Application of statistics in engineering

Snezhana G. Gocheva-Ilieva
University of Plovdiv Bulgaria, *snow@uni-plovdiv.bg*

7. Application of multivariate cluster analysis (CA)

7.1. Introduction

Cluster analysis is used to group observations (experiments) into clusters by some measure of similarity. It is also possible to group variables as in factor analysis, but in CA this is based on a selected similarity criterion. The aim is to find an optimal grouping so that the observations within each cluster are similar, but the clusters are dissimilar to each other.

Usually, in CA (hierarchical CA) the number of groups or the groups are unknown and have to be determined by the researcher.

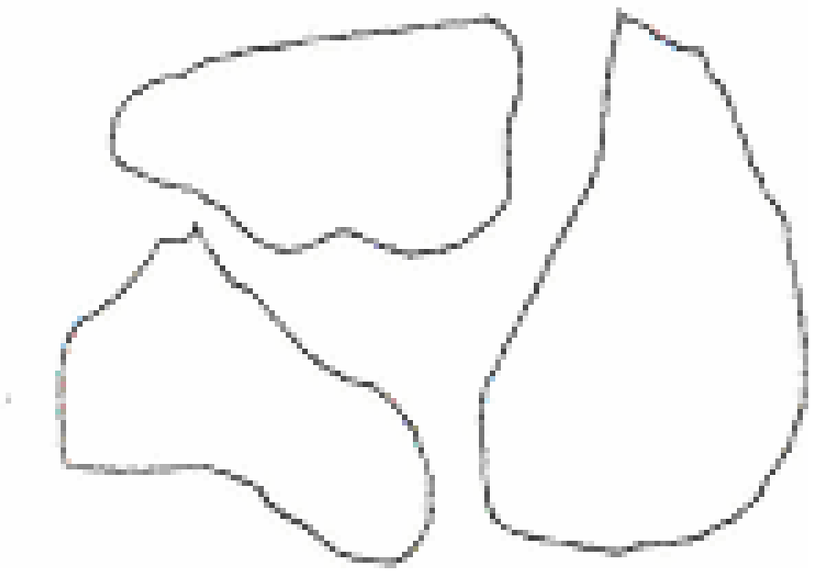
As a rule, the similarity is considered to be some measure of distance between all pairs of cases (observations, experiments or objects). The techniques of cluster analysis is widely used to data in many fields, such as

medicine, psychiatry, sociology, archaeology, market research, economics, and engineering.

It must be differentiate two common clustering approaches: ***hierarchical clustering*** and ***partitioning*** (or non-hierarchical clustering). If our data consist of *n* observations in hierarchical clustering one starts with *n* clusters, one for each observation, and end with a single cluster containing all *n* observations. At each step, an observation or a cluster of observations is grouped with another cluster. It is also possible to reverse this process, that is, start with a single cluster containing all *n* observations and end with *n* clusters of a single item each. In partitioning (K-means cluster analysis), we simply divide the observations into *k* clusters. This can be obtained by starting with an initial partitioning or with cluster centers and then rearranging the observations according to some distance measure.

As a rule hierarchical clustering is used when you have a relatively small number of observations and variables (50, 100 or 200). In the case of a huge amount of data the partitioning methods are more appropriate.

Hierarchical clustering



Non-hierarchical clustering

7.2. General assumptions in CA

- The data used in cluster analysis can be interval, ordinal or categorical. However, having a mixture of different types of variable you need to have some way of measuring the distance between observations and the type of measure used will depend on what type of data you have. In K-means CA the data can be quantitative (interval or at ratio level).
- Different measures are used to measure ’distance’ for binary and categorical data. For interval data, mostly arising in physics and engineering, the most common distance measure used is the Euclidean distance.
- If some variables have very large interval of data, they should be scaling, by standardized them in z-variables (with mean=0 and standard deviation=1), or by some other type of scaling.
- **Warning:** CA has no mechanism for differentiating between relevant and irrelevant variables. If you have chosen an inappropriate number of clusters or omitted important variables, your results may be misleading. This

is very important because the clusters formed can be very dependent in the variables included.

7.3. SPSS statistics

- For hierarchical CA: Agglomeration schedule, proximity (or similarity) matrix, and cluster membership for a single solution or a range of solutions. Plots: dendrograms.
- For K-means CA: initial cluster centers, ANOVA table, distance from cluster center.

7.4. Measures of similarity and dissimilarity

There exist many techniques to present the similarity or proximity between each pair of observations. In CA a convenient measure of proximity is the distance between two observations (objects). Actually, a small distance is considered as similarity and large distance is a measure of dissimilarity.

We shall consider quantitative variables.

Let the data matrix is presented in the form:

$$\mathbf{X} = \begin{pmatrix} x_1' \\ x_2' \\ \vdots \\ x_n' \end{pmatrix} = \left(x_{(1)}, x_{(2)}, \ldots, x_{(p)} \right),$$

where x_i' is a row (observation vector) and $x_{(j)}$ is a column (corresponding to a j -variable). We generally wish to group the n x_i' (rows) into k clusters. We may also wish to cluster the variables (columns) $x_{(j)}$, $j = 1, 2, \ldots, p$.

The usual distance measures for interval data are:

1) Euclidean distance between two vectors $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})'$ and $x_j = (x_{j1}, x_{j2}, \ldots, x_{jp})'$:

$$d(x_i, x_j) = \sqrt{\sum_{m=1}^{p} (x_{im} - x_{jm})^2}, \qquad i, j = 1, ..., n \; . \tag{1}$$

2) Squared Euclidean distance

$$d(x_i, x_j) = \sum_{m=1}^{p} (x_{im} - x_{jm})^2, \qquad i, j = 1, ..., n \tag{2}$$

3) Chebyshev distance

$$d(x_i, x_j) = \max \left| x_{im} - x_{jm} \right|, \qquad i, j = 1, ..., n \tag{3}$$

4) Minkowski metric

$$d(x_i, x_j) = \left(\sum_{m=1}^{p} \left| x_{im} - x_{jm} \right|^r \right)^{\frac{1}{r}}, \qquad i, j = 1, \ldots, n \,. \tag{4}$$

For the *n* observation vectors $x_1, x_2, \ldots, x_n$ we can compute an $n \times n$ matrix $D = d(x_i, x_j)$ of distances (or dissimilarities). The matrix D typically is symmetric with diagonal elements equal to zero.

For example, suppose three items (points, observations) have the following bivariate measurements (x_1, x_2): (2, 5), (4, 2), (7, 9). Here p=2, n=3. Using the definition of Euclidean distance we calculate: $d_{12} = \sqrt{(2-4)^2 + (5-2)^2} = \sqrt{13} \approx 3,6$, $d_{13} = \sqrt{(2-7)^2 + (5-9)^2} = \sqrt{41} \approx 6,4$ and $d_{23} = \sqrt{(4-7)^2 + (2-9)^2} = \sqrt{58} \approx 7,6$. This way the distance matrix D is

$$D_1 = d(x_1, x_2) = \begin{pmatrix} 0 & 3,6 & 6,4 \\ 3,6 & 0 & 7,6 \\ 6,4 & 7,6 & 0 \end{pmatrix}.$$

However, if we multiply the first variable x_1 by 100 as, for example, in changing from meters to centimeters, the matrix becomes

$$D_2 = d(x_1, x_2) = \begin{pmatrix} 0 & 200 & 500 \\ 200 & 0 & 300 \\ 500 & 300 & 0 \end{pmatrix}$$

and the largest distance is now d_{13} instead of d_{23}. The distance rankings depends on scaling measures.

This problem can be solved by a previous z-standardization, or other appropriate scaling of the variables.

7.5. Methods for combing clusters in an hierarchical CA

In an agglomerative hierarchical procedure the two closest clusters are combined at each step on the basis of some measure of similarity or dissimilarity, usually some kind of distance. At each step the number of clusters is therefore reduced by 1. After two clusters are merged, the procedure is repeated for the next step: the distances between all pairs of clusters are calculated again, and the pair with minimum distance is merged into a single cluster.

The results of a hierarchical clustering procedure can be displayed graphically using a tree diagram, also known as a dendrogram, which shows all the steps in the hierarchical procedure, including the distances at which clusters are merged.

Different approaches to measure distance between clusters exist in hierarchical methods. We will consider the most popular of them:

1) Between-groups linkage (*average linkage*)

In this method the distance between two clusters *A* and *B* is defined as the average of the $n_A.n_B$ distances between the n_A points in *A* and the n_B points in *B*:

$$D(A,B) = \frac{1}{n_A.n_B} \sum_{i=1}^{n_A} \sum_{j=1}^{n_B} d(x_i, x_j) \qquad (5)$$

where the sum is over all x_i in A and all x_j in B. Here $d(x_i, x_j)$ is the Euclidean distance or some other distance between the vectors x_i and x_j . At each step, we join the two clusters with the smallest distance.

2) Within-groups linkage

This method is a variant of the average linkage, but all possible distances between all points of the two clusters A and B are calculated, including the distances between the points in the same cluster:

$$D(A,B) = \frac{1}{(n_A + n_B).(n_A + n_B - 1)} \sum_{i,j} d(x_i, x_j) \qquad (6)$$

where the sum is over all x_i in A and all x_j in B. Here $d(x_i, x_j)$ is the Euclidean

distance or some other distance between the vectors x_i and x_j. At each step, we join the two clusters with the smallest distance.

3) Nearest Neighbor (*single linkage*)

In this method, the distance between two clusters *A* and *B* is defined as the minimum distance between a point in *A* and a point in *B*:

$$D(A,B) = \min\left\{d(x_i, x_j),\ x_i \in A,\ \ x_j \in B\right\}$$

where $d(x_i, x_j)$ is the Euclidean distance or some other distance between the vectors x_i and x_j. At each step, we merge the two clusters with the smallest distance.

7.6. Defining the number of clusters in hierarchical CA

It is not a well formalized procedure. Usually one uses the dendrogram by cutting across the branches at a given level of the distance measure. We wish to determine the number *k* of clusters that provides the best fit to the data. One approach is to look for large changes in distances at which clusters are formed and to compare the change in distance between the two solutions. See further examples.

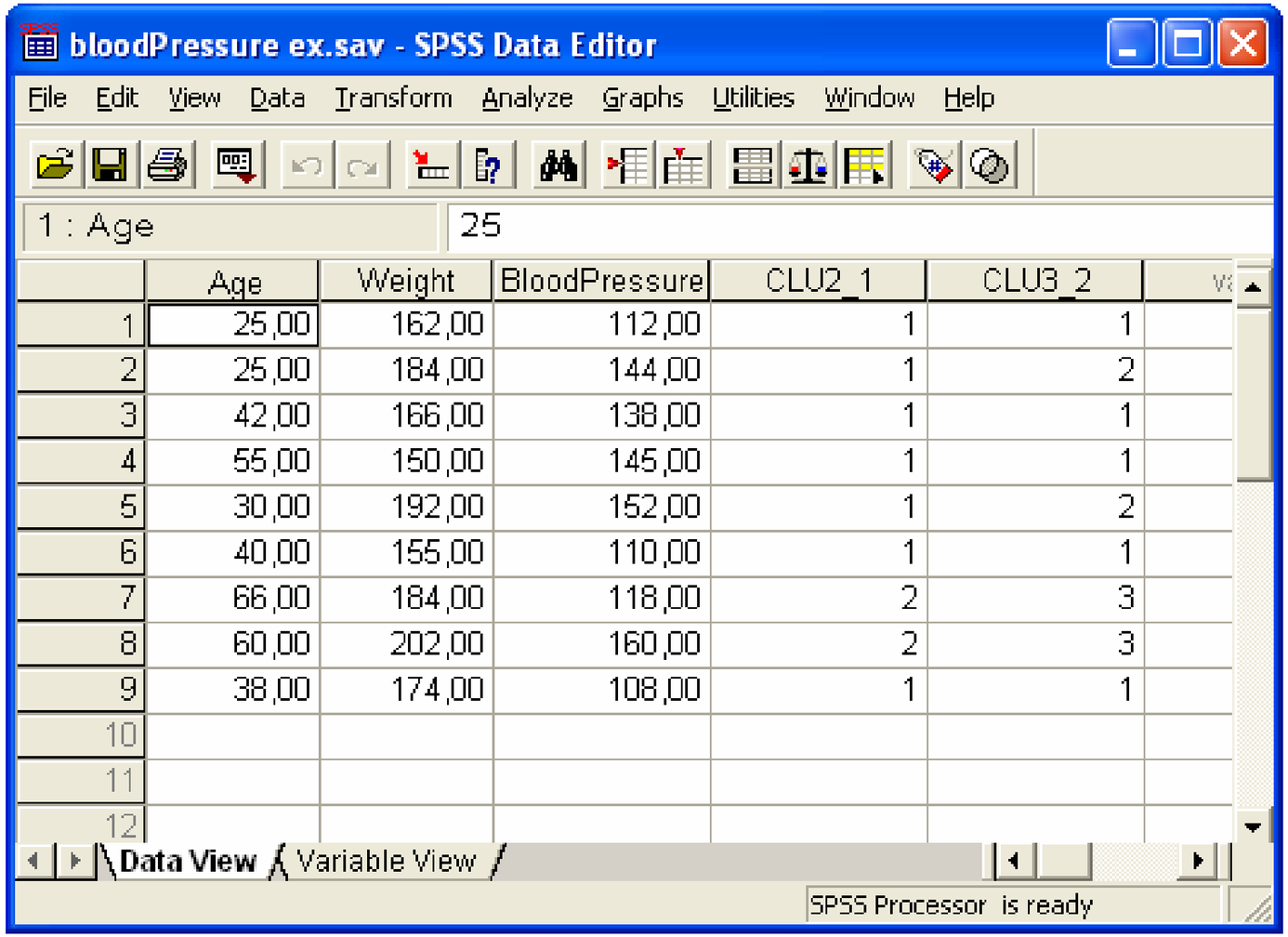
7.7. How to carry out CA in SPSS? ---→ See the files:

SPSS_CA_3_EN.pdf – Cluster analysis with SPSS: Hierarchical Cluster Analysis

SPSS_CA_2_EN.pdf – Cluster analysis with SPSS: K-Means Cluster Analysis

Example. BloodPressure example – Cluster analysis

Data two with saved solutions – clustered by 2 and by 3 clusters.



	Age	Weight	BloodPressure	CLU2_1	CLU3_2
1	25,00	162,00	112,00	1	1
2	25,00	184,00	144,00	1	2
3	42,00	166,00	138,00	1	1
4	55,00	150,00	145,00	1	1
5	30,00	192,00	152,00	1	2
6	40,00	155,00	110,00	1	1
7	66,00	184,00	118,00	2	3
8	60,00	202,00	160,00	2	3
9	38,00	174,00	108,00	1	1
10					
11					
12					

Proximity Matrix

Case	Squared Euclidean Distance								
	1:Case 1	2:Case 2	3:Case 3	4:Case 4	5:Case 5	6:Case 6	7:Case 7	8:Case 8	9:Case 9
1:Case 1	,000	4,148	3,025	7,180	7,061	1,164	9,099	16,415	1,257
2:Case 2	4,148	,000	2,425	7,752	,480	6,638	9,126	7,114	4,325
3:Case 3	3,025	2,425	,000	1,705	3,334	2,379	4,606	6,875	2,537
4:Case 4	7,180	7,752	1,705	,000	8,641	4,149	6,135	9,497	6,591
5:Case 5	7,061	,480	3,334	8,641	,000	9,334	8,837	4,465	6,197
6:Case 6	1,164	6,638	2,379	4,149	9,334	,000	5,892	15,247	1,206
7:Case 7	9,099	9,126	4,606	6,135	8,837	5,892	,000	5,642	4,042
8:Case 8	16,415	7,114	6,875	9,497	4,465	15,247	5,642	,000	11,481
9:Case 9	1,257	4,325	2,537	6,591	6,197	1,206	4,042	11,481	,000

This is a dissimilarity matrix

Agglomeration Schedule

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	2	5	,480	0	0	7
2	1	6	1,164	0	0	3
3	1	9	1,231	2	0	5
4	3	4	1,705	0	0	5
5	1	3	4,310	3	4	7
6	7	8	5,642	0	0	8
7	1	2	5,985	5	1	8
8	1	7	8,488	7	6	0

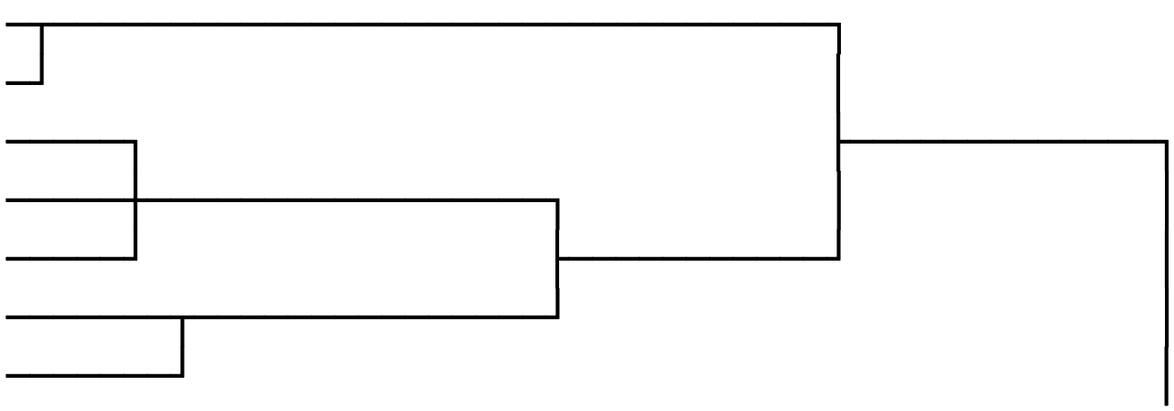
Dendrogram

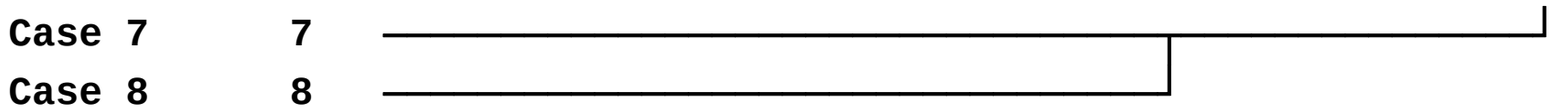
* * * * * * H I E R A R C H I C A L C L U S T E R A N A L Y S I S * * * * * *

Dendrogram using Average Linkage (Between Groups)

Rescaled Distance Cluster Combine

C A S E Label	Num	0	5	10	15	20	25
Case 2	2						
Case 5	5						
Case 1	1						
Case 6	6						
Case 9	9						
Case 3	3						
Case 4	4						





Conclusion:

Clustering by 3 clusters gives 3 groups of respondents that can be named:

1 cluster – healthy people (respondents 1, 3, 4, 6, 9)
2 cluster – medium stage (respondents 2, 5)
3 cluster – critical stage (respondents 7, 8)

Clustering by 2 clusters gives 2 groups of respondents that can be named:

1 cluster – good stage (respondents 1, 2, 3, 4, 5, 6, 9)

2 cluster – bad stage (respondents 7, 8)

The clustering in 3 groups is more appropriate.

8. Application of cluster analysis for exploring copper bromide laser variables

DESCRIPTION OF DATA

11 basic physical parameters of the CuBr laser:

D– the inside diameter of the laser tube,
dr– the inside diameter of the internal rings in the tube,
L – the length of the active area (electrode separation),
Pin – the input electric power,
PL – the input electric power per unit length (25% losses),
Prf – the pulse repetition frequency,
Pne – the neon gas pressure,
PH2 – the hydrogen gas pressure,
C – the equivalent capacity of the capacitor bank,
Tr – the temperature of the CuBr reservoirs,
Pout – the output laser power.

All the results are based on a 25% sample of all 300 experiments for the above mentioned eleven variables.

We performed hierarchical CA of variables.

Cluster analysis results

Initially we will conduct a partial cluster analysis for the first six variables (*D*, *dr*, *L*, *Pin*, P_L and P_{H2}), participating in the previous consideration. Our task is to compare the results with the calculations from factor analysis.

The first stage is to construct a matrix containing the results from the comparison of the objects (table 3). In our case the squared Euclidean distance is used as the indicator for similarity (or difference). It has to be noted that in table 3 are given only the comparative results for stage one when each object is considered as a cluster. Using the between group linkage method, independent variables are grouped into three clusters as seen in table 4. The first cluster includes variables *D*, *dr*, *L* and *Pin*, the second - P_L, and the third – P_{H2}. There is complete correspondence with the results from factor analysis (see the previous presentation 2).

TABLE 3. Proximity matrix of the first six variables.

Variable	D	dr	L	Pin	P_L	P_{H2}
D	0	23.9	36.7	44.7	227.5	103.0
dr	23.9	0	12.2	20.1	223.8	86.0
L	36.7	12.2	0	21.2	247.8	67.3
Pin	44.7	20.1	21.2	0	189.3	91.7
P_L	227.5	223.8	247.8	189.3	0	217.9
P_{H2}	103.0	86.0	67.3	91.7	217.9	0

TABLE 4. Cluster membership in 3 clusters.

Variable	3 clusters
D	1
dr	1
L	1
Pin	1
PL	2
PH2	3

Dendrogram

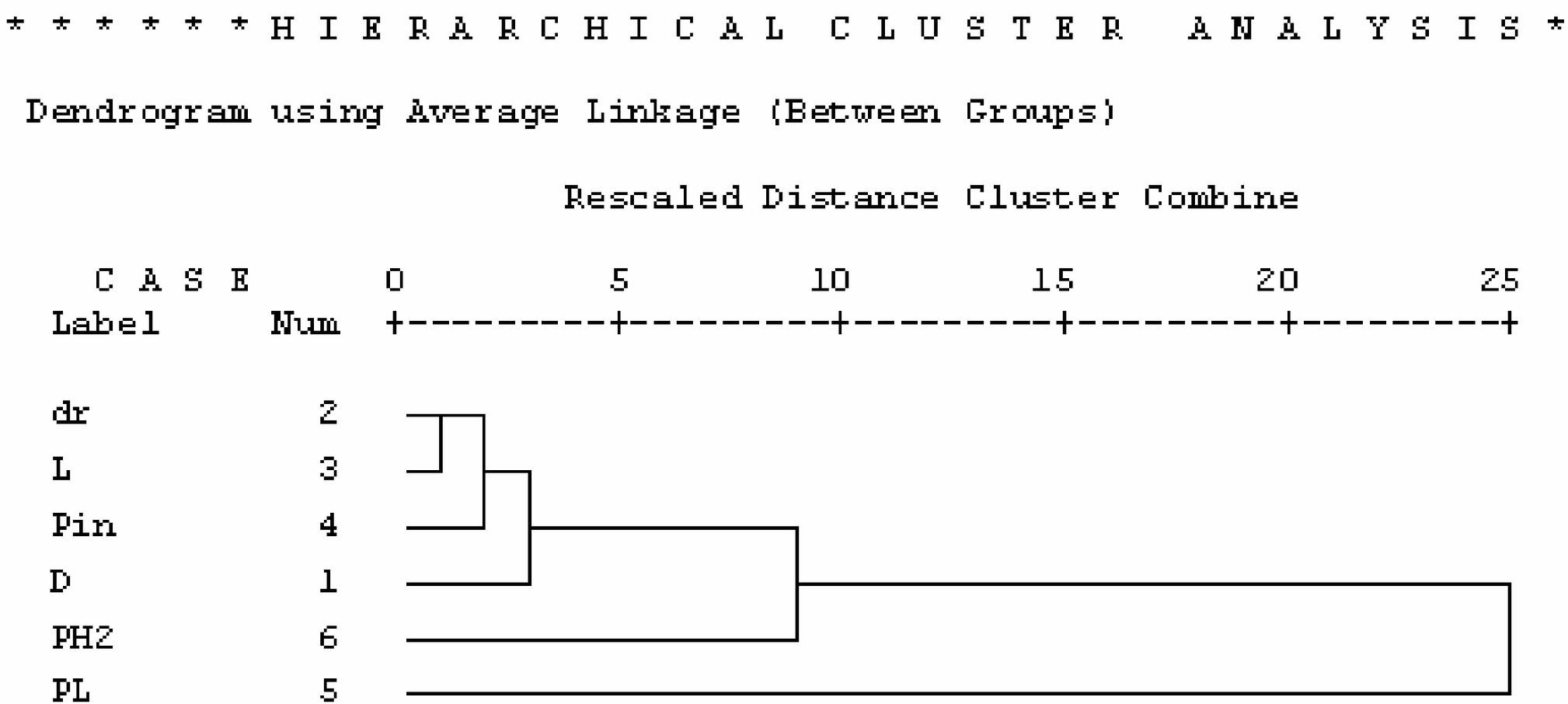


FIGURE 1. Dendrogram of variables from Tables 3 and 4.

The next stage is classifying all ten variables. Table 5 shows their initial similarity matrix. Basic table 6 shows the simultaneous clustering in two, three, four and five clusters. In every column opposite each independent variable is given its corresponding cluster number.

The optimal number of clusters has to be established. This problem can be solved by means of the dendrogram in Figure 2. It is the result of the same method which was used in Figure 1.

TABLE 5. Proximity matrix of ten input variables.

Variable	D	dr	L	Pin	P_L	P_{H2}	Prf	Pne	C	Tr
D	**0**	**19.5**	**33.0**	**39.4**	**202.4**	**95.9**	**145.0**	**158.4**	**101.5**	**115.9**
dr	**19.5**	**0**	**9.5**	**16.9**	**201.1**	**79.7**	**155.3**	**161.0**	**105.8**	**101.6**
L	**33.0**	**9.5**	**0**	**19.4**	**223.3**	**58.9**	**159.0**	**149.5**	**120.2**	**115.0**
Pin	**39.4**	**16.9**	**19.4**	**0**	**169.3**	**83.1**	**155.1**	**145.0**	**107.6**	**116.3**
P_L	**202.4**	**201.1**	**223.3**	**169.3**	**0**	**196.0**	**113.4**	**88.3**	**141.2**	**132.6**
P_{H2}	**95.9**	**79.7**	**58.8**	**83.1**	**196.0**	**0**	**164.8**	**140.7**	**158.2**	**174.3**
Prf	**144.7**	**155.3**	**159.0**	**155.1**	**113.4**	**164.9**	**0**	**97.4**	**139.8**	**120.5**
Pne	**158.4**	**161.0**	**149.5**	**145.0**	**88.3**	**140.7**	**97.4**	**0**	**172.6**	**126.5**
C	**101.5**	**105.8**	**120.2**	**107.6**	**141.2**	**158.2**	**139.8**	**172.6**	**0**	**93.4**
Tr	**115.9**	**101.6**	**115.0**	**116.3**	**132.6**	**174.3**	**120.5**	**126.5**	**93.4**	**0**

A careful review of the sequence of the clustering procedure reveals that all ten independent variables form three clusters. The first cluster includes *D*, *dr*, *L*, *Pin* and P_{H2}. The second includes variables P_L, *Pne* and *Prf*, and the third - *C* and *Tr*. This grouping corresponds to the column of three clusters in table 6. In the end we get three clusters for clustering all ten independent variables.

The next stage is to determine the position of the dependent variable (*Pout*) among the independent variables. The proximity with them is visible in Figure 3. As expected *Pout* is nearer to variables *D*, *dr*, *L*, *Pin* and P_{H2} forming together with them the first cluster. The latter serves to confirm the influence of these variables on *Pout*.

TABLE 6. Cluster membership in 2 to 5 clusters of all input variables.

Variable	5 clusters	4 clusters	3 clusters	2 clusters
D	1	1	1	1
dr	1	1	1	1
L	1	1	1	1
Pin	1	1	1	1
P_L	2	2	2	2
P_{H2}	1	1	1	1
Prf	3	3	2	2
Pne	2	2	2	2
C	4	4	3	1
Tr	5	4	3	1

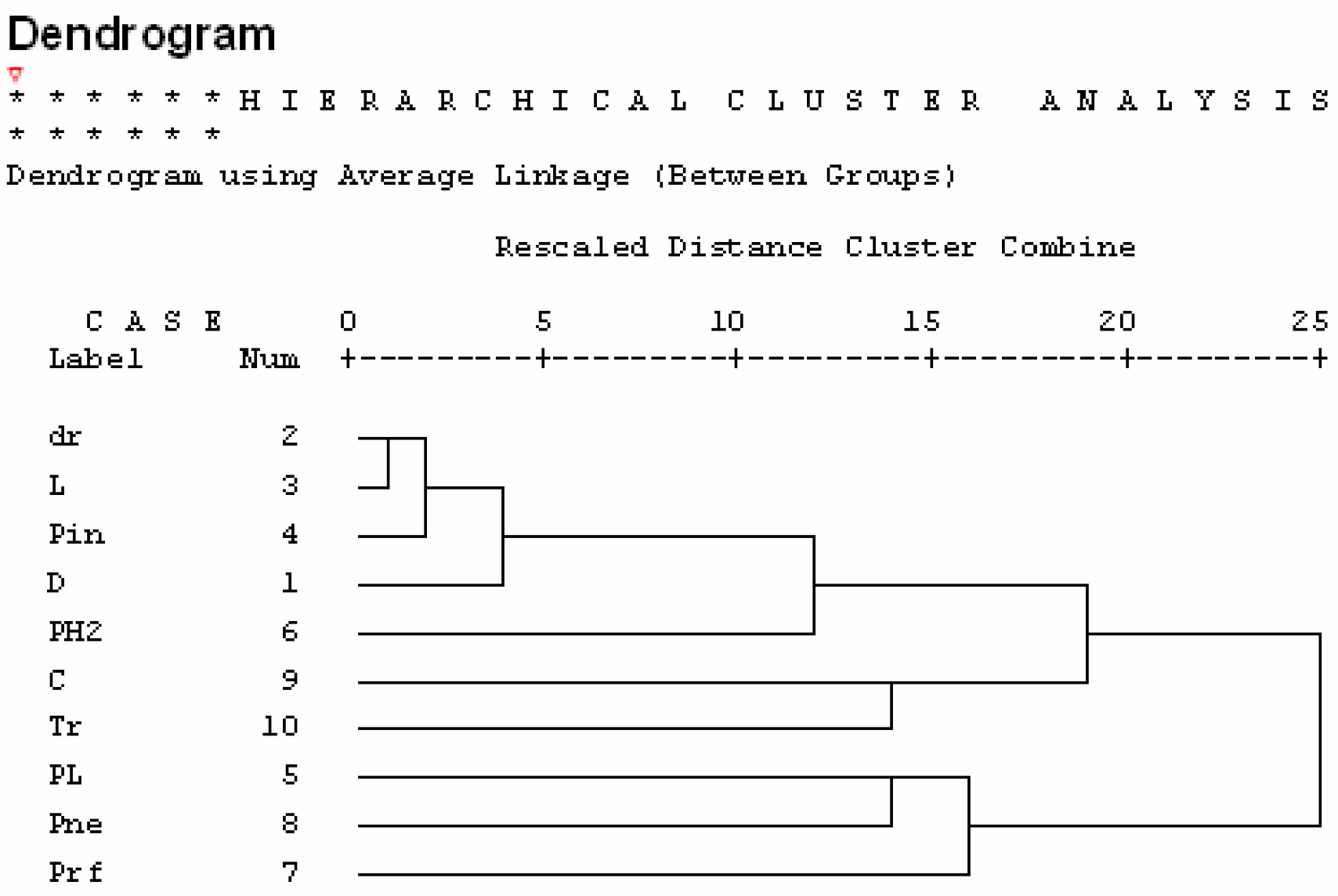


FIGURE 2. Dendrogram of ten input variables.

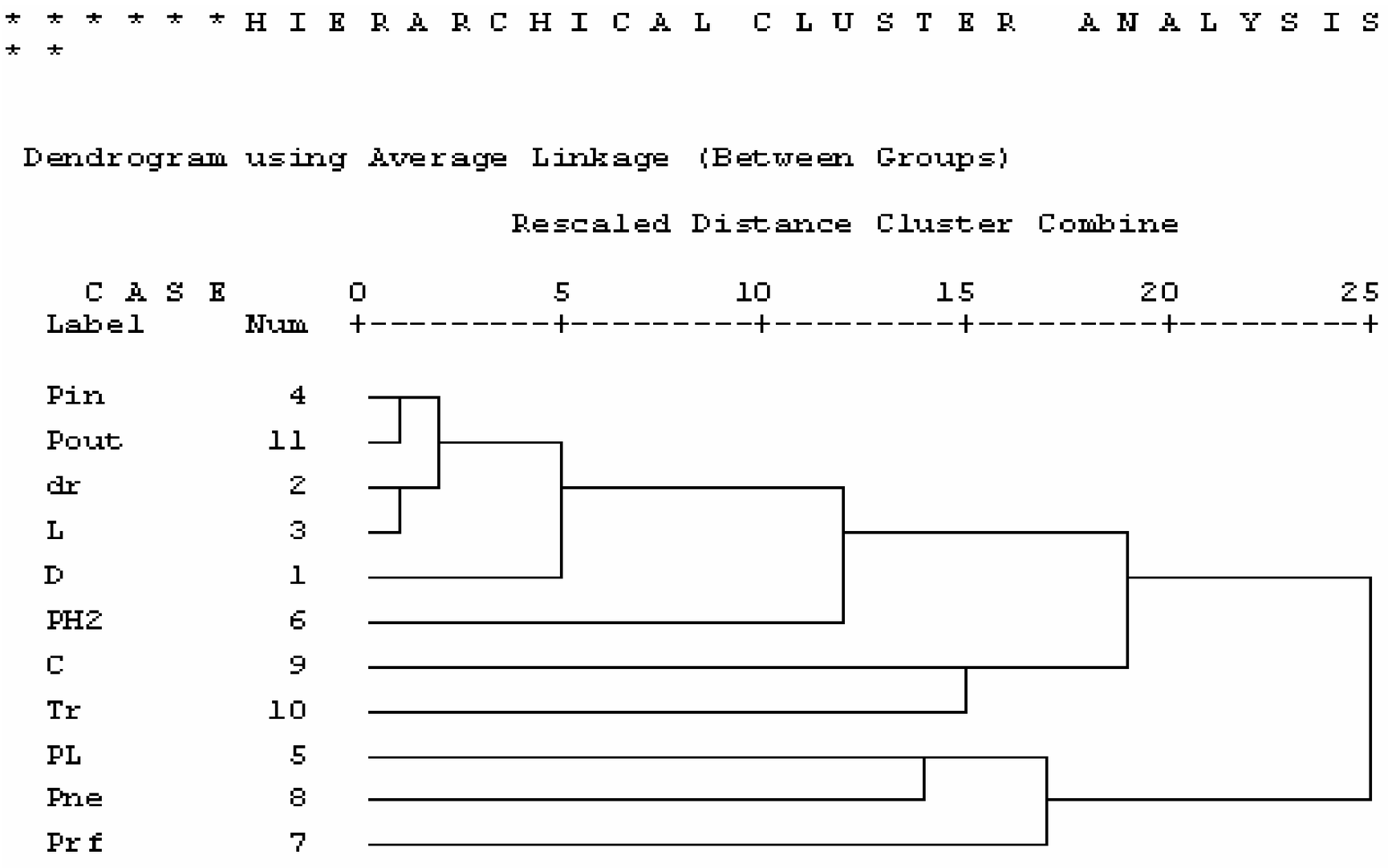


FIGURE 3. The dendrogram of all input variables and the laser power *Pout*.

Discussion on the Cluster Analysis results

The results from the clustering procedure can be used when planning a filtering experiment in which out of the totality of independent quantities the main group of quantities has to be separated, so as to be used later on for a more detailed examination. Also under the conditions of an extremal experiment with the goal of optimizing the object being studied, it is necessary to begin varying essential variables in accordance with their homogeneity (similarity), i.e. *dr*, *L*, *Pin*, *D* and P_{H2}.

Conclusion

This example is an application of cluster analysis of variables in the field of metal vapor lasers. Ten independent variables are examined, nine of them being actual physical quantities and one dependent variable – laser power *Pout*. In the previous application of Factor analysis, conducted for the same sample of data a general totality of all available experiments, following the correlation principle, two groups of variables are differentiated - significant and insignificant. The significant ones are classified into three groups (factors). In this example

they are classified using cluster analysis following the principle of homogeneity. The result is a hierarchical linking of variables. Three clusters and the order of similarities are established.

The obtained results could be used as a basis of multiple regression analysis and prognosis of further experiment in order to enhance the laser power generation.