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

## Transcript video 3

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

Ian Thomas: So in the previous video I went through how to create a numerical representation of how different sequences are from each other, and in this one we’re actually going to engage in a cluster analysis of that distance matrix which is the point at which we’re actually creating a typology because the point of cluster analysis is to identify groupings of similar things, similar objects, which in this case would be groupings of similar sequences.

 As with anything in R there are numerous ways in which you can engage in a cluster analysis of a distance matrix and here we’re using the hclust function which is from the cluster library that we brought in earlier, and there are numerous other libraries that you can use to engage in cluster analysis.

 So the hclust function allows you to engage in any number of what’s called hierarchical clusterings and I’ll produce a visualisation soon which explains a bit more about why it’s called hierarchical. There are other methods, non-hierarchical methods, but for illustrative purposes we’re just going to do this.

 And so the hclust function draws on the distance matrix that we created earlier. Because the matrix isn’t in a format that hclust is expecting, we have to do this bit of data wrangling to make it conform to a format that the hclust function can use and with this method option we’re telling function to do a specific kind of hierarchical analysis which is called Ward’s method. So if we run up to get a piece of code, again running that, the hclust function produces an array of different outputs which are all saved as this thing called Ward hclust which we can draw on through numerous other functions later on.

 Now one of the main things, well, a visualisation that will hopefully help you understand why it’s called hierarchical, is something called a dendrogram, which is essentially a visualisation of how each of the sequences can be combined into clusters that are then combined into larger ones and so on and so forth, and this is the sort of visualisation so at the bottom would be your observations, so your cases. Because there are 2,000 people in this particular dataset then the clarity of this visualisation is a bit low but you can see how these individual sequences combined into larger groups which then combine and so on and so forth, into this hierarchical tree.

 Now with this particular form of hierarchical cluster analysis the way you assign people to clusters is by what’s known as cutting the tree, which is in terms of this diagram, would be visualises drawing a horizontal line across this diagram and wherever that line cuts or intersects a branch, the people underneath that branch or, you know, the people linked with that branch, are associated to the same cluster.

 For the purposes of just this video we’re going to say we’re going to go with a five cluster solution so I will close that. So in order to cut that tree, in order to produce five clusters, we use the function known as cut tree, so what this is saying is… I mean how this function works is drawing in the output from our initial hierarchical cluster analysis, cut the tree so that it produces five clusters and if you wanted a different number of clusters then you would change the value of this that appears after this K equals option, so if I run that this is going to create an object that contains for each person what cluster are they attributed to.

In the worksheet I go through a few more functions that would allow you to attribute that cluster membership to your original dataset and also how to create descriptive statistics for the number of people in each of those clusters but for now that’s all we’re going to do in this video and in the next video we’re going to explore how to create visualisations in order to interpret what those clusters mean.

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