitly on the pattern of interchange between class locations, looking, as it were, at each *step* of the journey but at the expense of losing all information on overall trajectory and on durations.

A third approach considers duration spent in various class as the most important attribute (an early example of such analysis is provided by Rogoff Ramsøy (1975), more recently Mayer (1991), Gershuny (1993)). We can analyse duration directly, for instance, by modelling cumulative duration in each category in terms of the mature class position (Halpin, 1993, ch. 5). This gives insights on the sorts of background people in different mature classes have. However, it loses a great deal of information too, as very many different career trajectories could generate the same distribution of cumulated duration.

Another very powerful technique, and one more amenable to the introduction of explanatory variables, is hazard modelling (for a relevant application see Carroll and Mayer, 1986). It represents a very searching look at the dynamics of transition from one state to another, and can therefore take good account of both duration and the individual steps of the trajectory. It can also take account of prior history (and concurrent events) via time-dependent variables. For instance, in their analysis of how schooling and career experience affect entry rates into marriage and motherhood, Blossfeld and Huinink (1991) incorporate a time-dependent variable which measures the level of career resources. This variable is constructed to take account of the number and length of previous spells of work, job changes, career interruptions, and so on. Blossfeld and Huinink demonstrate how hazard modelling can be skillfully exploited to take account of past history. But note that they were guided by prior knowledge and theoretical models in their construction of the career resources variable. In situations where prior knowledge is relatively thin, it will be difficult to construct precise covariates that tap past history. In other words, open-ended, exploratory attempts to understand patterns of sequence will not be as well served by hazard modelling: if we try to include more than a small number of variables representing history prior to the current state, we find it becomes quickly unwieldy, and difficult to estimate (Halpin, 1993, ch. 6).

Each of these techniques is powerful, nevertheless. Furthermore, they are supported by conventional statistical theory, in which terms the results can be precisely defined and interpreted, in regard to fit and significance. This is no small advantage.

But how do we take holistic account of the career? How can we treat it as a sequence? OMA seems to offer a way of doing this, though not, perhaps as yet, with the full power that conventional statistics offers. Its greatest immediate utility is the generation of empirical typologies of sequences: sequences by their nature are highly detailed. The set of all logically possible sequences is enormous, and it is impractical to attempt a systematic typology of the entire logically possible set. However, it is clear that the sequences that actually exist are drawn from a highly patterned subset of the possible set because there is a ‘logic’ to the progress along the sequence. In terms of class careers this logic can be thought of in terms of context-specific transition rates. This logic is the object of our study, and therefore not something we can use as an input, so the possibility of generating an empirical typology of sequences, for further analysis, is very attractive.

Other approaches to sequence analysis

OMA is not the only method for the analysis of sequential data. In the context of life-course and the labour market, various other techniques are currently in use. These techniques tend to have simpler algorithms, and so they allow for easier implementation or easier access to suitable software. But they are less general than the optimal matching algorithm.

For example, Buchmann and Sacchi (1995) analyse occupational life-histories of two cohorts of Swiss workers. They first apply factor analysis to a battery of occupational-level variables, reducing what would have been an intractably large set of categories to a tractable five-dimensional space. However, their method assumes there is a one-to-one correspondence between elements occupying the same position or time point in two different sequences. Given this assumption, they define the overall distance between two sequences as the mean of the distance between the sequences at each time point. The assumption of one-to-one correspondence may be justifiable in certain situations, but it precludes taking account of the possibility that similar patterns may display at different places in the sequences. Consider a simple situation where there are only two states, `A` and `B` (say, 'in employment' and 'not in employment'). Suppose there are two sequences with the following patterns:

- Sequence 1: `ABABABAB`
- Sequence 2: `BABABABA`.

We may think that sequences 1 and 2 are substantively the same in that both are highly unstable, alternating rapidly between state `A` and state `B`. But given the assumptions of one-to-one correspondence, the two sequences are maximally different from each other.

Degenne, Lebeaux and Mounier (1996) analyse the early labour market experience of a sample of young French women, with a tractably small state space (unemployment, training, inactivity, different sorts of employment contracts, *etc.*). Their approach is to summarise the career histories at discrete intervals, in terms of (a) cumulated duration in the several states during that interval and (b) month-by-month transition matrices. They then apply a factor analysis to the summary variables, and use this as the basis for a cluster analysis. This method is attractive because it makes use of information on both durations and transitions, but it is not yet adequately explored.

Degenne et al. also propose a 'blind' technique that is interesting: it is blind in that it makes no assumption about distances between states in the state space. They define inter-sequence distance to be the sum, across observation points, of the angle between the vectors of cumulated duration. That is,

$$D_{ij} = \sum_t \cos^{-1}(\mathbf{X}_{ti}, \mathbf{X}_{tj}),$$

where $\mathbf{X}_{ti}$ is a vector of the cumulated duration in the several states, of person *i* at time *t*. This is attractive in that it will to some degree reflect the similarity of sequences which have similar subsequences at different locations (if sequence *i* has a late subsequence which occurs early in sequence *j*, its pattern of cumulated duration will tend to converge with that of sequence *j*).

Dijkstra and Taris (1995) propose a method more closely related to optimal matching, which uses a more complicated algorithm than Buchmann and Sacchi or Degenne et al., but is simpler and less general than the full OMA algorithm. This is implemented in a Macintosh program (Dijkstra, 1994). In a reply to Dijkstra and Taris, Abbott (1995) explores the relationship between their method and his, and sketches out a more general domain of sequence alignment methods within which specific techniques can be located and compared. While some of Abbott's criticisms of the Dijkstra–Taris method are justified, in many respects the loss of generality is compensated for by the simpler algorithm, and by the availability of a social-science oriented implementation, albeit restricted to one type of computer. However, the fact remains that their algorithm is less general than OMA.

Class careers: data and definitions

We use two separate data sets. For the MDS and the cluster analysis we use the Irish Mobility Study (IMS) of 1973/4 (Jackson, 1974) and also for the cluster analysis we use the British Household Panel Study (BHPS, first wave 1991, Buck, Gershuny, Rose and Scott, 1994). The Irish Mobility Study interviewed only men, aged 18–64, and collected complete retrospective work-life histories. The BHPS is, as its name suggests, a panel study, but over the first three waves enough data were collected to construct retrospective work-life histories for the panel members. Because the IMS did not collect information from women, and for compatibility across the two samples, female BHPS respondents are excluded.

A further restriction is imposed: in order to have completely comparable data for each respondent, only information up to age 35 is included, and respondents aged less than 35 at the date of interview are excluded. This truncation is not required by OMA: the algorithm is entirely capable of calculating meaningful pairwise distances for sequences of different length. However, it simplifies the present analysis, especially in relating OMA to other measures, such as cumulated duration, where the restriction is more relevant. In other contexts it seems entirely appropriate to include short sequences, or to look at mobility after age 35. Each remaining respondent contributes a sequence representing his career from age 15 to age 35, one datum per three months, coded according to the seven-category 'EGP' variant of the Goldthorpe class scheme with one variation: an eighth category, indicating 'not-yet-in-labour-market' is added. See Table 1 and Erikson and Goldthorpe (1992), pp. 35ff, especially figure 2.1.

[Table 1 about here.]

We choose to formulate our study in terms of class mobility partly because of a substantive interest in social mobility; we could have framed our study differently, say, in terms of job mobility. The important thing is that the state space should be relatively small, so as to keep the matrix of substitution costs manageable.[2] The state space we end up with satisfies this criterion, and provides a set of categories which are reasonably familiar to sociologists. The substantive justification for including the extra pre-entry state (which is added to the sequences on the left where entry to the labour market is after age 15) is to avoid having sequences of different length, or which start at radically different ages. This is a relatively arbitrary choice, as OMA copes equally well with sequences of uneven length.

Determining the substitution costs

The heart of the optimal matching algorithm is the use of a set of 'elementary operations' (substitution, insertion, deletion) to calculate how to change one sequence into another. It is 'optimal' in that it finds the least expensive set of elementary operations, but in order to make this calculation, it must know the cost of each operation. Thus, putting a value on substitution and *indel* (*i.e.*, insertion/deletion) costs is a necessary but problematic part of doing OMA. Substitution cost in this sense is a measure of the difference between states in the state space. What differences the investigator chooses to identify, and the values put on them, are external to the optimal matching algorithm. They will have an effect, and possibly a strong effect, on the resulting inter-sequence distances. Setting costs can be regarded as one of the main points of theoretical or substantive input into the algorithm.

For the present exercise we borrow from the 'hierarchy' component of Erikson and Goldthorpe's core model of intergenerational social mobility in assigning substitution costs (1992). Erikson and Goldthorpe argue that the three hierarchical levels capture broad differences among social classes in terms of the resources they offer as class of origin and their desirability as class of destination. We take these hierarchical differences as crucial forces shaping work-life mobility as well.[3] Given this scheme, the substitution costs are as assigned as follows: the greatest closeness is between identical elements, and as Table 2 shows, this zero-distance substitution results in a closeness score of 4. The next greatest closeness is between classes at the same hierarchical level, and thus for instance the score for a III–IV*ab* substitution (*i.e.*, routine non-manual to self-employed) is 3. Substitutions across one level boundary (*e.g.*, from I–II to III, salariat to routine non-manual) score 2, those across two boundaries (*e.g.*, I–II to VII*b*, salariat to agricultural labourers) score the minimum closeness of 1. (It is important to bear in mind the distinction between the state space through which the sequences move – the 7 EGP categories plus the waiting state of 'not-yet-in-labour-market' – and the 3-category grouping we use to relate the eight states. Mobility is constituted by movement between the eight categories; the use of the three hierarchy categories simply allows us to judge certain pairs of states as more different than others.) We set the cost of *indel*s relatively high, at 3.

[Table 2 about here.]

As for the eighth category of 'not-yet-in-labour-market', since we are not particularly interested in this category for its own sake, we treat it as completely 'like' each of the seven class categories (with a closeness score of 4). This de-emphasises it, and give more weight to the portion of the sequence in the labour market. The effect is to judge sequences '`XABC`', and '`AABC`' more similar than '`XABC`' and '`ABCC`', where '`X`' is the pre-entry state.

Multi-dimensional scaling: what is this space?

OMA generates a similarity (or distance) score for every pair of sequences under consideration. As an alternative to using this information in a cluster analysis it is interesting to examine the space implied by these scores: do the intuitively reasonable elementary operations used in aligning pairs of sequences result in a coherent, stable and meaningful space?

To this end we partitioned the Irish sample into six five-year-wide pseudo-cohorts, applied OMA to the cohorts, and thus generated for each cohort similarity scores for all sequence pairs. For each matrix we extracted the first three principal components (this constitutes a simple form of classical multi-dimensional scaling; we stopped at three dimensions simply because that is the limit for convenient visual inspection). We were mainly interested to see if the six cohorts showed similar patterns, and whether the placement of the sequences in space seemed meaningful. If the subsamples showed unstable patterns it would suggest that the method of determining inter-sequence distances was overly sensitive to the actual sample, whereas if they seemed stable then at least the space, if not the clustering within it, could be expected to be stable across samples.[4]

In the event, we were reassured by what emerged. Each sample generated a three-dimensional scatter roughly in the shape of a tetrahedron, with different types of sequences well separated and, most importantly, with similar sequences occupying similar locations in each plot. All six plots are reproduced in the appendix (see figure 4), but the plot for cohort 6 is reproduced in figure 2 with annotations.[5]

[Figure 1 about here.]

[Figure 2 about here.]

In the analysis we had the substantial advantage of access to interactive animated three-dimensional scatterplots (XLispStat's spinning scatterplots: Tierney, 1990). Coupled with the ability to use the mouse to read a label for each data point (the labels concisely summarized each sequence), this made it much easier to find patterns than it is to reproduce them on paper. The closest we can come is a scatter-plot matrix, which presents three views of the data cloud, in plan, elevation and side-elevation. To read these, it is perhaps best to start by looking at figure 1, which presents a familiar structure viewed in an analogous manner. Figure 2 presents the first three principal components of the inter-sequence distances for cohort 6 in the same orientations: these three triangles represent a tetrahedron (it may help to keep the heads in mind when viewing this plot).

The most notable features about the tetrahedron are its apices, each one is occupied by a sequence that represents 100 per cent of the career, *i.e.*, twenty years, in a given class. The four apices thus represent classes I–II, V–VI, VII*a* and VII*b* (respectively the salariat, skilled and supervisory manual work, semi- and unskilled manual work and agricultural labour). Data points very near the apices are typically careers consisting of slightly late entry followed by unbroken careers in the appropriate class. Moving further away from the apex, the careers show increasingly long prior spells in other classes. Tracking along an edge of the tetrahedron from the apex, the sequences typically change from 100 per cent of the relevant class to mixtures of that class and the one towards whose apex you are moving. However, there is a lot of empty space, as there are limited numbers of sequences in the analysis.

Classes III, IV*ab* and IV*cd* (routine non-manual work, self employed and small employers, and farming) do not feature on apices of the tetrahedron, but have specific and separate locations within it. Very likely, if we were able to represent higher-order dimensions these would also emerge as apices. As with the 3-dimensional apices, the point indicating 100 per cent careers in these classes is surrounded by very similar careers, with similarity reducing as distance increases.

[Figure 3 about here.]

Figure 3 focuses in on one of these apices in detail, that dominated by class I–II (this is the top apex of the two bottom cells of figure 2). This feature is visible in the body of, and in the same orientation as, the top–left cell of figure 2. This has two clear features, a well-populated line running approximately northwest–southeast in the diagram, and another less well-populated line more-or-less at right angles to it. All the data points in the main line are composed of complete careers in class I–II, after late entry (in this subsample there are no complete class I–II careers starting at age 15). The amount of time spent in the class varies from 17 years (at the top–left) to 7.25 years, and it varies monotonically along the line. The data points to the side also represent long spells in class I–II, with prior experience in other classes; the further from the main line, the longer the prior experience. This sort of pattern is also visible in the other apices.

This exercise reassures us. First, the pairwise similarity scores generated by OMA populate the 3-dimensional space in a coherent and patterned manner: similar sequences are placed adjacent to each other while unlike sequences are placed further apart. Secondly, the observed patterns are, broadly speaking, stable across the 6 cohorts. Thus, it seems that the core algorithm of optimal matching are not overly sensitive to sampling variation. Thirdly, it becomes clear that, as an alternative to cluster analysis, one can use the extracted dimensions from MDS to characterise the sequences.[6]

Optimal Matching and larger samples

Having satisfied ourselves that the algorithm generates a reasonably stable multi-dimensional space, we now take a second approach to the problem of the sample-sensitivity of cluster analysis. This approach consists in using as large a sample as possible at one time, for two reasons: (i) to reduce the indeterminacy of cluster analyses, which is particularly acute in small samples and (ii) in order to be able to compare the distribution of subgroups across a *single set* of clusters, rather than to compare several sets of clusters, each generated within subgroups of the sequences.

The two samples, IMS and BHPS, represented respectively about 1,300 and 1,600 entire career sequences from age 15 to age 35, coded into eighty three-month periods. However, there are significant numbers of duplicates in the samples, and excluding these reduces them to the order of 1,000 distinct sequences.[7] There is no point in including duplicates, as the pairwise distances will be identical, and the clustering algorithm does not weight according to cluster size (in all analyses subsequent to the clustering the duplicates are replaced).

Sample size is a problem for analyses like optimal matching, primarily because the resources needed tend to rise with the square of the sample size, but also because longitudinal data is necessarily ‘larger’ and more complex than cross-sectional. The ‘square’ problem arises because we compare each case with each other case. The 1,000-case analyses took of the order of forty-five minutes of CPU time on a fast VAXStation. But despite the fact that it is a problem, it is our opinion that, *pace* Abbott and colleagues (Abbott and Forrest, 1986; Abbott, 1988), it is a problem that is necessary to face. For a given set of states, the set of all possible sequences is extremely large. And even though the set of empirically likely sequences is a small subset of this set, it is still a very large set. Any sample will of necessity contain only a tiny fraction of the possible sequences, and while the bulk of them will be typical (*e.g.*, in the career context, entire sequences of one state, or sequences which switch once at or near a particular point) and will occur reliably from sample to sample, a large proportion of them will be relatively unusual, simply because they are detailed. Thus a significant proportion of each sample, if repeated samples were collected, would consist of sequences relatively unlikely to correspond closely to sequences in other samples. This is simply because sequences have more potential to vary than conventional variables do. Thus analyses, and *a fortiori* cluster analyses, may be overly affected by small numbers of atypical cases. From this point of view, even 1,600 sequences may be an inadequate sample.

Looking for historical change

Our interest in looking at career data as sequences arose from an existing interest in historical change in social mobility during the work life. In particular, different summary measures of work-life mobility yield different conclusions about change in the underlying mobility regime: in the IMS the association between entry point and class at age 35 is stable, once changes in class distributions are controlled for, while measures which take more information from the career sequence than its start and finish (analysis of all spells, pooled and tabulated by class of spell and outcome of spell; analysis of cumulated duration in each category by class-at-35; see section above) show some evidence of change across cohort in the underlying mobility structure. What can OMA, a technique which analyses sequence *qua* sequence, tell us about historical change in work-life mobility pattern?

Generating and identifying the clusters

In answering this question, we encountered various problems. First, as mentioned above, there is the problem of computing power. But after that comes the problem of interpreting output. It is not immediately obvious how to interpret a dendrogram which spreads over several pages. There may be better, automated, strategies, but we simply examined the dendrogram in order to identify clusters. There is no guarantee that tight easily identifiable clusters will emerge either: inspection of the three-dimensional scatterplots above, for instance, shows that though there clearly is a structure there are no tight clumps. Thus the borders between clusters, and the number of clusters that emerge, are arbitrary, depending on how the investigator decides to delineate them.

We used a simple rule of thumb: taking the dendrogram, draw a horizontal line halfway between the top of the diagram (indicating the distance between the final two clusters) and the bottom (indicating the distance between identical data points, *i.e.*, zero) and count the vertical lines crossing it, treating each one as the root of a cluster. Were we to move this line up, the number of clusters would fall as adjacent clusters became linked in the tree, and *vice versa* were we to lower it. Though arbitrary, this rule has the virtue of simplicity, and in the two samples we used it on, of generating a reasonable number of distinct clusters, but it has shortcomings. First, it will probably generate several tiny clusters, including singletons, of sequences relatively unlike others. Secondly, it will also tend to pick out one or more very large clusters, and these clusters may well be composed of quite distinct sub-clusters. In the former case, at least in the extreme, singletons and very small clusters can be amalgamated into a residual category (which must not be regarded as a 'cluster' since its members will be more like other clusters than like each other). Where very large categories are clearly divisible into smaller clusters, this is an attractive thing to do, but it is not always cleanly possible: very often to remove a sub-cluster from a large cluster leaves behind a relatively disparate set, or the sub-cluster may be composed of sub-sub-clusters which are nearly as different from each other as the sub-cluster is from the remainder of the cluster. But this is part and parcel of cluster analysis – its use is far more satisfactory where there is a strong theoretical rationale for it, as there is in comparing DNA sequences in evolutionary terms.

Though this procedure is messy, it does generate a set of useful and informative categories, and it does separate the career sequences into sensible groups. Inspection of the clustered and aligned sequences shows how similar the clusters are, and it is often easy to recognise sub-clusters within the clusters, by the relatively minor dissimilarities between adjacent sequences.[8] However, while the clusters are very easy to characterise in a general way, it is impossible to characterise them formally and exhaustively, that is, to define rules which will replicate the clusters exactly or close to exactly. This is largely because the clusters will contain small numbers of sequences whose similarity to the core sequences is less obvious. One cluster may be typified by large numbers of sequences such as the following:

```
EEEEEEEEEEEEEEEEEEEE EEEEEEEEEEEEEEEEFFFF FFFFFFFFFFFFFFFFFFFF FFFFFFFFFFFFFFFFFFFF
```

(a long spell in category E followed by an approximately equal spell in category F) which may be accompanied by sequences like

```
HHHHEEEEEEEEEEEEEEEE EEEEEEEEEAAAAAAAAFFF FFFFFFFFFFFFFFFFFFFF FFFFFFFFFFFFFFFFFFFF
```

and

```
HHHHHHHHHHHHBBBBBBBB BBBBBBBBAAAAAAAAFFFF FFFFFFFFFFFFFFFFFFFF FFFFFFFFFFFFFFFFFFFF.
```

The second and third are quite like each other, and therefore will cluster together, but only the second is like the typical, first, sequence. Thus the third is 'chained' into the cluster. This is correct behaviour in terms of the clustering algorithm, but it is not possible to replicate in a set of simple rules.

It is more useful to take the clusters 'as given' and use them as an input to further analysis. In the remainder of the paper we present and analyse the cluster sets we derived for the full IMS and BHPS samples.

The cluster sets

The IMS data generates 16 distinct clusters, plus a residual, while the BHPS breaks down into 9 plus 1. The fact that more clusters are identified for the IMS than for the BHPS may reflect greater historical change in the Irish sample: agriculture underwent a significant decline over the period while industry began to get established; in the period the BHPS data covered, the magnitude of the change may have been less (there the shift is out of heavy industry, and into white-collar work). Visual inspection also suggests that the Irish data contains more careers with many short spells, often but not only representing seasonal work in agriculture. These patterns of sequences in these clusters are described briefly in Tables 3 and 4.

There are too many categories to discuss in detail, but it is worth noting certain features. In the Irish data the largest single cluster is that containing long, usually permanent, careers in agricultural labour, class VII*b*. This is approximately six times the size of the most similar cluster in the BHPS data, a difference due entirely to the different sizes of the two countries' agricultural sectors. That we can identify corresponding pairs of clusters in the two samples is due to the fact that the typical sequences involved are extremely simple, tending towards 100 per cent in one state. This is in turn due to the fact that agricultural labour is, for some people at least, a relatively absorbing location (the exception would be farmers' sons, who may well become farmers themselves). Another location which tends to produce 'simple' sequences is the salariat, class I–II (professional and managerial workers). Positions in the salariat tend to be secure, so the hazard of exit from the class is relatively low, and therefore long spells are common. It is also often entered directly from higher education, because of the strong role of qualifications in controlling entry, and therefore the pattern of late entry followed by an unbroken spell in the class is typical. In the IMS this pattern emerges as a distinct cluster (8) but in the BHPS, while the most similar cluster is much larger, experience prior to class I–II is more various. This is partly due to the fact that the salariat is simply larger in Britain over the period covered by the BHPS, but there is also distinctly more variety in routes into it.

The clustering identifies features of the data that conventional summaries of the sequences will miss. For instance, the distribution of cumulated time in different classes will be similar for individuals in IMS clusters 6 and 14, or BHPS clusters 2 and 9, but the fact that the order of the classes is reversed will be lost: clusters 6 (IMS) and 2 (BHPS) represent downward mobility from skilled/supervisory work to semi-skilled work, while clusters 14 (IMS) and 9 (BHPS) represent the opposite.

Having generated the clusters and inspected them at length, we find with some satisfaction that it is relatively easy to characterise the clusters verbally, and distinctly. That is, the contents of the clusters are relatively homogeneous and the clusters are relatively distinct, when looked at as class careers. In other words, the optimal matching and clustering does, at first blush, generate a reasonable 'empirical typology' of career sequences. It is not, of course, a definitive typology:

different matrices of substitutions costs would generate different clusterings. But it does make sense at an intuitive level. We regard this as a second reason to be reassured about optimal matching.

[Table 3 about here.]

[Table 4 about here.]

Simple inspection shows that the clusters are distinct, but we can add a little precision by examining them in terms of more conventional measures of work-life mobility. We can look at how the clusters spend their time, and where they end up. Tables 5 and 6 present the mean cumulated duration in each of the eight states, for the two samples. The rows sum to twenty years, and present the 'average' class time-budget for a member of each cluster, between ages 15 and 35.[9] The first thing to notice is how different the two samples are: in the British sample over 80 per cent of the person–years are spent in classes I–II, III, V–VI and VII*a*, compared with just over 50 per cent for the Irish sample. In Ireland the two agricultural classes account for almost 40 per cent of the time, compared with just 4 per cent in Britain. Thus we cannot expect the clusters to be similar across the two surveys, simply because of the gross differences in the distribution of the 'material' out of which the sequences are constructed.

Even at the level of cumulated duration (which is a lossy, or information-discarding, representation of the sequence information) we see that the clusters are all distinct (though certain ones do have similarities, as remarked above, page 13). For instance, while clusters 2, 7 and 8 of the BHPS sample all account for substantial proportions of time in class V–VI, skilled and supervisory workers, how they spend the rest of their time is very distinct: respectively, unskilled manual (VII*a*), nowhere much (*i.e.*, most time is in V–VI) and the salariat (I–II).

Class 'destination' or mature class position is an even more lossy way to represent a sequence, since it reduces each sequence to a single data point. However, it is a very important summary, because the mature position is most likely a very stable one, and one which has a very large bearing on the individual's life-time situation. From this point of view also, the clusters are very distinct. Of the 16 proper clusters in the IMS, 12 are clearly dominated (*i.e.*, over 75 per cent) by a single class-at-35. Since there are only 7 class categories, this implies some overlap: clusters 7 and 8 in particular. This pair are 100 per cent in class I–II at age 35 and therefore completely indistinguishable in these terms. Clusters 4 and 6 are dominated by class VII*a*, semi- and unskilled labour, clusters 9 and 10 by farming. By reference to the table of duration, or by the descriptions of the clusters in Table 3 we can see how distinct are the histories which lead to these similar or identical outcomes: cluster 4 spends over 17 years in class VII*a* while cluster 6 spends less than 12, making up the difference in class V–VI, supervisors and technicians. However, only by looking at the clusters do we really see how the outcomes are arrived at, and see that cluster 4 is typified by long absorbing spells in class VII*a*, cluster 6 is typified by downward mobility into VII*a* from V–VI.

[Table 5 about here.]

[Table 6 about here.]

[Table 7 about here.]

[Table 8 about here.]

Cross cohort change

Our purpose in generating an empirical typology of sequences is to see how its distribution varies with respect to other variables. That is, we want to use the clusters as an input to further analysis in which we ask, in what way does the pattern of sequences change in relation to other variables. In principle, any other variable could be used: sex, geographical region, educational qualifications, and so on, but since we are interested in historical change we opt to use cohort.[10] When considering sequences holistically we cannot really access historical time other than as cross-cohort contrast, as the sequences are by their nature not located at single time points but exist over a period. Tables 9 and 10 present the distribution of clusters across the three cohorts in both samples. (In the IMS, the cohorts are those born approximately 1908–17, 1918–27 and 1928–37; in the BHPS 1927–36, 1937–46 and 1947–56.) Both tables show substantial change from cohort to cohort. In the Irish data, the biggest change is in cluster 1, the main cluster for agricultural labour, which declines by almost two thirds. IMS cluster 8 (and also the smaller cluster 7) show systematic rises as the importance of class I–II increases. Cluster 16, typified by long spells in class V–VI, supervisors and technicians, also rises substantially. In the BHPS, clusters 1 (largely VII*b*) and 7 (largely V–VI) fall systematically while clusters 2 (V–VI switching to VII*a*) and 4 (I–II) rise.

[Table 9 about here.]

[Table 10 about here.]

Thus at one level we immediately see historical change across the clusters. However, have we found anything new? We already knew from all sorts of more conventional measures that changes like these were taking place: IMS cluster 1 falls simply because agricultural labour was in sharp decline over the period the data cover; IMS clusters 7 and 8 rise with the historical rise of the salariat, and cluster 16 rises with industrialisation and the increase in class V–VI jobs. BHPS cluster 7 falls with the decline of heavy industry in Britain, and cluster 4 rises with the relentless increase in class I–II. Perhaps the rise in cluster 2, representing downward mobility from class V–VI to class VII*a* (skilled to unskilled work in industry) is interesting. Certainly it would show up in a table of spells tabulated by class-of-spell, outcome and cohort, but inspection of the clusters and their tabulation across cohort makes it more intuitively clear that something systematic is happening.

A class-career is composed of person–months (or person–days, person–quarters or whatever) in various categories, and is situated in real historical time. At any moment in historical time, the distribution of classes is fixed – each class position is contributing exactly one person–moment to an individual's class career. Over time, the individuals move according to their desires and opportunities, and the class distribution changes as a result as their career sequences develop. Thus there is a strong relationship between the period-wise distribution and the distribution of sequences, though because the sequences are longitudinal this relationship is complex. Within cohorts, there is a similar relationship between cumulated duration and sequences. If we consider the cohort-wise distribution of cumulated duration as a pool of person–quarters out of which the sequences are built, and we take note that the clusters have a particular distribution of person-quarters, we see that at least some of the difference in the distribution of clusters across cohorts is driven by the difference in the pooled distribution of person-quarters across cohorts. Tables 11 and 12 show that the distribution of time in the various states is quite different across the cohorts and thus that a large amount of the change in the distribution of clusters is driven by this more general change. The IMS shows more overall change, with classes I–II, III, IV*ab* and V–VI showing systematic rises, and VII*b* falling sharply; the BHPS shows a rise in class I–II, a fall in class V–VI; both samples show rises in time before entry to the labour market.

[Table 11 about here.]

[Table 12 about here.]

However, it is likely that there is some variation in the distribution of clusters across cohort which is not driven by this change in time-budget across cohort. We can regard this as net difference in sequences in some respect. Can we get at this ‘pure sequence’ change, and identify what portion of it is a result of people doing different things over and above the differences forced by change in the class distribution? And if we could, what would it mean? If there were no pure-sequence change, that would suggest that the distribution of sequences changed only to the minimum extent forced by the change in the distribution of person–moments, and that insofar as possible, people took the same type of routes. If there were some pure-sequence change, this would mean that some individuals were taking different routes through the possible space.

We have experimented with some statistical models of cross-cohort change in an effort to control for cohort change in the overall distribution of time. This work is incomplete and therefore we do not report it in detail. We discuss it here primarily because it appears to be a sensible direction for further research on sequence alignment methods, and an essential direction to take if such methods are to have more than exploratory application. Our core idea is to analyse a table of cluster by cohort by time-budget, with each observation representing a person–quarter. That is, we pool all person–time-units and tabulate them according to cohort, class, and the cluster the sequence belongs to. The marginals of the table thus represent the cohort/cluster table, the cohort/class-time-budget table and the cluster/time-budget table, though the first of these is multiplied by the length of the sequence as each sequence contributes one observation per element. Fitting a loglinear model containing all three two-way interactions and then assessing whether the three-way interaction (or a subset of it) is needed is a means of ‘controlling for’ the gross change and searching for change across cohort in the ‘pure sequence’ pattern. However, loglinear models are not satisfactory because each individual or sequence contributes eighty observations, and these observations tend to be highly dependent: most class spells last much longer than three months and therefore if an individual contributes one observation in a given category, he is likely to contribute many. This inflated number of observations means that most additions to the model reduce the deviance far too much and thus it is next to impossible for a variable or interaction to be judged insignificant. A more satisfactory alternative is to treat the cell counts as values of a continuous variables, and fit a linear regression model instead, with cell count as the *y*-variable. In our limited experiments, this shows some power to discriminate changes in the underlying structure, after controlling for change in the distribution of time. However, this work is preliminary, and we are open to suggestions as to how to advance it.

[Table 13 about here.]

[Table 14 about here.]

Conclusion

In broad terms, we find that Optimal Matching Analysis works as a means of grasping sequences holistically, and that its application to class careers is worthwhile. It does seem to mutually place career sequences in a way that makes sense, and the clusters it generates do constitute a useful empirical classification of the sequences. As a means of looking at longitudinal data it is excellent: when sequences can be grouped on the basis of similarity it is much easier to get an overview of the patterns, either visually or by tabulating the empirical typology generated by the clustering against other variables. For this feature alone it is worth applying to longitudinal data sets, as an ‘Exploratory Data Analysis’ technique; because of the extra complexity of longitudinal data it is quite hard to get an overview of it otherwise.

Other techniques for analysing longitudinal data can link much more effectively into the power of conventional statistical method, and without doubt provide very solid insights into longitudinal processes (and it must be remembered that the sequence is just the trace of the longitudinal operation of these processes, an epiphenomenon). The cost of their greater power is a narrower focus. The means of grasping sequence holistically is an important complement to these techniques.

Correspondingly, it would enhance the value of OMA if it could be linked directly into statistical modelling. We have only just begun to think about this but would encourage work in this direction.

Appendix: Software

The software we used for the sequence alignment and the cluster analysis is PileUp, a program in the Wisconsin Package of the Genetics Computer Group Inc. PileUp conducts a progressive pairwise alignment of sequences, a simplification of the method described by Feng and Doolittle (1987). Since it does not compare every possible pair of sequences, but rather proceeds by aligning the two most similar sequences, and then aligning further sequences with the cluster, it is significantly quicker than other methods, but produces a multiple alignment which is not strictly optimal. We are grateful to Liz Cowe of the Department of Pathology in Oxford for giving us access to this software, and for her time spent in helping us use it.

[Figure 4 about here.]

Acknowledgements

Earlier versions of this paper were presented at a workshop in Nuffield College, Oxford, and at the 90th annual meeting of the American Sociological Association, August 1995. We are grateful for the comments of the participants in these meetings, in particular for the kind cooperation of Liz Cowe of the Sir William Dunn School of Pathology, University of Oxford, and also of Ineke Maas.

Notes

[1] A useful general discussion of sequence matching techniques is contained in Sankoff and Kruskal (1983).

[2] The algorithm can in principle deal with calculating substitution cost quantitatively: if each state has a score or value, or a vector thereof, the substitution cost could be calculated dynamically as some function of the difference between pairs of states' scores, such as absolute difference, or Euclidean distance.

[3] While attractive, this is not definitive: it leaves out several other dimensions, for instance sector: the resulting clusters would certainly be different if we acknowledged that, for instance, agricultural labour is closer to farming than to unskilled work in industry. Different substantive ends will be served by different substitution matrices.

[4] This claim depends on the space being a function of the data state-space, the substitution matrix and the optimal matching algorithm, and not of the actual sample of sequences. Of course, different samples will differ in how well they permit the MDS exercise to map out the space.

[5] The earlier cohorts' plots in the appendix look rather more like flat-irons than tetrahedra: this is largely because these cohorts' class distributions are different, with far fewer spells in I–II (the salariat) in particular, and thus the space is populated differently.

[6] Of course, this is a very superficial analysis and we have not identified how the space is structured. Nor have we looked in detail at how problematic sequences are placed (for instance, AAABBB relative to BBBAAA, AABBB relative to AACBB or AAABB, *etc.*). Nor have we looked at how the implied space varies when the input distance matrix is changed (this should be particularly enlightening).

[7] PileUp, the matching and clustering program we used, had to be recompiled to cope with this number, but since its default is a maximum of 300 extremely long sequences (more in later versions; 'long' means some thousands of elements) it is well able to deal with relatively large numbers of short sequences (eighty elements in our case).

[8] Visual inspection is vastly eased by colour coding the sequences: we represented our twenty-year sequences as rows of eighty characters in a text file, in the order of alignment, and using either a simple program to print the sequences to screen interspersed with appropriate ANSI colour codes, or a colour-capable editor (*e.g.*, GNU Emacs), it is convenient to browse the entire sample and get a good impressionistic overview of the distribution of sequences. As an exercise this is worth doing even if there no further interest in optimal matching, but where the main concern is with more conventional analyses of longitudinal data such as haz-

ard models, because such an overview is informative and gives a good feel for the data.

[9] Of course, ‘average’ is a fiction: most individuals have very small numbers of distinct spells so no one will have non-zero cumulated durations in all categories.

[10] In the MDS analysis we used six five-year wide cohorts in order to have a small enough sample, but here we use three ten-year wide cohorts for the opposite reason: to ensure adequate cell sizes in the tabulations.

References

Abbott, A. (1988) Transcending General Linear Reality, *Sociological Theory*, **6**, 169–186.

Abbott, A. (1991) The Order of Professionalization: an Empirical Analysis, *Work and Occupations*, **18**(4), 355–384.

Abbott, A. (1995) A Comment on "Measuring the Agreement Between Sequences", *Sociological Methods and research*, **24**(2), 232–243.

Abbott, A. and DeViney, S. (1992) The Welfare State as Transnational Event: Evidence from Sequences of Policy Adoption, *Social Science History*, **16**(2), 245–274.

Abbott, A. and Forrest, J. (1986) Optimal Matching Methods for Historical Sequences, *Journal of Interdisciplinary History*, **XVI**(3), 471–494.

Abbott, A. and Hrycak, A. (1990) Measuring Resemblance in Sequence Data: An Optimal Matching Analysis of Musicians' Careers, *American Journal of Sociology*, **96**(1), 144–85.

Aldenderfer, M. and Blashfield, R. (1984) *Cluster Analysis*. Newbury Park, Sage.

Blossfeld, H.-P. and Huinink, J. (1991) Human Capital Investments or Norms of Role Transition? How Women's Schooling and Career Affect the Process of Family Formation, *American Journal of Sociology*, **97**(1), 143–68.

Buchmann, M. and Sacchi, S. (1995) Mehrdimensionale Klassifikation Beruflicher Verlaufsdaten: Eine Anwendung auf Berufslaufbahnen zweier Schweizer Geburtskohorten, *Kölner Zeitschrift für Soziologie und Sozialpsychologie*, **47**(3), 413–442.

Buck, N., Gershuny, J., Rose, D. and Scott, J. (1994) *Changing Households: The British Household Panel Survey 1990–1992*. Colchester, ESRC Research Centre on Micro-Social Change.

Carroll, G. R. and Mayer, K. U. (1986) Job Shift Patterns in the Federal Republic of Germany: the Effects of Social Class, Industrial Sector, and Organizational Size, *American Sociological Review*, **51**(3).

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

Degenne, A., Lebeaux, M.-O. and Mounier, L. (1996). Typologies d'itinéraires comme instrument d'analyse du marché du travail. Troisièmes journées d'études Céreq-Cérétim-Lasmas IdL, Rennes, 23–24 May 1996.

Dijkstra, W. (1994) Sequence – a Program for Analysing Sequential Data, *Bulletin de Méthodologie Sociologique*, (43), 134–142.

Dijkstra, W. and Taris, T. (1995) Measuring the Agreement Between Sequences, *Sociological Methods and Research*, **24**(2), 214–231.

Erikson, R. and Goldthorpe, J. H. (1992) *The Constant Flux: A Study of Class Mobility in Industrial Societies*. Oxford, Clarendon Press.

Featherman, D. L. and Selbee, L. (1988). Class Formation and Class Mobility: a New Approach with Counts from Life-History Data, in M. Riley and B. Huber (eds), *Social Structure and Human Lives*, Sage, Newbury Park.

Feng and Doolittle (1987) Progressive Alignment of Sequences, *Journal of Molecular Evolution*, **35**, 351–360.

Gershuny, J. I. (1993). Post-industrial Career Structures in Britain, in G. Esping-Andersen (ed), *Changing Class: Stratification and Mobility in Post-Industrial Societies*, Sage, London, 136–170.

Goldthorpe, J. H. (1987) *Social Mobility and Class Structure in Modern Britain*, 2nd edn. Oxford, Oxford University Press.

Halpin, B. (1993). *Life-History Data and Social Mobility: Analysing Change in Mobility During the Work Life*, DPhil thesis, Nuffield College, Oxford.

Jackson, J. A. (1974). Determinants of Occupational Mobility in Northern Ireland and the Irish Republic. Computer files, ESRC Data Archive.

Jonnson, J. (1993). Education, Social Mobility and Social Reproduction in Sweden: Patterns and and Changes, in E. J. Hansen, S. Ringen, H. Uusitalo and R. Erikson (eds), *Welfare Trends in Scandinavian Countries*, M.E. Sharpe, Armonk, New York.

König, W. and Müller, W. (1986) Educational Systems and Labour Markets as Determinants of Worklife Mobility in France and West Germany: A Comparison of Men's Career Mobility, 1965–1970, *European Sociological Review*, **2**.

Mayer, K. U. (1991). Berufliche Mobilität von Frauen in der Bundesrepublik Deutschland, in K. U. Mayer, J. Almendinger and J. Huinink (eds), *Vom Regen in die Traufe: Frauen zwischen Beruf und Familie*, Campus Verlag, Frankfurt/New York.

Rogoff Ramsøy, N. (1975). On Social Stratification in a Temporal Framework. Mimeo, paper presented at OECD Seminar on Education, Inequality and Life Chances, Paris.

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

Tierney, L. (1990) *Lisp-Stat : an Object-Oriented Environment for Statistical Computing and Dynamic Graphics*. New York, Wiley.

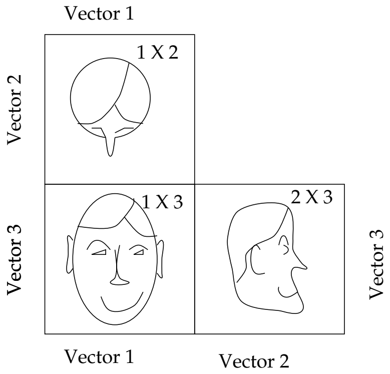

Figure 1: An aid to read the 3-D scatterplot matrix in fig. 2. The three views of the head correspond to the three views of the three-dimensional data cloud.

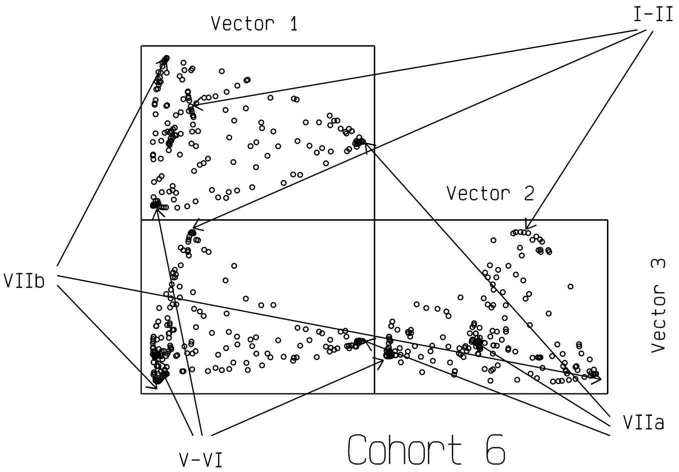

Figure 2: The first three principal components of the inter-sequence distances for cohort 6, Irish Mobility Study.

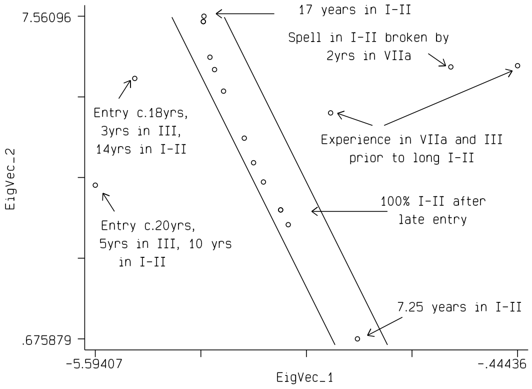

Figure 3: Zooming in on the I–II apex (IMS, cohort 6).

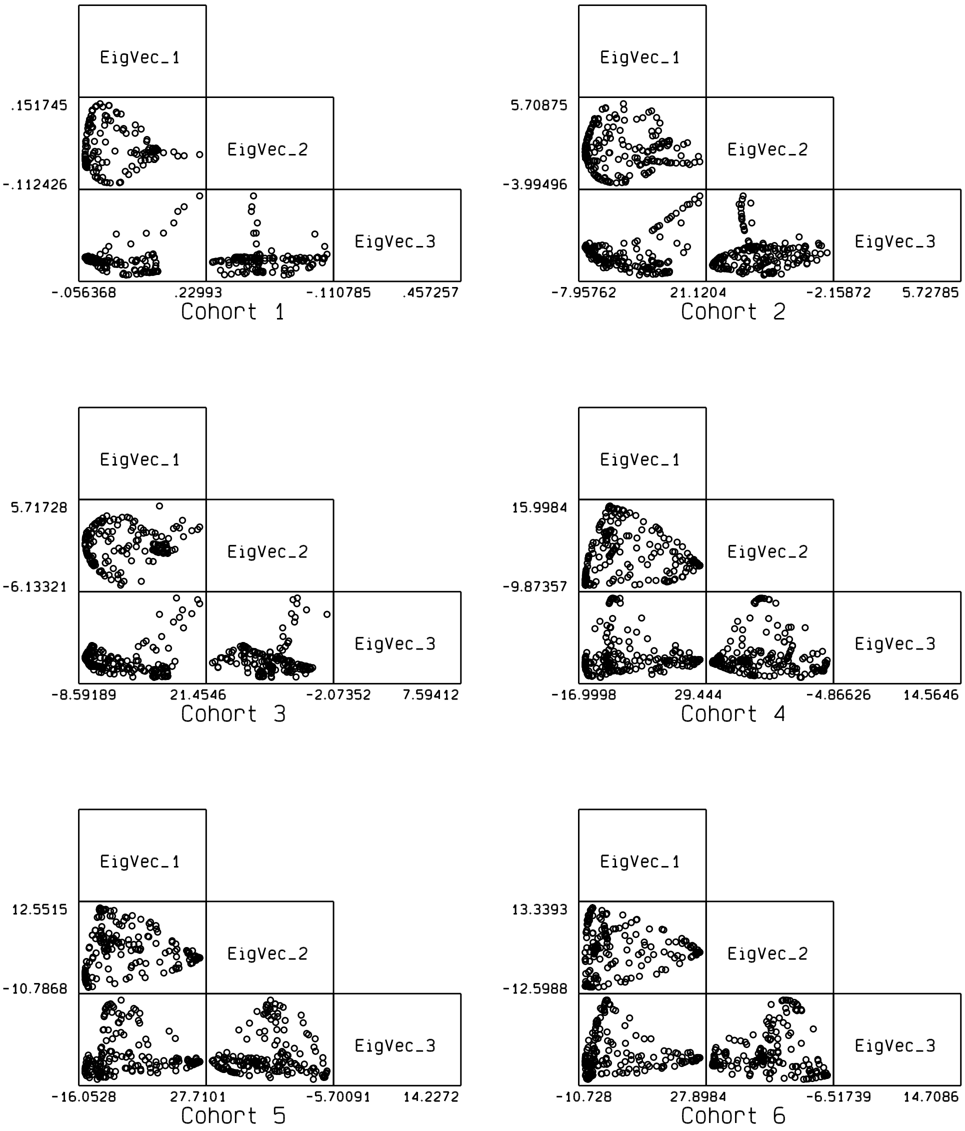

Figure 4: The first three principal components of the inter-sequence distances, for all six IMS cohorts.

Table 1: Outline of the state space, which is based on the EGP class scheme.

Class	Description
I–II	Professional and managerial
III	Routine non-manual
IV*ab*	Self-employed and small employers
IV*cd*	Farmers
V–VI	Supervisory and skilled manual
VII*a*	Semi- and unskilled manual
VII*b*	Agricultural labour
X	Awaiting entry to labour market (not an EGP category)

Table 2: Determining substitution costs.

The three groups

Group 1	I–II			
Group 2	III	IV*ab*	IV*cd*	V–VI
Group 3	VII*a*	VII*b*		

Pairwise ‘closeness’ matrix

I–II	III	IV*ab*	IV*cd*	V–VI	VII*a*	VII*b*	X
4	2	2	2	2	1	1	4
	4	3	3	3	2	2	4
		4	3	3	2	2	4
			4	3	2	2	4
				4	2	2	4
					4	3	4
						4	4
							4

Table 3: The IMS clusters.

Cluster	*N*	Brief Description
1	283	Long, usually absorbing, spells in VII*b*[a], some include short spells (5–7 years) in VII*a*. Exits from VII*b* tend to be late.
2	130	First 10–15 years in VII*b*, leading mostly to IV*cd*, but also IV*ab* and/or V–VI.
3	60	Two sub-clusters: (i) about 10 years in VII*b*, followed by 10 years in VII*a*; (ii) long spells in VII*a* leading to very short spells (2–3 years) in III, IV*ab*, IV*cd* or V–VI.
4	198	Long spells in VII*a*, sometimes preceded by very short spells (3–4 years) in V–VI, III or, more typically and slightly longer, VII*b*.
5	31	10 to 12 years in VII*a*, followed by spells in III, IV*ab*, IV*cd* or V–VI.
6	29	6 to 8 years in V–VI, leading to spells in VII*a*.
7	28	Work-life mobility into I–II after 5–8 years in III.
8	66	Direct late entry into I–II.
9	43	Long spells in IV*cd*, preceded by short spells in III, IV*ab*, V–VI, VII*a* or VII*b*.
10	30	8 to 10 years in VII*b*, leading to IV*cd*.
11	43	Spells in III lasting 10–12 years, followed or preceded by spells in I–II, IV*ab* or IV*cd*.
12	51	Direct entry into III.
13	44	Long spells in IV*cd*, preceded by spells in I–II, III, V–VI, VII*a* or VII*b*.
14	40	10 to 12 years in V–VI, preceded by spells in VII*a* or VII*b*.
15	35	Very long spells in V–VI, leading to IV*ab*, IV*cd* or VII*a*.
16	137	Very long spells (16–18 years) in V–VI, preceded by VII*a* or VII*b*.
17	61	Residual category.

Note: (a) Class categories are described in Table 1

Table 4: The BHPS clusters.

Cluster	*N*	Brief Description
1	51	Dominated by long, usually absorbing, spells in VII*b*[a]
2	61	Long V–VI, followed by long VII*a*. There are two large subgroups apparent, differing in when the transition takes place – about 25 and about 30.
3	237	Dominated by long late VII*a*, with varying backgrounds, including V–VI. Typically 15yrs+ in VII*a*
4	406	Seems to be the main I–II cluster. Second half or more is I–II, though there is a subgroups of 14 cases with late entry direct to I–II after age 32. Heterogeneous subgroup of 21 at the end has transition to V–VI, III, IV*ab* around age 30.
5	159	Typified by long spells in III (10-20 yrs), generally staying, though with several subgroups exiting to I–II, VII*a* and IV*ab*. Two substantial subgroups exit to I–II at about 30 and about 33/34. Subgroups with prior spells in VII*a*, V–VI and I–II, generally settling by early/mid 20s.
6	43	The petty bourgeoisie: entry by 25, usually. Only 3 exits. Main entry is from V–VI in 20–25 age band. Subgroups from I–II and III.
7	498	Skilled manual. Apart from 21 which switch to IV*ab* around 30, dominated by long absorbing spells in V–VI. Typically at least 10 years, large inflow from VII*a* around 18–20, large numbers of entire 20 years or late entry. Some other small subgroups with outflow to other classes fairly late. Exit to I–II happens towards 34 in a handful of cases.
8	57	V–VI to I–II after age 25. A subgroup go through III on the way.
9	52	First 10 years in VII*a*, leaving almost all to V–VI, some to IV*ab*.
Residuals	27	One cluster starts in I–II, spends time in VII*a*, exits to V–VI or IV*ab*, a second enters IV*cd* around age 25, a third switches from VII*a* to I–II around age 27–31.

Note: (a) Class categories are described in Table 1

Table 5: IMS: The class time-budget ‘signature’ of the clusters.

	Person–years in								
Cluster	I–II	III	IV*ab*	IV*cd*	V–VI	VII*a*	VII*b*	Pre-entry	Total
1	.00	.03	.02	.08	.10	1.06	18.36	.35	20.0
2	.00	.02	.63	4.21	.55	1.00	12.99	.60	20.0
3	.06	.39	.55	.24	.66	10.74	6.72	.64	20.0
4	.00	.20	.02	.01	.38	17.34	1.47	.57	20.0
5	.00	2.73	.98	.92	1.72	11.67	.91	1.07	20.0
6	.00	.47	.10	.00	6.86	10.92	.98	.68	20.0
7	10.94	3.73	.11	.18	.92	1.02	.69	2.41	20.0
8	13.29	.17	.00	.00	.04	.03	.18	6.28	20.0
9	.01	.58	.27	15.18	.41	.66	2.01	.88	20.0
10	.00	.01	.00	11.08	.00	.28	8.28	.34	20.0
11	2.28	12.25	.70	.44	.91	1.07	.41	1.94	20.0
12	.00	17.32	.17	.04	.12	.08	.15	2.13	20.0
13	.28	.77	12.54	.06	2.47	2.29	.59	1.00	20.0
14	.06	.71	.19	.03	10.89	5.30	1.95	.87	20.0
15	.27	.67	2.77	.59	13.14	1.87	.09	.61	20.0
16	.00	.08	.06	.02	18.14	.47	.30	.92	20.0
17	1.96	3.92	.98	.41	2.83	5.77	2.92	1.20	20.0
Total	1.09	1.57	.72	1.28	3.22	4.66	6.38	1.07	20.0

Table 6: BHPS: The class time-budget ‘signature’ of the clusters.

	Person–years in								
Cluster	I–II	III	IV*ab*	IV*cd*	V–VI	VII*a*	VII*b*	Pre-entry	Total
1	0.12	0.08	0.28	0.24	1.01	1.81	15.59	0.87	20.0
2	0.31	0.63	0.59	0.00	9.35	7.59	0.04	1.50	20.0
3	0.15	0.53	0.14	0.00	1.79	16.34	0.18	0.88	20.0
4	13.21	1.01	0.09	0.00	0.64	0.44	0.07	4.55	20.0
5	0.96	14.45	0.09	0.04	1.41	0.88	0.07	2.10	20.0
6	0.66	0.91	12.42	0.01	3.63	1.01	0.04	1.32	20.0
7	0.22	0.27	0.34	0.02	17.14	0.75	0.15	1.12	20.0
8	5.91	0.80	0.12	0.00	11.19	0.53	0.06	1.40	20.0
9	0.21	0.34	2.03	0.24	6.29	9.82	0.34	0.72	20.0
10	4.25	0.62	0.43	3.08	1.24	7.05	1.36	1.97	20.0
Total	3.88	1.97	0.60	0.08	7.05	3.70	0.64	2.08	20.0

Table 7: IMS: The class destination 'signature' of the clusters.

Cluster	Class at age 35, %ages							
	I–II	III	IV*ab*	IV*cd*	V–VI	VII*a*	VII*b*	Total
1	0.0	0.4	0.7	7.1	1.4	9.9	80.6	283
2	0.0	0.0	6.9	81.5	8.5	2.3	0.8	130
3	3.3	3.3	6.7	8.3	8.3	68.3	1.7	60
4	0.0	0.5	1.0	0.5	2.0	93.4	2.5	198
5	0.0	29.0	19.4	16.1	22.6	12.9	0.0	31
6	0.0	6.9	0.0	0.0	13.8	79.3	0.0	29
7	100.0	0.0	0.0	0.0	0.0	0.0	0.0	28
8	100.0	0.0	0.0	0.0	0.0	0.0	0.0	66
9	0.0	0.0	0.0	100.0	0.0	0.0	0.0	43
10	0.0	0.0	0.0	96.7	0.0	3.3	0.0	30
11	37.2	39.5	14.0	7.0	0.0	2.3	0.0	43
12	0.0	94.1	5.9	0.0	0.0	0.0	0.0	51
13	2.3	0.0	90.9	2.3	2.3	2.3	0.0	44
14	7.5	5.0	2.5	2.5	77.5	5.0	0.0	40
15	2.9	0.0	45.7	5.7	25.7	20.0	0.0	35
16	0.0	1.5	1.5	0.7	94.2	1.5	0.7	137
17	16.4	13.1	6.6	9.8	8.2	41.0	4.9	61
Total	9.7	7.0	7.3	17.0	16.0	24.7	18.3	1309

Table 8: BHPS: The class destination ‘signature’ of the clusters.

Cluster	Class at age 35							
	I–II	III	IV*ab*	IV*cd*	V–VI	VII*a*	VII*b*	Total
1	0.0	3.9	3.9	5.9	5.9	15.7	64.7	51
2	3.3	0.0	0.0	0.0	0.0	95.1	1.6	61
3	2.1	1.7	4.2	0.0	3.0	88.6	0.4	237
4	92.1	3.0	2.0	0.0	2.0	1.0	0.0	404
5	17.6	76.7	3.1	0.0	0.0	2.5	0.0	159
6	0.0	2.3	93.0	0.0	2.3	2.3	0.0	43
7	3.4	1.8	5.2	0.4	86.9	2.2	0.0	498
8	94.7	1.8	3.5	0.0	0.0	0.0	0.0	57
9	3.8	1.9	25.0	1.9	67.3	0.0	0.0	52
10	40.7	0.0	11.1	25.9	11.1	11.1	0.0	27
Total	30.9	9.6	6.9	0.8	30.8	18.8	2.2	1589

Table 9: IMS: Distribution of the clusters across cohort.

Cluster	Cohort			Total
	col. percentages			
	A	B	C	
1	32.5	20.0	12.4	283
2	9.7	9.3	10.7	130
3	5.3	4.9	3.5	60
4	13.5	17.1	14.7	198
5	2.8	2.4	1.9	31
6	0.2	4.0	2.3	29
7	0.9	1.8	3.7	28
8	4.4	4.9	5.8	66
9	3.9	1.6	4.4	43
10	2.3	3.3	1.2	30
11	1.9	4.0	4.0	43
12	2.6	4.0	5.1	51
13	1.6	3.8	4.7	44
14	2.1	3.1	4.0	40
15	3.2	2.2	2.6	35
16	8.4	9.8	13.3	137
17	4.6	3.8	5.6	61
Total	431	450	428	1309

Table 10: BHPS: Distribution of the clusters across cohort.

Cluster	Cohort			Total
	col. percentages			
	A	B	C	
1	3.8	3.5	2.7	51
2	3.2	3.9	4.1	61
3	15.4	13.5	15.8	237
4	18.6	24.7	29.9	406
5	8.9	11.1	9.7	159
6	1.9	3.0	2.9	43
7	39.6	30.8	27.1	498
8	2.7	4.6	3.2	57
9	4.6	3.0	2.8	52
10	1.3	2.0	1.6	27
Total	371	542	678	1591

Table 11: IMS: Cohort class time-budgets.

Cohort	Person-years in								
	I–II	III	IV*ab*	IV*cd*	V–VI	VII*a*	VII*b*	Pre-entry	Total
A	.74	1.09	.51	1.47	2.62	4.39	8.31	.88	20.0
B	1.06	1.57	.78	1.09	3.19	4.94	6.42	.96	20.0
C	1.48	2.06	.86	1.29	3.87	4.64	4.40	1.39	20.0
Total	1.09	1.57	.72	1.28	3.22	4.66	6.38	1.07	20.0

Table 12: BHPS: Cohort class time-budgets.

Cohort	Person-years in								
	I–II	III	IV*ab*	IV*cd*	V–VI	VII*a*	VII*b*	Pre-entry	Total
A	2.82	2.00	0.62	0.09	8.56	4.00	0.76	1.14	20.0
B	4.04	2.10	0.55	0.12	7.28	3.42	0.77	1.72	20.0
C	4.34	1.84	0.63	0.04	6.04	3.76	0.46	2.88	20.0
Total	3.88	1.97	0.60	0.08	7.05	3.70	0.64	2.08	20.0

Table 13: IMS: Cohort class destination patterns.

Cohort	Class at age 35							
	I–II	III	IV*ab*	IV*cd*	V–VI	VII*a*	VII*b*	Total
A	29	21	20	77	59	109	116	431
B	40	35	35	72	68	120	80	450
C	58	36	40	74	83	94	43	428
Total	127	92	95	223	210	323	239	1309

Table 14: BHPS: Cohort class destination patterns.

Cohort	Class at age 35							
	I–II	III	IV*ab*	IV*cd*	V–VI	VII*a*	VII*b*	Total
A	86	32	26	2	142	74	9	371
B	164	55	35	8	171	94	15	542
C	241	65	48	3	177	131	11	676
Total	491	152	109	13	490	299	35	1589