Generalized Ascending Hierarchical Classification

By Dr. Karapistolis Dimitrios

Do not walk on a predetermined path, it will lead you where others have already gone.

Pythagoras

7.0 Generally

The proposed classification method by Dr. D. Karapistolis, named «Generalized Ascending Hierarchical Classification (GAHC)», differs from the VACOR method, which is based on the concept of «**nearest neighbor»** by J.P. Benzecri. In the process of merging two classes of "subjects," it utilizes the Ward method, with the criterion of minimal loss of ordinal inertia and the metric "X^2 distance" introduced by J.P. Benzecri.

The difference lies in the GAHC process proposed by Dr. Dimitrios Karapistolis, which uses the term «**Nearest Kinsman»** to create the classification of "subjects" regarding the variables they are most closely associated with. In the process of merging two classes of variables, the WARD method is used with the criterion of minimum variance within the classes, based on the Lance-Williams algorithm using Euclidean metric on the coordinates matrix of variables on the p-1 axes resulting from the application of Correspondence Factor Analysis on the elements of the data matrix for classification.

The purpose of the proposed method is to classify the variables, while simultaneously classifying the «subjects» within the classes of the formed typology of variables.

Through this process, the simultaneous presence of the ascending hierarchical classification of variables and the classification of "subjects" within the formed classes is visualized with the dendrogram. These classes contain "subjects" with homogeneous values, and their coherence is based on the KARAP method.

Regarding the classification process of variables using the VACOR method, it can be carried out after transforming the data matrix into a logical 0-1 matrix to create the corresponding BURT matrix. This BURT matrix will then be subjected to the process of classifying the gradients of the variables.

7.1 The step of GAHC

Step 1 : The data matrix T(n,p) is analyzed by Factor Analysis of Correspondences

Step 2: Using the KARAP method, the "nearest relative" of each «subject» is identified based on the metric. X^2 .

Step 3: The extraction of the file containing the p-1 coordinates of the GA variables is performed in .txt format. It will be used for constructing the classification of variables based on the Euclidean metric (due to the orthogonal system they create) and the Ward method for minimum variance within classes (Anglo-Saxon cluster process), instead of the process of minimum loss of ordinal inertia (French process using the X metric x^2)

Step 4: Utilizing the ranking of "subjects" in combination with the classification of variables, a complete dendrogram is obtained. This dendrogram presents the simultaneous presence of the connection between "subjects" and the representative variables they are most associated with, creating homogeneous classes of "subjects" in terms of their values.

7.2 Application of GAHC

With the following numerical example, we will present the utility of the KARAP method in the GAHC classification method.

Coded array of 3 variables

INT	A	B	C
I1	1	4	3
I2	2	2	1
I3	1	4	2
I4	2	2	3
I5	1	1	4
I6	4	3	4
I7	2	4	3
I8	1	4	2
I9	3	2	4
I10	4	1	4

Logical 0-1 table of 3 variables

Ο λογικός πίνακας 0-1 των 3 μεταβλητών

INT	A1	A2	A3	A4	B1	B2	B3	B4	C1	C2	C3	C4
I1	1	0	0	0	0	0	0	1	0	0	1	0
I2	0	1	0	0	0	1	0	0	1	0	0	0
I3	1	0	0	0	0	0	0	1	0	1	0	0
I4	0	1	0	0	0	1	0	0	0	0	1	0
I5	1	0	0	0	1	0	0	0	0	0	0	1
I6	0	0	0	1	0	0	1	0	0	0	0	1
I7	0	1	0	0	0	0	0	1	0	0	1	0
I8	1	0	0	0	0	0	0	1	0	1	0	0
I9	0	0	1	0	0	1	0	0	0	0	0	1
I10	0	0	0	1	1	0	0	0	0	0	0	1

Initially, the logical 0-1 table is analyzed using the KARAP method. The characteristic roots of the table are presented in Table 7.1

Table 7.1 The characteristic roots

1	0,8583072	28,61	28,61	|**********
2	0,7711879	25,71	54,32	|**********
3	0,4511810	15,04	69,36	|**********
4	0,3538555	11,80	81,15	|**********
5	0,3426094	11,42	92,57	|**********
6	0,1067500	3,56	96,13	|*****
7	0,0750680	2,50	98,63	|****
8	0,0410410	1,37	100,00	|**
9	0,0000028	0,00	100,00	|*
10	0,0000001	0,00	100,00	|*
11	0,0000000	0,00	100,00	|*

The table 7.2 presents the coordinates GA of the 12 graduations of the 3 variables, which must be converted into a .TXT file due to the R language code used in the program for constructing the dendrogram at the location C:\MAD\TXT

Table 7.2

B17

	A	B	C	D	E	F	G	H	I	J	K	L
1		GA1	GA2	GA3	GA4	GA5	GA6	GA7	GA8	GA9	GA10	GA11
2	A 1	391	-973	277	234	111	-470	-165	-103	-1	0	0
3	A 2	771	1048	-620	184	-84	276	-128	346	2	0	0
4	A 3	-558	1039	2390	-1169	-539	23	471	128	1	-2	0
5	A 4	-1662	-149	-822	-161	171	512	285	-380	-4	0	0
6	B 1	-1137	-395	158	1534	-385	107	-70	104	0	0	0
7	B 2	276	1323	517	-130	137	78	-359	-263	-3	-1	0
8	B 3	-1788	-82	-1396	-1669	698	-615	-360	261	2	0	-1
9	B 4	807	-775	-119	-253	-86	41	393	79	0	-1	0
10	C 1	665	1805	-290	891	1934	-522	627	-96	-1	1	0
11	C 2	761	-1330	451	-302	935	629	-295	26	0	0	0
12	C 3	809	228	-642	-173	-1066	-120	-4	-186	-2	-1	-1
13	C 4	-1155	42	328	57	-153	-95	-7	149	1	-1	0

The linking of variables and «subjects» based on the «**nearest kinsman»** follows. Table 7.3 describes the connections.

Table 7.3: The linkage of variables and "subjects" based on the KARAP method.

A1	A2	A3	A4	B1	B2	B3	B4	C1	C2	C3	C4
	I4	I9		I5		I6	I1	I2	I3	I7	
				I10					I8		

From table 7.3, it is observed that four variables (A1, A4, B2,C4) do NOT constitute the «n**earest kinsman»** of any «subject»

In Figure 7.1, the representation at the factorial plan 1x2 of the formed dependencies of the «subjects» based on their «**nearest kinsman»** is presented visually.

Also in Figure 7.1, continuous contours describe the groupings based on the analysis of the logical table 0-1 using the KARAP method.

The KARAP method provides 100% of the information by assigning the primary linkage of «subjects» with variables that are more strongly connected, leaving out

some variables that do not play as significant a role in the interactions of the «subjects», such as variables A1, A4, B2, and C4 (see Table 7.2).

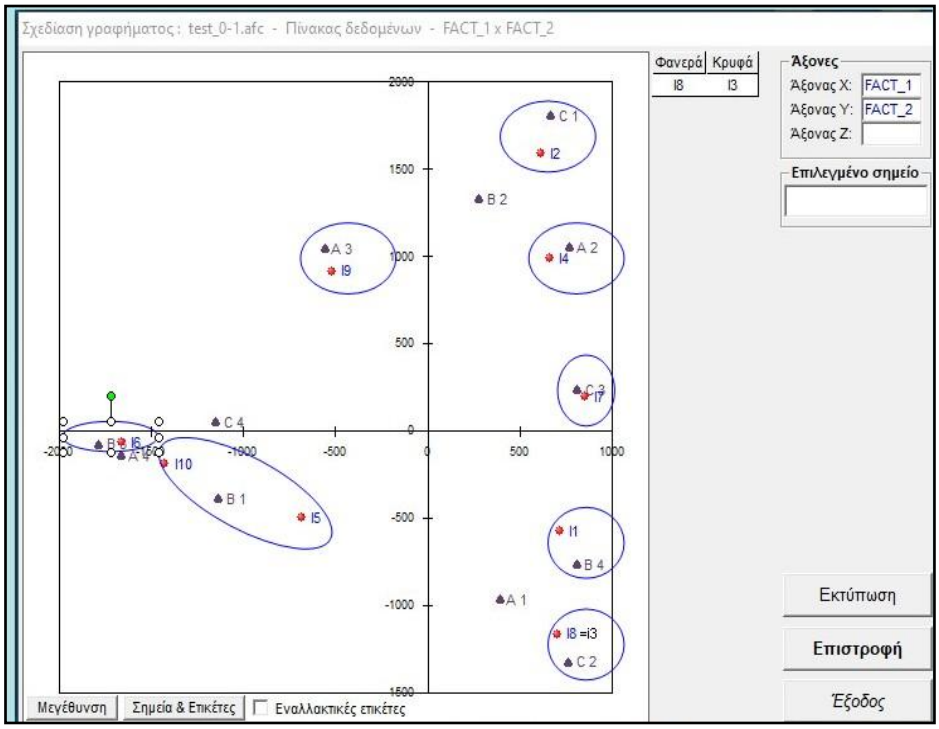


Figure 7.1

In Figure 7.2, with dashed contours at the factorial plan 1x2, the 4 classes formed by the 12 graduations of the 3 variables A, B, and C are described.

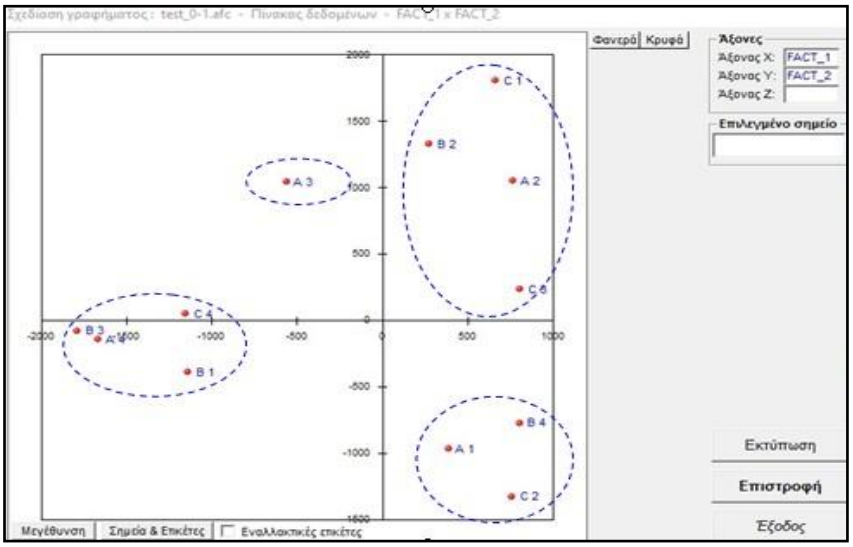


Figure 7.2

Applying «Generalized Ascending Hierarchical Classification (GAHC)», we obtain the dendrogram representing the hierarchy of the variables.

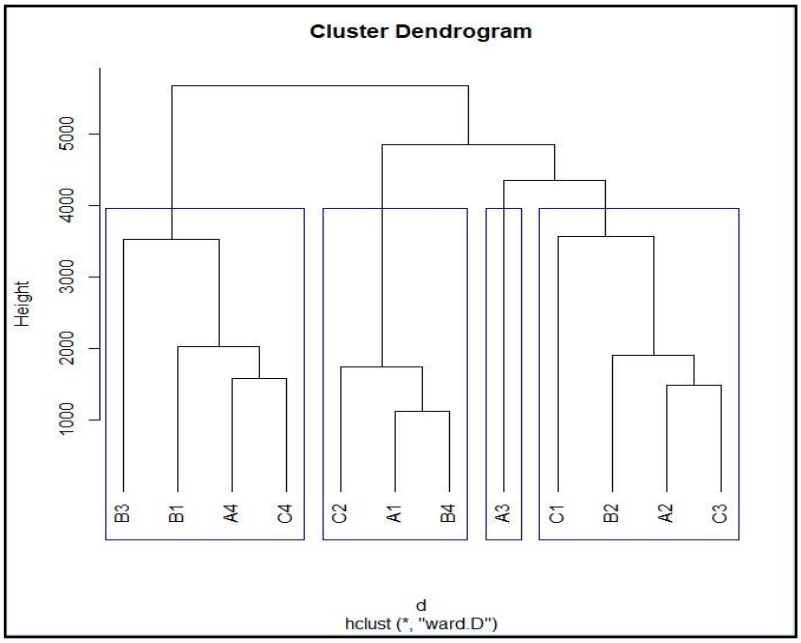


Figure 7.3

Other classification diagrams offered by the MAD program include the following (with «subjects» of each class colored differently as shown in the upper-left table):

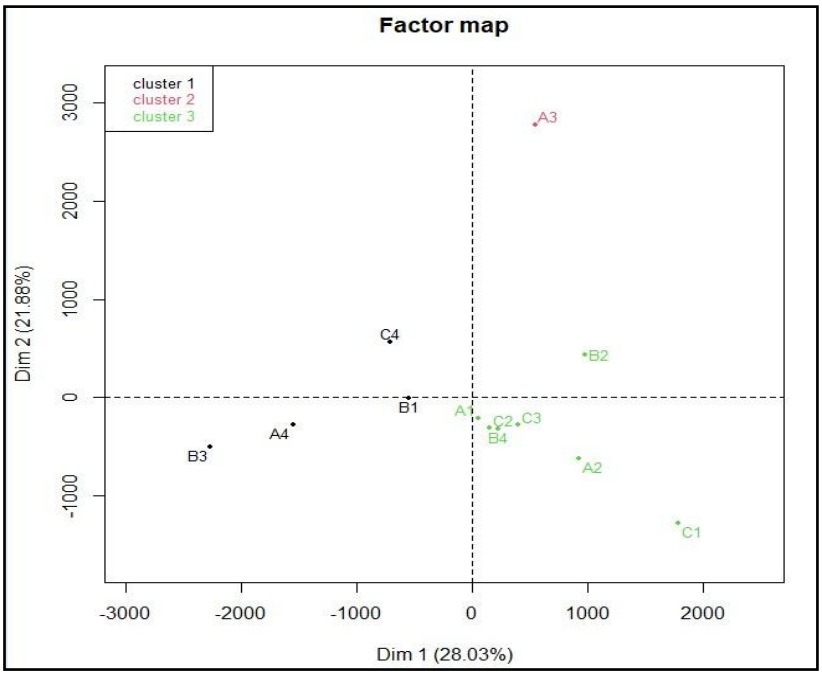


Figure 7.4

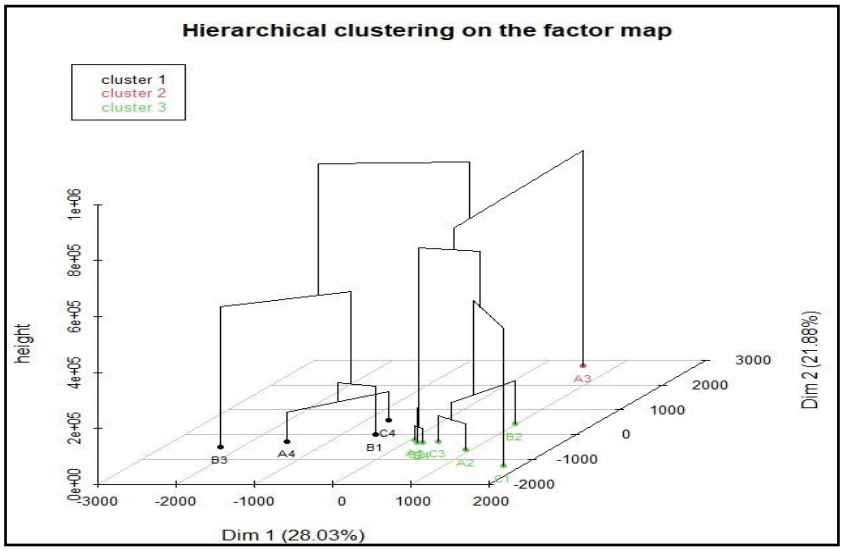
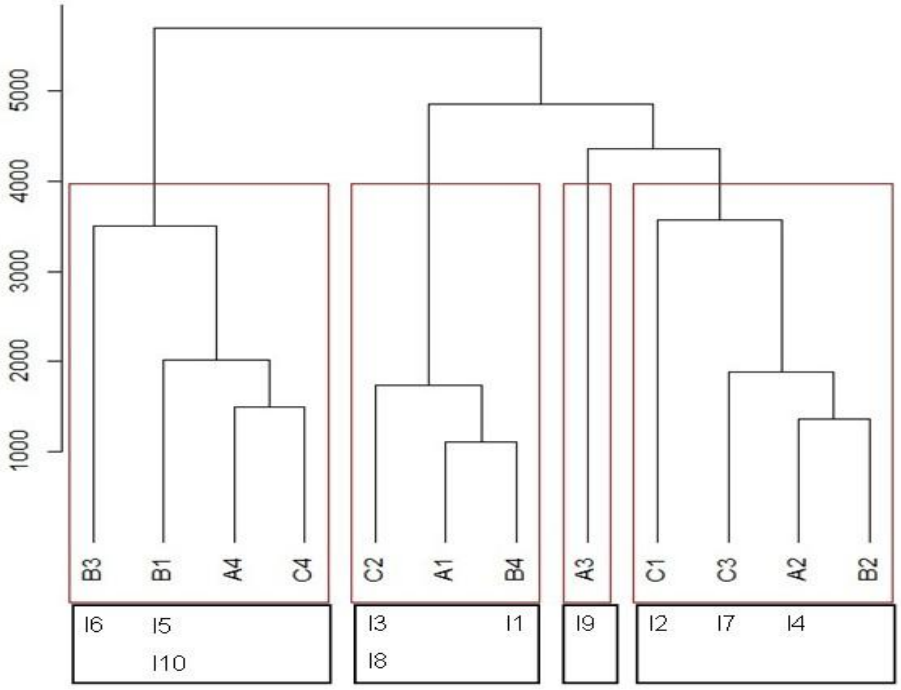


Figure 7.5

7.3 Final result of the method GAHC

By combining the data from Table 7.1 with that of Figure 7.3, we obtain a complete image of the classification between «subjects» and their «nearest kinsman» in each of the four classes of the ascending hierarchy of variables (Figure 7.6).



Διάγραμμα 7.6

7.3 Conclusions

Finally, based on the dendrogram, we have achieved the visualization of the simultaneous presence of the ascending hierarchical classification of the variables into four classes and the classification of the «subjects» within the formed four classes, which contain «subjects» with homogeneous values, the coherence of which emerged using the KARAP method..