# Introduction to Sequence Analysis in R to create typologies of longitudinal trajectories

## Transcript video 4

Full resource (and video), see: <https://www.ncrm.ac.uk/resources/online/all/?id=20853>

Ian Thomas: So in the previous video we’d ended off by assigning each of the 2,000 people in the biofam dataset to one of five clusters based on how dissimilar the sequences of family household formation or family trajectories were, and in this final video I’m just going to take you through some of the possible plots that could be used to visualise sequence data in order to help interpret what those five clusters mean.

 I should say that this is only a selection of some of the possible visualisations, the TraMineR user guide has a far more detailed discussion of all of the possible variations that you can use.

 The first plot I’m going to go through is called meantime plot and in essence this gives you an indication of how much time on average did the different clusters spend within each of the possible states, family formation states, so if I run this. So the function, the meantime plot function, calls in the sequence object and it produces a visualisation that’s grouped by our cluster membership and here’s the visualisation.

 So the height of these bars is a representation of the average time spent within that particular state, time relates to… well, in this case years because our states were measured at each age, each year between the age of 15 to 30, so as an example of how you’d interpret this, cluster three is a fairly easy cluster and unique in that the majority of time within this cluster was spent in state zero and in cluster one for example, people spent roughly, I don’t know, equal amounts of time in state zero and state six.

 At the moment the labelling of the states makes this pretty difficult to interpret but in the worksheet I actually show you how to attribute values and labels to the states which makes interpreting these visualisations a lot easier.

 And I’m just going to move on to the next visualisation and I’ll just run it and explain it, again it uses the same process of splitting the visualisation by cluster membership. This is a different approach, so it shows you… essentially gives you a breakdown of what states were that particular cluster in during each of the time periods that are measured in our sequences, so it’s essentially just a stacked bar graph multiplied by however many observation points you have and it’s quite useful for understanding general sort of migrations into certain states, so as an example, the majority or all of the people in cluster one were in state zero when they were 15 years old and they all seemed to migrate into state six by the time they’re 30.

 Similarly, cluster four all seemed to be in state zero at age 15 but by the time they’re 30 they are all in state number two and for state cluster number two there seems to be an intermediate period spent in state zero, so everyone’s in state zero to begin with, they all seem to migrate into state one and then end up going into different states by the end of the observation period.

 The one thing to be mindful of is that this is aggregated information split by these different time points and so these aren’t actual sequences and so you kind of lose the uniqueness of the sequences through this aggregation.

 The next visualisation I’m going to go through kind of get around that sort of loss of information because it actually plots an actual sequence of information rather than aggregating it and it’s called a medoid or representative sequence plot, and I’m going to run it and explain it.

 So again, we’re using the sequence object, grouping the visualisation by our cluster solution. Because this is a plot of the most representative sequence we need to feed it to the distance matrix because, you know, it needs something to calculate the most central sequence and then this option is just saying ‘extract one sequence’ so you can extract multiple sequences in order to get greater coverage of representativeness within each cluster.

 And if I open the visual, admittedly this is a bit cluttered and there are methods that you can use just to focus on one of these visualisations for a cluster rather than presenting them in this sort of tessellated thing, but we’ll just go with this for now for the purposes of the video. So the way to interpret this is this is an actual sequence that belongs to someone and it gives a breakdown of what this person reported they were at each of these 16 timepoints, so cluster one, the most central sequence, was being in state zero between the ages of 15 to 21 and then going in to state six from age 22 all the way up to last observation at age 30.

 There is some additional information that you can sort of see here which is the coverage of this sequence, so it’s essentially saying how many other sequences within this cluster are within the vicinity of this most central sequence, and I’m just going to move on to the last visualisation.

 So this visualises the transitions between the different states and if you recall, this is sort of similar to the methodology used to calculate the costs behind our distance matrix, so as before I’m just going to run it and explain.

 So I’ve taken a slightly different approach and I’ll open it up so you can see why. So this is quite a busy diagram and it obviously doesn’t lend itself well to trying to cram five other transition matrices into the same plot, so I’ve used a piece of code in order to just focus on cluster one, so this is the transition matrix for cluster one, and if you wanted to produce the transition matrices for other clusters then you just change the number after this particular piece of code there.

 This option within the transition matrix plot function is just to make sure that the labels within the diagram don’t overlap, and also this particular function means that the transition matrix only plots transitions into a different state, so obviously if you’re comparing two timepoints you could remain within the same state between those two timepoints, using this option means that you only look at instances where someone changes a state.

 And if I open back that, so essentially the transition matrix compares the state at one given timepoint, so timepoint T, with the following timepoint T plus one, so you sort of interpret this as if you were in state one and you left then 58% of people left state one to go into state three and 41% of people left to go into state six. These figures, some across rows, so this is the proportion of people who leave this state at time T and then where they go at time T plus one.

 So that’s all the visualisations I’m going to go through, as I said the TraMineR manual has loads more. You would also probably do some other analysis, so in the worksheet I provide you with a piece of code that allows you to attach cluster membership to the biofam dataset so you can then engage in a series of bivariate analyses. For example, you could look at the distribution of gender by cluster membership or, you know, whatever else is contained within the biofam dataset, so you could look at if there’s different nationalities follow different clusters or different languages follow different clusters, so that’s one potential avenue for sort of describing what’s going on within those clusters.

 You may also want to engage in multinomial logistic regression to predict cluster membership, so the interest there would be controlling for a range of characteristics, what makes you more likely to follow or to belong to one particular cluster over another cluster?

National Centre for Research Methods (NCRM)
Social Sciences
Murray Building (Bldg 58)
University of Southampton
Southampton SO17 1BJ
United Kingdom

**Web** www.ncrm.ac.uk
**Email** info@ncrm.ac.uk
**Tel** +44 23 8059 4539
**Twitter** @NCRMUK