Employment trajectories in the South Caucasus during the transition from communism to post-communism: a comparison of OMA and non-sequence based methods of producing typologies*

Gary Pollock,

Manchester Metropolitan University, UK

Ken Roberts,

University of Liverpool, UK

Corresponding author:
Dr. Gary Pollock
Department of Sociology
Manchester Metropolitan University
Manton Building
Manchester
M15 6LL
UK

Phone: +44 (0) 161 247 3466
Fax: +44 (0) 161 247 6321
Email: g.pollock@mmu.ac.uk

*The research reported in this paper was funded by INTAS, project 05-1000008-7803

May 2012

Abstract

A survey carried out in six locations in the South Caucasus during 2007 is used to explore the employment experiences of people born between 1970 and 1976. Retrospective longitudinal data maps out these experiences and is used to develop typologies which seek to capture meaningful and distinct employment trajectories. We present a comparison of two contrasting methods of developing typologies; firstly using time-in-state where category membership is determined using a sorting algorithm. Sequence analysis (SA) is then used to create a typology using the complete set of each individuals' employment experiences as a single variable. For both analyses employment statuses for each respondent between the ages of 16 and 30 are used. Our results show: (i) both SA and non SA typologies are useful as predictor variables, (ii) the SA typology is able to capture certain longitudinal experiences more accurately than time-in-state, and (iii) there is a correspondence between the typologies produced by both methods which suggests that, depending on the aims of the analysis, for the majority of respondents a time-in state analysis is often sufficient but that for a significant minority, often with mixed employment experiences, the sensitivity of SA better represents their life experiences.

Keywords: OMA, sequence analysis, South Caucasus, typology, transition to post communism, youth, employment

1. Introduction

Sequence analysis (SA) is fast becoming a standard tool for dealing with particular types of longitudinal data. Interest in the order in which a series of events occur has long been a feature of social science. Whether it be the analysis of origins and destinations using mobility tables, or pathways into and through working life, the need to take into account the sequence in which events occur is clear. The increasing availability of longitudinal data has encouraged exploration into how we can best represent and analyse an ordered set of experiences. While Andrew Abbott (Abbott and Forrest 1986, Abbott and Hrycak 1990, Abbott 1995) was the pioneer who introduced SA using Optimal Matching Analysis (OMA) to the social sciences, and whose works are still the standard reference, alternative strategies were developed elsewhere at around the same time (Dex 1984, Djikstra and Taris 1995). Questions remain as regards how an SA should be executed as no standard methodology exists. Researchers must therefore identify from the literature the most appropriate SA tools on the basis of the growing body of work that has arisen (see Brzinsky-Fay and Kohler, 2010). Despite the growth in the literature, there remain questions over the analytic value that is added by undertaking any SA rather than an alternative method of allocating cases to a typology. While it may seem intuitively correct to attempt to cope with as much complexity as exists in the data, there is always a need to simplify at some point to make the findings intelligible, which invites questions about how much complexity should remain and at what stage should simplification occur. Decisions over the length of the time periods and the number of different states per time period impact on just one aspect of the process. Further steps allow the analyst to refine the operation of the algorithm in order to be sensitive to other variables. Sequential complexity is partly a function of the user defined state-space. By adding further dimensions, the number of theoretical possibilities mushrooms. Sequences which have multiple layers such as employment plus housing plus fertility and marriage are inevitably more complex than any of the individual sequences on their own (Pollock 2007). Coefficients of complexity have recently been suggested as a suitable direction in which to take sequence analysis (Elzinga 2010). Typology construction, however, existed long before the social sciences had the tools of OMA and it is not unreasonable to suggest that it is possible to incorporate sufficient complexity into a typology without resorting to SA.

This paper reports an analysis of young people's labour market experiences in the South Caucasus countries of Armenia, Azerbaijan and Georgia beginning with the breakdown of the Soviet Union up until Spring 2007. These countries are at the periphery of Europe and inhabit a space that has for long been highly contested. Their Soviet experience has meant that they are relative newcomers to international circles. Their territorial disputes, on the other hand, render them better known than they otherwise would be. They are truly at the crossroads of contrasting regions and have yet to become fully aligned with any wider political system. Their economies have come through troubled times and have yet to demonstrate the capacity to underpin anything approaching the employment levels that existed in Soviet times. The restructuring of the economies has created new opportunities for some, particularly in the urban centres, but for most the chances of achieving stable and secure employment remain low, and very much connected with the transmission of advantage from one generation to the next.

We have typologized the employment experiences of the young people in our survey, and we can show that there are distinct trajectories which are related to regional and social contexts. In addition we have subjected the process of typology construction to a comparative analysis in order to evaluate the relative efficacies of sequence analysis using OMA and a non-sequential method of typology construction and allocation. To the best of our knowledge, this is the first attempt to assess the value-added by sequence methods using the results achieved with more conventional techniques as a benchmark, operating on biographical, life history data. Our evidence is from a retrospective life history survey conducted in 2007 in six regions across the South Caucasus which collected data on the experiences of a cohort of 31-37 year olds from the late-1980s. Assessing the social consequences of the breakdown of the Soviet system would ideally be done with prospective longitudinal data but this is out of the question as such data simply does not exist for these countries. Retrospective life history data presents the best alternative as it allows us to capture data from the generation who witnessed the breakdown in communism and who were at the forefront of the newly developing societies.

Our overall aim was to derive a suitable typology of employment trajectories and to establish the extent to which these distinct experiences were related to place, education, family background and so forth. Underneath this substantive aim were three further methodological aims directed towards exploring the relative benefits of using OMA for typology allocation versus a non-sequential methodology. These were:

(i) To undertake an analysis of the data using tabular methods and derived variables in order to develop a working typology of career types without using SA.

(ii) To develop a typology using OM and clustering methods.

(iii) To compare these typologies in terms of how well they could be predicted by background and other contextual variables, and how well they could predict employment and occupational status at the time when the survey was carried out.

On the basis of this work we are able both to explore the lives of the respondents to our survey using traditional techniques as well as measuring the value added by undertaking an OM-based sequence analysis.

2. The methodology of OMA

At the heart of all SA is the belief that the complete sequence of data represents a significant and unified entity, greater than the sum of its constituent sub-sequences. It is a truly holistic analysis. One's employment trajectory, the series of employment statuses and occupations, can be analysed as a totality and compared with other people's trajectories in order to establish measures of dissimilarity. It is therefore possible to include in an analysis individuals' entire employment history data at one and the same time. There is no need to select a particular type of employment experience, or a particular period in an individual's life, or a particular life event. All of these more focused analyses are, of course, equally valid and useful depending on the research question being asked. However, only with SA is the researcher able to ask if there are patterns which run across different dependencies, over different time frames and in relation to a variety of life events.

SA is not yet a widely accepted methodology despite recent momentum. The April 2010 edition of SMR both confirms the growing status of SA and at the same time demonstrates its contested nature. This special edition has shown that significant advances have been made in our understanding of the ways in which OMA can be used on different types of data. It also suggests that this technique is widely adaptable, but possibly more suited to exploratory than confirmatory analyses. The debate around SA has been lively if not always constructive. The categories of engagement can be arranged under four headings. First, there is the question over what SA can offer which cannot be done with other techniques (see for example: Levine 2000, Wu 2000). This is the question that we address below. Second, there are the respective merits of alternative (ie non OMA) methods of undertaking sequence analysis (Djikstra and Taris 1995, Elzinga 2003). Third, there are those who deal with specific questions of OMA which seek to refine our understanding of the method (Gauthier et al 2009, Hollister 2009, Halpin 2010). Finally, there are those who simply use SA with little or no methodological reflections on the SA processes (Buhlmann 2010). As the number of papers in this last category increases the broader acceptance of SA will be demonstrated. A challenge for those interested in using SA up until recently has been the lack of expertise amongst researchers and the lack of widely available statistical software which includes a SA suite. With the inclusion of SA in both STATA (Brzinksy-Fay 2006) and 'R' (Gabadinho 2008) there is now a rapidly developing user base from which future developments are bound to come.

The basic question for any SA is: how similar are two sequences? From this we need a measure so that each sequence in the data set can be given a score which indicates how similar (or dissimilar) it is to others. The OMA algorithm does just that, albeit requiring a certain amount of user input. In OMA the scope for user influence is no different than in any other analysis where the processing of the data, its operationalisation, and the way a statistical operation is executed require informed decisions.

There are three steps where user defined parameters directly influence the results. First, there is the option of doing a full pairwise sequence comparison comparing all cases with all others or using an 'ideal type' set of sequences against which each sequence is compared. In the first instance a dyadic matrix of distances results where there are dissimilarity scores for each pair of sequences. When ideal types are used, instead of the dyadic matrix, the output produces as many variables as there are ideal types where each variable contains the dissimilarity between the actual and the 'ideal' sequence. Undertaking a full pairwise analysis is more computationally intensive, but it is nonetheless relatively unproblematic to accomplish. Where there are no *a priori* grounds for suspecting any particular pattern in a data set, a full pairwise analysis is clearly warranted. However, where there are good grounds to believe that the sequences have particular properties, then the use of ideal types is at least justified, and arguably more appropriate. In terms of employment trajectories there are some widely found results that cannot be ignored to do with the ages at which young people leave education and enter the labour market and the relative stability that many then experience of being in employment (or indeed unemployment).

Second, there are the parameters for the OMA itself, the way in which the algorithm is 'programmed' to compare one sequence to another using rules which define what to do when sequences contain particular elements. This is the 'optimal' part of OMA in that once the insertion, deletion and substitution costs are set, the algorithm will compute the ‘cheapest’ way of transforming one sequence into another. The consequences of using different insertion, deletion and substitution costs have been known since the outset to produce different results (Abbott and Hrycak 1990) and they are at the centre of one side of the future development of OMA (Gauthier et al 2009).

Last, the method chosen to reduce the resulting distance matrix (or distance variables if ideal types are used) contains a similar level of user input. There is a need to simplify the dissimilarity scores in order to make them intelligible. The standard method of doing this is some form of cluster analysis. There are questions over which method of clustering to choose (hierarchical or otherwise) and how to decide how many ‘real’ clusters exist. Should a hierarchical method be used in the process of experimenting with different cluster solutions, the fact that a case cannot jump between clusters in different branches at higher levels might be regarded as problematic. A phylogenetic approach to social phenomena is, arguably, not justified. The separation of sequences of unemployment and employment is arguably not comparable to the differentiation of different species or other genetic groupings. Methods such as K-means clustering do not make any such assumptions and may be more appropriate when clustering social data. On the question of how many ‘real’ clusters exist, there have been various suggestions which can be broadly separated into those which favour technical and those which favour theoretical solutions. There are problems in cluster analysis in that there are no robust statistical tests which apply across the board. The very nature of cluster analysis – putting together similar data in a category - means that tests aiming to assess the significance of the cluster dissimilarities will generally find significance. Searching for the ‘true’ number of clusters has been compared with searching for the crock of gold at the end of a rainbow (Johnson 1970). This is a good quip but it highlights a serious flaw in any analysis which claims to have identified that which truly exists. We must instead rely on theory in both the early and latter stages of the process to guide the analysis and give foundations for conclusions.

Whichever method is used one can consider how far the SA typology maps onto one produced using a different sorting mechanism. This allows an analysis of the similarities and differences produced by different methodologies and helps to focus on the distinctive contribution that each has to offer. More optimistically, we can assess the efficacy of each in terms of its predictive power. There may be a need to assess at the outset of a SA what the likely advantages might be for the analysis. Alternative, non-OMA ways of processing the sequential data may be sufficient as well as less complex. This paper contributes to a deeper understanding of the relative merits of undertaking a SA. The analysis presented below shows the similarities and differences in alternative approaches to analysing sequential longitudinal data on the employment histories of a representative sample of South Caucasians.

3. The South Caucasus

The South Caucasus consists of Armenia, Azerbaijan and Georgia. All three were part of the Soviet Union and are now independent states. While the region is currently relatively stable there remain significant territorial disputes and frosty relationships with neighbouring countries, most recently highlighted by the 2008 Georgia-Russia war over the breakaway region of South Ossetia. Georgia and Russia sandwich a further disputed region, that of Abkhasia which was drawn into the 2008 conflict. Armenia and Azerbaijan have long had a territorial disagreement over the region of Nagorno Karabakh, an area which has always had a significant population of Armenians but which Azerbaijan claims as its own.

During Soviet times the region was relatively prosperous. Georgia, with its wine and its Black Sea coast had long been a desirable destination for vacations. Azerbaijan's oil was a key part of Soviet industrial expansion. When the Soviet Union collapsed, the region was sent into turmoil. Civil wars and territorial disputes meant that the transition to post-socialism was anything but smooth. Each country suffered a massive reduction in GDP which had a devastating effect on the economies. Over time there has been a recovery and prior to the banking collapses of 2008, each country experienced high economic growth with consequent rebuilding of the socio-economic infrastructure. These countries, however, remain poor in comparison with other post-socialist states although by 2007 life had returned to a higher level of stability and predictability than in the early and mid-1990s.

The substantive aims of our research were to examine the experiences of young people growing up during the transition from socialism to post-socialism. We expected that the turbulence of the events around 1990 and thenceforth would be manifest in the longitudinal experiences of those whose education was coming to an end in the late-1980s and early-1990s. We anticipated that sequence methods would help in making sense of young people's complicated life stage transitions. In the event, however, we have found that there had been impressive continuities in the lives of the young people who grew up during the breakdown of communism. However, the concern of this paper lies in the ways and extent to which sequence methods add to what could be learnt about the young people's lives using more conventional techniques.

4. Data

Our data comes from a survey carried out during Spring 2007 in each of the capital cities plus a contrasting region in Armenia, Azerbaijan and Georgia. The final sample totals 1215, all born between 1970 and 1976, and is representative of the six areas through the use of sampling lists from the 2005 Data Initiative Survey undertaken by Caucasus Research Resource Centres (CRRC). The questionnaire used retrospective questioning methods derived from instruments developed in the UK for the British Household Panel Survey and the National Child Development Study. We asked respondents questions relating to their life experiences since the age of 16. Separate question blocks on employment, education, family experiences, housing and leisure were included. In addition, we asked questions about family/household composition over the years, parental backgrounds and the economic position of the respondent and household. We are aware that recall error is a problem with this type of research and used a Life History Calendar which was independent from the questionnaire schedule but logged key dates for each respondent such as birth days, important personal events and so on to help respondents better recall when events took place (or episodes began/ended). The full questionnaire has been published by the Council of Europe (Pollock 2010).

The data was collected and stored in order to allow various types of analysis. A relational database of nine files efficiently holds the data and allows the data to be reconstructed in a number of ways to allow cross-sectional and longitudinal analysis. Here we focus on the samples' employment experiences.

Employment experiences were split into two separate modules; employment status and job status. Respondents were first asked to state their employment statuses since the age of 16 using an eleven category show card which was simplified into: employed / self-employed / unemployed / non-active / other. The months and years of the start and end of each episode were recorded. Job experiences were explored immediately following. A 10 category show card was used which we simplified into the following: manager / professional / clerical / farm work / manual / petty trader. Again, episode start and end dates were recorded.

5. Alternative typologies

5.1. The basis of the 'S' typology

We undertook a sequence analysis using OMA. After testing a range of alternative methods of using OMA we chose a comparison of each sequence to a set of pre-determined 'ideal types' where these represented extended experiences in each of the five major categories (non-manual employment / manual employment / self-employment / unemployment / non-activity). This is an example of researcher decisions guiding sequence analysis. Another was to treat changes in types of occupations (manual versus non-manual) as more significant than job changes within either of these categories. The resulting distance scores were then clustered using K-means clustering. An eleven cluster 'solution' produced distinct and meaningful categories, which are shown in table 1.

These categories are sensitive to order and content. We were able to identify sequences through a visual analysis of 'carpet graphs' which show in colour code the string of employment statuses for each individual, sorted by cluster membership. While useful in visual analyses, these carpet graphs do not translate well to monochrome and are not required for the purpose of typology comparison.

Table 1: Employment typology based on sequence analysis

Non manual – early starters	8
Non-manual – mid starters	11
Non-manual – late starters	6
Manual – early starters	5
Manual – mid starters	4
Self – early starters	8
Self – mid starters	5
Unemployed – early start and/or long	14
Unemployed – mid to late starters	9
Inactive	12
Other and mixed	18
Total	100% (N=1215)

The eleven category clusters appeared to be the most meaningful. When higher K values were used it led to the creation of clusters with no obvious interpretation (labels). Fewer clusters led to categories which were too broad. The Calinski-Harbasz pseudo F index for 11 clusters was the highest which, while not an indication of the 'true' number of clusters, lends weight to this as a choice for further investigation. The importance of unemployment and inactivity (35% in total) demonstrate that during the observed time there was a great deal of non-employment in the young people's life histories. This fits with our knowledge about the economic situation during the transition to post-communism, the catastrophic drop in GDP implying that employment was scarce. The lack of manual work is also of interest; only 9% can be easily classified in this sequence. Factory closures and a general decline in the availability of such jobs meant that where employment existed, it was most likely to be non-manual (usually in the public sectors). The non-manual category contained three distinct clusters; those who began at an early age, mid-starters and those who began later in life. It is not unusual for transitions from higher education to result in this sequence, and the South Caucasus proved no exception. Self-employment is a fairly heterogeneous category which can be sub-divided by age of entry, it includes farm small holdings and petty trading as well as ownership of firms. Most, however, were of the former and these constituted the main component of the early entrants to self-employment . This leaves 18% where it was not possible to classify the respondents, largely due to lack of data or high level of heterogeneity.

We can note here that while there is a sequence basis for non-manual, manual, self-employment and unemployment. That they are consistent in the pattern with age of entry the key defining feature. The absence of well defined sequences which show any mixing of labour market experiences is the result of a powerful tendency for individuals to 'stick' in positions entered early in their labour market careers. Thus the search for sequences, using methods designed specifically to identify sequences, might be regarded as fundamentally misconceived. We shall return again to this issue.

5.2 The 'T' Typology

We also created a typology of employment careers where category membership was determined by two elements neither of which used any form of sequence analysis. First, we computed the proportion of time since completing full-time education that had been spent in different employment statuses. We used this data to allocate each respondent to a single category. We experimented with different methods and ultimately chose to allocate using a threshold of 50% of time in a single state since completing full-time education. The allocation was done incrementally so that we initially filled the unemployment category, then the self-employed, the non-active, and finally the employed. We further sub-divided the employed category using the data from the jobs module of the survey in order to differentiate manual and non-manual jobs. The employment status inputs were therefore the same as for the computation of the S typology. On the basis of our categorising methodology the outputs for T were always going to be fewer and directly correspond to the input categories ie the typology is almost analogous to the employment status that each individual has spent the most time in.

The T typology is shown in table 2. Unemployment had been a common experience with 32% of respondents, more than any other single group, experiencing 50% or greater of their labour market careers in this state. The eleven category SA solution is informative in differentiating qualitatively distinct longitudinal experiences, but it can easily be collapsed into six categories which have a close conceptual association with the T typology. This collapsed S typology is shown alongside the T typology in table 2. Note that when a cluster analysis with K=6 was performed, the collapsed S typology was not produced. The K=6 typology was close (Cramer's V of 0.827 when tabulated with the collapsed K=11 typology) but it conflated the non-manual and ‘other’ experiences while retaining two separate non-manual categories, it was largely the same in other respects. We take the view that a purely mechanical approach to cluster selection is inappropriate and instead chose to find a cluster solution which we could adequately interpret, and then aggregate it as best we could to allow a comparison with T.

Table 2: Six category version of the S typology and the T typology

	S	T
Non manual	25	24
Manual	10	11
Self employed	12	14
Unemployed	23	32
Inactive	12	10
Other	18	10
Total	100%(N=1215)	100%(N=1215)

The T typology emphasises unemployment substantially more than S, the likely explanation for this is that there are many who have had relatively short labour market availability and where a substantial (greater than 50%) has been in unemployment. The T methodology will place these into the unemployment category, but the OMA algorithm may draw out those who have had some form of employment alongside the unemployment and thus place them in a different category. This explanation plus other differences can be further explored through the cross-tabulation in table 3 which arranges the common categories on the diagonal.

Table 3: comparison of employment typologies – percentages of total distribution (N=1215)

T → / S ↓	Non manual	Manual	Self	Unemp	Inactive	Other
Non manual	20	1	*	3	*	*
Manual	0	7	*	1	2	*
Self employed	0	*	12	*	0	*
Unemployed	*	*	1	24	0	0
Inactive	1	1	*	1	7	2
Other	3	2	1	5	1	4

*** <0.5%**

The comparison of the T and S typologies shows that 74% of respondents are on the diagonal and are classified similarly by both methods. This represents a remarkable agreement between the two typologies. They were constructed using radically different techniques yet they concur in accounting for by far the majority of the employment experiences. Together they show that both T and S appear to measure almost the same phenomena. The importance of a sustained experience of unemployment is reinforced through 24% of the entire sample being similarly classified. Likewise most of the other categories on the diagonal show high levels of agreement. It is possible that other parameters may have constructed typologies with an even better fit, but this would not have meant that the typologies would have been any better (or worse) than the ones we have here at representing the lives of the respondents. The point is that each is theoretically informed and might easily have been used without comparison with the other in a substantive analysis. A central concern of this paper is the consequence of using one rather than the other, and in any particular advantages of SA, given that it is widely seen as being able to capture a style of longitudinal data – combined sequences – which other methods simply cannot process.

The 26% of off-diagonal cases represent those not matched by T and S. This is not to say that they are inherently anomalous, but they have in common that they do not appear in the same categories of T and S. Most of those classified differently are contained in the other category of S (12%) and the unemployed category of T (10%). Close inspection of the differently classified cases enables us to explain how these disagreements arose. Using the T-method, individuals were allocated to the non-manual type, for example, if they had spent at least 50% of their time in the labour market in such jobs. However, most of those concerned had spent close to 100% of their time in such jobs. Outliers, who had held non-manual jobs for only 51% of the relevant time, were actually more optimally matched with another career group.

The broad agreement of S and T suggests that there is an underlying phenomenon that both represent, thus we have evidence of construct validity. In order to assess which of the two might be taken to best represent the employment histories we shift the focus towards the predictive power that each possesses, and test the relationship between each typology and employment and occupational status at the time of the interviews.

6. Using the typologies to predict employment status at the time of the interviews

Here we are interested in the extent to which each typology is able to accurately predict employment status and occupation at the time of interviews. Our assumption is that past experience will strongly associate with future prospects. A good typology should represent the experiences of each of the individuals, taking into account both time spent in a state and (possibly) the order in which different states are experienced. This, however, is a retrospective analysis focused upon depicting what has happened in the past. We have been examining how experiences up to age 30 were grouped, but our interest is also in how far past experiences can help us predict future outcomes. State-dependency, the inertia or continuity of social life, is widespread, but where there are discontinuities we need to be able to understand the dynamics of the changes. We are therefore concerned to identify the factors that are important where individuals have improved their positions.

Rather than take the typology as the end point of the analysis. We can assess its value in predicting outcomes at a later point in time. Our data lends itself to exploring how well the different typologies are able to do this. The typologies created were, for all respondents, based on experiences from age 16 to 30. At the point of the survey (2007) the respondents were between the ages of 31 and 37. We are thus able to use S and T to evaluate how well each is able to predict employment and job status at the age of 30. Here each typology is used as an independent variable where the aim is to predict an outcome for the respondent by the time they were aged 30. Any cohort effect arising from the age range of our sample is easily explored in the analysis. We felt that fixing the 'outcome' to employment at age 30 was more suitable than employment at the time of survey to avoid any age related effects for the older respondents. We therefore examine the ways in which each typology associates with employment status and occupation at age 30. One would expect a fairly close correspondence on these measures given that both typologies are constructed from the same past employment data, but our main interest here is the comparison between T and S and the extent to which time-in-state is a sufficient measure or whether OMA provides greater insights.

Table 4: Employment status at age 30 (in columns), comparison of T and S using 'inflows' (T= top left, S = bottom right of each cell)

	Employed	Self employed	Unemployed	Inactive	Other
Non manual	**55** / **59**	**2** / **6**	**4** / **11**	**19** / **10**	**12** / **13**
Manual	**22** / **25**	**3** / **3**	**3** / **4**	10 / **2**	**6** / **2**
Self	**4** / **2**	**86** / **82**	**1** / **1**	**3** / **2**	**2** / **1**
Unemployed	**11** / **5**	**7** / **3**	**91** / **81**	30 / **13**	**4** / **2**
Inactive	**6** / **5**	**1** / **1**	**0** / **1**	**30** / **39**	**1** / **0**
Other	**3** / **4**	**1** / **6**	**0** / **2**	8 / **33**	**75** / **82**
Total	N=370	N=161	N=264	N=312	N=108

*** <0.5%**

Typology	Cramers'V':	Chi2	Likelihood ratio
T	0.6619	2129.442	1545.133
S	0.6822	2262.055	1736.191

Table 4 begins to reveal significant differences, though there aren't many, according to the ways in which each typology has classified the respondents. The inflows in table 4 contain T in the top left and S in the bottom right of each cell. They show where those in the categories of employment at the time of interview have come from. For both S and T over 80% of those unemployed at the age of 30 are classified as such by both typologies. This shows good predictive performance of both typologies. In predicting employment at age 30, S appears to be superior to T in that 84% correspond (versus 77%). T, however, seems slightly better at predicting self-employment with in flows from non-manuals and 'other' for S. The correspondence between both typologies and inactives is the lowest though S performs better than T.

The percentages in table 4 have been highlighted where the adjusted standardised residuals are greater than 1.96. This shows the locations of the cells which contribute most to the overall lack of fit between the cross-tabulated variables. However, virtually every cell for both S and T shows a significant departure to what would be expected under conditions of independence. Indeed, there is a strong linear pattern shown on the diagonal for both typologies. Where there is disagreement between T and S, this is shown using a thick border around the cell. In these instances the cell contains a value with a significant adjusted standardised residual for one typology but not for the other. These three instances all relate to the classification of those who were 'inactive' at the age of 30. This reinforces the interpretation above that the inflows to inactivity are significantly better predicted by S compared to T. Finally, the Cramers V values for each of the tabulations in table 4 suggest that S has a slightly better fit than T.

Turning to occupation at the time of survey we undertook a similar analysis, again looking at inflows and goodness-of-fit. The numbers are obviously much smaller than for employment status (N=531), it corresponds directly to those classified as employed or self-employed.

Table 5: Occupation at age 30 (in columns), comparison of T and S using 'inflows' (T= top left, S = bottom right of each cell)

	Manager	Prof	Clerical	Farm	Manual	Petty
Non manual	46 / 57	**77** / **83**	**79** / **83**	**0** / **1**	**2** / **3**	**2** / **11**
Manual	11 / **4**	**4** / **1**	**8** / **4**	**5** / **9**	**47** / **59**	13 / 18
Self	25 / 18	**3** / **2**	**4** / **4**	**89** / **89**	23 / 20	**67** / **56**
Unemployed	7 / 4	12 / 6	7 / **0**	**2** / **0**	**16** / **9**	9 / 4
Inactive	0 / 0	2 / 5	**1** / 4	5 / 0	**9** / 6	7 / 2
Other	11 / 18	2 / 2	2 / 5	0 / 1	2 / 2	2 / 9
	N=28	N=130	N=113	N=87	N=128	N=45

Typology	Cramers'V':	Chi2	Likelihood ratio
T	0.439	511.665	541.182
S	0.468	581.758	600.303

The non-manual jobs are fairly strongly associated in both S and T according to table 5. However, managers at age 30 are better predicted by S in that 57% (as opposed to 46% for T) come from the non-manual category of the typology. The results for professionals and clerical workers are much closer. The closest association, true for both S and T, is for farm workers with 89% coming from the self-employed. The picture is much less clear for manual workers and petty traders. This shows what is to be expected about the more precarious nature of these categories. There is a greater likelihood that manual workers and petty traders will experience periods of unemployment and self-employment. The overall predictive power is therefore lower than for those with non-manual jobs. Nonetheless it is worth noting that both S and T show a fairly similar pattern such that these job categories can be seen as rooted in mixed experiences of manual, self-employment and unemployment.

When the analysis of the adjusted standardised residuals is done it is clear that there is some divergence between T and S. The manual trajectory in S significantly negatively associated with becoming a manager at age 30, unlike T which does not depart from what would be expected under conditions of independence. Clerical workers are less likely to come via the unemployed trajectory in S and less likely to come from the inactives in T. Inactives in T correspond with the manual workers at the age of 30 more than for S. The picture is more mixed in respect to the mapping of occupations from each of the typologies compared to employment status. This is to be expected given that the typologies are largely based on employment status and only marginally relate to occupation through the manual/non-manual distinction. Finally, the Cramers V values show that S is again a better fit than T. Both are fairly good but S clearly has the edge.

7. Class formation

We have reported elsewhere our evidence that by 2007 a new middle class had been formed in the South Caucasus countries, based on secure non-manual employment (Roberts and Pollock 2009). Outside this class there appears to be a large and heterogeneous ‘class’ which cannot be easily characterised. Persistent unemployment, precarious and intermittent manual or agricultural employment, and a general detachment from mainstream society is how we have described it. We know from analysing our data in conventional ways that young adults’ chances of entering their countries’ new middle classes have been related to living in the capital cities as opposed to the regions, being reared in a higher SES family in terms of the parents’ education and occupations, progressing through higher education, and remaining childless up to age 30. Here we seek to explore the factors which associate with the T and S types and employment/occupation at age 30 variables, using a wider range of evidence, including other predictors of membership of the countries’ new middle classes. Our T and S typologies both identify routes leading to middle class positions, with both S and T performing almost identically, it being far too close to call one way or the other, in predicting holding a non-manual job at the age of 30. Here we also compare how closely these types nest on the background predictors – family SES, higher education, place of residence, and parental status at age 30 – here included as an interaction variable with gender as it is known that having children has opposite effects on men and women in terms of labour market outcomes. The following tables help us to understand class formation by examining factors which associate with being an employee at the age of 30 (models 1 to 5), while models 6 to 10 examine the factors which associate with being in a non-manual job at the age of 30.

Table 6: logistic regression on being employed versus all others at the age of 30*

		S		T	
Variables	Model 1	Model 2	Model 3	Model 4	Model 5
Woman & child(ren)**	**0.202**	**0.284**	**0.281**	**0.166**	**0.163**
Man & child(ren)	0.886	0.854	0.850	0.937	0.920
Woman no child(ren)	**1.865**	**1.745**	1.685	**2.628**	**2.623**
Highly educated	**3.610**	**2.263**	**2.145**	**2.678**	**2.378**
Born 70-73	1.011	0.993	0.998	1.030	1.018
SES (high)	**1.735**	1.644	**1.853**	1.272	1.327
SES(intermediate)	1.226	1.049	1.102	0.933	0.937
Yerevan***	**4.079**	**1.962**	**-**	**1.974**	**-**
Kotayk	1.689	1.120	**-**	0.972	-
Baku	**2.395**	1.534	-	1.200	-
Aran Mugan	**2.617**	1.725	-	1.896	**-**
Tbilisi	**2.482**	1.733	**-**	0.920	-
non-manual****	**-**	**6.906**	**26.107**	**18.791**	**18.227**
Manual	**-**	**14.138**	**45.242**	**10.498**	**10.926**
Self	-	0.692	0.679	0.616	0.569
Inactive	-	**2.169**	**2.429**	**2.656**	**2.857**
Other	**-**	0.718	0.846	0.498	0.549
N	1183	1183	1183	1183	1183
LR	219.62	642.98	637.02	527.87	513.87
Pseudo R-squared	0.1509	0.4418	0.4377	0.3627	0.3531
Hosmer-Lemeshow chi2	7.41 (p=0.4935)	18.40 (p=0.0184)	10.92 (p=0.2064)	12.27 (p=0.1394)	7.92 (p=0.4416)
% correctly classified (p>=0.5)	73.71	86.31	86.31	82.76	82.76

*Odds ratios for each variable, significance at 0.05 shown in bold.
**Reference category = man with no child by age 30
***Shida Kartli is the regional reference category
****Unemployed is the employment typology reference category

There is a close agreement between S and T in table 6. This confirms the earlier analysis which has consistently shown that both S and T operate almost identically in relation to employment status outcomes in 2007. Model 1 is clearly the least adequate with a pseudo R square of 0.1509. Once the typology variables are included the fit of models 2 to 5 are substantially better. For these models the inclusion of region, while slightly improving the fit, is of far more marginal importance than the other variables. Model 3 compared with model 2 is superior, and the same is true of model 5 compared with model 4. Being an employee by the age of 30 is, for both, related to high SES family backgrounds, being highly educated. S has slightly larger R-squared values than T, but the Hosmer-Lemeshow chi-squared values for model 2 show that it is not a good fit so model 3 is the best fit with the least variables. Women with children are, unsurprisingly, significantly less likely to be employed at the age of 30 than men without children. It is interesting, however, that in model 3 there is no significance in the contrast between women and men without children. Highest social class, on the other hand, is identified by model 3 as significantly associated with being employed at the age of 30. While in many respects S and T predict equally well as is shown by the largely similar patterns of odds ratios and statistical significance, there is some evidence here that S may be slightly better. The magnitude of the odds ratios for the employment typological in model 3 is considerably larger than in the other models which suggests that the sequence analysis may actually be better than the T method at detecting state dependency.

Table 7: logistic regressions on being in a non-manual job at the age of 30*

		S		T	
Variables	Model 6	Model 7	Model 8	Model 9	Model 10
Woman & child(ren)**	**3.922**	2.536	2.559	**2.976**	2.697
Man & child(ren)	1.102	0.875	0.991	1.228	1.265
Woman, no child(ren)	**0.142**	**0.279**	0.304	**0.289**	**0.287**
Highly educated	**11.024**	**4.039**	**3.769**	**7.591**	**6.643**
Born 70-73	0.801	0.719	0.826	1.153	1.232
SES (high)	**4.562**	3.052	**4.244**	**4.376**	**5.129**
SES(intermediate)	**2.617**	1.932	**2.315**	2.088	**2.257**
Yerevan***	**4.193**	**3.585**	**-**	2.088	**-**
Kotayk	1.798	2.963	**-**	0.884	-
Baku	**3.964**	3.388	-	1.900	-
Aran Mugan	**4.527**	2.722	-	2.075	**-**
Tbilisi	**3.354**	2.692	**-**	1.155	-
non-manual****	**-**	**28.977**	**29.505**	**42.649**	**43.331**
Manual	**-**	**0.152**	**0.175**	**0.242**	**0.256**
Self	-	**0.280**	**0.241**	**0.183**	**0.162**
Inactive	-	1.607	1.550	0.252	0.323
Other	**-**	2.706	2.545	0.886	1.178
N	517	517	517	517	517
LR	268.66	475.18	469.73	473.59	469.80
Pseudo R-squared	0.3749	0.6631	0.6555	0.6609	0.6556
Hosmer-Lemeshow chi2	4.80 (p=0.7788)	11.33 (p=0.1839)	12.86 (p=0.1167)	11.36 (0.1822)	19.16 (p=0.0140)
% correctly classified (p>=0.5)	81.62	92.84	92.84	92.46	92.07

* Odds ratios for each variable, significance at 0.05 shown in bold.
** Reference category = man with no child by age 30
***Shida Kartli is the regional reference category
****Unemployed is the employment typology reference category

We then modelled occupation at time of survey, focusing on non-manual versus all other positions, and found that S and T performed remarkably similarly as when predicting employment status. Firstly, models 7 to 10 all showed significant improvement on model 6 which included neither of the typologies. In terms of the overall model statistics, there is little to separate S and T. The Hosmer Lemeshow statistics suggest that model 10 is not a good fit, otherwise we would be hard pressed to argue that one is superior than another. Model 8 has both the simplicity and the fit to suggest that it is the most suitable here. In most respects the pattern of odds ratios in model 10 are similar to model 8, the only difference is that there is no significance in the gender and children variables in model 8. Education, class and employment trajectory together contribute much to the prediction of job type at the age of 30.

In total, our evidence suggests that there is little between T and S when it comes to predicting both employment status and occupation in 2007, but that S has the edge on both occasions. This is the value-added by using optimal matching sequence methods. These methods add measurable value, but just how valuable is this value-added?

8. Discussion

We feel that we have been unable to demonstrate convincingly that sequence analysis adds truly significant value (for either theory-building or policy purposes) to what can be accomplished with more conventional methods when analysing labour market careers. There are higher labour costs involved in sequence analysis, and in our own work the outcomes have not really repaid the investment. The T career types were created using a theoretically informed sorting algorithm which, while it took some time to generate the code to process the data, was computed almost instantly. The production of the S career types was also informed by theory but there were more steps in the process: the decision to use ideal types as benchmark comparisons, the setting of the OMA costs, and the parameters of the cluster analysis all introduced theoretically informed subjectivities, and at each point, once the algorithm was started, the results were generated by much more complex and time consuming procedures. The costs of creating the S typology were thus a lot higher.

Our T and S methods allocated some cases to different types (of careers), and this was despite aligning the T and S typologies as closely as possible. The differences will have occurred mainly through the sequence method for achieving optimal matches. It might be said the sequence method has demonstrated greater construct validity as it uses a fully ordered set of experiences for each individual rather than an aggregated simplification. However, this would be to accept the optimal match definition of what was to be measured. Time methods will be a superior allocator to constructs in which time spent in a position is what counts. There is no benchmark, no court of appeal, that is not a researcher construct.

Criterion validity never proves construct validity, but the ability to predict is a reasonable test of typologies, assuming that the aims include accurate prediction. Here we found that our T and S methods performed similarly when predicting both employment status and occupation at the age of 30. That said, the sequence method was superior on both counts. We therefore feel that the sequence typology suits predictive purposes more than the time method. We should not forget that the sequence typology that we used was a simplification of an initial eight category typology. The simplification was in order to permit a direct comparison with the time-based career types. The superior performance of S was despite this simplification. However, the superior predictive power of the S method is most likely to have been due to the lesser likelihood of persons who had withdrawn from the labour market by age 30 (typically women who had married and become mothers) being allocated to one of the employment career groups (manual, non-manual or self-employed). Thus we must note that the superior predictive power of the sequence method was at least partly due to the 'rule' of state dependence rather than its ability to anticipate future movements between different, sequential positions.

State dependence (the tendency for individuals to remain in the same positions) is evident in all dimensions of life histories. In our view, this imposes a serious limitation on the potential value of sequence methods. These methods have hitherto been regarded as possessing huge potential value when examining life histories during which individuals do indeed move between positions, and there are *a priori* grounds for treating these movements as non-haphazard, implying that there is a finite number of sequences to be discovered. However, movements at 'critical junctures' are infrequent, and can be separated by long periods of state dependence. Our evidence shows that this applied to young people's labour market careers even during the turbulent transitions from communism experienced by the South Caucasus countries. It also shows that it is possible for individuals' lives to remain orderly and stable amid major macro-economic and political transformations: state dependence still prevailed in labour market careers. In other youth careers, outside the labour market, there was a single dominant sequence of status movements, ending with relatively prolonged state dependence. This applied to family careers. Individuals changed from being single, to married, then married with children, and then tended to remain in this position. In terms of housing, young people typically changed (if at all) from living with their parents to living in households where they were the senior occupier(s), and then retained this status indefinitely. In other words, there was no ‘problem’ for sequence analysis to solve. In the labour market, individuals could, in theory, switch between types of jobs, being employed and unemployed, then self-employed, maybe interspersed with periods of inactivity. In practice, however, our evidence shows that most labour market histories from age 16 to 30 were characterised by occupancy of a single position. This applied whether career types were generated using time or sequence methods.

Up to now sequence analysis experts have tended to share a benign assumption that their techniques will, at a minimum, make a useful addition, and potentially, when fully operational, a powerful addition to the social research armoury. The experts have applauded each others’ demonstrations that the techniques work in a technical sense, that is, in revealing sequences. They have argued among themselves about what should be the preferred way of conducting sequence analysis. They have demonstrated, as we ourselves have above, their methods’ ability to corroborate what we already know about the typical regularities in people’s lives. Our view is that, if it is to repay the costs, sequence analysis needs to do more, and to the best of our knowledge, ours is the first attempt, using benchmarks created using non-sequence methods, to assess the value-added in analysing labour market careers, or any other dimensions of life histories. Corroborating what is more easily demonstrable using other techniques is insufficient. There are no so far unsolved work-life history problems for sequence analysis to address in future work. This also applies to family and housing careers. Lifelong status attainment and social mobility flows can be examined effectively without resorting to sequence techniques. The challenges for sequence analysis, which we intend to address in future work, where the costs might be repaid, are problems that cannot be tackled effectively using other methods. These will include identifying multiple sequences: for example, the intersections between labour market, family and housing careers. Leisure careers are a further example: the dependencies and independencies between changes over time in participation in different kinds of leisure activities. Until it can demonstrate success with so-far unsolved puzzles (and there are many such puzzles), sequence analysis will remain promising but still of unproven value.

References

Abbott, A. (1995) Sequence analysis: new methods for old ideas. *Annual Review of Sociology.*, 21, 93–113.

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

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

Brzinksy-Fay, C., Kohler, U. (2010) 'New Developments in Sequence Analysis' *Sociological Methods and Research* 38(3), pp359-364.

Brzinksy-Fay, C., Kohler, U., and Luniak. M., (2006) 'Sequence Analysis in Stata' *Stata Journal* 6, 435-60.

Buhlmann, F. (2010) 'Routes into the British Service Class: Feeder Logics according to Gender and Occupational Groups' *Sociology,* 44(2), pp195-212.

Dex, S. (1984) 'Work History Analysis, women and large scale data sets' *The Sociology Review,* , 637-661.

Djikstra, W. and Taris, T. (1995) 'Measuring the agreement between sequences' in *Sociological Methods and Research,* 24, 214-231.

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

Elzinga, C.H. (2010) 'Complexity of Categorical Time Series' *Sociological Methods and Research,* 38(3) pp463-481.

Gabadinho, A., Ritschard, G., Studer, M., and Müller, N. S. (2008) *Mining sequence data in R with the TraMineR package: A user's guide* University of Geneva.

Gauthier, J-A., Widmer, E.D., Bucher, P. Notredame, C. (2009) 'How much does it cost? Optimization of costs in sequence analysis of social science data' *Sociological Methods and Research* 38, 197-231.

Halpin, B. (2010) 'Optimal matching analysis: the importance of duration'

Hollister, M. (2009) 'Is optimal matching suboptimal?' *Sociological Methods and Research* 38, 235-264.

Johnson, L.A.S (1970) Rainbow's end: The quest for an optimal taxonomy. *Systematic Zoology* 19, 203–239.

Levine, J. H. (2000) But what have you done for us lately? *Sociological Methods and Research*, 29, 34–40.

Pollock, G. (2007) 'Holistic Trajectories: a study of combined employment, housing and family careers using multiple sequence analysis' *Journal of the Royal Statistical Society A: Statistics in Society*, Vol. 170, pp167-183.

Pollock (2010) 'Appendix1: Questionnaire show cards and instructions' in J. Evans and W. Shen *Youth Employment and the Future of Work,* Council of Europe, Strasbourg.

Roberts, K., Clark, S.C., Fagan, C., and Tholen, J. (2000), *Surviving Post-Communism: Young People in the Former Soviet Union.* Edward Elgar, Cheltenham.

Roberts, K. and Pollock, G. (2009) 'New Class Divisions in the New Market Economies: Evidence from the Careers of Young Adults in Post-soviet Armenia, Azerbaijan and Georgia' *Journal of Youth Studies* 12, 579-596.

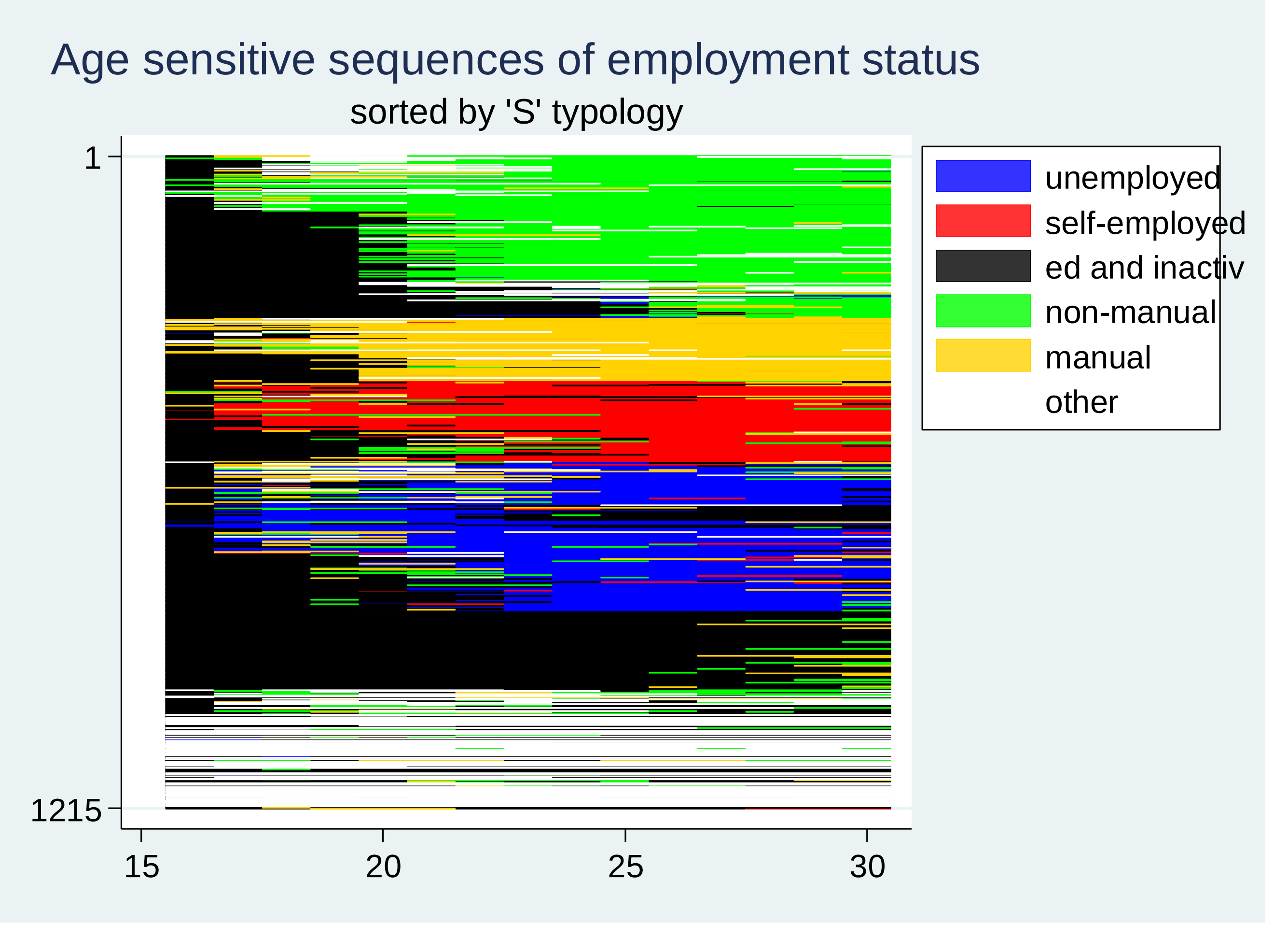
Roberts, K. Pollock, G. *et al* (2009) ‘Young adults’ family and housing life-stage transitions during post-communist transitions in the South Caucasus’ *Journal of Youth Studies.* Vol 12 (2) pp151-166.

Roberts, K. Pollock, G. Manasyan, H. and Tholen, J.(2008) ‘School-to-Work Transitions after Two Decades of Post-Communist Transition: What’s New?’ *Eurasian Journal of Business and Economics.* 1, 103-129.

Roberts K, Tholen J, Khachatryan G, Velidze R, and Pollock, G. (2008) Transitions to adulthood in rural villages during the transition from communism in the South Caucasus. *Social Problems* vol 4 pp 117-140.

Tholen, J. Huseynzade, D. Ibrahimov, A. Pollock, G. and Roberts, K. (2010) ‘Graduate careers during the post-communist transition in the South Caucasus’ *Educational Research and Reviews* Vol. 5(3), pp. 108-118.

Wu, L. L. (2000) Some comments on “Sequence analysis and optimal matching methods in sociology: review and prospect”. *Sociological Methods and Research* ., 29, 41–64.



Age sensitive sequences of employment status
sorted by 'S' typology
1
1215
15
20
25
30
unemployed
self-employed
ed and inactiv
non-manual
manual
other