Sequence analysis proceeds by calculating distances between pairs of sequences, often generating matrices of distances between all pairs or distances to reference sequences. The sequence analyses will utilise economic activity data from sweep 23 of the NCDS that measures the economic activity of individuals from the age of 16-23.

The following sequence analyses will employ seven different distance measures and where a substitution matrix is required, uses a data-driven matrix excluding the diagonal for a larger range compared to a data-driven substitution matrix that includes the diagonal. The seven distance measures are: the Optimal Matching Algorithm, Hamming Distance, Hollister’s localised OM, Halpin’s duration-adjusted OM, time-warp edit distance (TWED), and Dynamic Hamming. Each one will now be described below:

The Hamming distance is the simplest way of mapping information about differences in a state space to differences between trajectories within that state space. The distance between the sequences is the sum of the period-by-period distance between the states. Optimal Matching Algorithm improves upon this by allowing ‘alignment’, in other words, sliding parts on one sequence along the other if this improves the ultimate matching.

Both localised OM produced by Hollister and Duration-adjusted OM developed by Halpin are variants of OM but have a relative similarity. Both approaches focus on the insertion, deletion, and substation and how OM doesn’t take account of the local context in which they operate. Neither measure generates metric distances which produces issues for any desire in continuing with them when it comes to cluster analysis.

TWED operates very similarly to OM, as it uses a substitution matrix, and has operations analogous to substitution, insertion, and deletion. The key difference being that time warping does ‘continuous time-series to time-series correction’ whilst OM does ‘string to string correction’. TWED was designed to accommodate irregular time-sampling.

Finally, Dynamic Hamming changes its transition matrix over time and therefore uses transition-derived inter-state distances in an element-wise sequence comparison, where the distances change over time. It is most appropriate where there is a clear ‘’calendar’’ and has been used to great effect in the analysis of time-diary data.

The simplest way to compare the seven pairwise distance matrices is via correlation. The Pearson correlation for each distance below demonstrates that…

|  |  |
| --- | --- |
| Table 1: Examining Distance Measures | VECH Correlation |
| OMA & OMV | 0.9334 |
| OMA & HOL | 0.9955 |
| OMA & TWD | 0.9618 |
| OMA & HAM | 0.9667 |
| OMA & DYN | 0.9129 |

For many of these distance matrices, they must imply a metric space. For continuing to use these distance matrices for eventual clustering analysis, metric space is a requirement. This requires, inter alia, that the distances obey the triangle inequality: for all A and B, there is no C such that d(A, B) >d(A, C)+d(C, B). As mentioned prior to this, as well as confirming it via statistical testing, it is relevant to point out that neither OMv nor Hollister are consistent with a metric space and are thus dropped from analysis.

Moving on to cluster analysis to provide a more holistic descriptive picture, we continue with 5 of the previous 7 matrices. Each generates 8 clusters – somewhat arbitrary. Using the Adjusted Rand Index, we can reflect on the similarity of different clusters in comparison to our parent OMA cluster solution. This similarity reflects the extent to which the members of a pair of cases, if in the same cluster in one solution, are in the same cluster in another. As we can see in table 2 all three comparisons produce a very similar result. This isn’t particularly surprising for the comparisons of OMA to TWED and Hamming because all three used the same data-driven substitution matrix, however the remarkable similarity of the Dynamic Hamming distance also is worth comment. The permtabs give a more complete descriptive picture. The permutation maximises Cohen’s k as an index of agreement and reports the kmax to be again remarkably similar across the comparisons as seen in Table 3 (full permtab results in appendix below).

|  |  |
| --- | --- |
| Table 2: Adjusted Rand Index of o8 by other Distance Measures |  |
| OMA & TWD | 0.8189 |
| OMA & HAM | 0.7921 |
| OMA & DYN | 0.8091 |

|  |  |
| --- | --- |
| Table 3: Permtab Kappa max |  |
| OMA & TWD | 0.7649 |
| OMA & HAM | 0.7623 |
| OMA & DYN | 0.7635 |

From this point, it is fair to say that all four measures have a very similar composition, as witnessed through their VECH correlation, Adjusted Rand Index, and Kappa max. There has been no rationale at this point to desire any of the other distance measures over OMA, to continue the cluster analysis we will therefore drop the other distance measures.

Studer’s discrepancy measure brings a pseudo–analysis of variance perspective to distance matrices. If we partition the matrix using a cluster solution, or a pre-existing observed characteristic, we can compare the average distance to the centre of the partition with the average distance to the overall centre and generate a pseudo-R2 measure. This uses the distance to the center as an analogue of sums of squared deviations, and where the distances are squared Euclidean, it will generate results that are numerically equivalent. The approach uses bootstrapping to generate p-values. As we can see, the pseudo R2 is 65% - relatively high. The Psuedo F, a statistic is the ratio of the between-cluster variance to the within-cluster variation- our measure showing that this ratio is rather high.

|  |  |  |
| --- | --- | --- |
| Table 4: Discrepancy based R2 and F | | |
| Pseudo R2 | Pseudo F | p-value |
| .6475815 | 3288.926 | .01 |

I will now create string representations of the sequences, which provides a visual overview of the data and its subsequent patterns. For this data, I use the symbols “FPEsuon”, each of these are in reference to F= Full-time Employment, P= Part-time Employment, E= Education, s=school, u=unemployment, o=out of labour force, n=n/a. As can be seen from the table, a simple representation of the first 10 instances of sequences is given.

|  |  |
| --- | --- |
| Table 5: Sequence Text |  |
| 1 | s:2/F:71/u:6/o:10 |
| 2 | s:2/F:73/u:1/F:8/n:2/F:3 |
| 3 | s:2/F:87 |
| 4 | s:2/F:87 |
| 5 | s:2/F:87 |

As we can see from table 5, there are three variations of sequences apparent. The 3-5th sequences appear to show a very straightforward pathway that resembles a school-to-work transition. Sequence 1 shows something similar but with a period of unemployment at the end of their sequence, and sequence 2 shows an in and out of work pattern at the end of their sequence.

Before going further, we can now descriptively see how many instances individuals within these sequences exited from a particular pattern. Here I detail this for each of the symbols given – this has some relevance to the relative stability that sociologists have placed on this time period. Most important to highlight sociologically is the number of exists that sequences saw from Full-time Employment over the duration of their time from 16-23 years of age. Over half of all individuals in this timeframe experienced an entrance and exit from employment during this time. This does promote some doubt over the supposedly relatively smooth transitions that are said to have taken place around this time.

|  |  |
| --- | --- |
| Table 6: Count of sequences that observe and exit | No. |
| Schooling | 12,110 |
| Education | 5,120 |
| Full-time Employment | 6,389 |
| Part-time Employment | 663 |
| Unemployment | 4,977 |
| Out of Labour Force | 4,081 |

Whilst a visual representation of the sequence strings is descriptively beneficial to look at, it would provide much more insight if we were to look instead at the clusters of sequence strings. To look at the clusters more closely, medoids are constructed- sequences nearest the centre of the cluster. The table below provides these medoids and their string representations.

|  |  |
| --- | --- |
| Table 7: Sequence Text by cluster mediods |  |
| 1 | s:2/F:71/u:6/o:10 |
| 2 | s:3/F:50/o:36 |
| 3 | s:2/F:23/o:64 |
| 4 | s:3/n:1/E:22/F:63 |
| 5 | s:3/E:1/s:10/F:75 |
| 6 | s:3/E:1/s:23/F:56/u:1/o:3/F:1/n:1 |
| 7 | s:3/n:1/E:46/F:39 |
| 8 | s:3/E:1/s:23/E:41/F:21 |

There are three main sub-groups that make up these 8 clusters. The first is cluster one – which makes up just over half of all sequences. This is the ‘school-to-work’ cluster, where see people go straight from schooling into Full-time employment, with some periods of unemployment at the end of the sequence. The second, is the ‘hard times’ sub-group forming clusters 2 and 3. This group does similarly to cluster 1 but has a prolonged period out of the labour force nearer the end of their sequences. Sub-group 3 comprises of the ‘Education First’ groupings from cluster 4-8. This group sees people staying in school or going on to education for a prolonged period prior to entering the Full-time Workforce. Importantly this sub-group does not suffer from the large spouts of unemployment and out of the workforce periods that the other sub-groupings face.

To visually represent this, first figure 1 shows a wide format view of the entire dataset. With this it is straightforward to look at certain trends as time progresses – spikes in unemployment and out of the labour force indicate key economic downturns in the UK economy for example (just prior to 200 months, around 220 months, and finally the large stagflation crisis around 270 months onwards). We can also pull apart some very general trends – confirming with cluster one that the majority of individuals did indeed go straight from school-to-work, however there is a substantive amount of individuals that stayed on in schooling post-16, as well as sizeable among of people that decide to go on to higher education.

Chart

Description automatically generated with medium confidence

Figure

This period of British history was a pivotal moment for women entering into sectors of education and the labour market, thus figure 2 breaks down this chronogram by sex. Shockingly it appears that almost all the rising unemployment can be attributed towards women, apparently detailing a collapse in the labour market centring on women specifically. It is also worth noting that women appear to be less likely to jump straight into Full-time Employment, instead relying on staying on in school, and going on to Higher education.

Chart

Description automatically generated

Figure

To get a more detailed look at these patterns, figure 2 breaks this data down by cluster.

Graphical user interface, application, table

Description automatically generated

Figure

As we can see from figure 3, this visually represents the data we have in table 7 but also helpfully visualises the size of each cluster as well as their overall pattern in time. Whilst cluster 1 is by far the largest cluster- taking up a majority of sequences, the sub-group ‘Education First’ combined nearly matches its size.

Before going further with any descriptive analysis, we look at the cumulated duration, entropy, and number of spells to differentiate between the clusters, sequence complexity, and volatility respectively. These can all be found within the appendix (as is subsequent analysis).

The index plot allows us to convert data into long format, and plots each sequence as a line which allows us to reproduced the sequence data in full. Figure 4 demonstrates this.

Graphical user interface, application

Description automatically generated

Figure

Whilst not as visually appealing as the wide format chronogram, place the visual representation in long format does allow for the inspection of clusters in more detail. Especially with respect to cluster one, where can see once more, the periods of unemployment and out of the labour force appear to line up with periods of exogenous economic shock, with an increasing concentration around the end point of our time scale- a period in the UK that saw widespread economic turmoil.

Finally, we see a transition pattern graph in figure 4, which creates a composite 7\*7 grid representing the transition rates between states over time. This is also visualised by producing the matrix transition rates in table format, giving some examples below.

Letter

Description automatically generated with medium confidence

Figure

From figure 5 some trends also emerge that are important to reflect on. The obvious trend that has previously been stated is the rapid growth near the end fo the time scale – entering the 1980s – of both unemployment and out of the labour force. The more sociologically interesting trend however is the steady growth of Part-time employment – this growth trend starts from the age of 16 and continues to rise all the way through to the end of the time scale at age 23. Whilst the rise in unemployment and out of the labour force demonstrate economic exogenous shocks in the forms of recessions, the rise in Part-time employment is an indication at least, of a changing structure of the labour market.