

An example of doing a cluster analysis in a simple way with continuous data.

David Byrne

The data set is derived from the 1991 Census and consists largely of a series of percentages calculated in order to yield a set of social indicators for wards in the Bradford and Leicester areas. For the use made of the data set see:

Byrne, D.S. - 'Class and Ethnicity in Complex Cities' 1998
Environment and Planning A 30 703-720

Cluster Analysis

This is most easily done with continuous data although it can be done with categorical data recoded as binary attributes.

We begin by doing a hierarchical cluster from the **classify** option in the **analyse** menu in SPSS. The classifying variables are % White, % Black, % Indian and % Pakistani.

Inspect the Agglomeration Schedule to identify the stage at which significant types emerge. Here if we look at the Coefficient value we can see rather big jumps from 4 to 3 and 3 to 2 clusters. We can see this on the graph which I made by pasting from SPSS into MINITAB – we all have our little tricks.

Agglomeration Schedule:

The agglomeration schedule shows the amount of error created at each clustering stage when two different objects – cases in the first instance and then clusters of cases – are brought together to create a new cluster. A large jump in the value of the error term indicates that two different things have been brought together and that there is a significant typology at that level of

fusion. For the Bradford / Leicester Wards data set we can see a jump in which the value of the error coefficient nearly doubles when 4 clusters are reduced to three.

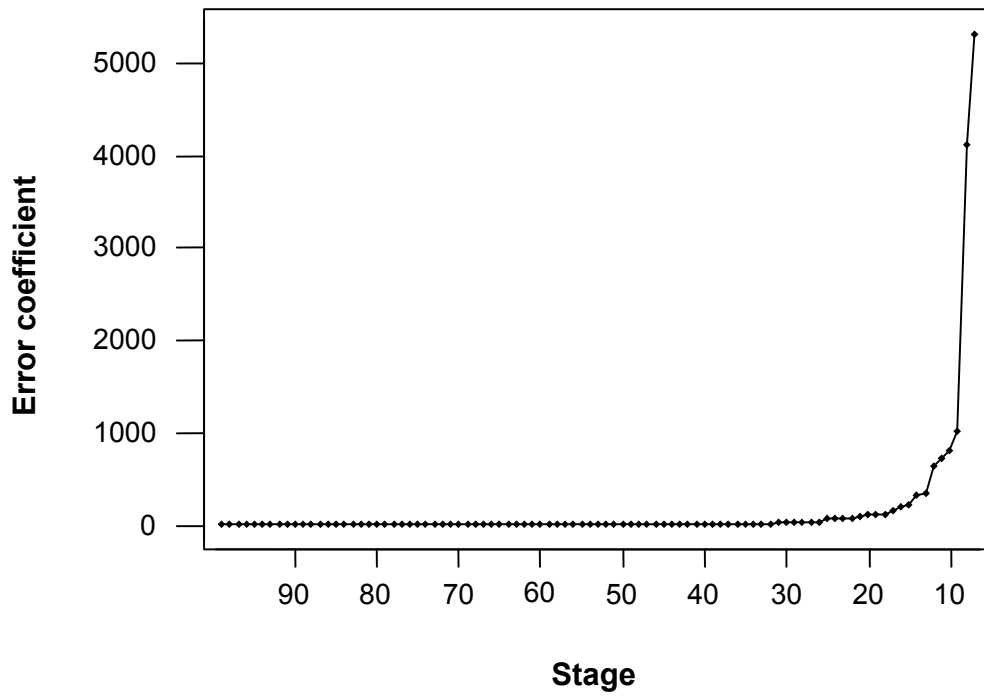
The clustering process starts with all the cases – 94 Wards in this case – and then fuses the two which are least different to make 93 clusters, at this stage most being single cases. This then continues until all the cases are fused into a single cluster. The Wards as cases are described by a set of indices, most of which are percentage e.g. percentage population Indian, percentage of households owner-occupied and so on.

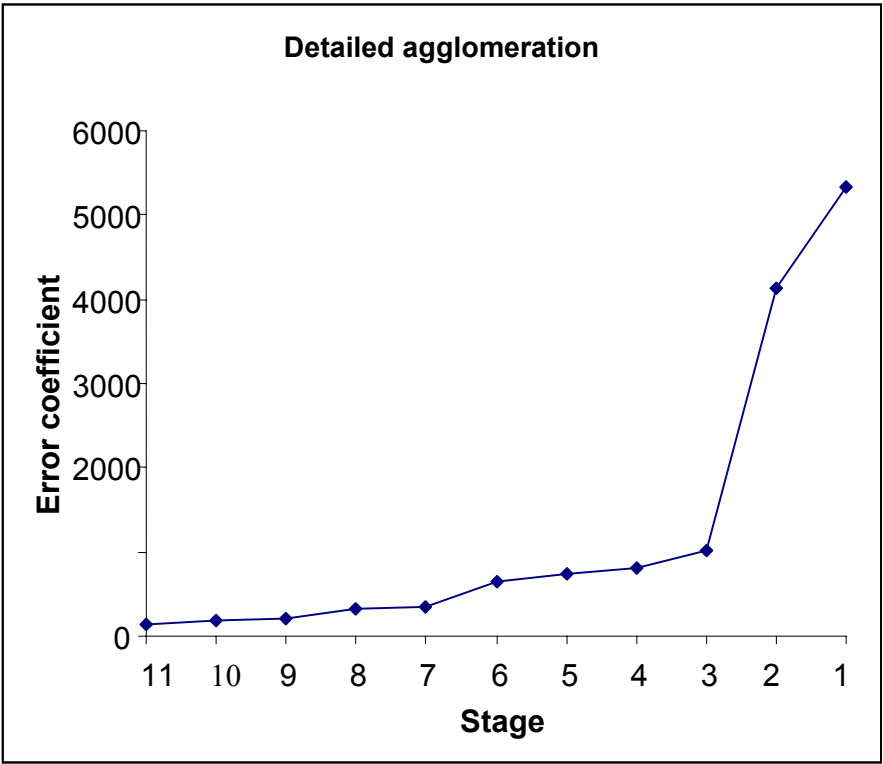
This table shows the end of the agglomeration schedule and indicates a jump in error value, and hence a significant typology, when three clusters are fused to make two.

Clusters Stage	Number Cluster	Error Co-efficient
	83	146.118
	84	190.286
	85	210.676
	86	317.640
	87	342.596
	88	641.646
	89	729.812
	90 Fourth --	815.925
	91 Third ---	1024.923
	92 Second --	4126.820
	93 All --	5327.993

We can copy the error coefficient column into any spread sheet and then graph it out with the value of the error coefficient as the Y axis and the stage of clustering as the X axis. Then a radical change in the slope of the graph indicates a fusion of two very different clusters. Obviously subsequent fusions are joining together very different objects so there is a big jump in error values after the first significant typology. Here significant means substantively significant.

Full agglomeration Schedule





So we repeat the clustering exercise first **saving** the three cluster solution. Inspection of the data spread sheet shows that this creates a new variable **clu4_1**. This indicates the cluster of which each case is a member.

We can inspect the clusters by using the procedure Compare Means – Means in the **analyse** menu. We can then see the mean values by cluster, not only for the classifying variables but also for any other variables of interest. The table below does this. We can see that the clusters differ so that we have a predominately white, mixed, large Pakistani component, and large Indian component cluster. We can then see the pattern of other variables for these clusters.

Report

	Average Linkage (Between Groups)							
	1	2	3	Total				
	Mean	N	Mean	N	Mean	N	Mean	N
% population white	91.6404	82	40.0938	3	38.1885	9	84.8776	94
% Black	1.0076	82	2.3427	3	3.3289	9	1.2724	94
% Indian	4.3040	82	8.2598	3	51.6431	9	8.9627	94
% Pakistani	1.7757	82	44.6023	3	2.1566	9	3.1790	94
% limiting long term illness	11.9328	82	13.3224	3	13.2007	9	12.0985	94
% born in UK	92.8633	82	68.2819	3	60.7847	9	89.0074	94
% males unemployed	9.8773	82	28.5518	3	18.7085	9	11.3188	94
% households owner occupied	74.3387	82	65.4386	3	64.3822	9	73.1014	94
% rented privately	5.4801	82	15.8041	3	11.9833	9	6.4322	94
% social housing	19.1071	82	17.5504	3	22.7305	9	19.4043	94
% no car	31.8482	82	56.8496	3	45.8470	9	33.9865	94
% child and no working adult	18.9553	82	38.1990	3	23.7481	9	20.0284	94

We could have used a K means cluster analysis and this produces somewhat tighter clusters. In other words we specify that we will be using four clusters at the beginning of the analysis.