

1 The Analysis of groups (2): Cluster Analysis¹)

Note: This is only an elementary introduction, explaining the basic ideas of cluster analysis and leaving out a number of technical details. Cluster analysis is a field of its own, complex enough for at least two two-week workshops.

Below we shall briefly explain how they work and then use some examples to demonstrate the use of cluster analysis with the EDA package.

The purpose of these techniques is to analyse a data set with many variables and cases and try to summarize its main features by forming groups of cases or variables, where hopefully similar cases or variables are grouped together. The basic result of a cluster analysis is a nominal variable, showing to which group a case (or variable) belongs [group membership]. Within the EDA program this variable is called GVAR (for cases) or TABLE (for the variables).

As it is possible to classify variables, observations, correlations, distances or other things, we often use the term 'object' i.e. any entity we can compare to another and tell whether they are similar or not.

There are basically two families of methods: (1) hierarchical methods and (2) non-hierarchical (agglomerative) methods.

1.1 Hierarchical clustering

Hierarchical clustering proceeds as follows:

- ◆ 1. At the start each object is a group by itself.
- ◆ 2. Seek for the *two most similar* objects.
- ◆ 3. The two objects are *merged* into a single new object and no longer distinguished as individual objects, i.e. the merger is a new object to be compared in the next step to the other objects.
- ◆ 4. Repeat (2) and (3) until all objects are in a single group.

There are several problems to be solved. We did not say:

- ◆ What "similar" means, i.e. the question is how we measure similarity or dissimilarity between objects.
- ◆ How individual objects are merged, i.e. what are the numerical values on all variables for the newly created object.
- ◆ How we create classes, as the the procedure does not produce the group memberships (the nominal variable) we want.

There are a large number of criteria defining similarity (or distance), as well as for merging objects to form clusters. (More on this later.) But is important to stress **1** *That the result will depend upon the similarities used and the way new objects are defined.* and **2** *There is no such thing as the best classification (except possibly "best yet" from a substantive point of view).* Each method differs it uses a different way of defining similarity and merging differnt objets into a new object.

This procedure leads to a interesting graphical representation: a *dendrogram* or *hierarchical tree*. The visual inspection of this tree will let you understand the logic of the classification process and will let you - ultimately - define the groups.

In the EDA program two commands perform hierarchical clustering: the HIERARCHY and VHIERARCHY commands. Let us consider an example using the world data, restricted to North and Central America:

```

WA:X1.EB World data North and Central America
Seq# TPD N tie Label descriptor
1 31 PGrow Population Growth
2 31 Urb Urbanization
3 31 InfMor Infant Mortality
4 31 PopHos Population/Hospital
5 31 PHBed Population/Hosp.Bed
6 31 PopDoc Population/Physician
7 31 GNPCap GNP per capita
8 31 GNPAgr %GNP for Agriculture
9 31 GNPInd %GNP for Industry
10 31 GNPServ %GNP for Services
11 31 Lit Literacy Rate

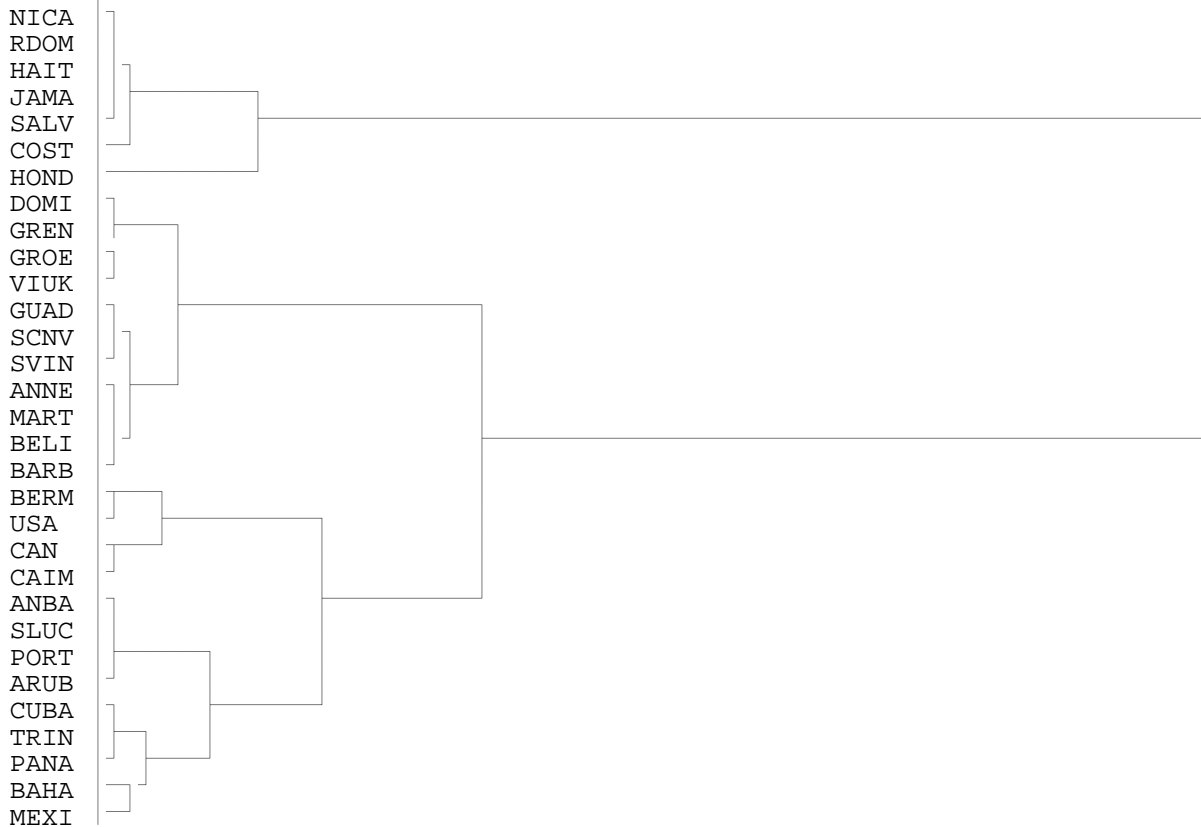
```

```
>HIERARCHY
```

```

Hierarchical clustering on WA:North and Central America
31 countries
Ward's method. 11 variables analysed.
coefficient minimum=347136.00 maximum=23263565800.00

```



Inspecting the tree will let you understand (hopefully) the group structure of the observations. The classification is then produced by cutting the tree at a meaningful level. This is done with the following command:

```
>TREE NCUT=3
```

```

Group sizes : 13 7 11
>>> GVAR stored.

```

It reports that a grouping variable has been stored, as well as the sizes of the groups.

The TREE command has a number of options for a closer inspection of the hierarchical tree.

```
>TREE DETAILS
```

level	who	with whom	last merges	next(left)	coefficient	
1	NICA	& RDOM		13	347136.000	
2	CUBA	& TRIN		4	791424.000	
3	DOMI	& GREN		19	1411609.000	
4	CUBA	& PANA	2	23	3210734.500	
5	GUAD	& SCNV		10	6200590.500	
6	ANBA	& SLUC		12	10531470.000	
7	GROE	& VIUK		19	15788294.000	
8	HAIT	& JAMA		13	21648134.000	
9	ANNE	& MART		11	28290502.000	
10	GUAD	& SVIN	5	22	36864604.000	
11	ANNE	& BELI	9	14	46209672.000	
12	ANBA	& PORT	6	18	58304924.000	
13	HAIT	& NICA	8	15	77628320.000	
14	ANNE	& BARB	11	22	99961968.000	
15	HAIT	& SALV	13	21	122852976.000	
16	BERM	& USA		24	158119664.000	
17	CAN	& CAIM		24	194446256.000	
18	ANBA	& ARUB	12	26	257019552.000	
19	DOMI	& GROE	3	7	25	331547968.000
20	BAHA	& MEXI		23	407870016.000	
21	COST	& HAIT	15	27	490495040.000	
22	ANNE	& GUAD	14	10	25	610837440.000
23	BAHA	& CUBA	20	4	26	811425152.000
24	BERM	& CAN	16	17	28	1052859200.000
25	ANNE	& DOMI	22	19	29	1487631740.000
26	ANBA	& BAHA	18	23	28	2031974140.000
27	COST	& HOND	21		30	3128473600.000
28	ANBA	& BERM	26	24	29	4478157310.000
29	ANBA	& ANNE	28	25	30	8024428030.000
30	ANBA	& COST	29	27		23263565800.000

The command sequence having produced the results above was

```
>HIERARCHY
>TREE NCUT=3
>TREE DETAILS
```

HIERARCHY produces the hierarchical tree and displays it. There is no variable list on the command line. This means for multivariate commands to take *all* variables in the WA, i.e. the behaviour is different from the other analysis commands.² If a variable list is present on the command line, only these variables will be clustered.

The TREE command is a special command: it can only be used immediately after a HIERARCHY or VHIERARCHY command. Immediately means that no other EDA command can be issued between the cluster analysis command and the TREE command.

The TREE command has additional options. The main commands however to analyse and interpret the result of a classification are all the commands you have learned to look at groups.³

At all times you can ask for a list of members...

```
>MEMBERS
```

```
GVAR is HIER (WARD) of World Data for North and Central America
Group 1 (1) has 13 members
ANBA ARUB BAHA BERM CAN CUBA USA CAIM MEXI PANA PORT SLUC TRIN
Group 2 (2) has 7 members
COST HAIT HOND JAMA NICA RDOM SALV
Group 3 (3) has 11 members
ANNE BELI DOMI GREN GROE GUAD VIUK BARB MART SCNV SVIN
```

2. Things are somewhat more complicated, as there is a special global switch called ALLVARS changing the default behaviour. If you SET ALLVARS OFF (default is ON), these commands behave like other commands, i.e. pick up the current list if no variable list is present on the command line.

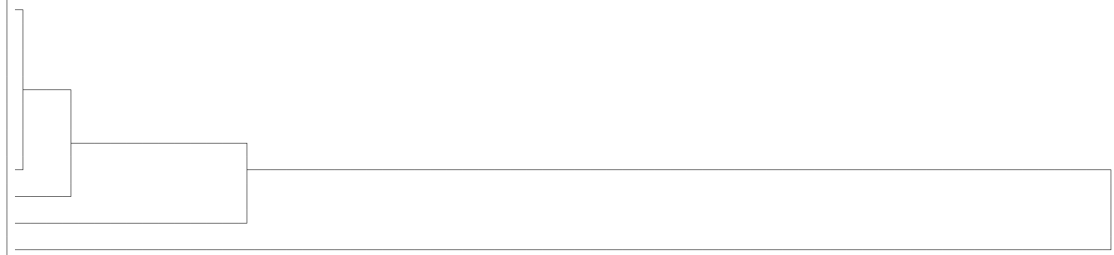
3. See the handout on *Analysis of groups (1)*.

Clustering variables

Here is a second example using the VHIERARCHY command, i.e. hierarchical clustering (by default) on variables. Classifications of variables produce groups or classes of variables, i.e. in EDA software terms it produces *ties* linking variables to groups. See below for a general explanation of ties.

```
>VHIERARCHY

31 countries
10x 10 distance MATRIX stored(N= 31)
Hierarchical clustering of variables on WA: World Data set N.&Cent. Am.
MATRIX analyzed:Distance p= 2 r= 2      D, # of variables 10
Single Linkage
coefficient minimum=71.63 maximum=49953.97
PGrow
GNPAgr
GNPInd
InfMor
Urb
GNPServ
Lit
PHBed
PopDoc
GNPCap
```



1.2 Non-hierarchical (agglomerative) clustering

EDA: CLUSTER command

- ◆ 1. Start by selecting the number of groups to define and select an equal number of cases as starting configuration (nuclei, centroids).
- ◆ 2. Each object is allocated to the nearest center.
- ◆ 3. Compute the new center for each group from the objects allocated to it.
- ◆ Repeat 2. and 3. until no objects change groups

We did not say

- ◆ How do we determine the number of groups?
- ◆ What is meant by “nearest” -> proximity/distance
- ◆ How to “compute the new center (centroid)”

but again, there a a nearly unlimited number of possible criteria...

```
>CLUSTER4
```

4. This is a screen shot. It shows the dialogue asking the user to enter a starting configuration.

```

31 countries
Non hierarchical clustering (casewise)
Forgy's method. 10 variables included.
Initial configuration:
Non hierarchical clustering (casewise)
Forgy's method. 10 variables included.
Enter up to 8 case-ids as starting configuration.
c:NICA USA TRIN
c:

```

```

Casids entered:NICA USA TRIN
31 countrie s moved iteration# 1, summed deviations= 76082.84
7 countrie s moved iteration# 2, summed deviations= 91762.08
3 countrie s moved iteration# 3, summed deviations= 77516.80
2 countrie s moved iteration# 4, summed deviations= 83412.04
0 countrie s moved iteration# 5, summed deviations= 78729.28

```

Membership list:

Cluster: 1 (21 members)

ANBA ANNE BELI COST CUBA DOMI GREN GUAD HAIT HOND JAMA MART MEXI NICA PANA
RDOM SCNV SLUC SVIN SALV TRIN

Cluster: 2 (4 members)

BERM CAN USA CAIM

Cluster: 3 (6 members)

ARUB BAHA GROE VIUK BARB PORT

1.3 Interpretation

```
>GSUMMARY 1-5
```

```
>GANAL #0
```

```

31 countries
Groups defined by GVAR:HIER (WARD) of Example data set1 (all countries)
Coded group differences.

```

Each symbol corresponds to 1/5 midspread deviation from the grand median.

Variable	Group 1	Group 2	Group 3	Overall
PGrow		++++	-	1.20 : Population Growth
Urb	++	--	-	57.00 : Urbanization
InfMor		\$++++++	+	18.00 : Infant Mortality
PopHos		+++++	--	30600.00 : Population/Hospital
PHBed		\$++++++	--	264.00 : Population/Hosp.Bed
PopDoc	-	++++		1197.00 : Population/Physician
GNPCap	+++	-	+	2762.00 : GNP per capita
GNPAgr	-	+++		9.00 : %GNP for Agriculture
GNPInd		+	-	22.00 : %GNP for Industry
GNPServ	+	--		64.00 : %GNP for Services
Lit		\$-----	+	94.00 : Literacy Rate
Group N	13	7	11	

```
>TRACE #GNPCap
```


Ties can be used in variable lists, as shown in the following example

```
>BOXPLOTS #3 PARALLEL
```

shows a parallel boxplot with all variables in group #3. This is an additional use of the # symbol in variable lists; note that here #3 means list number 3, in expressions however #3 means variable number 3 (the distinction is clear from the context; it is not possible to specify variable lists within expressions).

Table des Matières

1 The Analysis of groups (2): Cluster Analysis ⁵	1
1.1 Hierarchical clustering	1
Clustering variables	3
1.2 Non-hierarchical (agglomerative) clustering	4
1.3 Interpretation	5
1.4 Groups of variables	6