Completely elaborated example of data order instability in a Between-group Average hierarchical cluster analysis

Willem A. van der Kloot
Department of Psychology
Leiden University
The Netherlands

November 15, 2005

Given the matrix **D** of distances between six objects, *A* through *F*, a hierarchical agglomerative cluster analysis using the Between-group average[1] method (Baverage; also known as UPGMA) is performed.

D, matrix of distances between six objects, *A* through *F*

	A	*B*	*C*	*D*	*E*	*F*
A	0	25	17	115	102	64
B	25	0	19	90	76	38
C	17	19	0	101	89	48
D	115	90	101	0	22	51
E	102	76	89	22	0	49
F	64	38	48	51	49	0

Note. Matrix reproduced with kind permission of Springer Science and Business Media from Table 2, page 571, of Van der Kloot, W. A., Bouwmeester, S., & Heiser, W.J. (2003). Cluster instability as a result of data input order. In H. Yanai, A. Okada, K. Shigemasu, Y. Kano, & J. J. Meulman (Eds), *New developments in Psychometrics: Proceedings of the International Meeting of the Psychometric Society IMPS2001* (pp. 569-576). Tokyo: Springer.

Step 1. Find the smallest distance in **D**. This is $d_{AC} = 17$, the distance between Objects *A* and *C*. Merge Objects *A* and *C* into Cluster $\{A,C\}$. Compute the distances between Cluster $\{A,C\}$ and the remaining objects by averaging the distances between each object outside and the two objects in the cluster: $d_{B\{A,C\}} = (d_{AB} + d_{BC})/2 = (25 + 19)/2 = 22$; $d_{D\{A,C\}} = (d_{AD} + d_{CD})/2 = (115 + 101)/2 = 108$; $d_{E\{A,C\}} = (d_{AE} + d_{CE})/2 = (102 + 89)/2 = 95.5$; $d_{F\{A,C\}} = (d_{AF} + d_{CF})/2 = (64 + 48)/2 = 56$. Update distance matrix **D** into $\mathbf{D}_{(1)}$.

$\mathbf{D}_{(1)}$, updated matrix of distances between
one cluster and five remaining objects after Step 1 in the agglomeration process.

	$\{A,C\}$	*B*	*D*	*E*	*F*
$\{A,C\}$	0	22	108	95.5	56
B	22	0	90	76	38
D	108	90	0	22	51
E	95.5	76	22	0	49
F	56	38	51	49	0

Step 2. Find the smallest distance in $\mathbf{D}_{(1)}$. There are two such values: $d_{B\{A,C\}} = 22$ and $d_{DE} = 22$. Therefore, there are two options to form a new cluster:

Option 1: Merge *D* and *E*.

Compute distances as above: $d_{\{D,E\}\{A,,C\}} = (d_{AD} + d_{CD} + d_{AE} + d_{CE})/4 = (115 + 101 + 102 + 89)/4 = 407/4 = 101.75$, $d_{\{D,E\}B} = (d_{BD} + d_{BE})/2 = (90 + 76)/2 = 83$, and $d_{\{D,E\}F} = (d_{DF} + d_{EF})/2 = (51 + 49)/2 = 50$. Update $\mathbf{D}_{(1)}$ into $\mathbf{D}_{(2,\text{ Option 1})}$.

[1] Cf. Everitt, B.S. (1993). *Cluster analysis. Third Edition*. London: Arnold.

D$_{(2,\text{Option1})}$, updated matrix of distances between two clusters and two objects after Step 2 in the agglomeration process after choosing Option 2 of breaking two ties

	$\{A,C\}$	B	$\{D,E\}$	F
$\{A,C\}$	0	22	101.75	56
B	22	0	83	38
$\{D,E\}$	101.75	83	0	50
F	56	38	50	0

Step 3.2. Find the smallest distance in **D**$_{(2,\text{Option 1})}$. This is $d_{B\{A,C\}} = 22$. Merge Object B with Cluster $\{A,C\}$. Compute $d_{\{D,E\}\{A,B,C\}} = (d_{AD} + d_{BD} + d_{CD} + d_{AE} + d_{BE} + d_{CE})/6 = (115 + 90 + 101 + 102 + 76 + 89)/6 = 573/6 = 95.5$ and $d_{\{A,B,C\}F} = (d_{AF} + d_{BF} + d_{CF})/3 = (64 + 38 + 48)/3 = 50$. Update **D**$_{(2,\text{Option 1})}$ into **D**$_{(3,\text{Option 1})}$.

D$_{(3,\text{Option 1})}$, updated matrix of distances between two clusters and one object after Step 2 in the agglomeration process after choosing Option 1 of breaking two ties

	$\{A,B,C\}$	$\{D,E\}$	F
$\{A,B,C\}$	0	95.5	50
$\{D,E\}$	95.5	0	50
F	50	50	0

Step 4.2. Find the smallest distance in **D**$_{(3,\text{ Option 1})}$.

There are two smallest distances: $d_{F\{A,B,C\}} = d_{F\{D,E\}} = 50$ and therefore again two options:

Option 1.A: Merge Object F with Cluster $\{A,B,C\}$.

Now we have two clusters left: $\{A,B,C,F\}$ and $\{D,E\}$. The distance between those two clusters is

$$d_{\{A,B,C,F\}\{D,E\}} = (n_{\{A,B,C\}}\, n_{\{D,E\}} d_{\{A,B,C\}\{D,E\}} + n_F\, n_{\{D,E\}} d_{F\{D,E\}})/(n_{\{A,B,C\}}\, n_{\{D,E\}} + n_F\, n_{\{D,E\}})$$
$$= (3 \times 2 \times 95.5 + 1 \times 2 \times 50)/(3 \times 2 + 1 \times 2) = 673/8 = 84.125$$

(this is an equivalent and simpler formula that only needs the latest updated matrix **D** and the number of elements in each of the latest clusters).

The next and last fusion necessarily merges all objects into one cluster, with fusion coefficient 84.125.

Option 1.B: Merge Object F with Cluster $\{D,E\}$.

Again, we have two clusters left: $\{A,B,C\}$ and $\{D,E,F\}$.

$$d_{\{A,B,C\}\{D,E,F\}} = (n_{\{A,B,C\}}\, n_{\{D,E\}} d_{\{A,B,C\}\{D,E\}} + n_{\{A,B,C\}}\, n_F\, d_{\{A,B,C\}F\}})/(n_{\{A,B,C\}}\, n_{\{D,E\}} + n_{\{A,B,C\}} n_F)$$
$$= (3 \times 2 \times 95.5 + 3 \times 1 \times 50)/(3 \times 2 + 3 \times 1) = 723/9 = 80.333\ .$$

The next and last fusion necessarily merges all objects into one cluster, with fusion coefficient 80.333.

Option 2: Merge *B* with *A* and *C*.

Compute the distances between $\{A,B,C\}$ and the remaining objects and clusters:

$d_{D\{A,B,C\}} = (d_{AD} + d_{BD} + d_{CD}/3 = (115 + 90 + 101)/3 = 102$; $d_{E\{A,B,C\}} = (d_{AE} + d_{BE} + d_{CE})/3 = (102 + 76 + 89)/3 = 89$; $d_{F\{A,B,C\}} = (d_{AF} + d_{BF} + d_{CF})/3 = (64 + 38 + 48)/3 = 50$.

Update $\mathbf{D}_{(1)}$ into $\mathbf{D}_{(2,\text{Option 2})}$.

$\mathbf{D}_{(2,\text{Option 2})}$, updated matrix of distances between one cluster and three objects after Step 2 in the agglomeration process after choosing Option 1 of breaking two ties

	$\{A,B,C\}$	D	E	F
$\{A,B,C\}$	0	102	89	50
D	102	0	22	51
E	89	22	0	49
F	50	51	49	0

Step 3.1. Find the smallest distance in $\mathbf{D}_{(2,\text{Option 2})}$. This is $d_{DE} = 22$. Merge Objects D and E.

Compute the distances between {D,E} and the remaining objects and clusters:

$d_{\{D,E\}\{A,B,C\}} = (d_{AD} + d_{BD} + d_{CD} + d_{AE} + d_{BE} + d_{CE})/6 = (115 + 90 + 101 + 102 + 76 + 89)/6 = 573/6 = 95.5$ and $d_{F\{D,E\}} = (d_{DF} + d_{EF})/2 = (51 + 49)/2 = 50$. Update $\mathbf{D}_{(2,\text{Option 2})}$ into $\mathbf{D}_{(3,\text{Option 2})}$.

$\mathbf{D}_{(3,\text{Option 2})}$, updated matrix of distances between two clusters and one object after Step 2 in the agglomeration process after choosing Option 1 of breaking two ties

	$\{A,B,C\}$	$\{D,E\}$	F
$\{A,B,C\}$	0	95.5	50
{D,E}	95.5	0	50
F	50	50	0

Step 4.1. Find the smallest distance in $\mathbf{D}_{(3,\text{Option 2})}$. There are two smallest distances: $d_{F\{A,B,C\}} = d_{F\{D,E\}} = 50$. Therefore, there are again two options:

Option 2.A: Merge Object *F* with Cluster *A,B,C*}.

We have the same clusters left as in Option 1.A: $\{A,B,C,F\}$ and $\{D,E\}$.

$d_{\{A,B,C,F\}\{D,E\}} = 84.125$. The next and last fusion necessarily merges all objects into one cluster, with fusion coefficient 84.125.

Option 2.B: Merge Object F with Cluster *D,E*}.

We have have the same clusters left as in Option 1.B: $\{A,B,C\}$ and $\{D,E,F\}$.

$d_{\{A,B,C\}\{D,E,F\}} = 80.333$. The next and last fusion merges all objects into one cluster, with fusion coefficient 80.333.

The analysis of **D** thus may yield four different solutions. Which option is chosen, depends on the order in which the rows and columns of **D** are read into the computer program. Figure 1 shows the four trees that were obtained by SPSS after permuting the rows and columns of **D**. The corresponding agglomeration schedules are listed in Table 1. The SPSS data matrix and the SPSS syntax are listed in Appendix A.

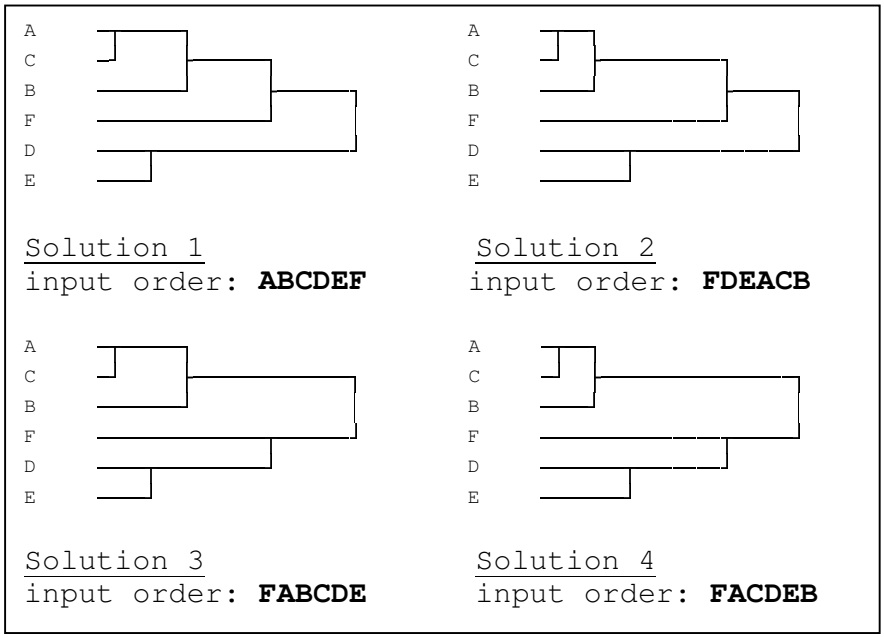


Figure 1. Four dendrograms of the Objects *A* through *F* obtained by SPSS' Baverage analysis of the distance matrix **D** in four different input orders. The connections between the horizontal lines indicate the stages at which the clusters were formed; usually, they indicate the values of the fusion coefficients.

Table 1.
Agglomeration Schedules Corresponding to the Dendrograms of Figure 1.

1. Input Order **ABCDEF**

Stage	Statistics					
	Cluster Combined: Cluster 1	Cluster Combined: Cluster 2	Coefficients	Stage Cluster First Appears: Cluster 1	Stage Cluster First Appears: Cluster 2	Next Stage
1	1 (A)	3 (C)	17.000	0	0	3
2	4(D)	5 (E)	22.000	0	0	5
3	1 (A,C)	2 (B)	22.000	1	0	4
4	1 (A,B,C)	6 (F)	50.000	3	0	5
5	1 (A,B,C,F)	4 (D,E)	84.125	4	2	0

2. Input Order **FDEACB**

Stage	Statistics					
	Cluster Combined: Cluster 1	Cluster Combined: Cluster 2	Coefficients	Stage Cluster First Appears: Cluster 1	Stage Cluster First Appears: Cluster 2	Next Stage
1	1 (A)	3 (C)	17.000	0	0	2
2	1 (A,C)	2 (B)	22.000	1	0	4
3	4 (D))	5 (E)	22.000	0	0	5
4	1(A,B,C)	6 (F)	50.000	2	0	5
5	1(A,B,C,F)	4 (D,E)	84.125	4	3	0

3. Input Order **FABCDE**

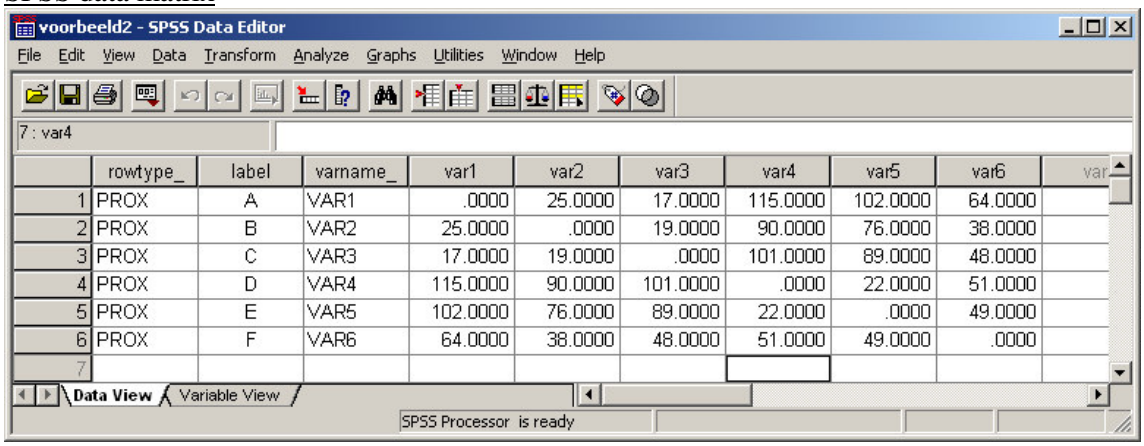
Stage	Statistics					
	Cluster Combined: Cluster 1	Cluster Combined: Cluster 2	Coefficients	Stage Cluster First Appears: Cluster 1	Stage Cluster First Appears: Cluster 2	Next Stage
1	1 (A)	3 (C)	17.000	0	0	3
2	4 (D)	5 (E)	22.000	1	0	4
3	1 (A,C)	2 (B)	22.000	1	0	5
4	4 (D,E)	6 (F)	50.000	2	0	5
5	1 (A,B,C)	4 (D,E,F)	80.330	3	4	0

4. Input order **FACDEB**

Stage	Statistics					
	Cluster Combined: Cluster 1	Cluster Combined: Cluster 2	Coefficients	Stage Cluster First Appears: Cluster 1	Stage Cluster First Appears: Cluster 2	Next Stage
1	1 (A)	3 (C)	17.000	0	0	3
2	1 (A,C)	2 (B)	22.000	1	0	4
3	4 (D)	5 (E)	22.000	0	0	5
4	4 (D,E)	6 (F)	50.000	3	0	5
5	1 (A,B,C)	4 (D,E,F)	80.330	2	4	0

Appendix A

SPSS data matrix



	rowtype_	label	varname_	var1	var2	var3	var4	var5	var6
1	PROX	A	VAR1	.0000	25.0000	17.0000	115.0000	102.0000	64.0000
2	PROX	B	VAR2	25.0000	.0000	19.0000	90.0000	76.0000	38.0000
3	PROX	C	VAR3	17.0000	19.0000	.0000	101.0000	89.0000	48.0000
4	PROX	D	VAR4	115.0000	90.0000	101.0000	.0000	22.0000	51.0000
5	PROX	E	VAR5	102.0000	76.0000	89.0000	22.0000	.0000	49.0000
6	PROX	F	VAR6	64.0000	38.0000	48.0000	51.0000	49.0000	.0000

SPSS syntax

```
* input order ABCDEF.
CLUSTER  var1 var2 var3 var4 var5 var6
  /METHOD BAVERAGE
   /ID=label
  /PRINT SCHEDULE
  /PLOT DENDROGRAM
 /matrix=in(*) .

* input order FDEACB.
CLUSTER  var6 var4 var5 var1 var3 var2
  /METHOD BAVERAGE
  /ID=label
  /PRINT SCHEDULE
  /PLOT DENDROGRAM
 /matrix=in(*) .

* input order FABCDE .
CLUSTER  var6 var1 var2 var3 var4 var5
  /METHOD BAVERAGE
  /ID=label
  /PRINT SCHEDULE
  /PLOT DENDROGRAM
 /matrix=in(*) .

* input order FACDEB .
CLUSTER  var6 var1 var3 var4 var5 var2
  /METHOD BAVERAGE
   /ID=label
  /PRINT SCHEDULE
  /PLOT DENDROGRAM
 /matrix=in(*) .
```