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

## Transcript video 1

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

Ian Thomas: In this first video I’m just going to give you a brief overview of the R studio layout, then I’m going to introduce the dataset that we’re going to analyse before I go on to showing you how to create sequences, which is obviously the first part of using sequence analysis to create a typology.

 In terms of the layout of R studio, on the bottom left hand side is what I’m going to call the console but it probably has a proper name, this is essentially the area where you can input directly different functions and it also is where feedback is also printed on the screen. Above that is the R script that I’m going to use today, there is a different version in the supplementary material associated with this where I go through in a bit more detail certain things, and then on the far right hand side of the screen - over here - is where the different datasets and dataframes that we’re going to create and use during this session will appear, and it’s just a place where you can double click on them and open things up a lot more easier.

 So the beginning of each R session you need to load the libraries that you’re going to use during that session, and in this case that’s the TraMineR, the cluster and this plot library. TraMineR library is the main one for sequence analysis, cluster, the cluster library, is going to be used to engage in cluster analysis and then this plot library is used to create visualisations that will help interpret clusters later on.

 So now I’m just going to run this particular script using the library function to call in those three libraries.

 The next thing I’m going to do is to call in the dataset that we’re going to use, and that is the biofam dataset. Now this dataset is already included with the TraMineR package so you can open this up yourselves.

 Okay, so I ran that and as you can see, the biofam dataset has appeared on the right hand side of the R studio console and I can double click it and it opens up the biofam dataset for me.

So a little bit about the biofam dataset, so it’s a subsample of 2,000 people from a much larger study that was looked at, the family formation histories of a group of people and essentially, in each year between the ages of 15 up to the age of 30 people were asked to give an indication of their family formation during those ages.

 And as you can see, the biofam dataset includes a number of variables, so sex, nationality, and if we scroll along, these are the variables that actually relate to family formation status in each single year of birth from age 15, so A15, all the way up to age 30, so A30, and I’m just going to close that down.

 So the first step of actually creating a typology of launch 2 null trajectories is to actually create those trajectories or to create those sequences, and in order to do that we use the function called seqdef and what this function does is it says look in the biofam dataset and specifically look at variables 10 to 25 and use those to create the sequences, which are then compiled into something that’s called a sequence object, so for each person it will combine variables 10 to 25 to create the sequences of family statuses.

 And if I run that particular function. So as you can see, another thing has popped up in the right hand corner of the R studio console, which is this thing called a sequence object and if we look at the console itself it provides us some summary information that’s quite useful to know, so there are eight distinct states, so eight distinct family formation types, that have been labelled zero to seven and at the moment no detail labelling has been added to those dates because they’re all labelled again zero to seven.

 There were 2,000 people in the biofam dataset and there are 2,000 sequences, and also we can see that the sequences are of length 16 and if you recall, we asked the seqdef function to use variables 10 to 25, which is 16 different observations, so that matches up with the expected number of variables, it matches up with the sequence length, so we know that that the function has actually done what we’ve wanted it to do.

 And there’s one last thing, I’ll just open up the sequence object so you can see what it looks like, again it’s not overly thrilling, it’s just essentially removed the variables that contained the states and put them in a separate R object so if I just open biofam data, so person 1167 their first observation at age 15 was 01167 in this sequence object, their first observation was also nought.

 In the worksheet I actually go through how to add different attributes to this sequence object, so you can include colours but also more detailed labels, but for the purposes of this video we’re just going to stick with a very basic sequence object.

And that is the first part of how to create sequences. In the next video we’re going to go through how to compare sequences.

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