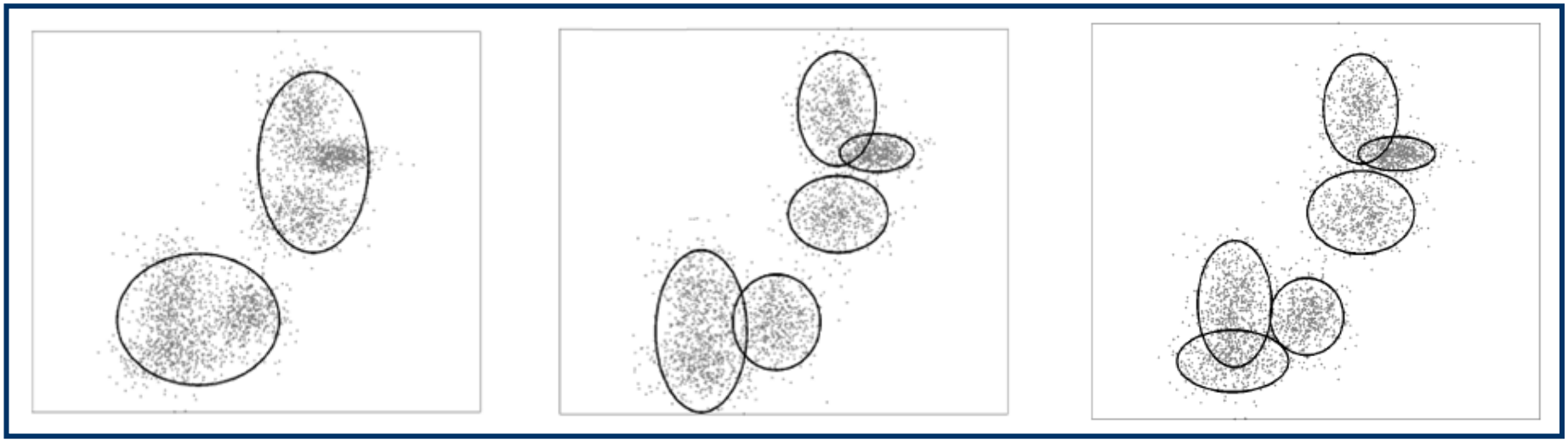
Cluster Analysis

Cluster Analysis (CA) is a multivariate method searching for a clustering of observations. Like Principal Component Analysis, CA is an unsupervised method, i.e., a method in which no preliminary information on the classification/clustering of samples is available.

In the context of Cluster Analysis a «cluster» can be defined as a group of contiguous elements in a statistical population.

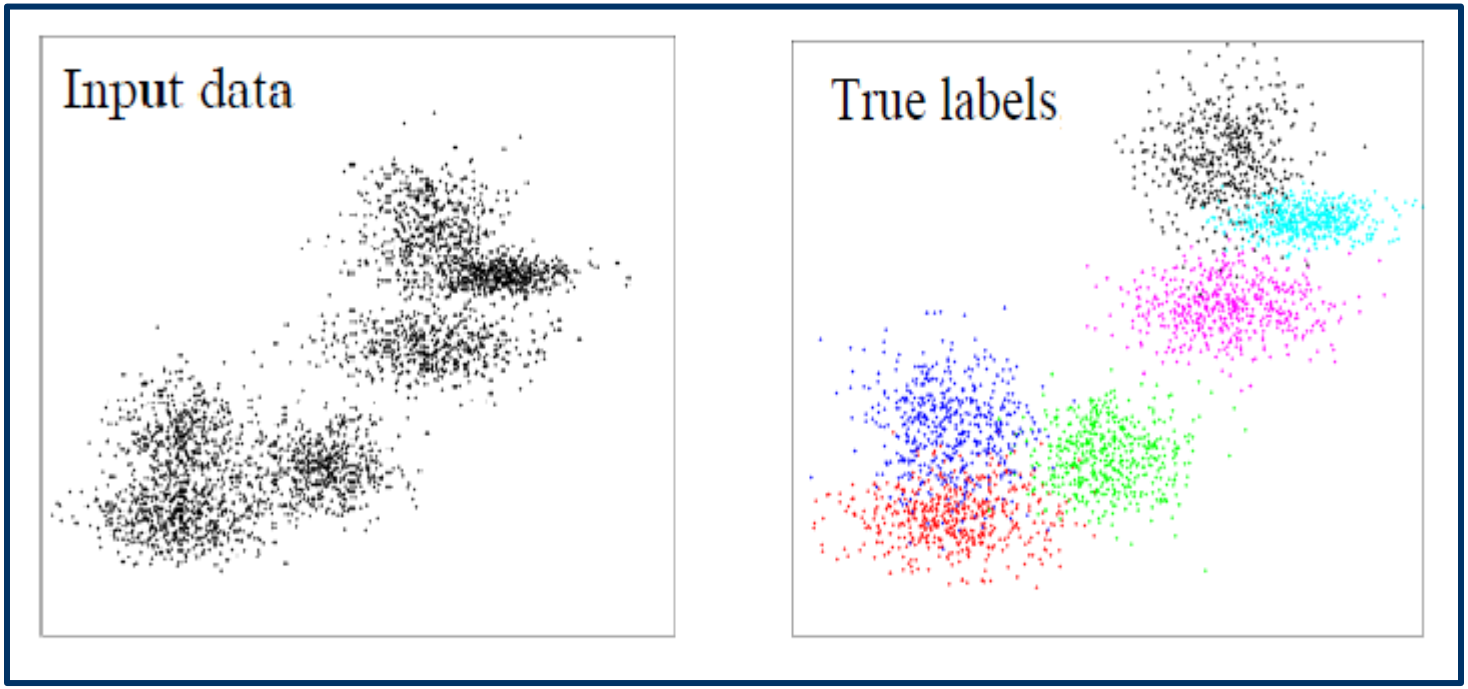
A more operative definition is based on the evaluation of internal cohesion for each cluster and of external separation between clusters.

As shown in the following figure, referred to bivariate data:



it is not surprising that the problem of the recognition of clusters may not have a single solution.

When the true labels for samples are reported, e.g., using a colour code, it is apparent that the solution including six clusters is the optimal one:



The general principle of the method is searching for non casual structures in data, relating the concept of non casual structure to that of group and looking for the presence of groups in the data space, in contrast with the hypothesis of complete homogeneity (isotropy).

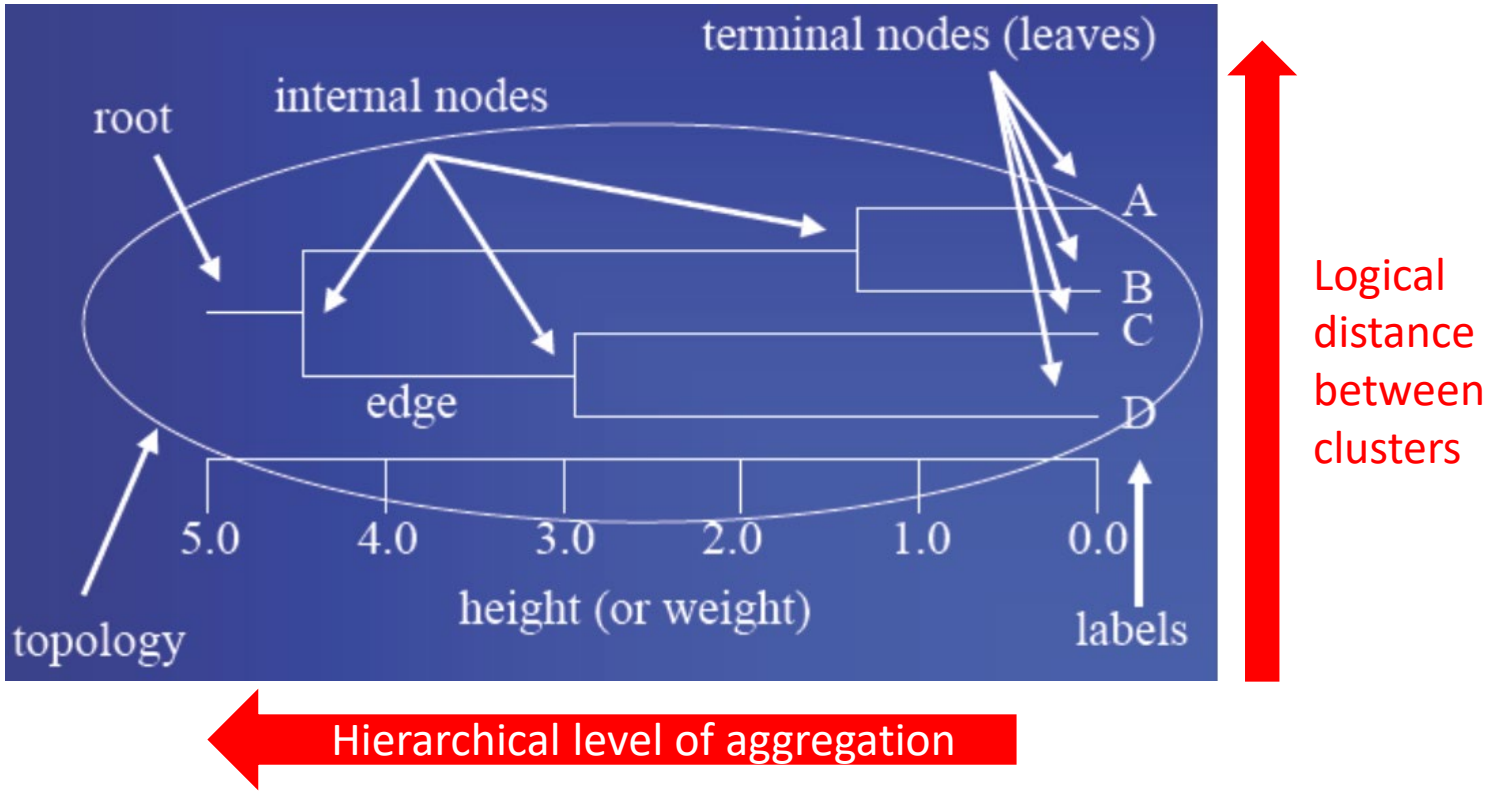
It is worth noting that Cluster Analysis can be used to find structures in data without necessarily providing explanations or interpretations.

Hierarchical Cluster Analysis (HCA)

In HCA objects (individuals, observations, etc.) are divided into a series of nested clusters, according to a hierarchical relation that can be represented through a graph known as *dendrogram*.

One of the axes of the dendrogram reports the logical distance between clusters/samples according to the defined metrics.
The other axis indicates the hierarchical level of aggregation.

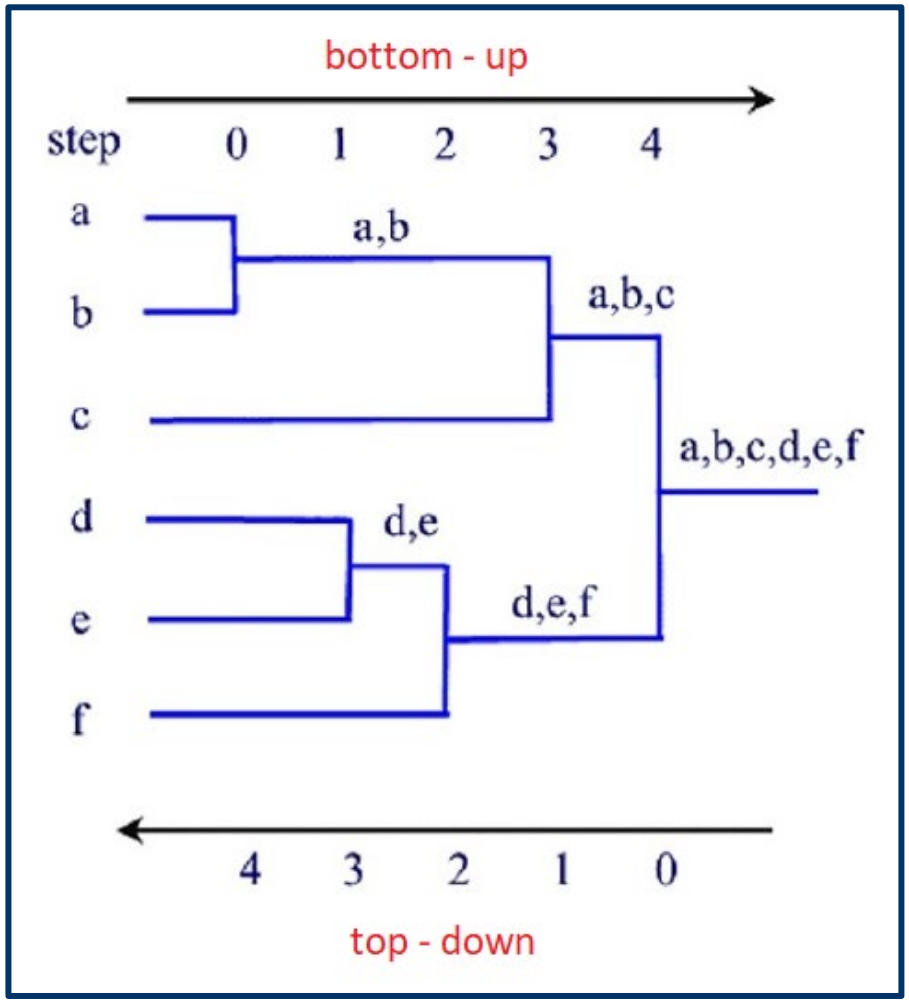
Root, edges and internal and terminal nodes (leaves) can be individuated in a dendrogram:



Two different strategies can be followed for hierarchical clustering:

1) agglomerative - it is a bottom-up approach, in which the starting point is the insertion of each element in a different cluster, followed by grouping between the resulting clusters, two at a time

2) divisive - it is a top-down approach, in which all elements are initially located in a single cluster, which is then recursively divided into sub-clusters.



Distances and similarities

In order to decide how single samples/clusters must be combined (in the agglomerative approach) or how a cluster must be divided (in the divisive approach) a measurement of (dis)similarity between clusters, based on the distance in the multivariate space has to be defined.

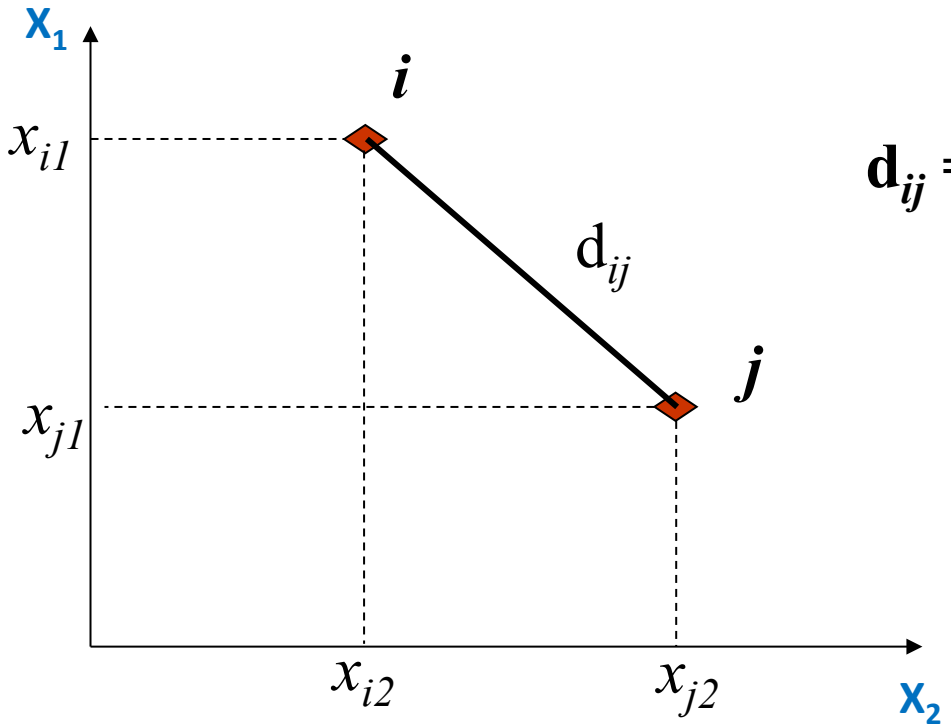
Examples of distance measurements are the following:

	Measurement	Distance
D1	Euclidean distance	$d_{ij}^{E} = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2}$
D2	Manhattan (City Block) distance	$d_{ij}^{CB} = \sum_{k=1}^{p} \lvert x_{ik} - x_{jk} \rvert$
D3	Minkowski distance	$d_{ij}^{M} = \sqrt[r]{\sum_{k=1}^{p} \lvert x_{ik} - x_{jk} \rvert^{r}}, \; r \geq 1$
D4	Canberra distance	$d_{ij}^{C} = \sum_{k=1}^{p} \frac{\lvert x_{ik} - x_{jk} \rvert}{\lvert x_{ik} \rvert + \lvert x_{jk} \rvert}$

where p is the number of variables describing each object, labelled by subscript k, whereas i and j indicate two different objects.

Euclidean distance

Euclidean distance, corresponding to the geometric distance between two objects in a multidimensional space, is likely the most used. In the figure it is shown for bivariate data:



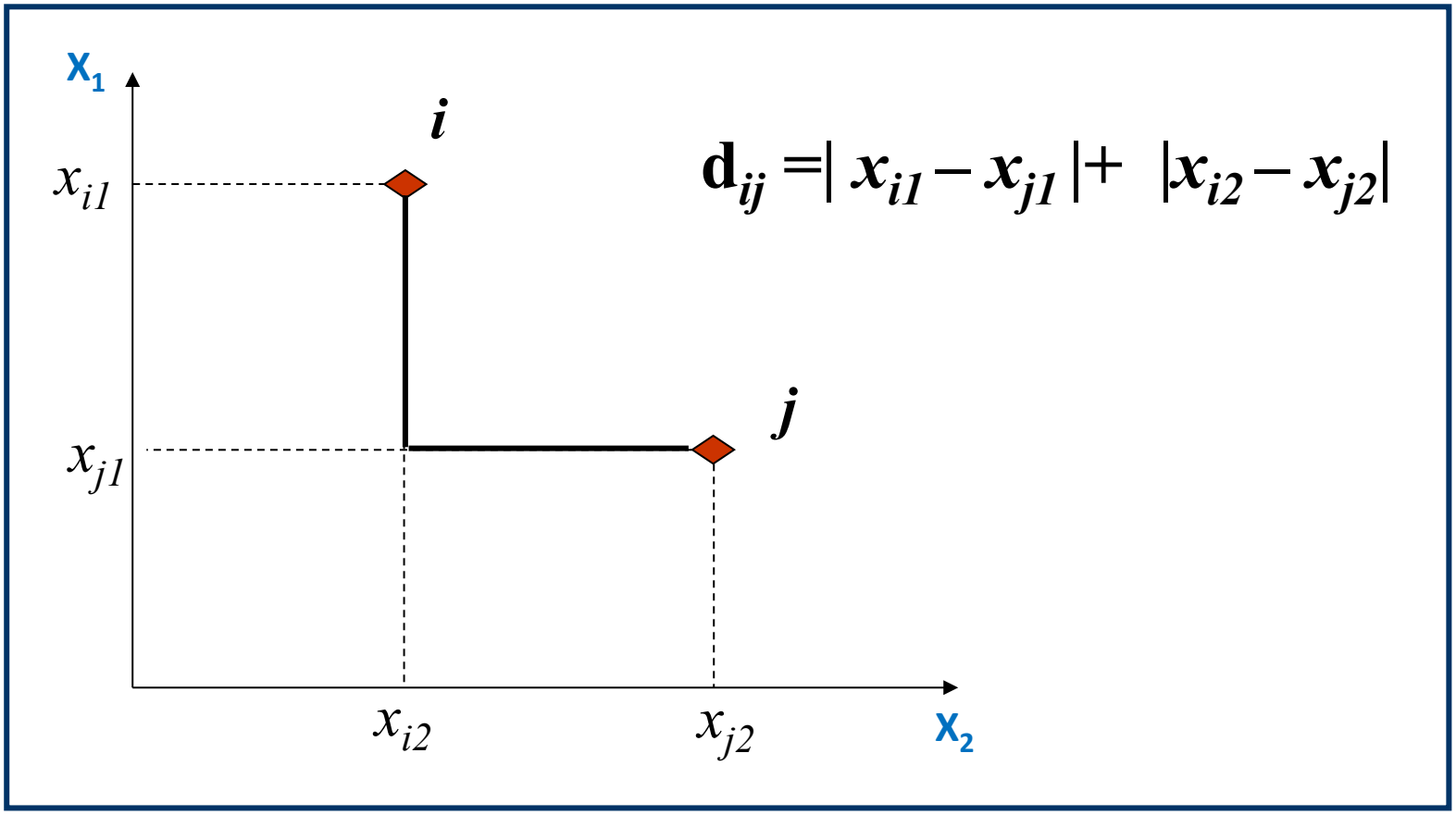
$$d_{ij} = [(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2]^{1/2}$$

One of the advantages of Euclidean distance is the independence on the addition of new objects, that could eventually be outliers.

On the other hand, Euclidean distance based on original variables can be strongly influenced by scale differences between different dimensions. It is thus a good practice to transform dimensions so that they have similar scales.

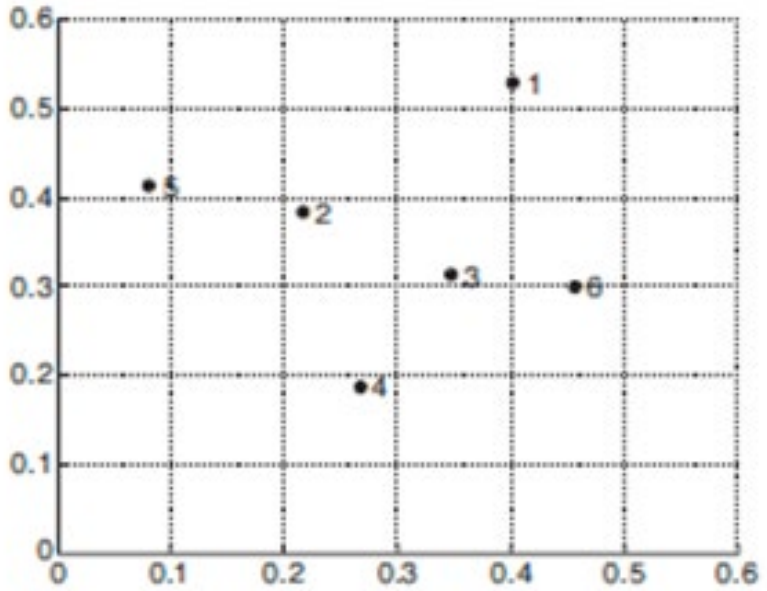
Manhattan or City Block distance

Manhattan distance considers the way two points on a city map would be connected by going around one of the city blocks. Its graphical representation for two dimensions is the following:



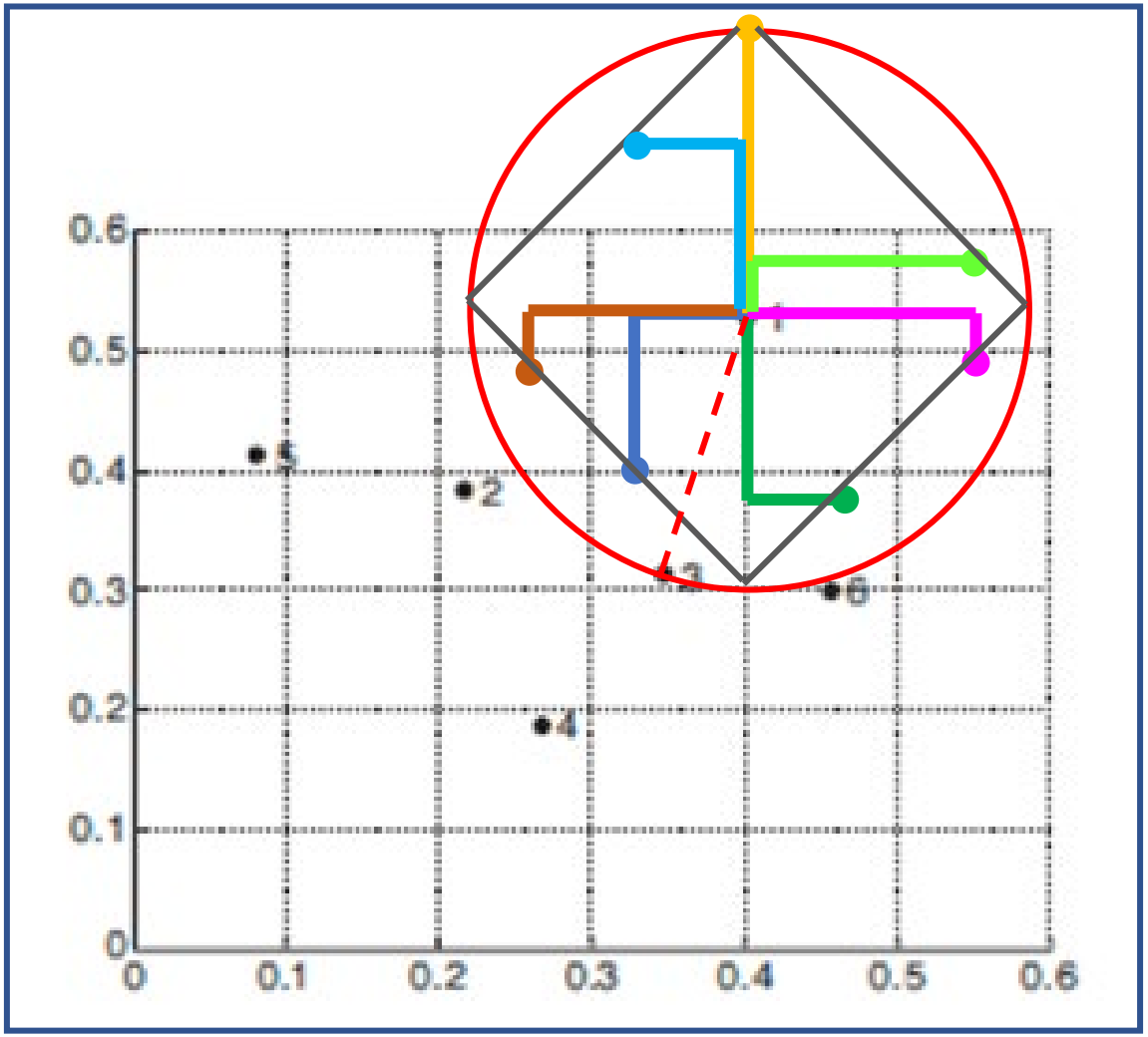
A comparison between distances obtained using different approaches can be performed by considering the following set of 6 objects (points) in a bi-dimensional space:

Point	x Coordinate	y Coordinate
p1	0.40	0.53
p2	0.22	0.38
p3	0.35	0.32
p4	0.26	0.19
p5	0.08	0.41
p6	0.45	0.30



If point p1 is considered, the red circumference in the following figure includes all points in the plane whose Euclidean distance from p1 is the same as that of p1 from p3.

On the other hand, points whose Manhattan distance from p1 is the same as the Euclidean p1-p3 distance are located on the perimeter of a rotated square whose side has Euclidean length equal to the p1-p3 distance multiplied by $2^{1/2}$.



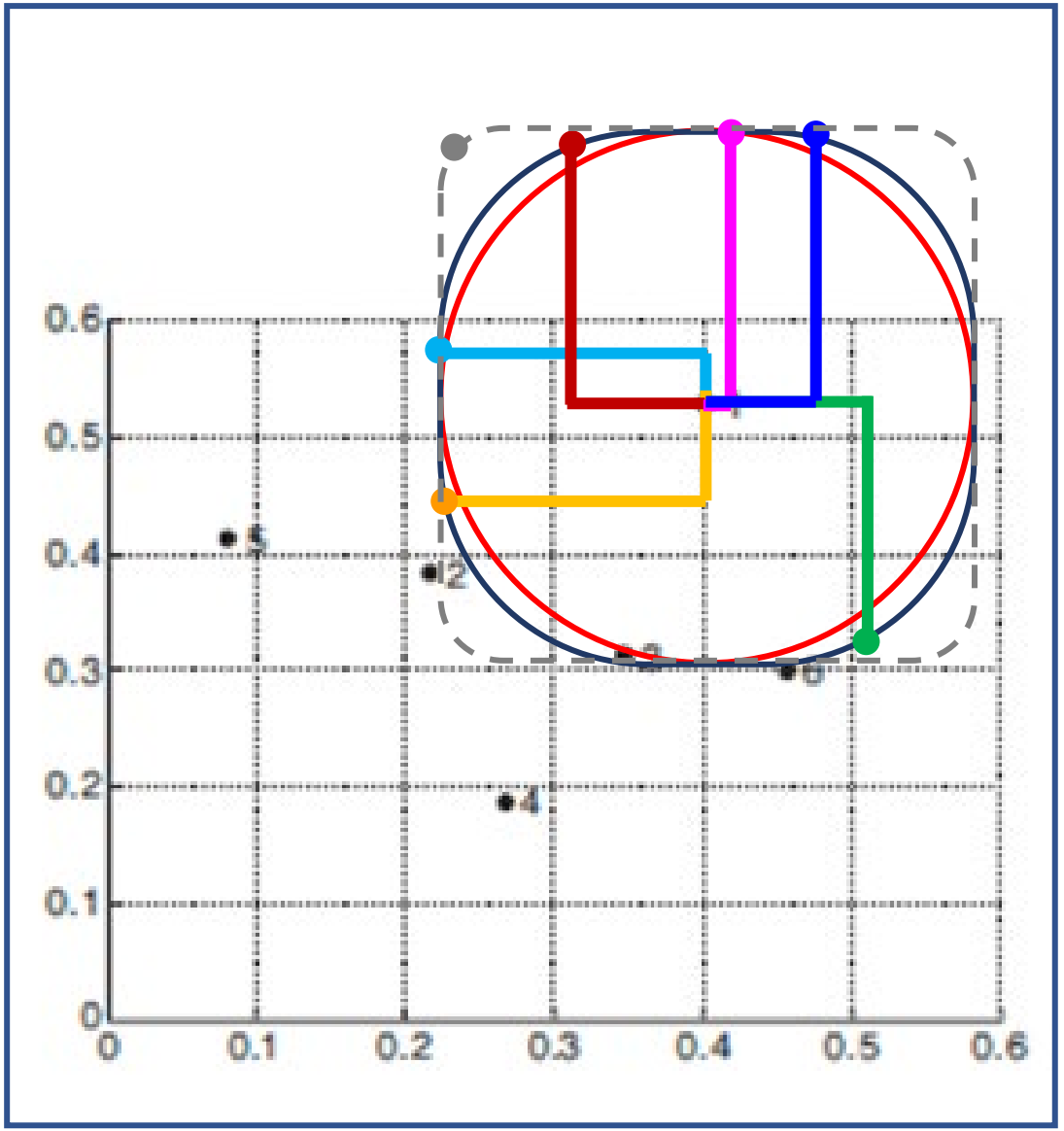
Minkowski distance with r = 3

If the Minkowski distance is considered with r = 3, points are located on a distorted circle, drawn with a black line in the figure on the right.

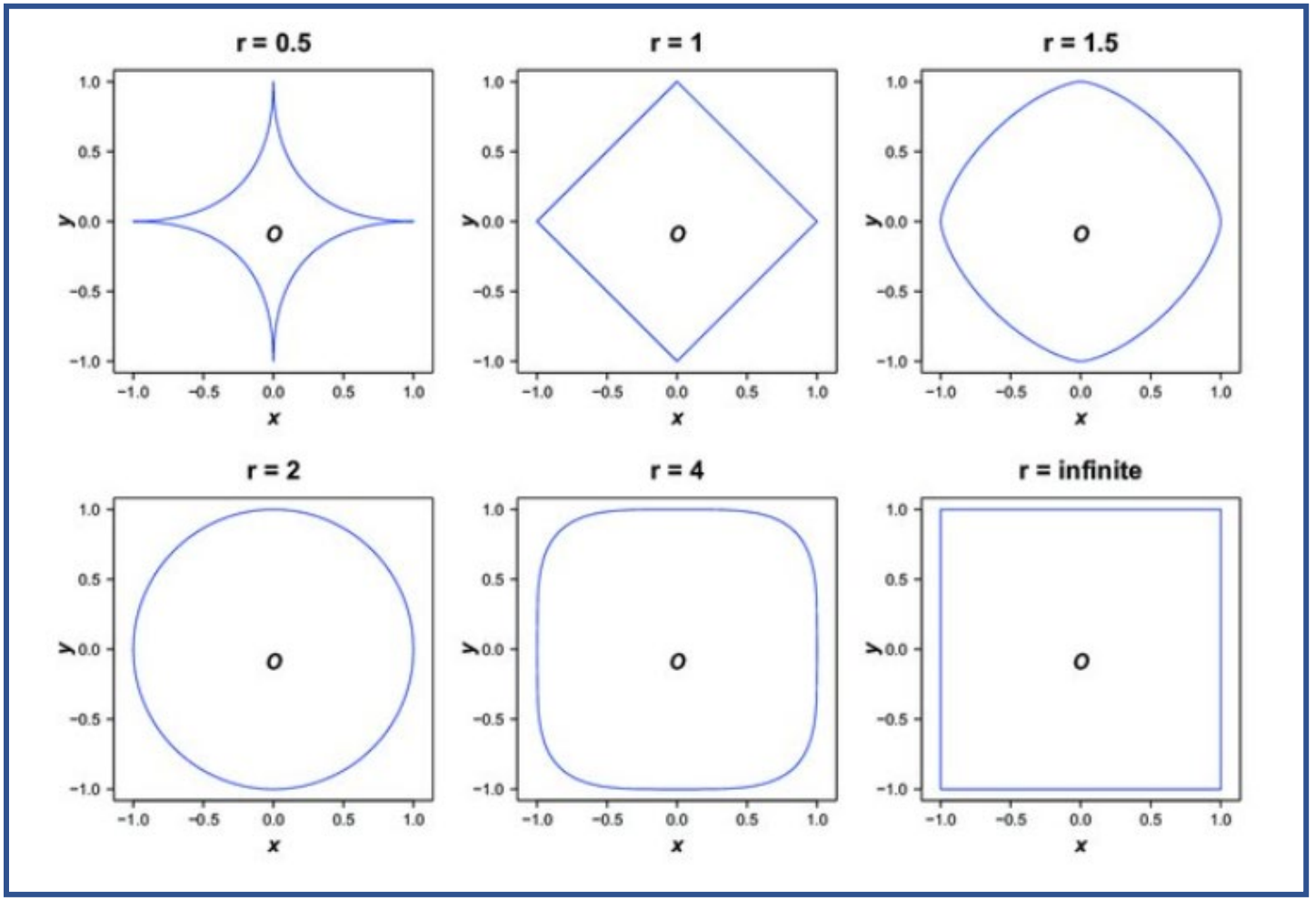
At the increase of r in the formula of Minkowski distance:

$$d_{ij}^{M} = \sqrt[r]{\sum_{k=1}^{p} \left| x_{ik} - x_{jk} \right|^{r}} \;,\; r \geq 1$$

the distorted circle is progressively enlarged at the corners, as evidenced by the grey point drawn in the figure, corresponding to the case of r = 18.



The loci of points in a 2-dimensional space at a distance of 1 from the center (O) using the Minkowski distance function with different values of the order r are shown in the following figure:

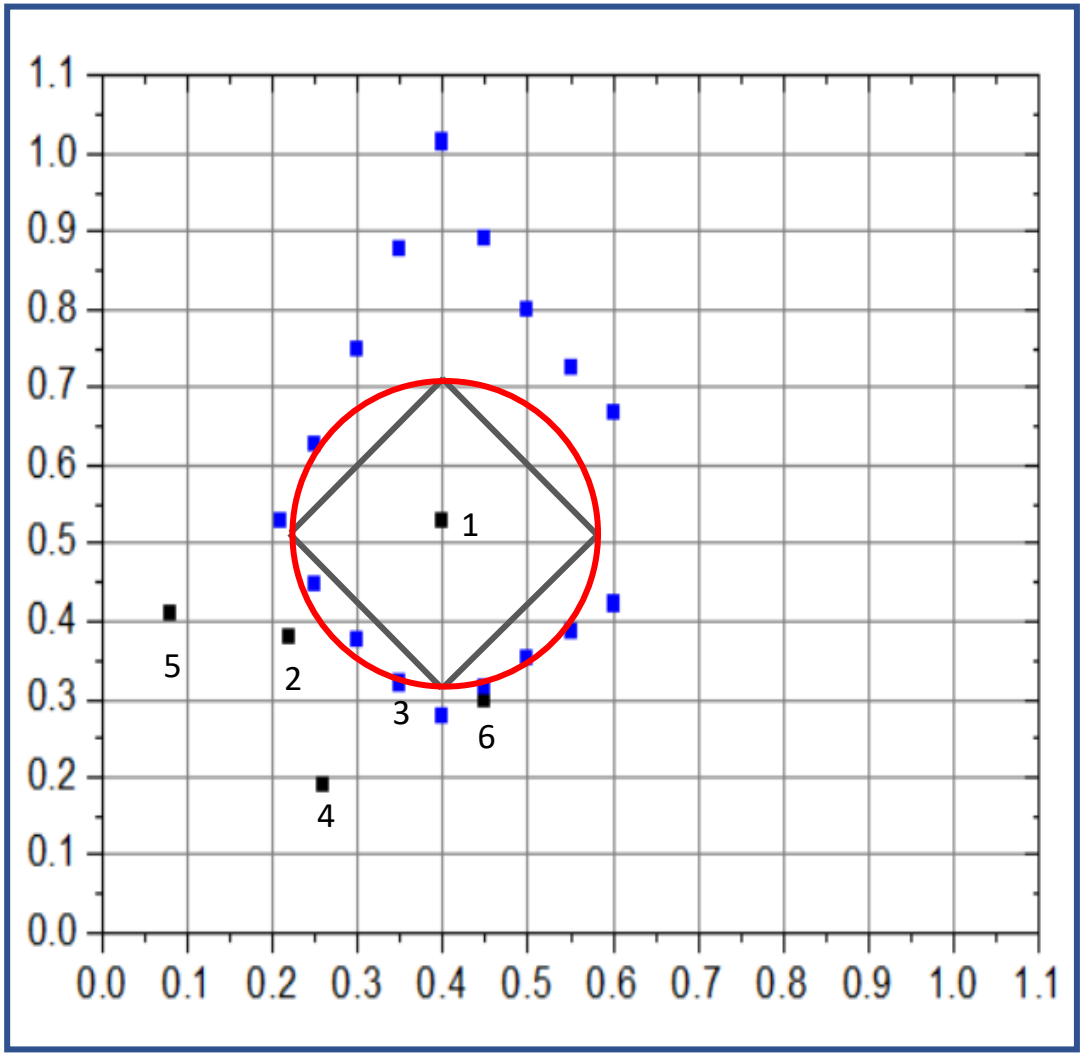


Along with Manhattan distance (r=1) and Euclidean distance (r=2), the Dominant distance ($r \to +\infty$) is another specific case of Minkowski distance. In the last case the locus of points is a square centered in O.

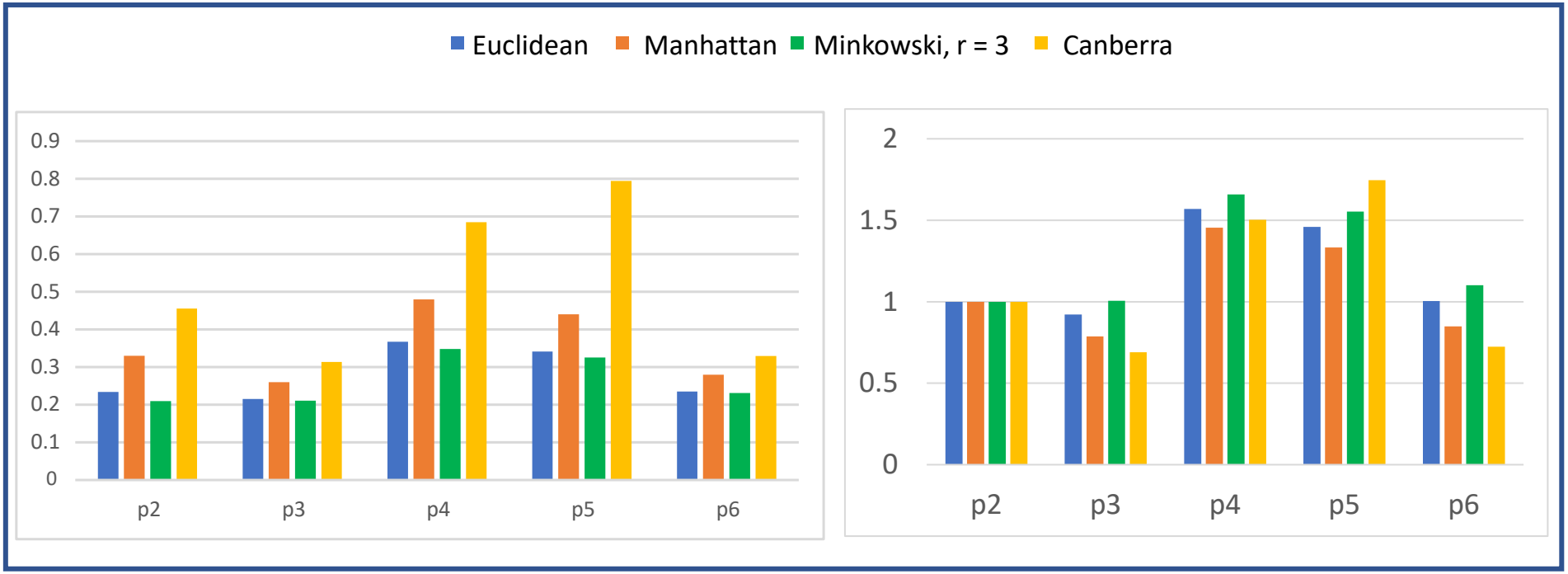
Canberra distance

$$d_{ij}^{C} = \sum_{k=1}^{p} \frac{\left| x_{ik} - x_{jk} \right|}{\left| x_{ik} \right| + \left| x_{jk} \right|}$$

If Canberra distance for the p1-p3 couple is calculated (0.31373), further points with the same distance from p1 can be drawn in the graph (blue points) and a comparison can be done with the loci of points whose Euclidean or Manhattan distance from p1 is equal to the p1-p3 distance.



Comparison between values of the distances between p1 and the other points shown before, as such or normalized by the p1-p2 distance, measured according to different approaches.



Properties of metric distance and standardization of variables

A metric distance between two objects **i** and **j** satisfies the following conditions:

$d_{ij} \geq 0$ non negativity

$d_{ij} = d_{ji}$ symmetry

$d_{ij} = 0$ if $i=j$ separability

$d_{ij} \leq d_{ik} + d_{kj}$ triangular inequality

Standardization of variables before proceeding with cluster analysis can be based on:

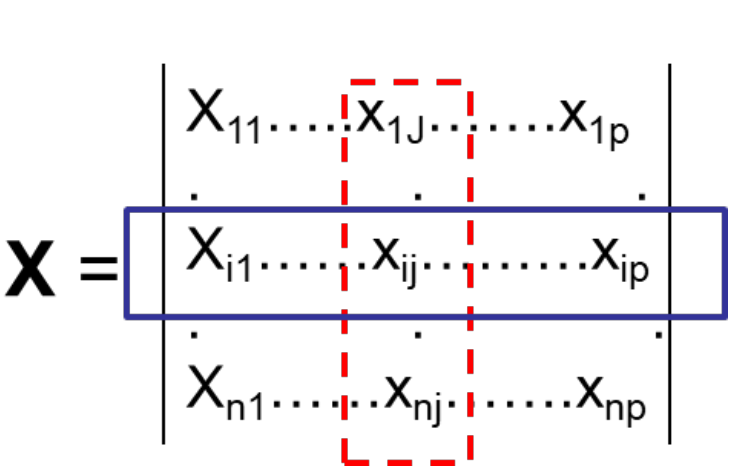
1) standard deviation for each variable $x_{ij} = \frac{x_{ij}}{s_j}$ this approach downweights variables with high standard deviation

2) z-scores $x_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}$ choice of mean and standard deviation is critical in this case

Matrix of distances and measures of similarity

Given the matrix of data **X**, in which each row represents one of the n objects and each column the p variables related to objects, the distance between each couple of objects is calculated according to one of the formulas shown before.

A dissimilarity matrix $\hat{D}_{n\times n}$, including all distances, is obtained:



$$\hat{D} = \begin{pmatrix} 0 & \dots & d_{1j} & \dots & \dots & d_{1n} \\ \dots & 0 & \dots & \dots & \dots & \dots \\ d_{i1} & \dots & 0 & \dots & \dots & d_{in} \\ \dots & \dots & \dots & 0 & \dots & \dots \\ \dots & \dots & \dots & \dots & 0 & \dots \\ d_{n1} & \dots & d_{nj} & \dots & \dots & 0 \end{pmatrix}$$

Elements along the main diagonal of matrix $\hat{D}$ are obviously all equal to zero, since they represent the distance between an object and itself. The matrix is also symmetric.

Once distances are calculated, measures of similarity, that are complementary to the former, can be obtained.

A general formula for the calculation of similarity between objects **i** and **j** is:

$$S_{ij} = 1 - \frac{d_{ij}}{d_{ij}(\max)}$$

Where $d_{ij}(max)$ is the maximum distance between two objects in the entire matrix of distances.

The following properties can be easily predicted for S_{ij}:

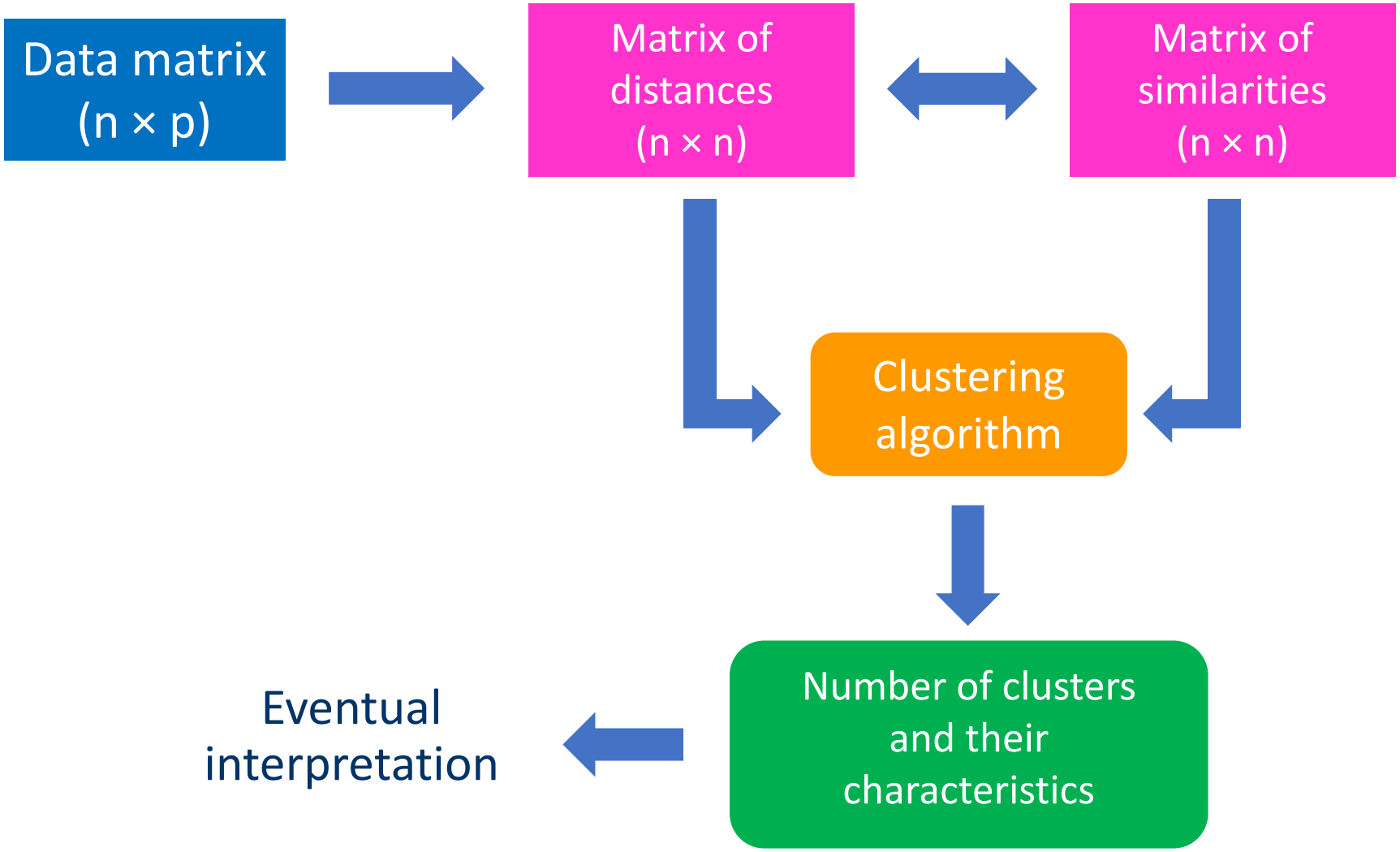
$$0 \leq S_{ij} \leq 1 \qquad\qquad S_{ii} = 1$$

The similarity of an object with itself is obviously equal to 1, since the distance of an object with itself is equal to 0.

Moreover, if $d_{ij} = d_{ji}$ also $S_{ij} = S_{ji}$.

Once matrices of distances (or, equivalently, of similarities) are calculated, a clustering algorithm has to be implemented to obtain clusters.

The general workflow for Cluster Analysis can be thus represented as follows:



Clustering algorithms in agglomerative Hierarchical Cluster Analysis

Clustering algorithms operate on the distance (or similarity) matrix according to the following general criteria:

1. Individuation of more similar objects or clusters to form couples
2. Gathering of the two clusters (or objects) in a unique new cluster, at a certain level of similarity
3. Calculation of the level of similarity of the new cluster with respect to the remaining ones.

Several algorithms can be used to agglomerate clusters:

1. Single linkage
2. Complete linkage
3. Average linkage
4. Centroid linkage
5. Ward method

Single linkage method

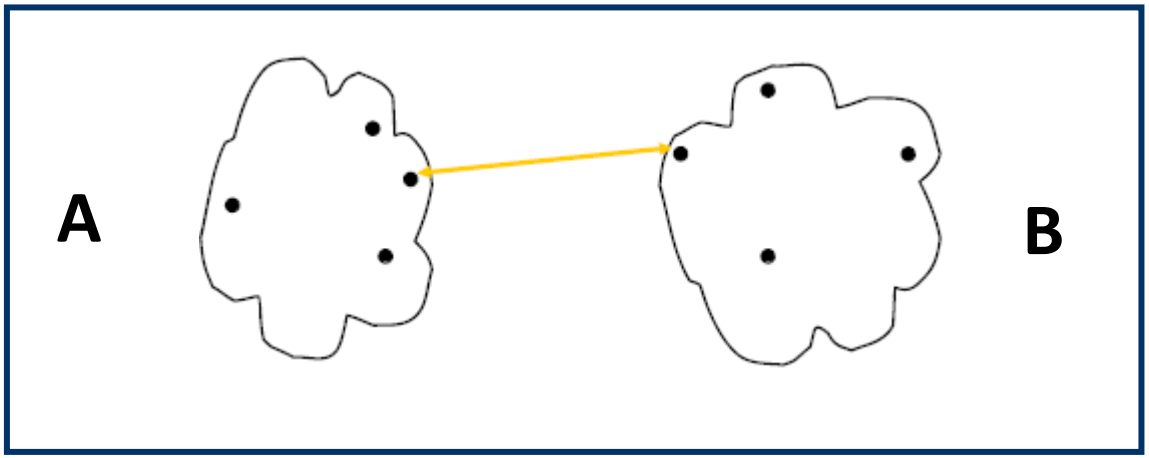
The single linkage method, also called nearest-neighbour method, is based on the definition of similarity or distance between clusters.

In particular, the level of proximity between two groups is evaluated by considering only information pertaining to the nearest objects belogning to the two groups, thus ignoring that referred to other objects.

Indicating as **A** and **B** two groups, the distance to be considered is:

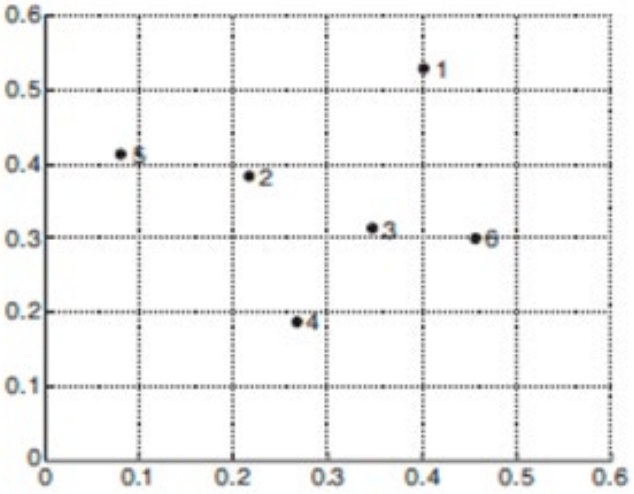
$$d_{AB} = \min_{i \in A, j \in B} d_{i,j}$$

In graphical terms, this distance can be represented as in the following figure:



As an example, let us re-consider the following bivariate dataset including 6 points and its graphical representation:

Point	x Coordinate	y Coordinate
p1	0.40	0.53
p2	0.22	0.38
p3	0.35	0.32
p4	0.26	0.19
p5	0.08	0.41
p6	0.45	0.30



Euclidean distances between objects are calculated as follows:

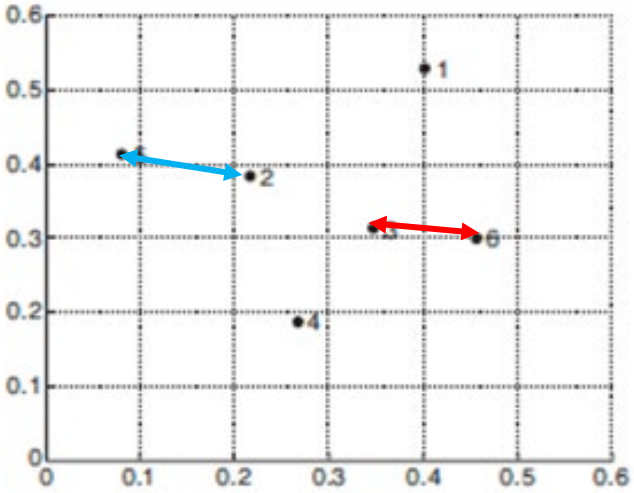
$$\mathbf{d}_{ij} = [(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2]^{1/2}$$

As an example, the distance between objects p2 and p5 is:

$$\mathbf{d}_{25} = [(0.22\text{-}0.08)^2 + (0.38\text{-}0.41)^2]^{1/2} = [0.0196 + 0.0009]^{1/2} = \mathbf{0.14}$$

The matrix of Euclidean distances is the following:

	p1	p2	p3	p4	p5	p6
p1	0.00	0.24	0.22	0.37	0.34	0.23
p2	0.24	0.00	0.15	0.20	0.14	0.25
p3	0.22	0.15	0.00	0.15	0.28	0.11
p4	0.37	0.20	0.15	0.00	0.29	0.22
p5	0.34	0.14	0.28	0.29	0.00	0.39
p6	0.23	0.25	0.11	0.22	0.39	0.00

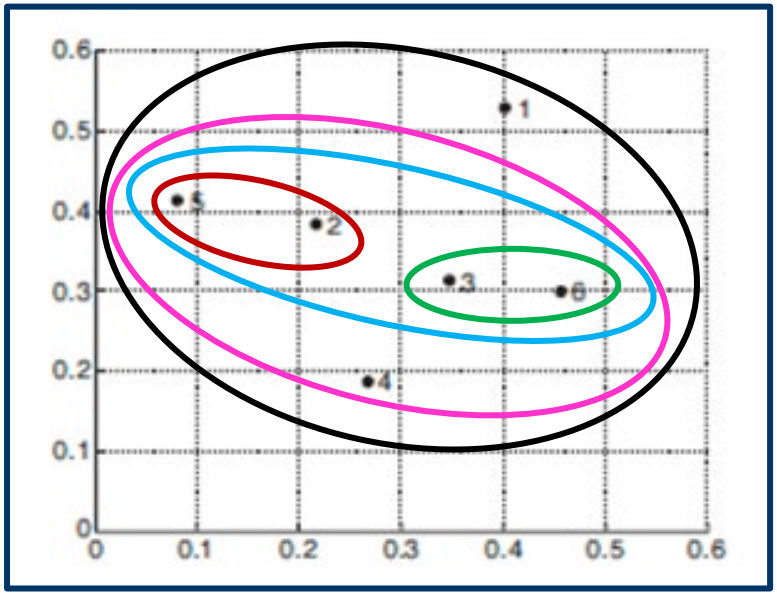


As apparent, minimum distances are observed between objects 3 and 6 and between objects 2 and 5. Two clusters, i.e., 3-6 and 2-5 can thus be considered as the basic ones.

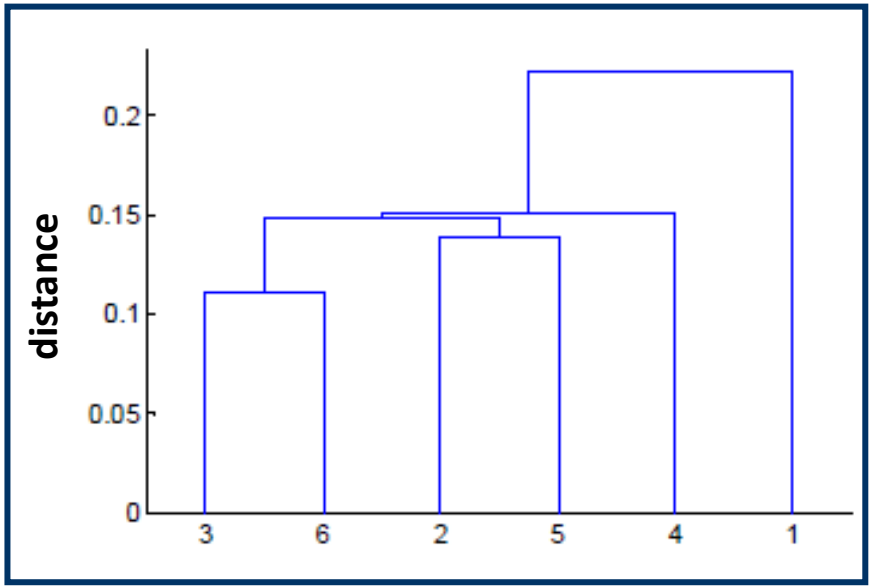
It is worth noting that distances between each object in one of these clusters and all the objects in the other one are all greater than the two distances evidenced above, thus the first clustering is confirmed.

Further objects are subsequently considered. In particular, object 4 is slightly closer to the two clusters than object 1.

Five nested clusters can thus be drawn in the plot of points:



The dendrogram resulting from Cluster Analysis is reported in the figure on the right:



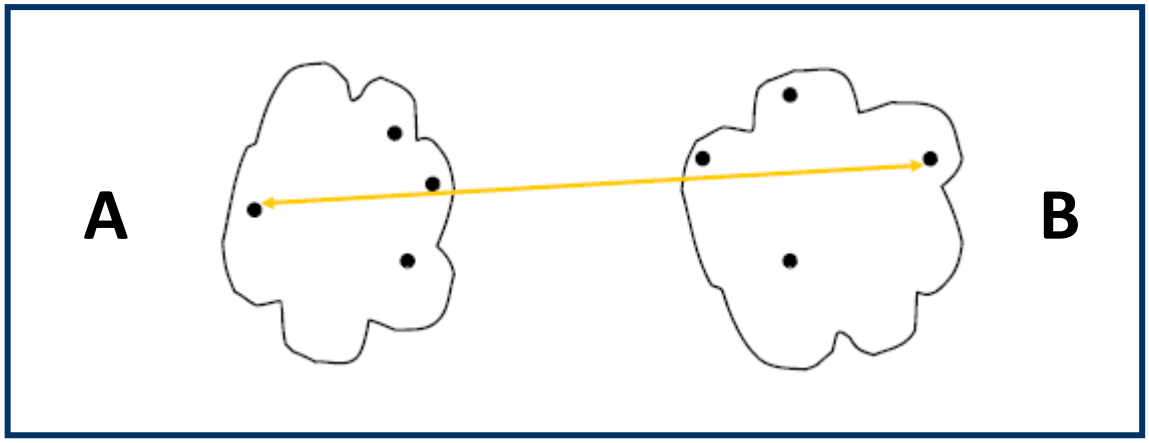
Complete linkage method

The complete linkage method, also called farthest-neighbour method, is based on the definition of similarity or distance between clusters based on the distance between their farthest points.

Indicating as **A** and **B** two groups, the distance to be considered is:

$$d_{AB} = \max_{i \in A, j \in B} d_{i,j}$$

In graphical terms, this distance can be represented as in the following figure:

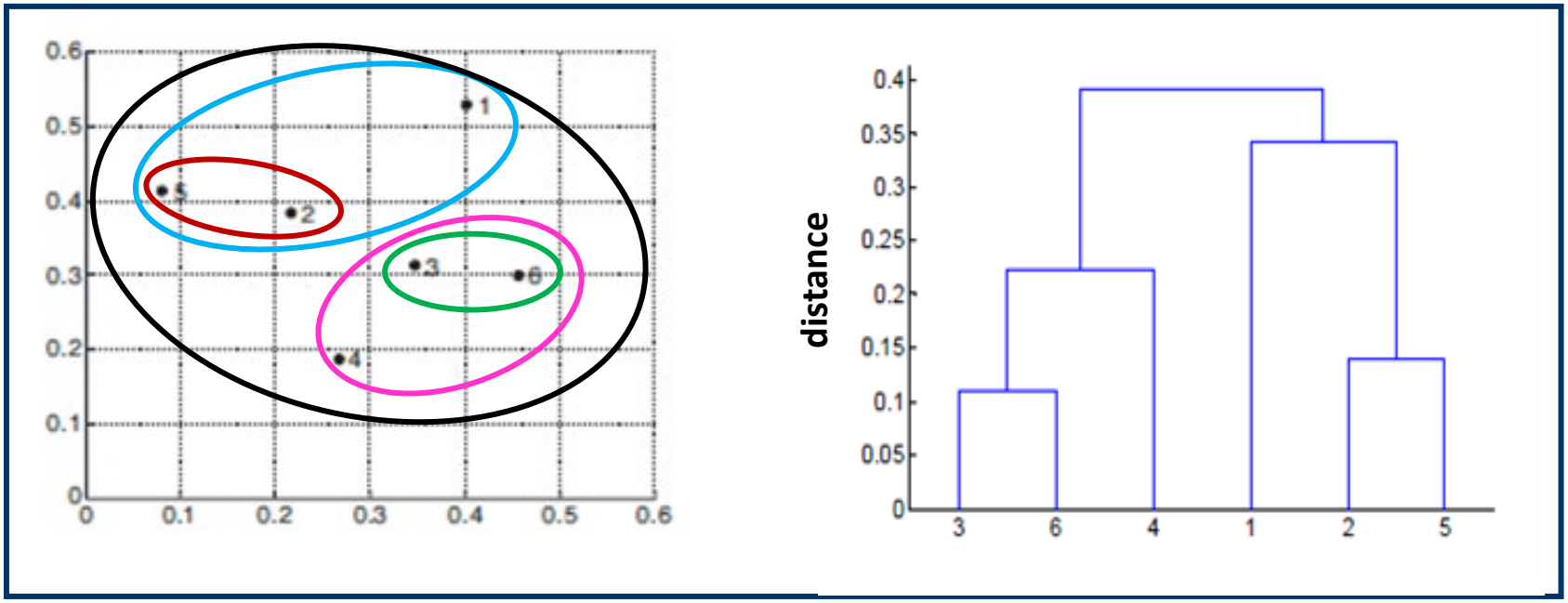


Considering the same dataset described before, the complete linkage method leads to a different clustering.

In fact, group 3-6 is merged with object 4 at the second level of clustering because the maximum of distances between cluster 3-6 and object 4 is 0.22, whereas the maximum of distances for clusters 2-5 and 3-6 is 0.39 (the distance between points 5 and 6), thus these two clusters cannot be grouped together.

Cluster 2-5 is merged with object 1 at the third level of clustering.

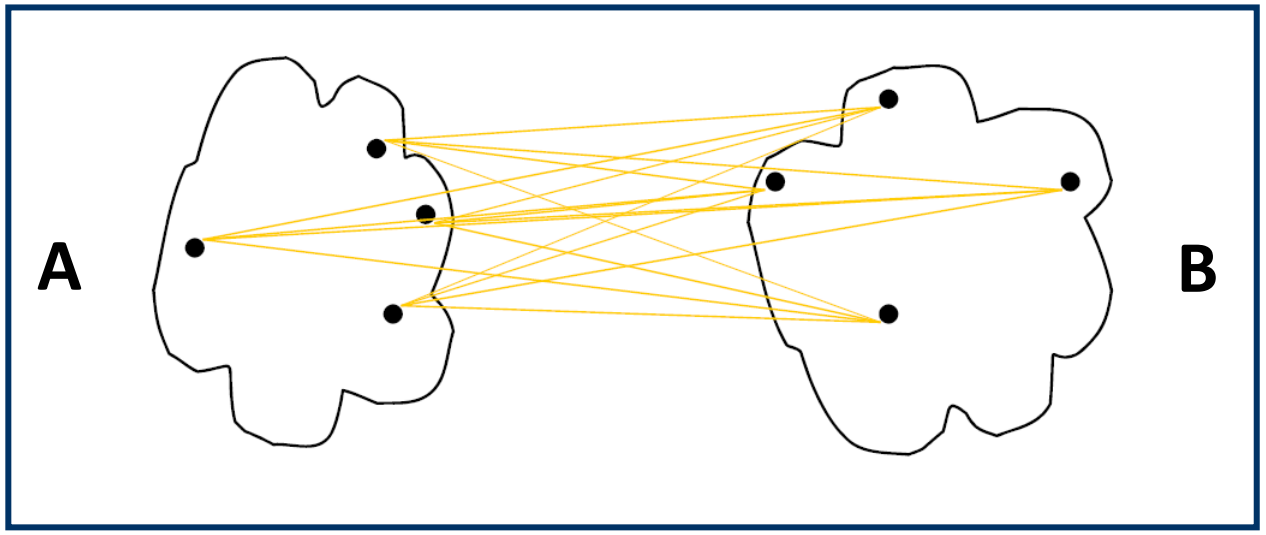
The results of CA using the complete linkage method can be thus represented graphically in the following figure:



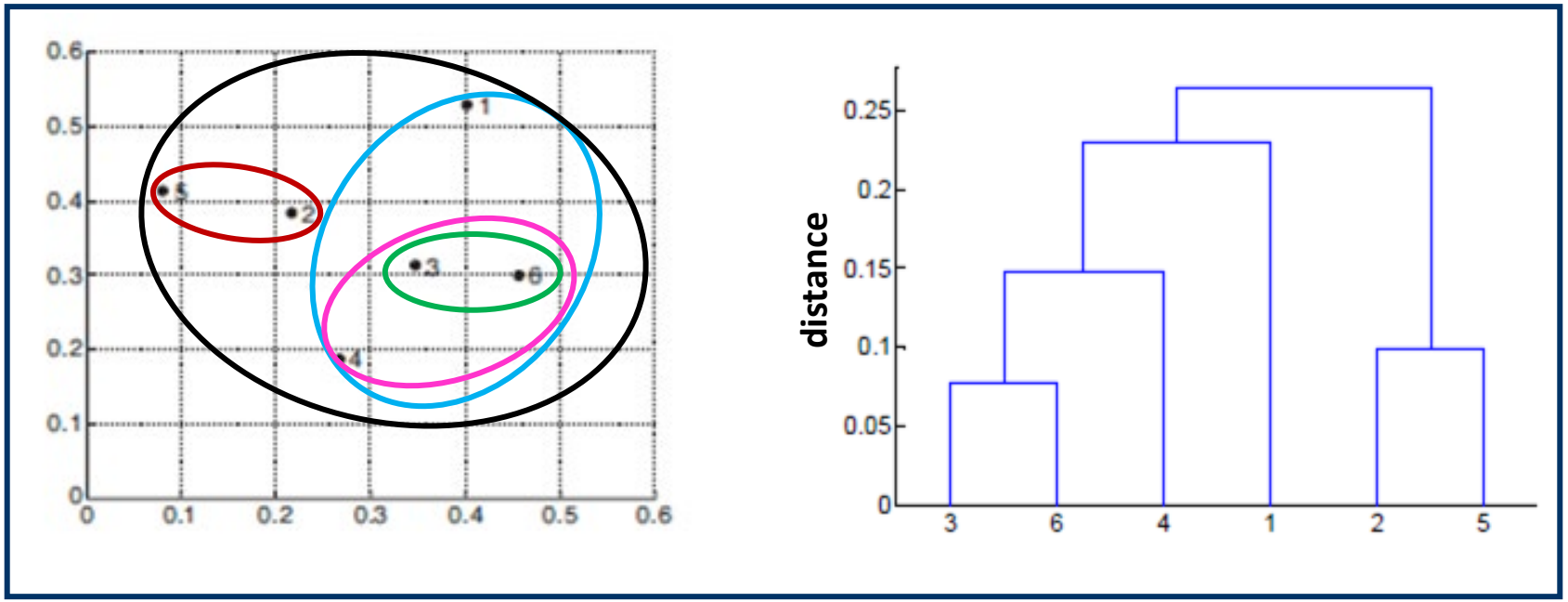
Average linkage method

The average linkage method considers all distances between the n_A objects of group **A** and the n_B objects of group **B** and then calculates their mean:

$$d_{AB} = \frac{1}{n_A \cdot n_B} \sum_{i=1}^{n_A} \sum_{j=1}^{n_B} d_{ij}$$

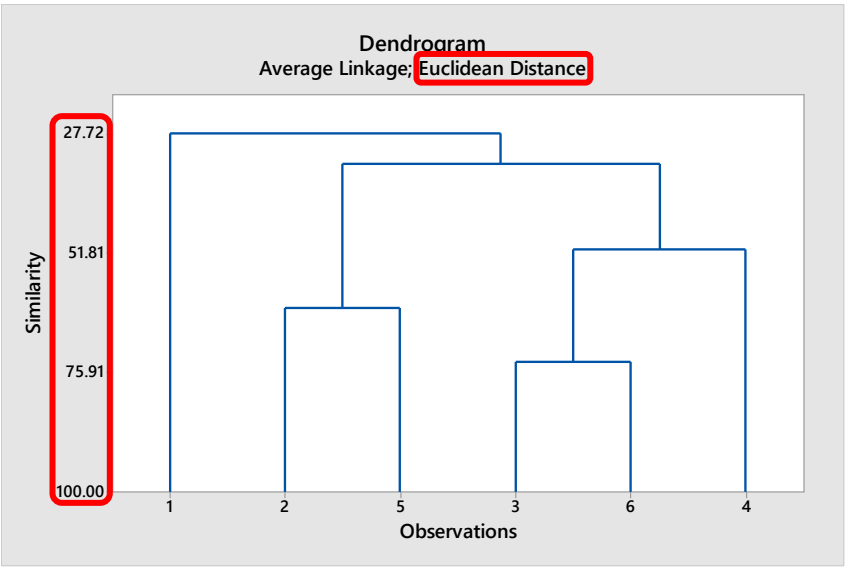


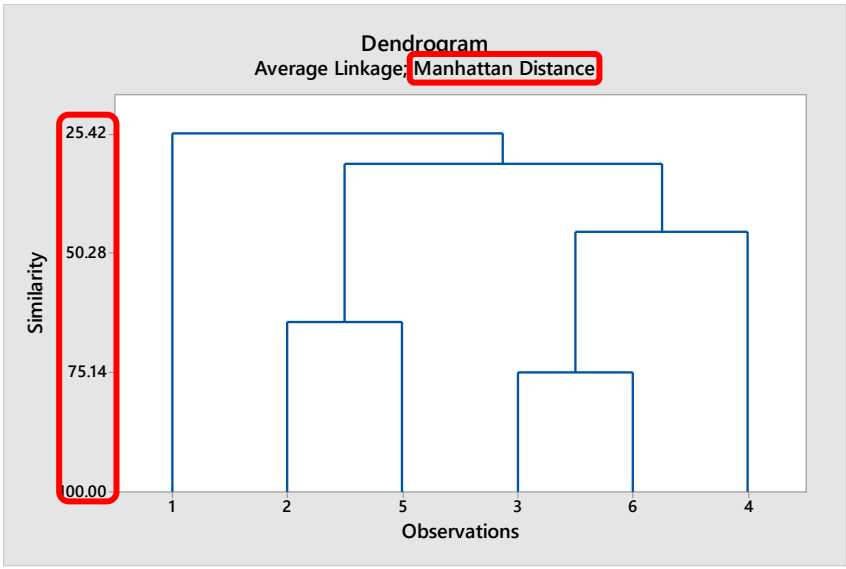
The outcome resulting from application of this approach to the already considered dataset can be represented graphically as follows:



The average linkage method represents a compromise between single and complete linkage methods.

Comparison between dendrograms obtained by Minitab 18 for the same dataset using the average linkage algorithm but changing the distance type:





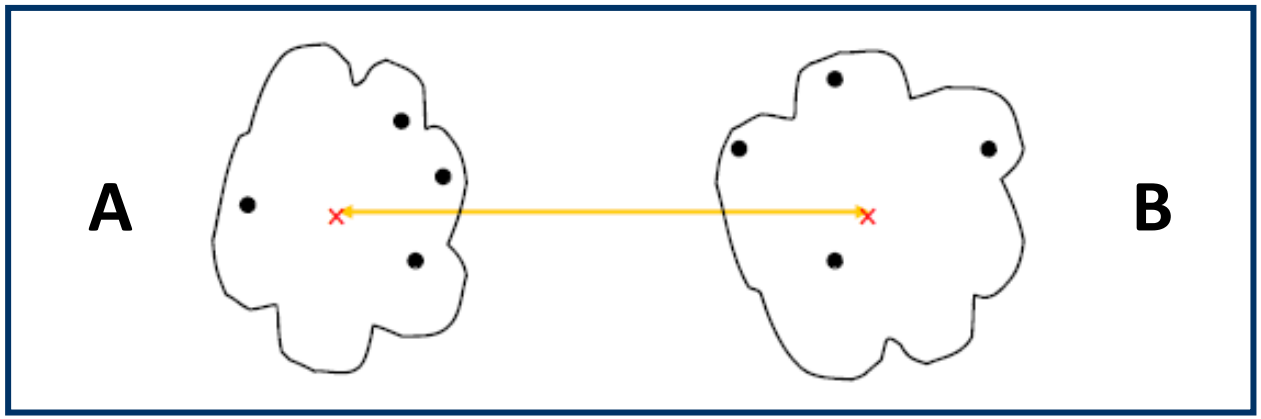
Note that, although the similarity values are slightly different, the clustering of the 6 objects is comparable in the two cases.

Centroid method

In the centroid method a centroid, *i.e.*, a point whose co-ordinates are the arithmetic means of those of all other objects belonging to the group, is defined.

The distance between two groups is thus coincident with the distance between the respective centroids:

$$d_{AB} = d(\bar{x}_A, \bar{x}_B)$$



When two groups are fused together, the centroid for the new group will be:

$$\bar{x}_{AB} = \frac{n_A \cdot \bar{x}_A + n_B \cdot \bar{x}_B}{n_A + n_B}$$

The Ward's method

The Ward's method, also called Ward's minimum variance method, introduced by the American statistician Joe H. Ward jr., aims at minimizing variance inside groups. Clustering is thus considered to be better when the resulting groups are more homogeneous internally and more different between each other.

The method starts by considering all n samples divided into n clusters of size 1 each, then n - 1 clusters are sequentially formed, one of size 2 and the remaining of size 1.

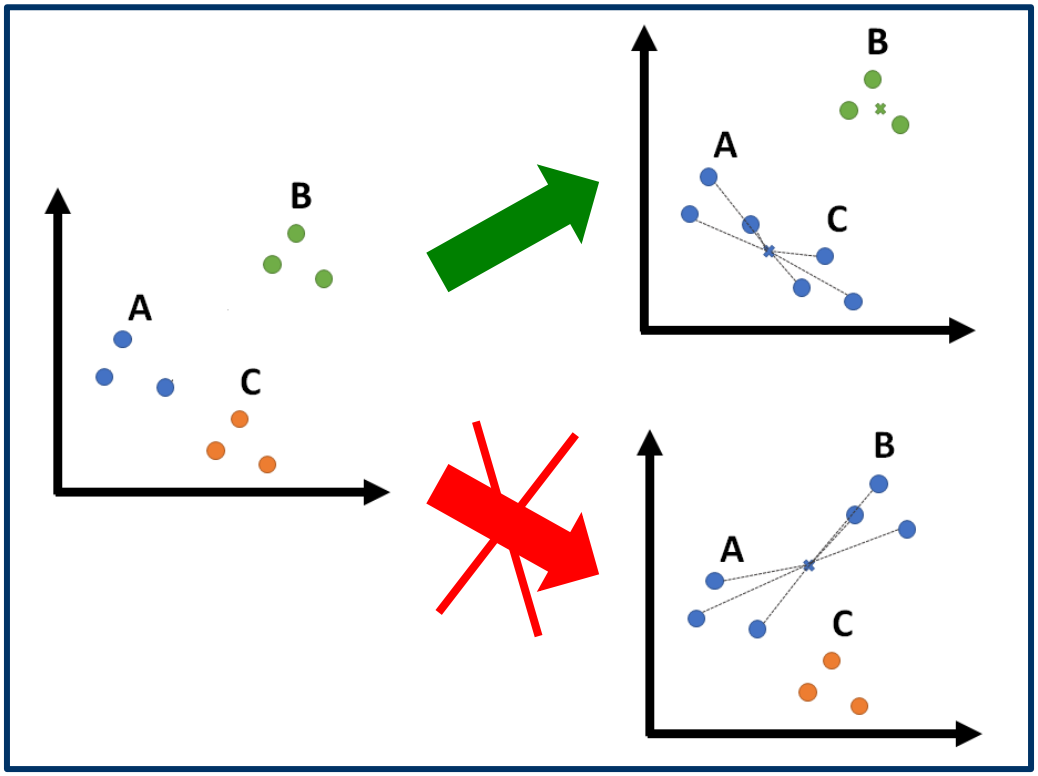
The error sum of squares (ESS), i.e., the sum of squared differences between the co-ordinates of each sample in the size 2 cluster and the corresponding mean, is calculated for each of the size 2 clusters. The pair of samples providing the smaller ESS will form the first cluster.

In the second step of the algorithm, n - 2 clusters are formed from the n - 1 clusters defined in the previous step. They may include two clusters of size 2, or a single cluster of size 3 including the two items clustered before. Again, the minimum value of ESS is searched for to decide how the new clusters are defined.

The algorithm stops when all samples are combined into a single large cluster of size n.

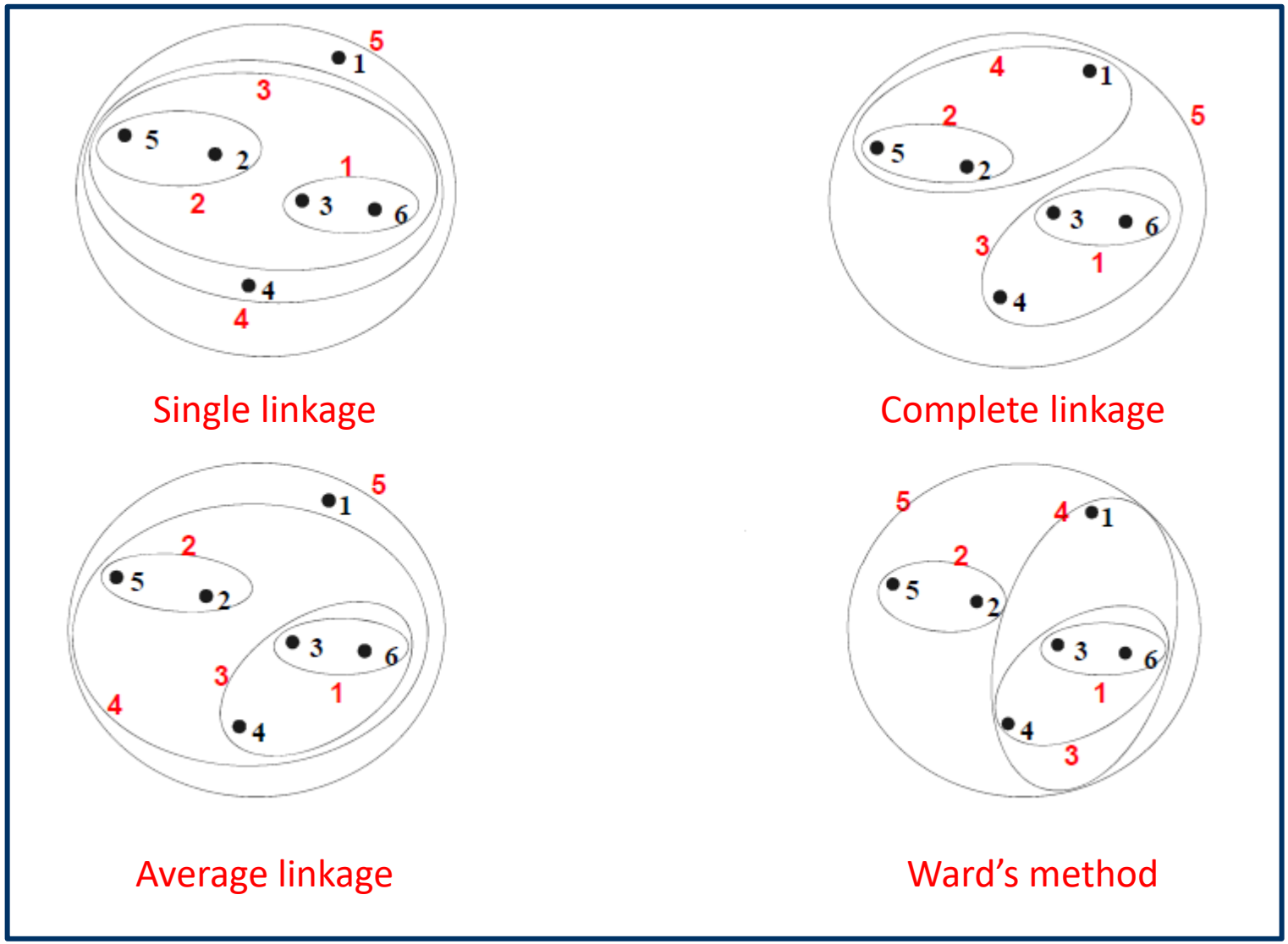
In the Ward's method the similarity between two clusters is based on the increase in squared error when two clusters are merged.

As an example, in the following figure three initial clusters (A, B and C) are shown in a bivariate dataset:



According to the Ward's algorithm cluster A has to be merged with cluster C since the resulting new cluster exhibits a lower dispersion of points around the centroid than the cluster resulting from the merging of clusters A and B.

A comparison between different types of hierarchical clustering applied to the same dataset shown before is reported in the following figure:



A step by step example of hierarchical clustering

Let us consider the following matrix including calcium and phosphate concentrations (mg/100 mL) observed in six individuals:

Sample	Calcium	Phosphate
1	8.0	5.5
2	8.25	5.75
3	8.7	6.3
4	10.0	3.0
5	10.25	4.0
6	9.75	3.5

Euclidean distance is chosen as the approach to measure distances.
As an example, the distance between samples 1 and 2 is:

$d_{12} = [(8.25-8)^2 + (5.75-5.5)^2]^{1/2} = [0.0625+0.0625]^{1/2} = 0.354.$

The matrix of distances is the following:

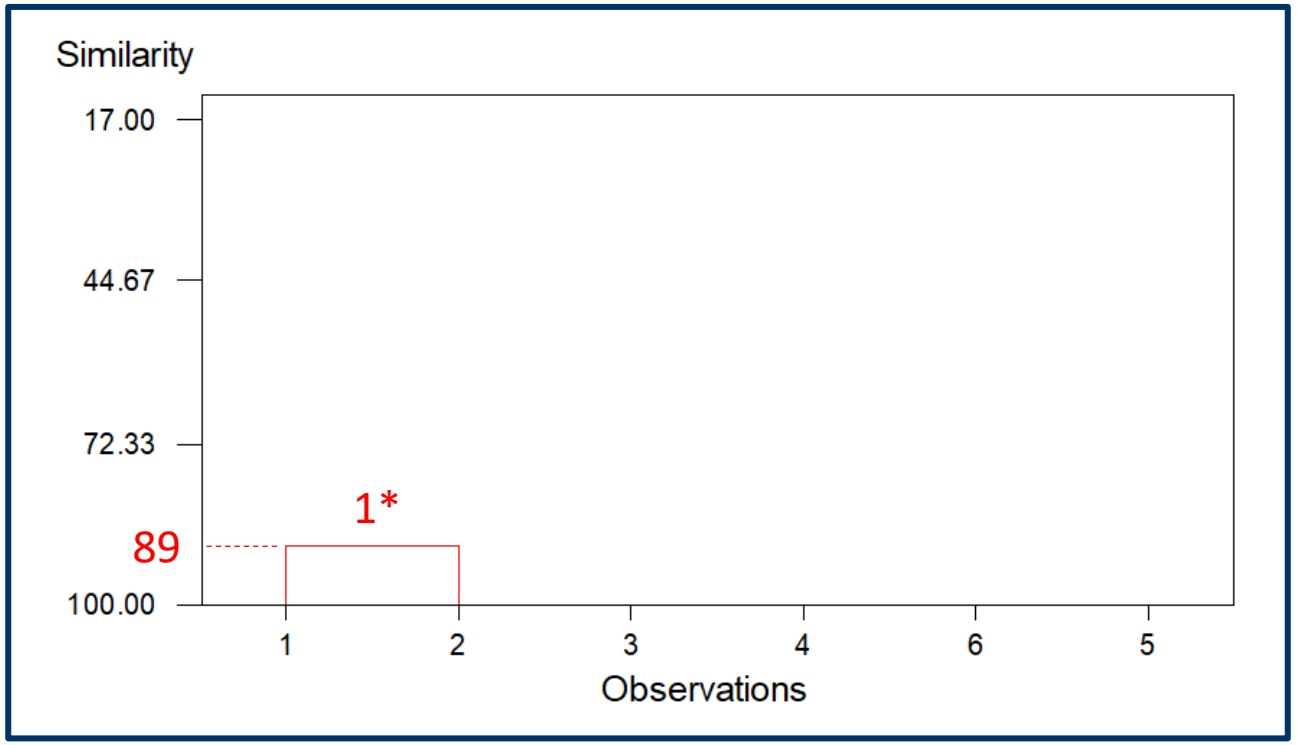
	1	2	3	4	5	6
1	0					
2	**0.354**	0				
3	1.063	0.711	0			
4	3.201	3.260	**3.347**	0		
5	2.704	2.658	2.774	1.301	0	
6	2.658	2.704	2.990	0.559	0.707	0

Let us suppose to use average linkage as the agglomerative clustering algorithm.

First, samples 1 and 2 are the closest ones. The level of similarity for these samples is:

$$S_{ij} = 1 - \frac{d_{ij}}{d_{ij}(\max)} = 1 - 0.354/3.347 = 0.89 = 89\%$$

The dendrogram can thus be constructed by considering samples 1 and 2 as a cluster (1*):



The distances between the 1* cluster and the other samples are:

$$\mathbf{d_{1*,3}} = (d_{1,3} + d_{2,3})/2 = (1.063 + 0.711)/2 = 0.887$$

$$\mathbf{d_{1*,4}} = (d_{1,4} + d_{2,4})/2 = (3.201 + 3.260)/2 = 3.231$$

$$\mathbf{d_{1*,5}} = (d_{1,5} + d_{2,5})/2 = (2.704 + 2.658)/2 = 2.681$$

$$\mathbf{d_{1*,6}} = (d_{1,6} + d_{2,6})/2 = (2.658 + 2.704)/2 = 2.681$$

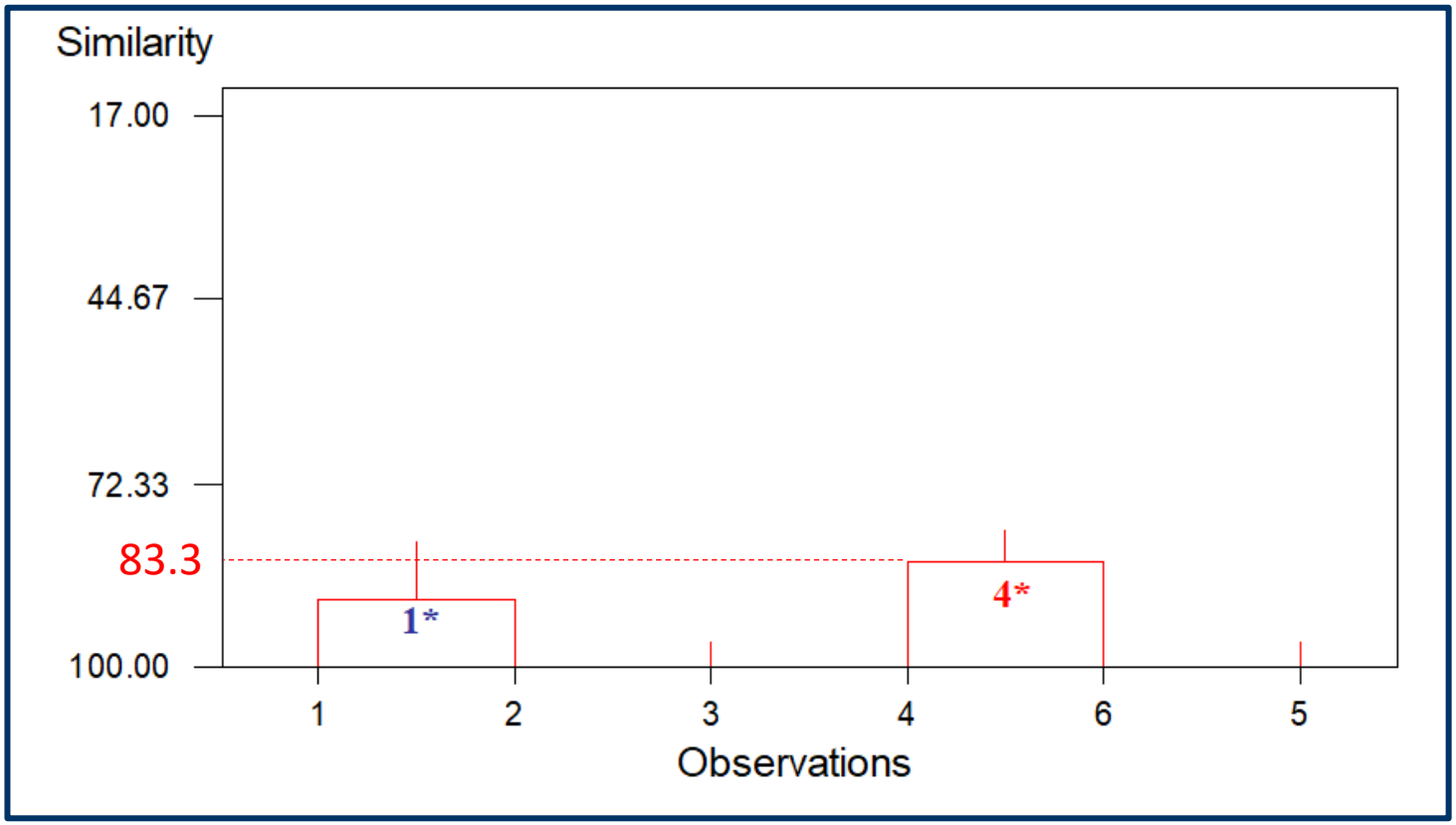
The first reduced matrix of distances is thus the following:

	1*	3	4	5	6
1*	0				
3	0.887	0			
4	3.231	3.347	0		
5	2.681	2.774	1.031	0	
6	2.681	2.990	0.559	0.707	0

As apparent, samples 4 and 6 are the closest ones among other samples, thus they can be fused in a unique cluster, 4*, with a level of similarity:

$$S_{ij} = 1 - \frac{d_{ij}}{d_{ij}(\max)} = 1 - 0.559/3.347 = 0.833 = 83.3\%$$

The dendrogram becomes:



Distances involving cluster 4* can now be calculated:

$$\mathbf{d_{1^*,4^*}} = (d_{1^*,4} + d_{6,1^*})/2 = (3.231 + 2.681)/2 = 2.956$$

$$\mathbf{d_{3,4^*}} = (d_{3,4} + d_{3,6})/2 = (3.347 + 2.990)/2 = 3.169$$

$$\mathbf{d_{5,4^*}} = (d_{4,5} + d_{5,6})/2 = (1.031 + 0.707)/2 = 0.869$$

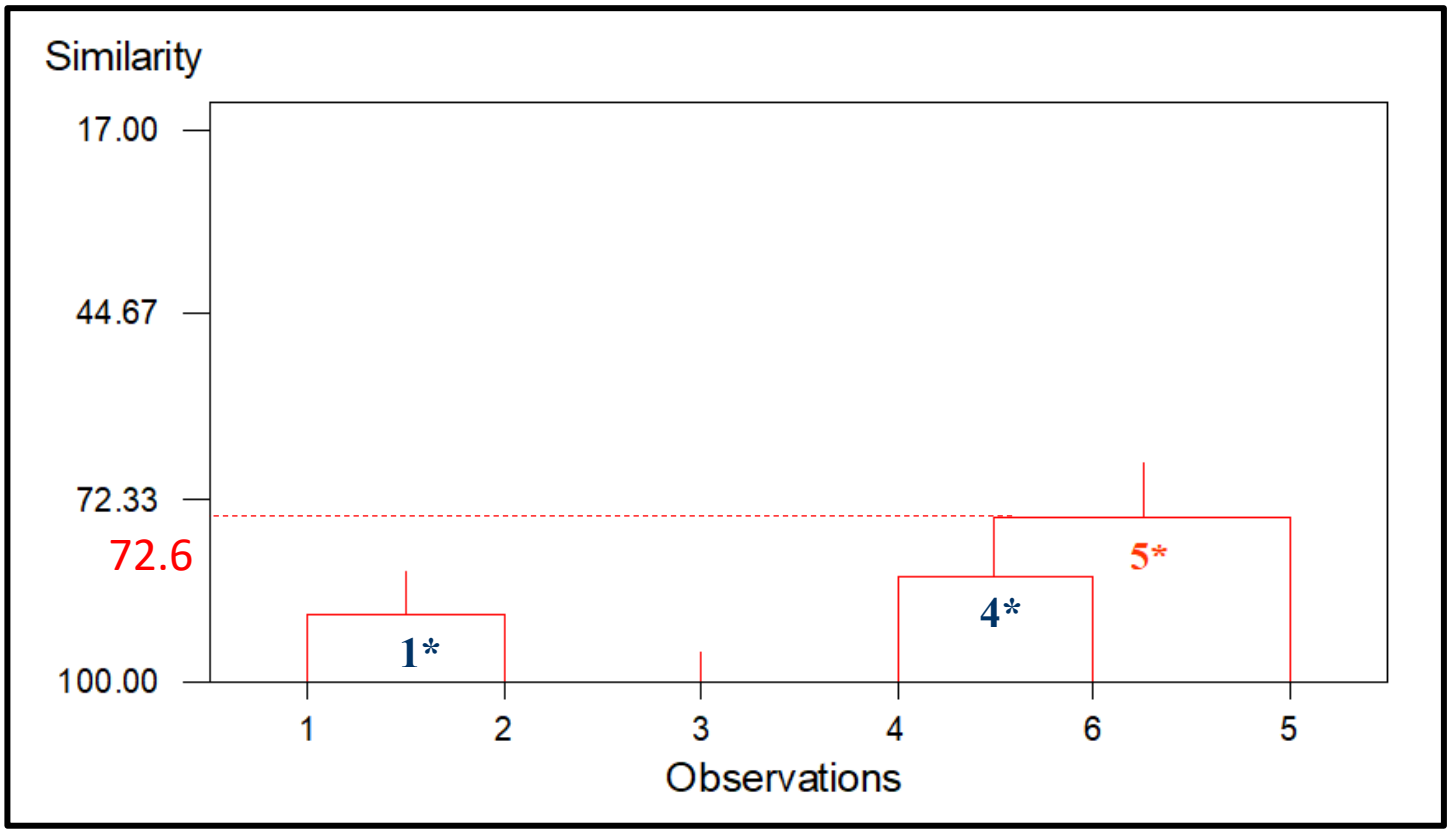
The second reduced matrix of distances can be obtained afterwards:

	1*	3	4*	5
1*	0			
3	0.887	0		
4*	2.956	3.169	0	
5	2.681	2.774	0.869	0

Cluster 4* and sample 5 are the closest objects, thus they can be fused in a new cluster, 5*, with a similarity:

$$S_{ij} = 1 - \frac{d_{ij}}{d_{ij}(\max)} = 1 - 0.869/3.169 = 0.726 = 72.6\%$$

The dendrogram becomes:



The distances between cluster 1* or sample 3 and cluster 5* are:

$$\mathbf{d_{1*,5*}} = (d_{1*,5} + d_{1*,4*})/2 = (2.681 + 2.956)/2 = 2.819$$

$$\mathbf{d_{3,5*}} = (d_{3,5} + d_{3,4*})/2 = (2.774 + 3.169)/2 = 2.972$$

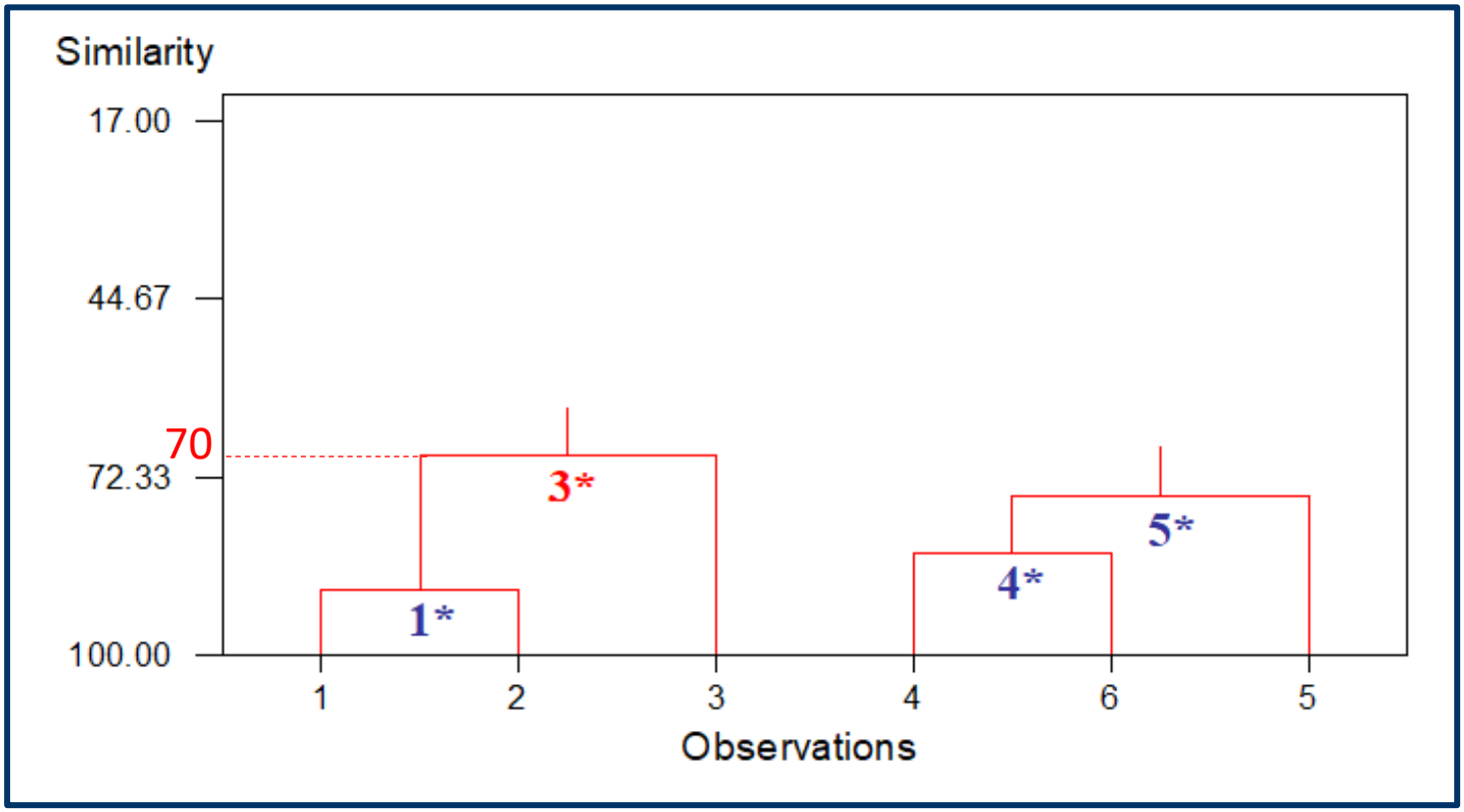
The third reduced matrix of distances can be obtained afterwards:

	1*	3	5*
1*	0		
3	0.887	0	
5*	2.819	2.972	0

Cluster 1* and sample 3 are now the closest objects, thus they are fused in the new cluster 3*, with a similarity index:

$$S_{ij} = 1 - \frac{d_{ij}}{d_{ij}(\max)} = 1 - 0.887/2.972 = 0.70 = 70\%$$

The dendrogram becomes:



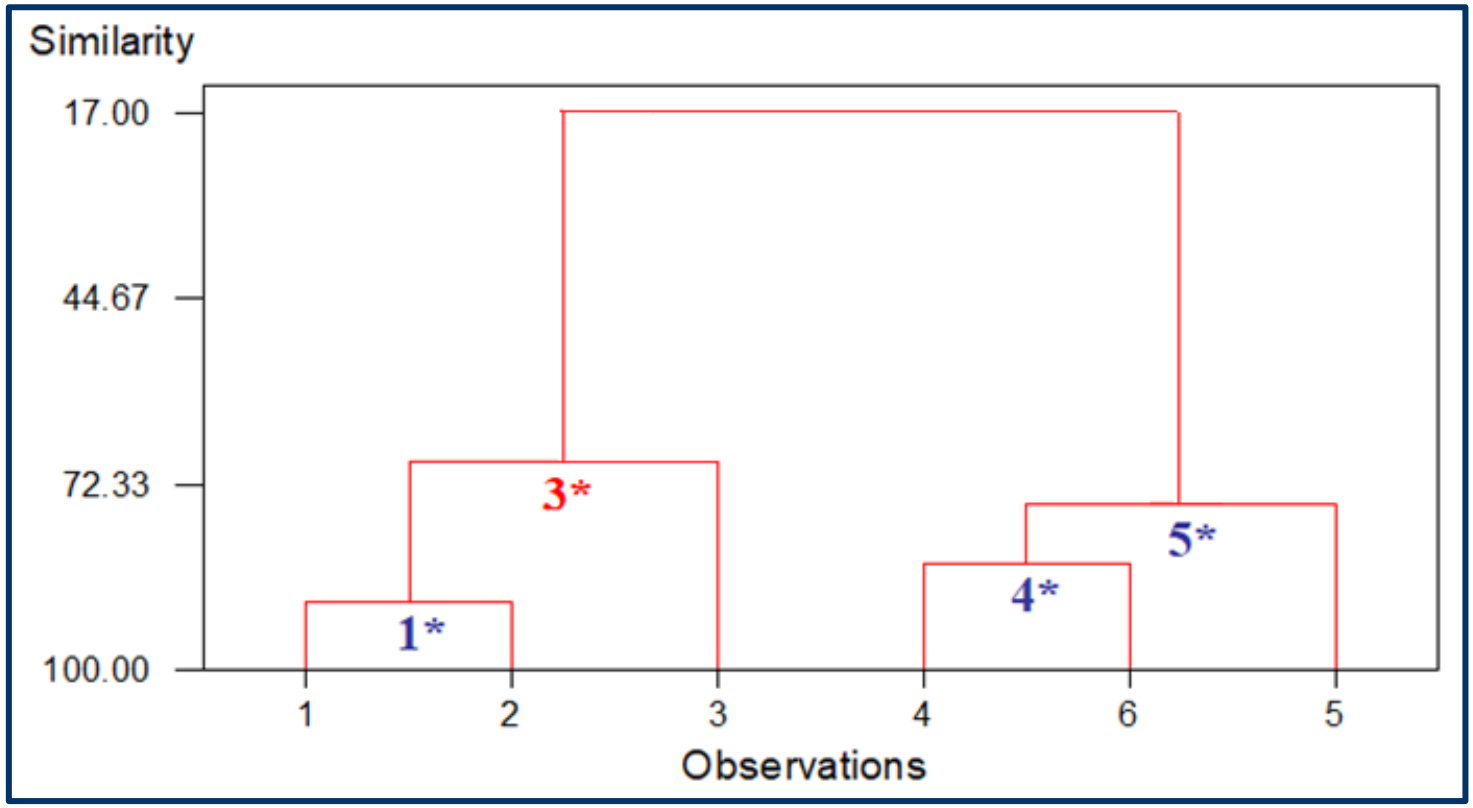
The distance between clusters 3* and 5* is:

$$\mathbf{d_{3^*,5^*}} = (d_{1^*,5^*} + d_{3,5^*})/2 = (2.819 + 2.972)/2 = 2.896$$

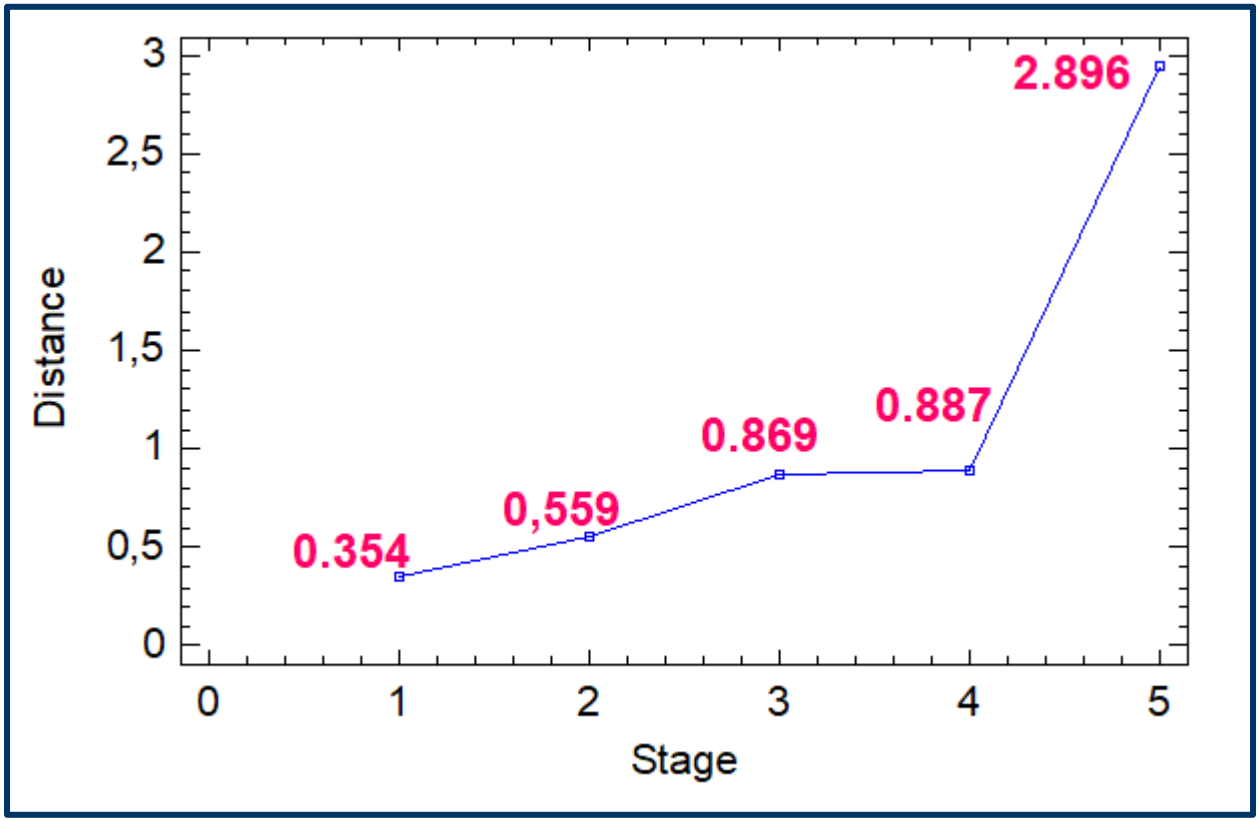
The fourth reduced matrix of distances is:

	3*	5*
3*	0	
5*	2.896	0

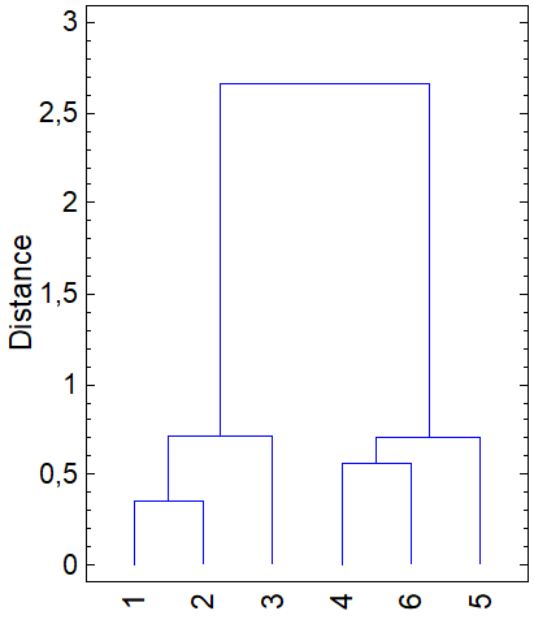
The final dendrogram is:



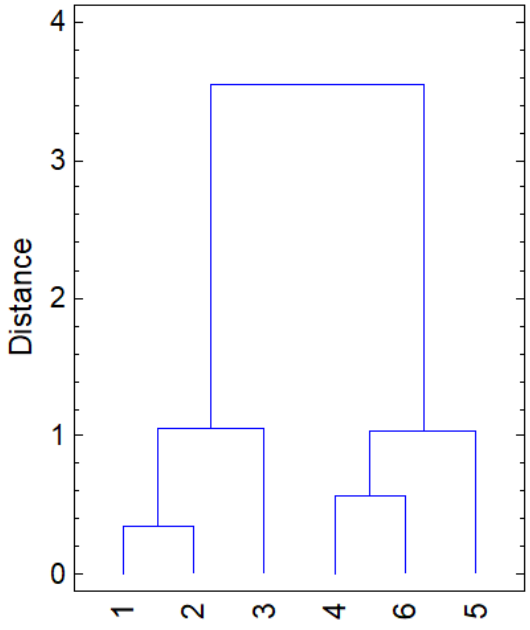
A plot of average distance as a function of clustering stage can be generated to emphasize the evolution of agglomerative clustering:



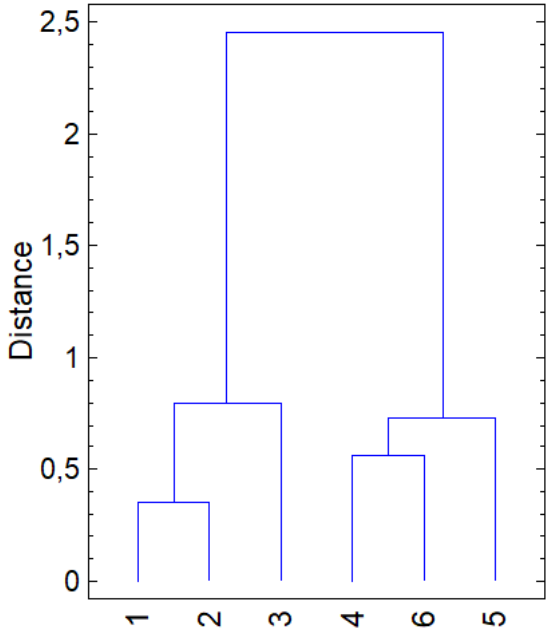
A comparison of dendrograms resulting from Cluster Analysis based on the Euclidean distance but different clustering algorithms, namely single and complete linkage and centroid methods, emphasizes that, although distances (or similarities) can be slightly different, a similar outcome is obtained, compared to the average linkage method:



Single linkage



Complete linkage

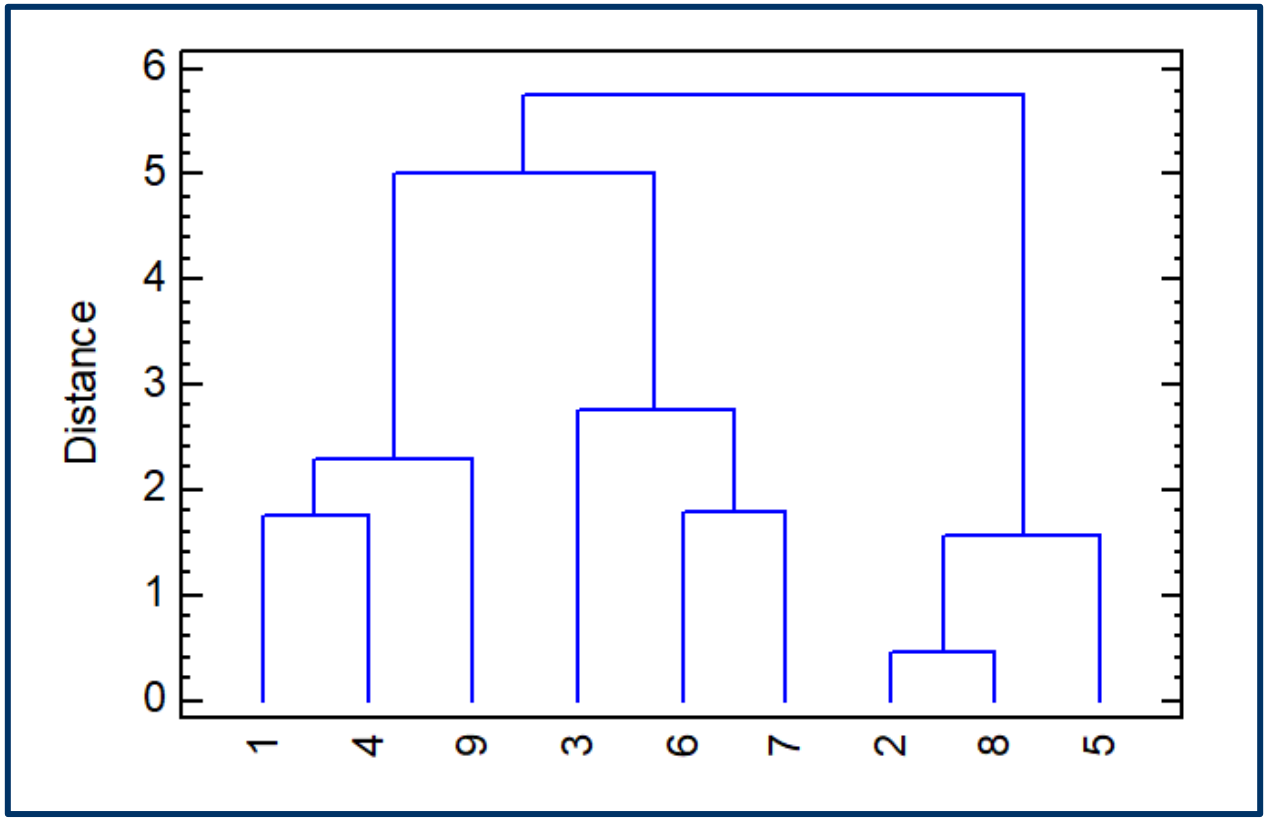


Centroid

A further example of hierarchical clustering

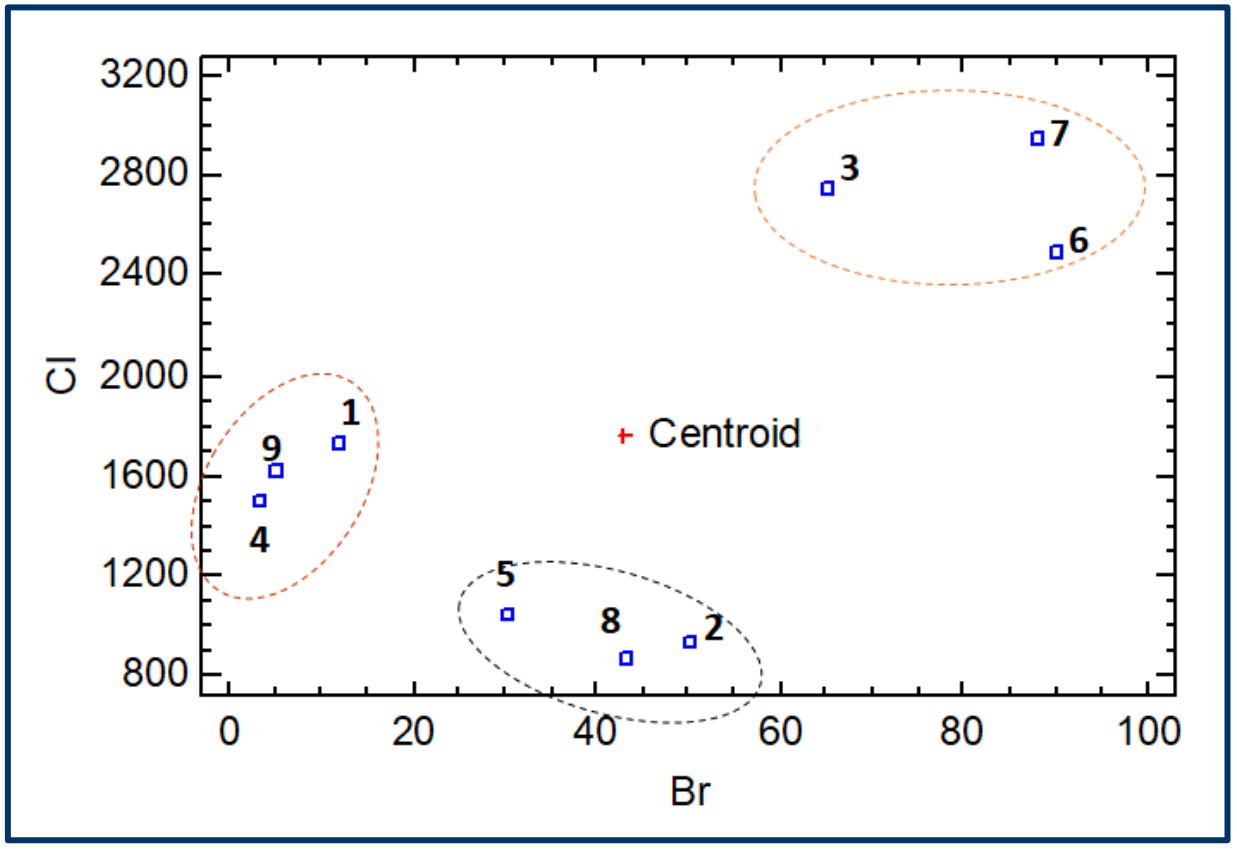
Let us re-consider the set of Cu, Mn, Cl, Br and I concentrations in 9 hair samples.

The dendrogram obtained using the Euclidean distance and the single linkage algorithm is the following:



The so called *cluster scatterplot* can be also generated using some programs, like Statgraphics, to emphasize the position of samples in a 2D graph based on values of two selected variables.

In the following example, the scatterplot based on Cl and Br concentrations is reported:



The three intermediate clusters of samples observed in the dendrogram are shown. Notably, the cluster scatterplot is quite similar, in terms of sample clustering, to the score plot referred to the first two principal components obtained using PCA.

Non hierarchical clustering methods

Non hierarchical methods are based on aggregative algorithms that produce a single partition, i.e., a division into separate clusters of the original dataset, starting from a set of initial centers.

At each step of the algorithm they reconsider the partition previously obtained; indeed, clusters obtained are erased and the aggregation process restarts from new centers.
Differently from hierarchical methods, the assignment of an object to a cluster is not irrevocable.

Non hierarchical methods are less demanding in terms of calculations than hierarchical ones, since the number of clusters is predetermined, based on the optimization of a parameter.

One of the most common methods among non hierarchical clustering approaches is known as K-means.

K-means algorithm

The K-means algorithm is based on the following steps:

1. the number, k, of clusters that will be individuated is chosen;

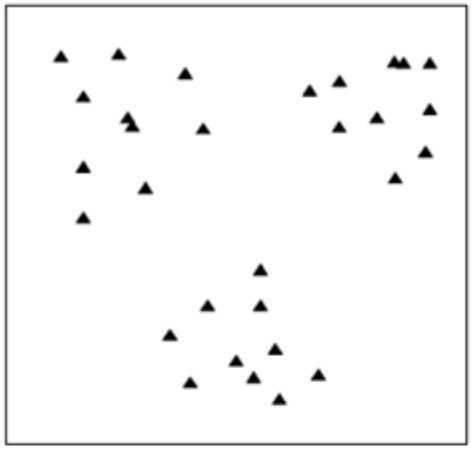
2. k values of the dataset are selected, usually randomly, and considered the centroids of the k clusters;

3. Euclidean distance with respect to the selected centroids is exploited to assign the remaining data to the clusters;

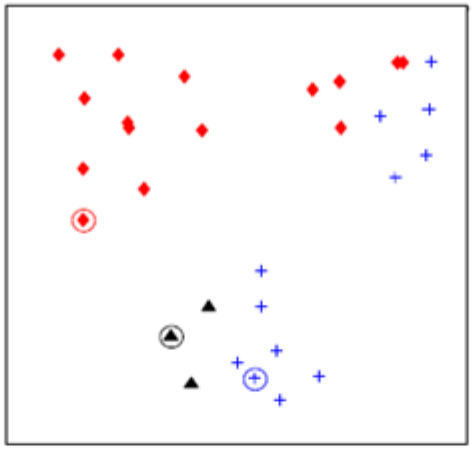
4. the co-ordinates of the new centroids are calculated from data referred to objects belonging to each cluster;

5. if the new centroids are equal to those calculated previously the procedure stops, otherwise steps from 3 to 5 are repeated and a new evaluation is made.

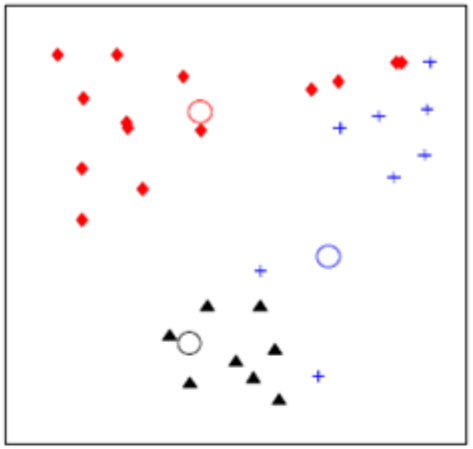
A graphical description of the k-means algorithm for objects represented by two co-ordinates is shown in the following figure:



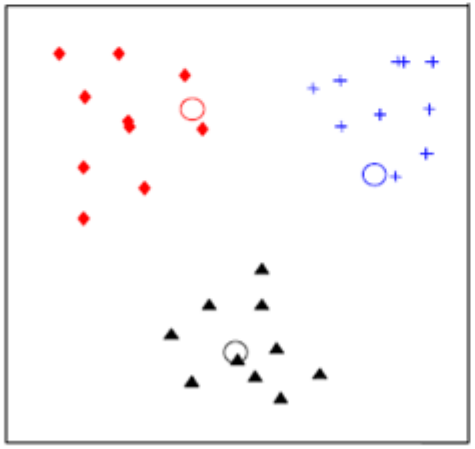
k = 3 is selected as the number of possibile clusters



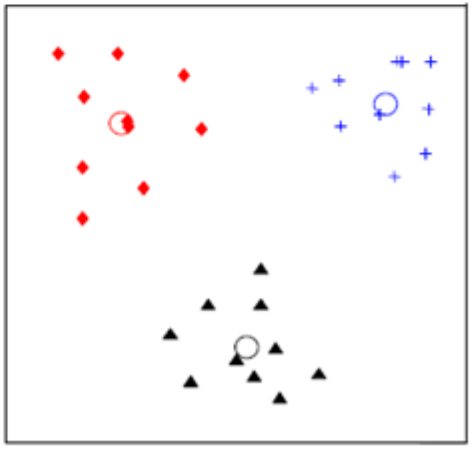
random selection of the first three centroids and first iteration



second iteration



third iteration

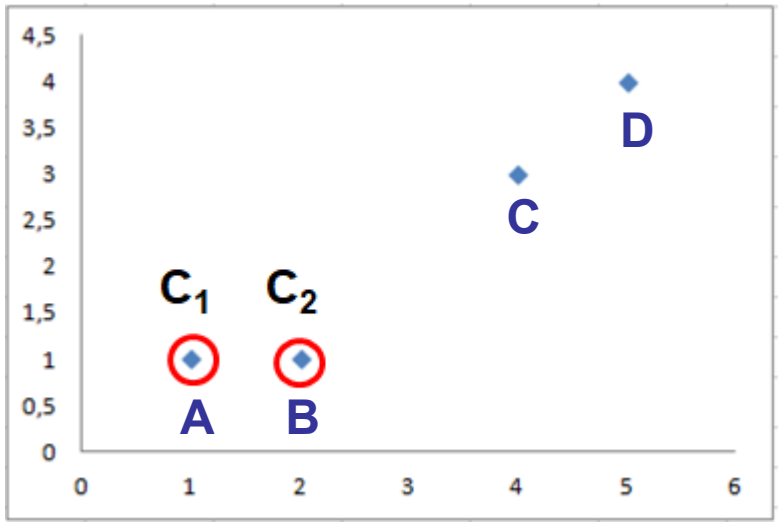


final clustering

A numerical example of k-means method: k = 2

Let us consider the following dataset, with its graphical representation:

	x	y
A	1	1
B	2	1
C	4	3
D	5	4



A value k = 2 is selected and points A and B are chosen as the initial centroids:

$\mathbf{C_1}$ = (1, 1) and $\mathbf{C_2}$ = (2, 1)

The matrix of distances from the centroids is:

	C1	C2
A	0	1
B	1	0
C	3,61	2,83
D	5	4,24

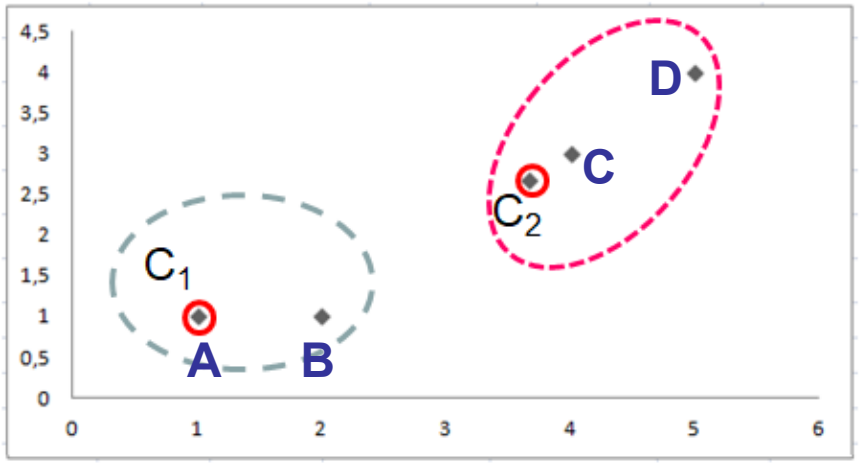
Consequently, points A is assigned to cluster 1, whereas points B, C and D are assigned to cluster 2.

The new centroids are:

$\mathbf{C_1}$=(1,1); $\mathbf{C_2}$= ((2+4+5)/3, (1+3+4)/3) = (3.67, 2.67)

The new matrix of distances and the graphical representation of clusters and centroids are:

	C1	C2
A	0	3,14
B	1	2,36
C	3,61	0,47
D	5	1,89



According to this matrix, points A and B are assigned to cluster 1 and points C and D to cluster 2.

The within-cluster sum of squares: $J(c_k) = \sum_{x_i \in c_k} \| x_i - \mu_k \|^2$

can be exploited to decide when iterations can be stopped, since it is a measure of the spread of points in a cluster around its centroid.

In the specific case $J(\mathbf{C_1}) = 1$ and $J(\mathbf{C_2}) = 3.79$, thus the total sum of squares is:

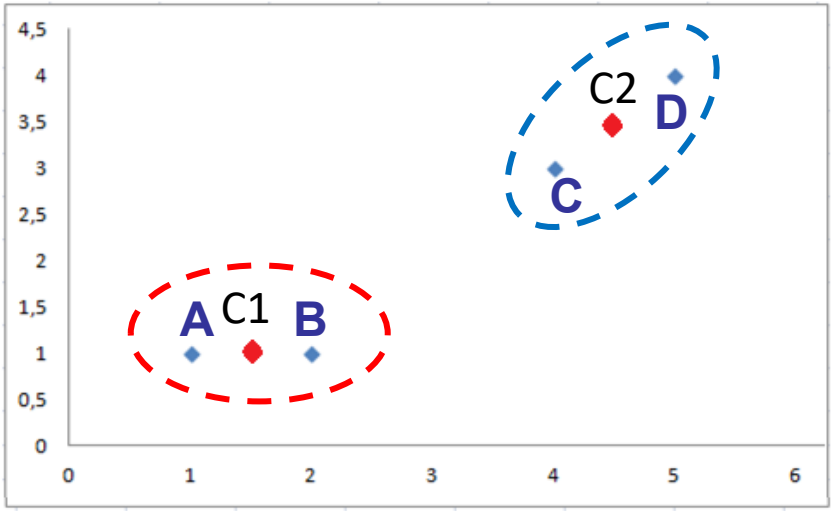
$$J(C) = \sum_{k=1}^{K} \sum_{x_i \in c_k} \| x_i - \mu_k \|^2 = 4.79$$

The new centroids are:

$\mathbf{C_1}$= ((1+2)/2, (1+1)/2) = (1.5,1.0); C2= ((4+5)/2, (3+4)/2) = (4.5,3.5)

The new matrix of distances and the graphical representation of clusters and centroids are:

	C1	C2
A	0.5	4.30
B	0.5	3.54
C	3.20	0.71
D	4.61	0.71

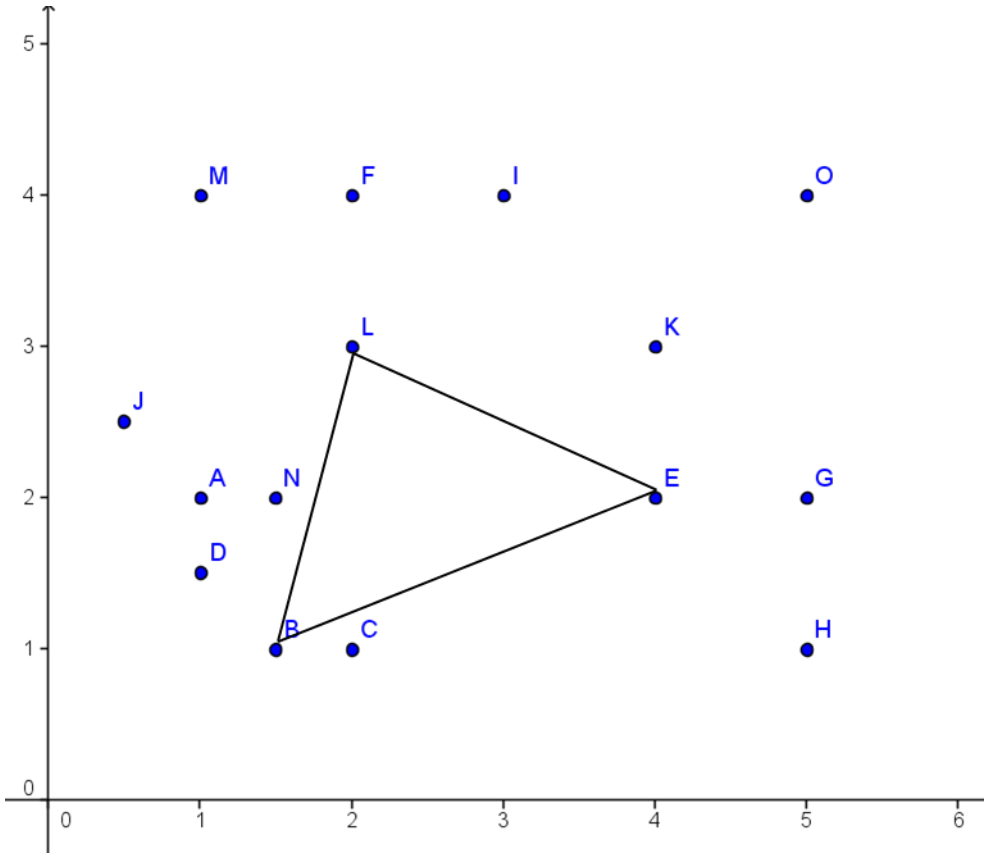


The same assignment obtained before is observed, moreover the J(C) value is 1.582, i.e., lower than the previous one. The procedure can then be stopped.

A numerical example of k-means method: k = 3

Let us consider the following dataset, with its graphical representation:

OBJECT	X	Y
A	1	2
B	1,5	1
C	2	1
D	1	1,5
E	4	2
F	2	4
G	5	2
H	5	1
I	3	4
J	0,5	2,5
K	4	3
L	2	3
M	1	4
N	1,5	2
O	5	4



Points B, E and L are chosen as the centroids of the initial clusters.

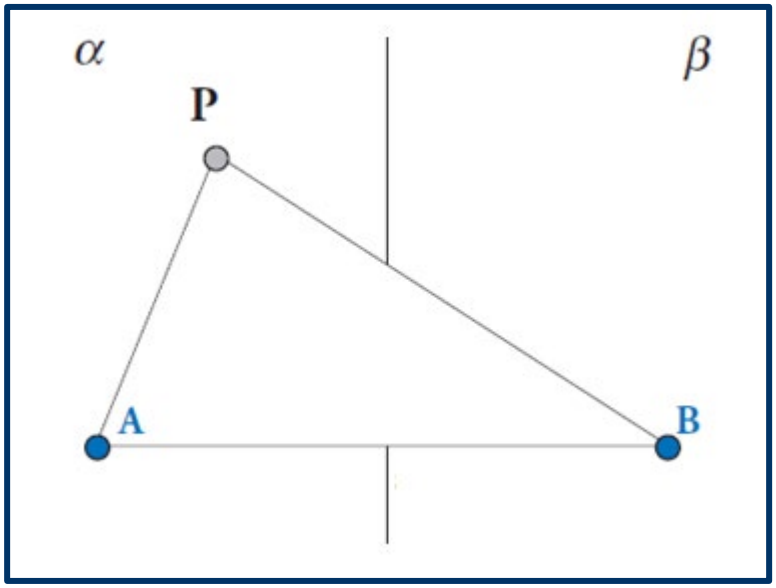
The assignment of all the other points to one of the three clusters is based on their Euclidean distances from the centroids.

A faster approach is based on the properties of the axis of an Euclidean segment.

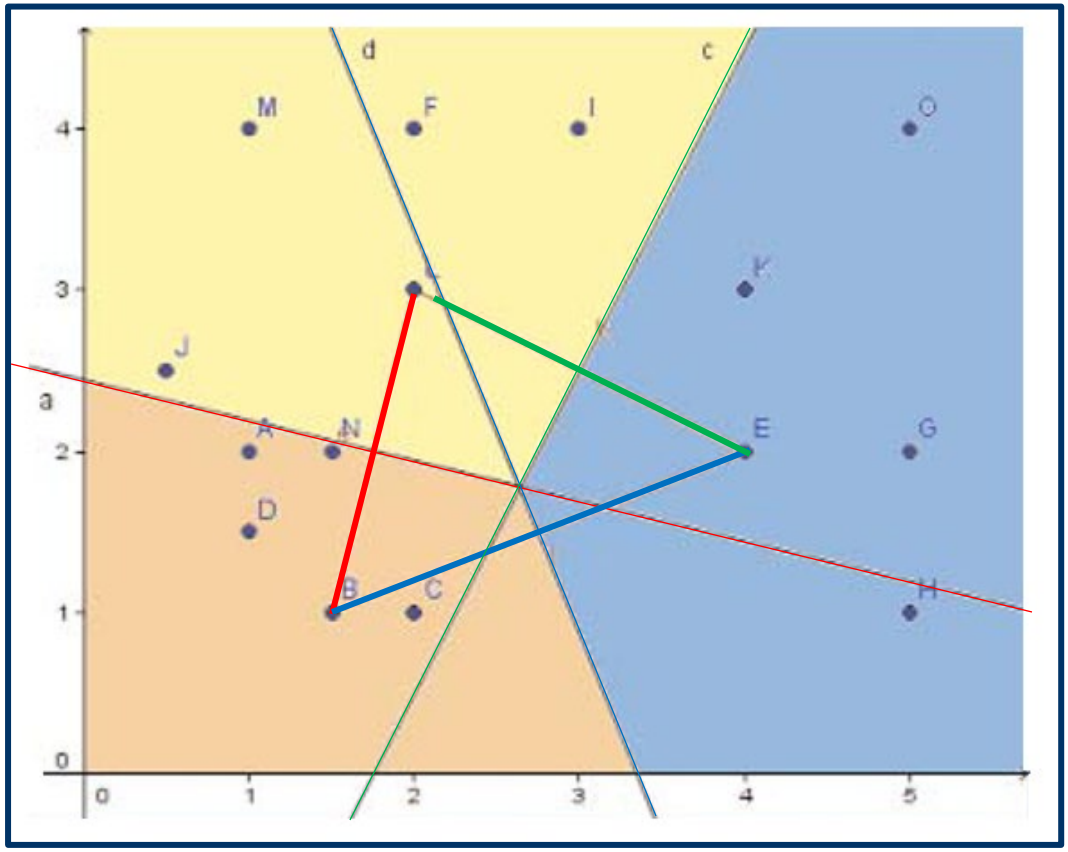
Indeed, whatever point is considered on the axis of a segment, its distance from each end of the segment end is the same.

Consequently, the axis divides the plane into two semiplanes, α and β.

A point located in the α semiplane (like point P in the figure) is thus closer to point A, whereas a point located in the β semiplane is closer to point B.



In the following figure, the axes (a, c and d) of the three segments connecting the initial centroids are drawn and the assignments of all the remaining points to one of the three possible clusters are emphasized by a colour code.



As an examples, point J is closer to centroid L since it is on the left semiplane with respect to axis a, perpendicular to segment BL; on the contrary, points A, N and D are in the right semiplane, thus they can be assigned to the cluster related to centroid B.

Points included in the three clusters, reported with their co-ordinates, are the following:

Cluster 1

	x	y
A	1	2
B	1.5	1
C	2	1
D	1	1.5
N	1.5	2

Cluster 2

	x	y
K	4	3
G	5	2
H	5	1
E	4	2
O	5	4

Cluster 3

	x	y
F	2	4
I	3	4
J	0.5	2.5
L	2	3
M	1	4

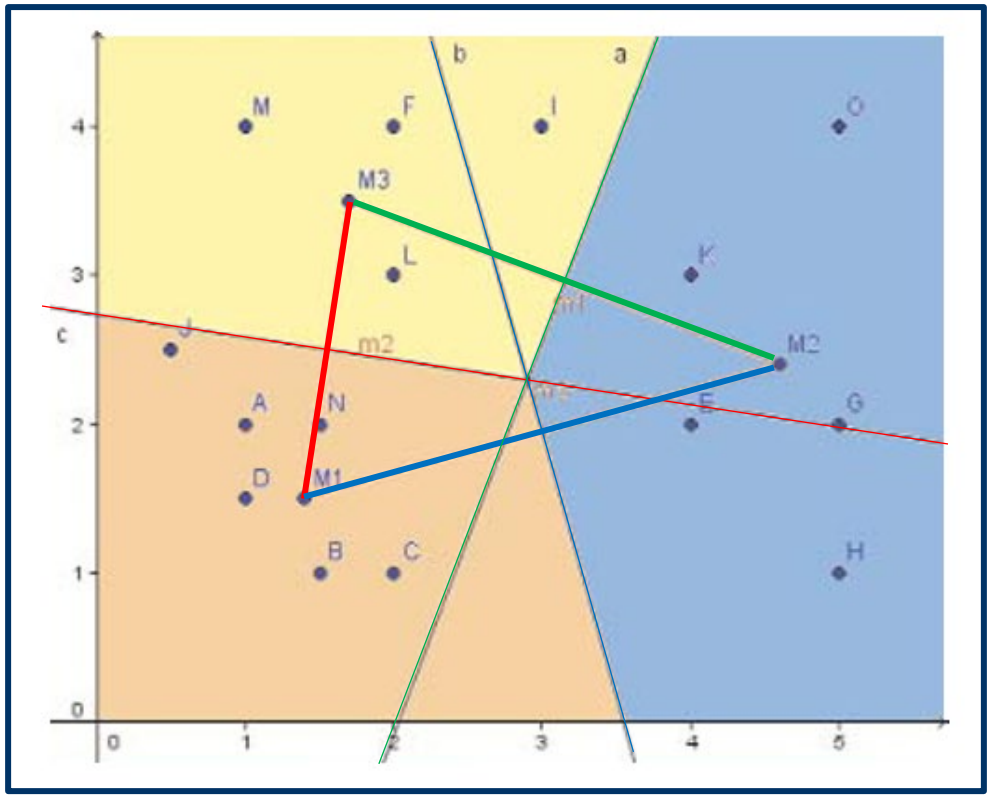
The new centroids can be obtained from the following calculations:

$$\begin{cases} M_{1,x} = (1+1.5+2+1+1.5)/5 = 1.4 \\ M_{1,y} = (2+1+1+1.5+2)/5 = 1.5 \end{cases}$$

$$\begin{cases} M_{2,x} = (4+5+5+4+5)/5 = 4.6 \\ M_{2,y} = (3+2+1+2+4)/5 = 2.4 \end{cases}$$

$$\begin{cases} M_{3,x} = (2+3+0.5+2+1)/5 = 1.7 \\ M_{3,y} = (4+4+2.5+3+4)/5 = 3.5 \end{cases}$$

Starting from new centroids M_1, M_2 and M_3 the new assignments of points are:



Cluster 1 - M_1 = (1.4, 1.5)

	x	y
A	1	2
B	1.5	1
C	2	1
J	0.5	2.5
D	1	1.5
N	1.5	2

Cluster 2 - M_2 = (4.6, 2.4)

	x	y
K	4	3
G	5	2
H	5	1
E	4	2
O	5	4

Cluster 3 - M_3 = (1.7, 3.5)

	x	y
F	2	4
I	3	4
L	2	3
M	1	4

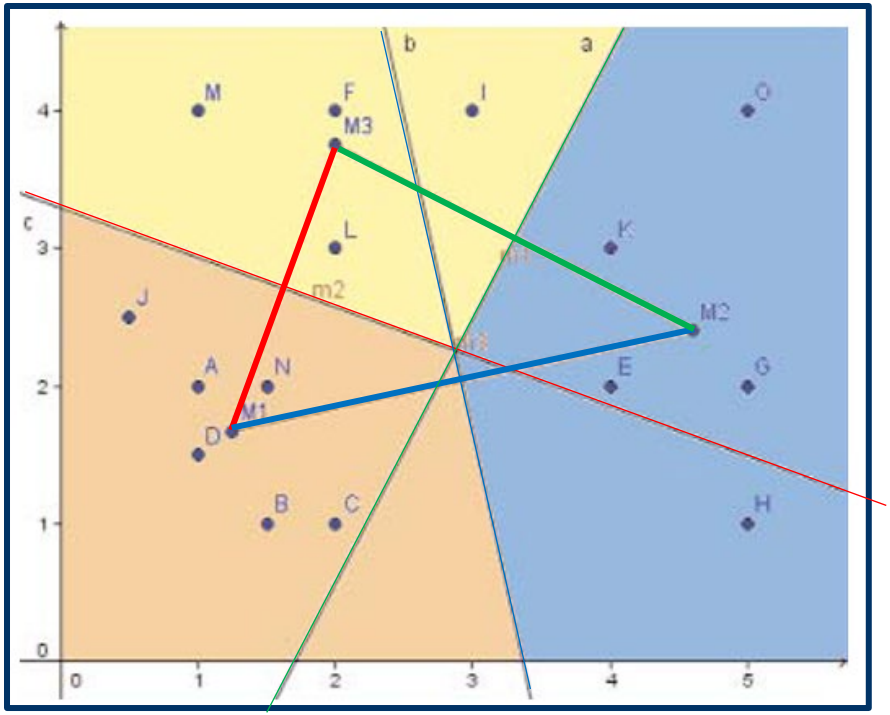
The new centroids can be obtained from the following calculations:

$$\begin{cases} M_{1,x} = (1+1.5+2+0.5+1+1.5)/6 = 1.25 \\ M_{1,y} = (2+1+1+2.5+1.5+2)/6 = 1.67 \end{cases}$$

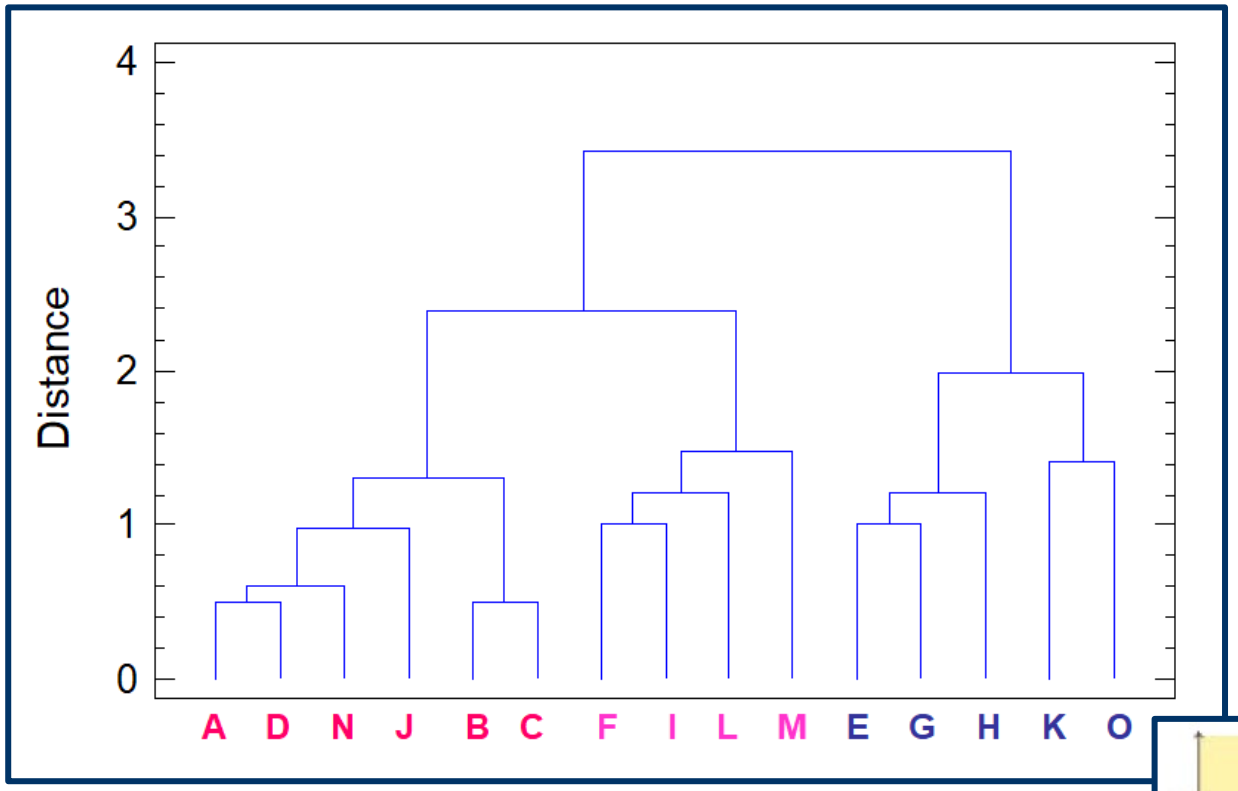
$$\begin{cases} M_{2,x} = (4+5+5+4+5)/5 = 4.6 \\ M_{2,y} = (3+2+1+2+4)/5 = 2.4 \end{cases}$$

$$\begin{cases} M_{3,x} = (2+3+2+1)/4 = 2 \\ M_{3,y} = (4+4+3+4)/4 = 3.75 \end{cases}$$

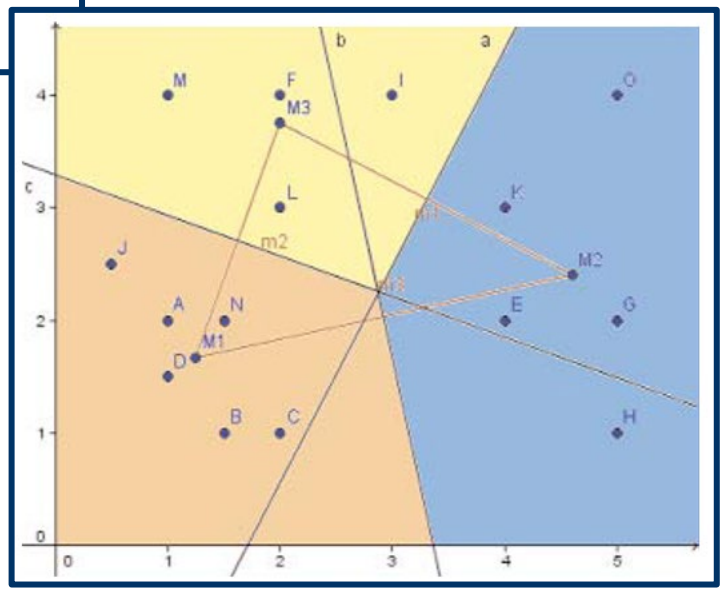
As shown by the figure on the right, the assignments of points are the same obtained in the previous iteration, so the procedure can be stopped.



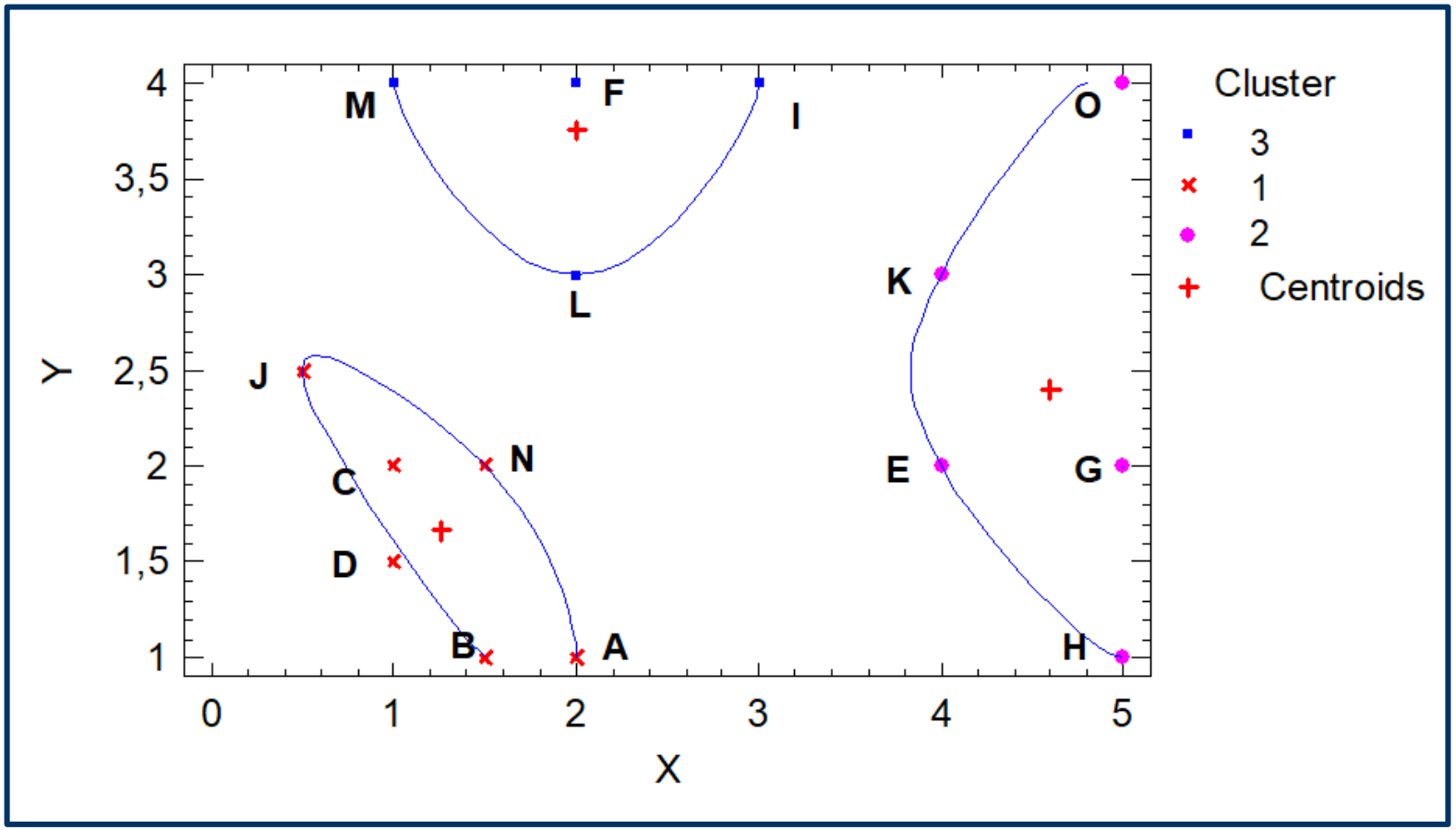
The dendrogram obtained from HCA based on group average method is the following:



As apparent, the two approaches provided the same final result in terms of clustering.



The cluster scatterplot arising from the application of the k-means method is the following:



An analytical application of Hierarchical Cluster Analysis

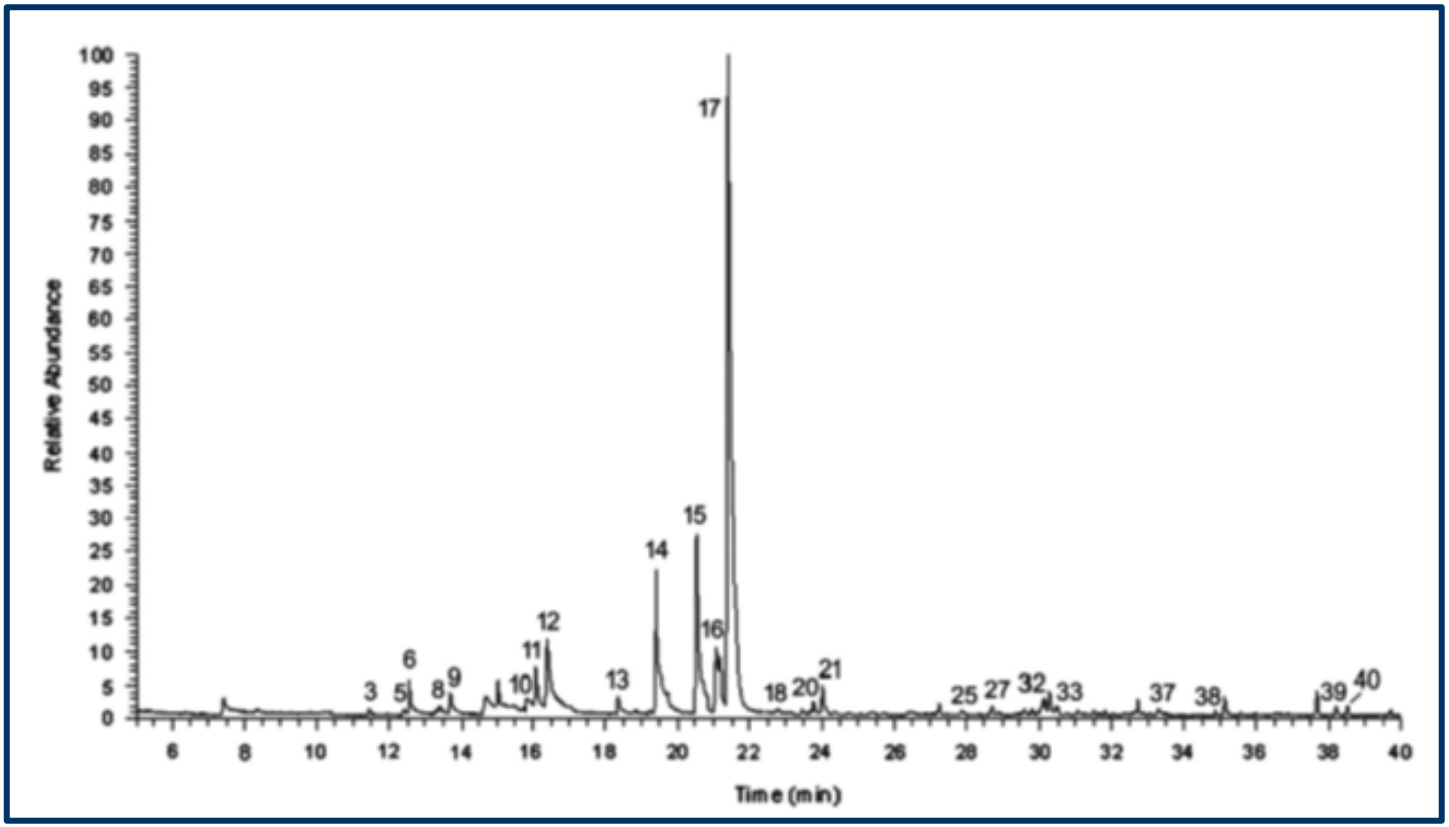
Cluster Analysis, together with Principal Components Analysis, was employed to characterize the geographical origin of amber samples.

The analytical approach adopted to collect data is Head Space – Solid Phase Microextraction – Gas Chromatography – Mass Spectrometry (HS-SPME-GC-MS).

In particular, amber samples were closed into glass vials with screw caps and then thermostated at 70°C for 10 min; afterwards, the fiber of a SPME device, coated with carboxen (carbon molecular sieve adsorbent resin) – polydimethylsiloxane, was inserted into the headspace and kept there for 15 min, in order to extract volatile components released from amber samples.

Volatile components were subsequently extracted thermally from the fiber, separated by GC and identified using Mass Spectrometry.

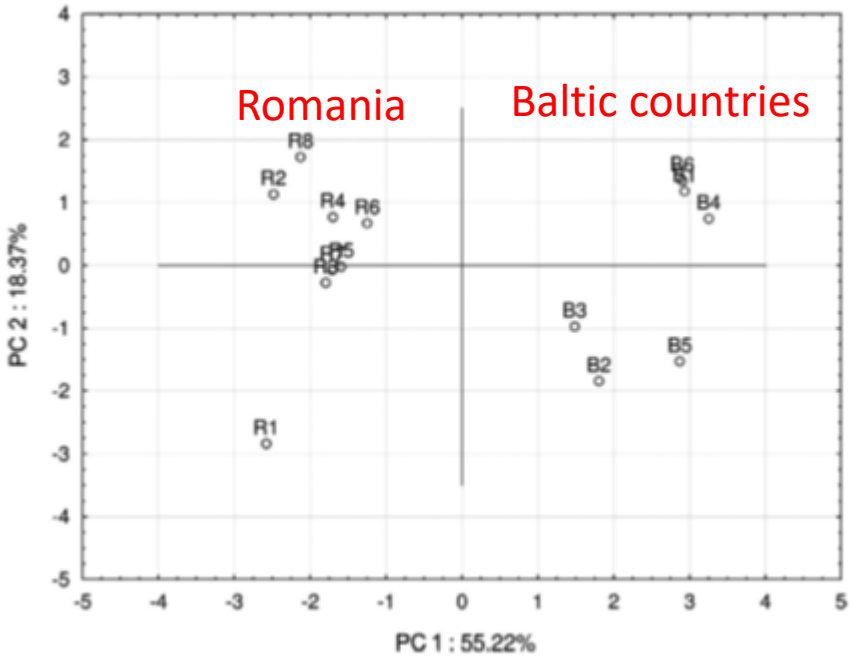
An example of GC-MS chromatogram obtained from an amber sample is shown in the following figure:



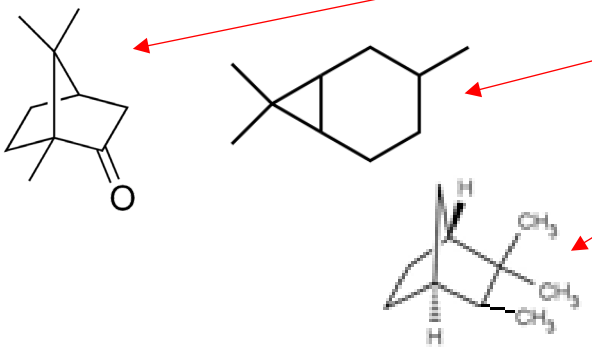
52 compounds were identified and the areas of GC peaks for 10 of them were adopted as variables for PCA and HCA.

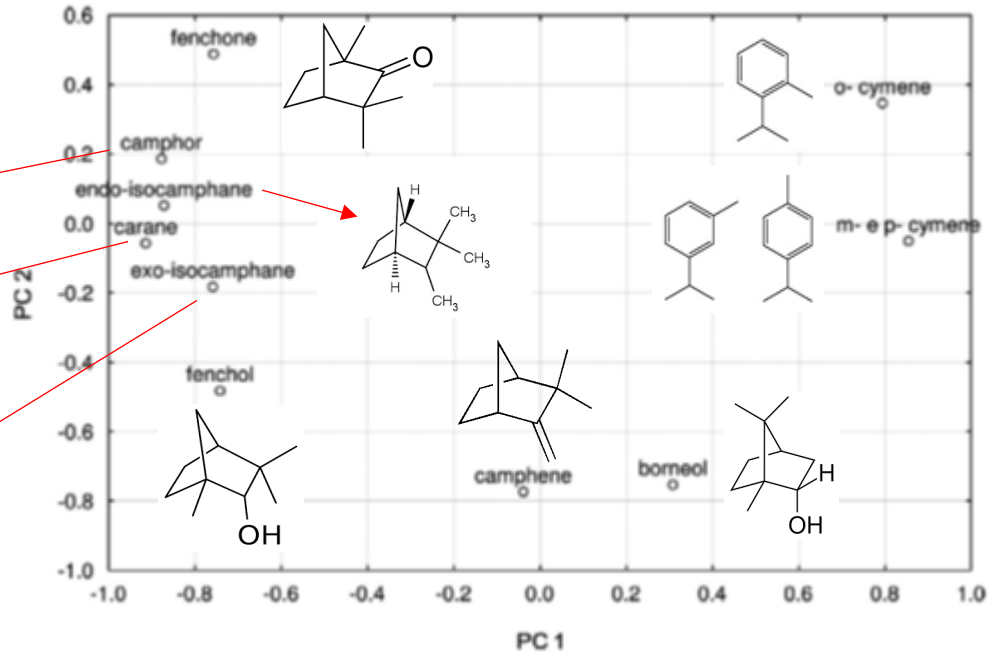
I. Van der Werf *et al.*, *Talanta*, 119 (2014) 435-439

The score plot referred to the first two principal components, accounting for *ca.* 73% of the total variance, clearly showed a distinction between amber samples originating from Romania (RX) and those originating from Baltic countries (BX).



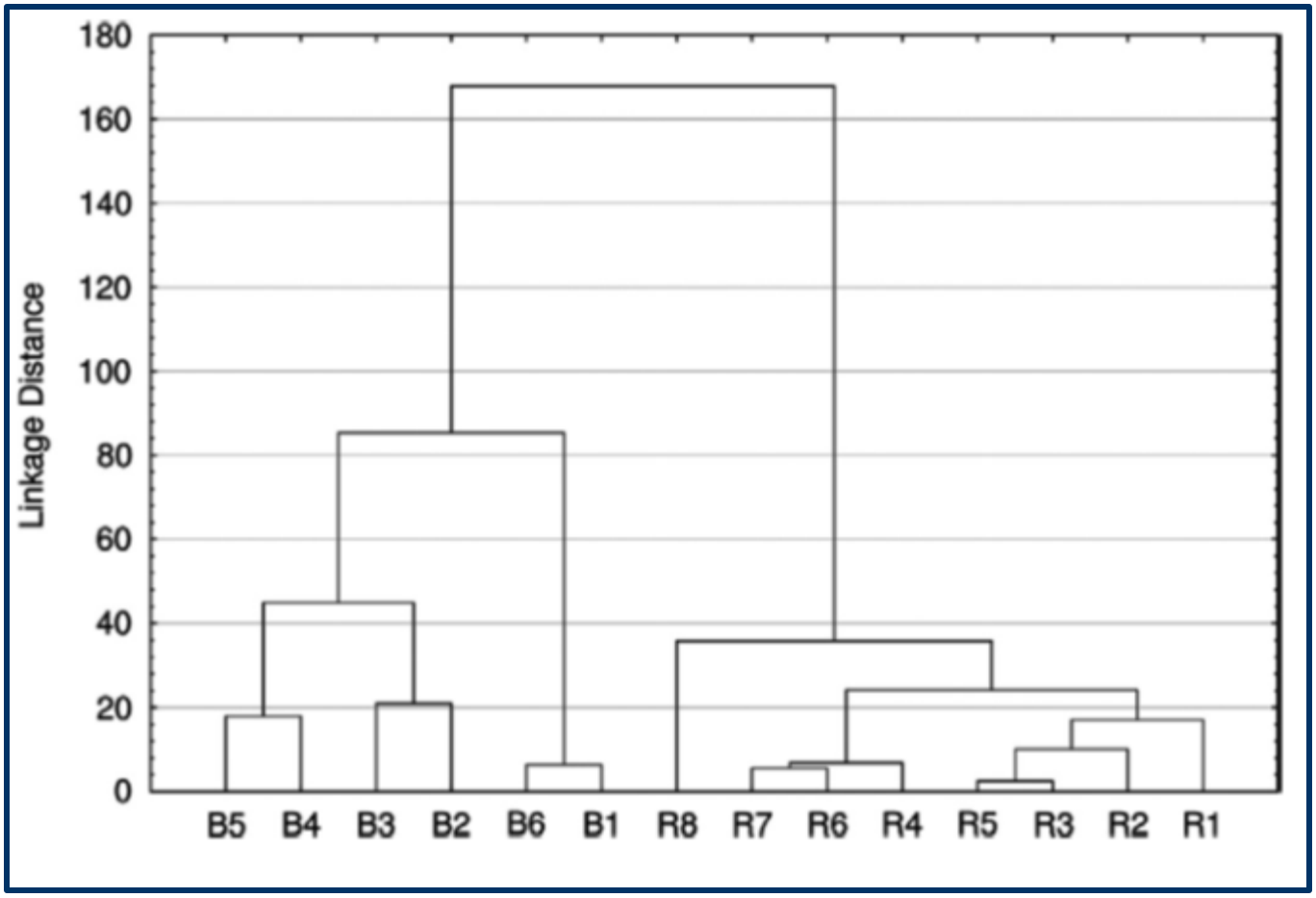
The loading plot clearly indicated which compounds contributed most to such a distinction.





I. Van der Werf *et al.*, *Talanta*, 119 (2014) 435-439

As shown in the following figure, this outcome was confirmed also by Hierarchical Cluster Analysis based on the Euclidean distance and on the Ward's method.

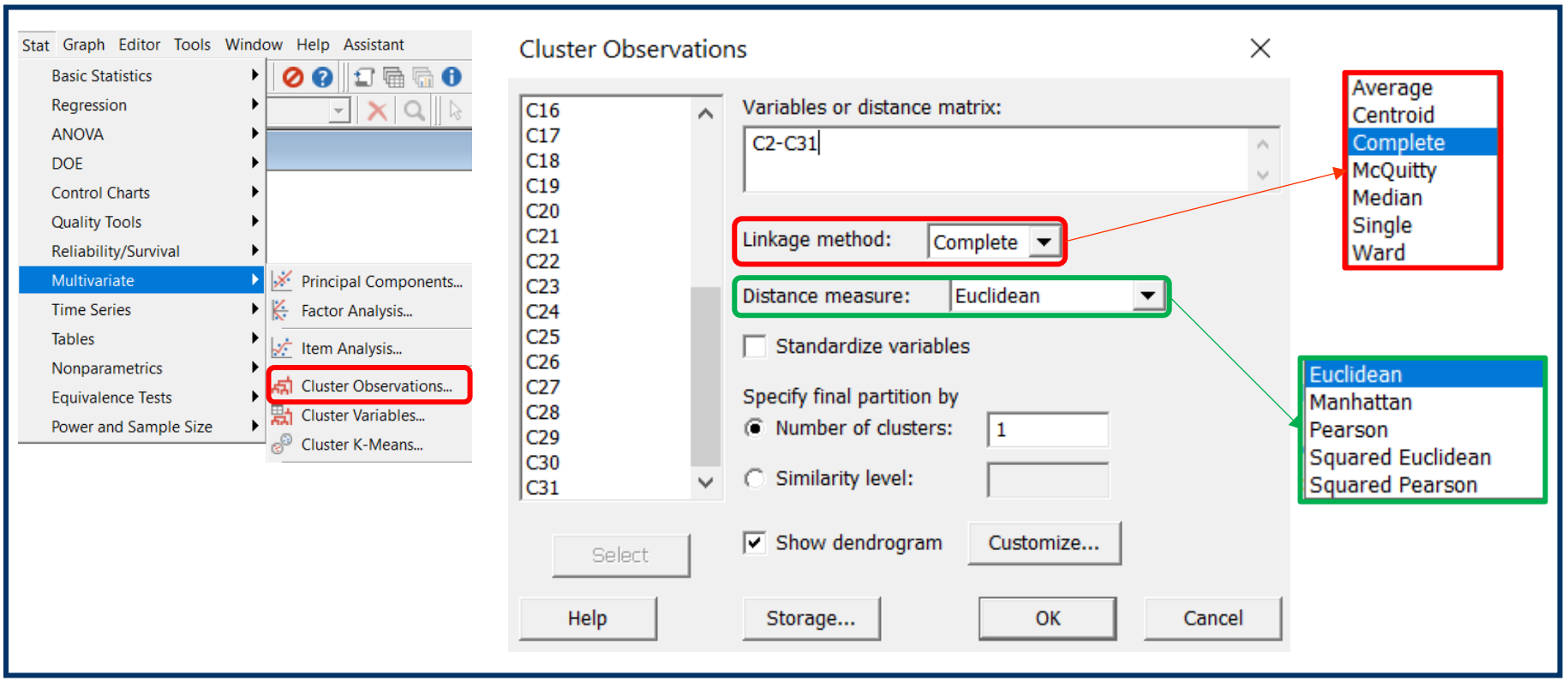


I. Van der Werf *et al.*, *Talanta*, 119 (2014) 435-439

Use of Minitab 18 for Cluster Analysis

Hierarchical Cluster Analysis can be performed using Minitab 18 by using the Stat > Multivariate > Cluster Observations… pathway.

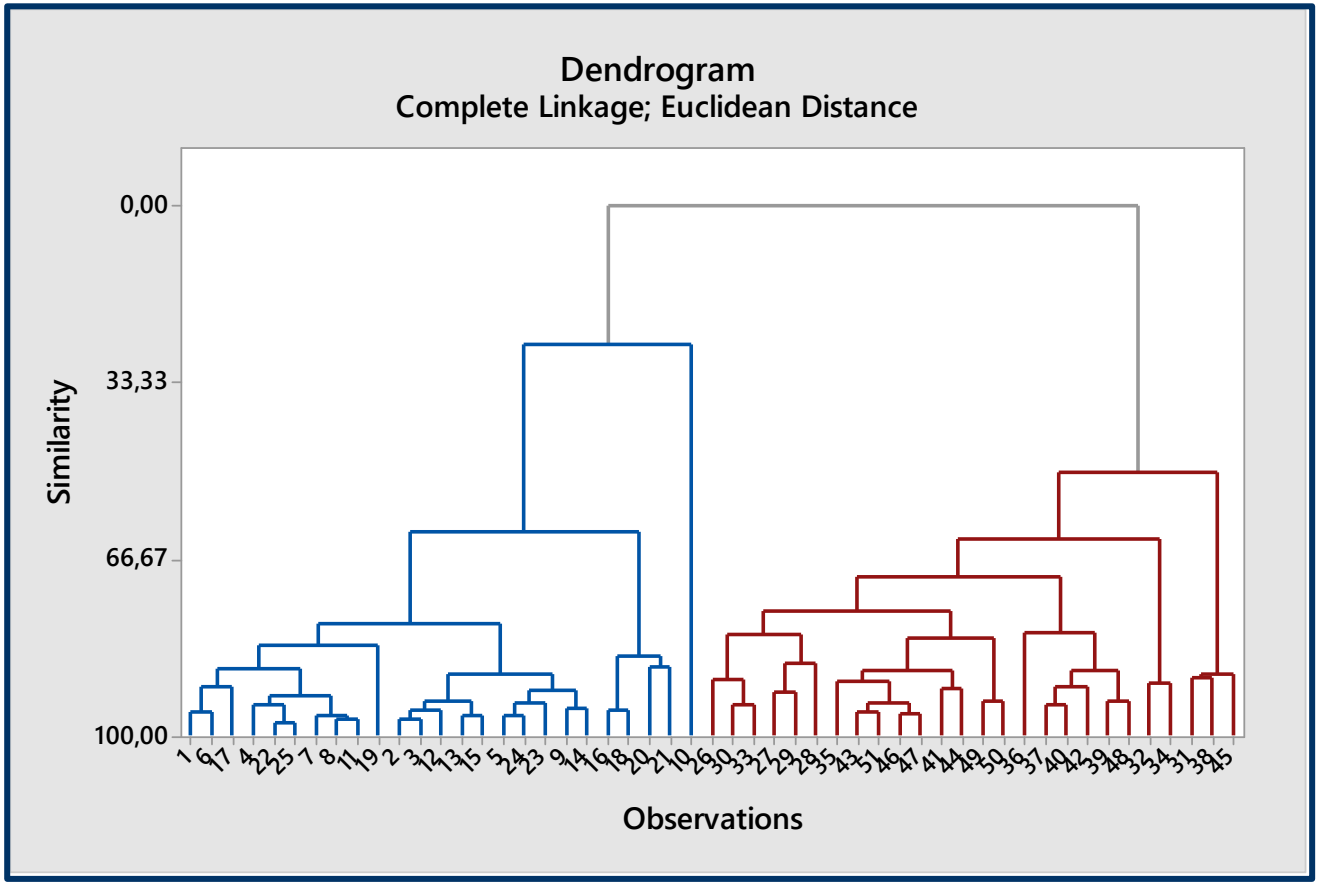
Different combinations of Linkage method and Distance measure can be selected inside the Cluster Observations window, in which the standardization of variables and the generation of a dendrogram can be also chosen.



Note that setting 1 as the number of clusters in the box referred to the specification of final partition indicates that one final cluster is reported at the top of the dendrogram.

The application described as an example is related to the distinction between 25 farmed and 26 wild Canadian salmons based on the profile of 30 fatty acids (FA) identified in their fillets using mass spectrometry.

The dendrogram obtained using Euclidean distance and Complete Linkage is the following:

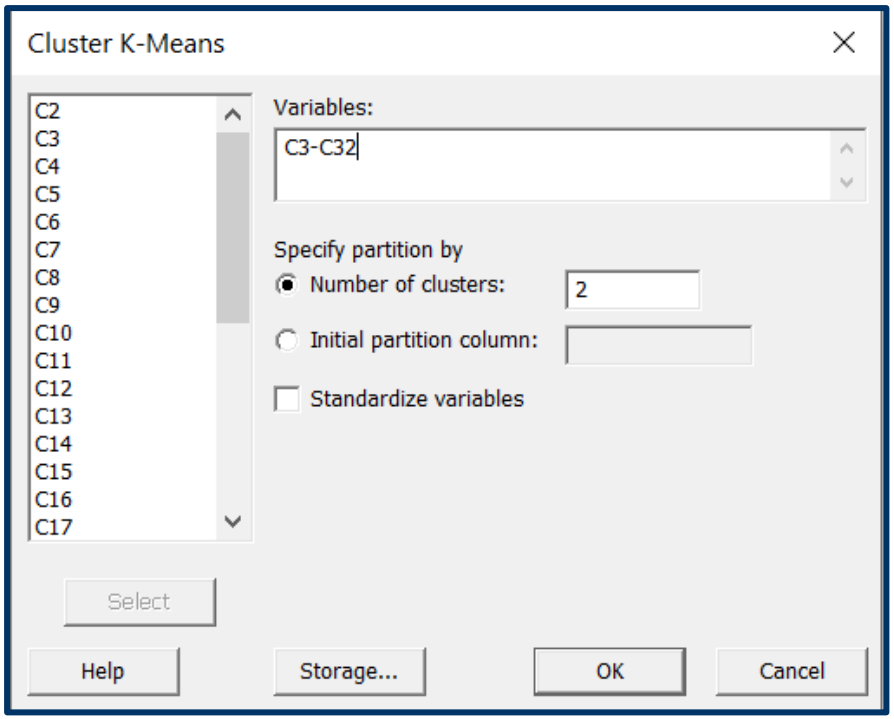


A nice separation between farmed (blue lines) and wild (red lines) salmon samples was obtained.

G.M. Fiorino *et al.*, *Food Res. Int.*, 116 (2019) 1258–1265

Cluster Analysis based on K-means was also performed on the same dataset, using the Stat > Multivariate > Cluster K-means… pathway.

Two clusters were specified in the Cluster K-Means window and no initial partition was indicated.



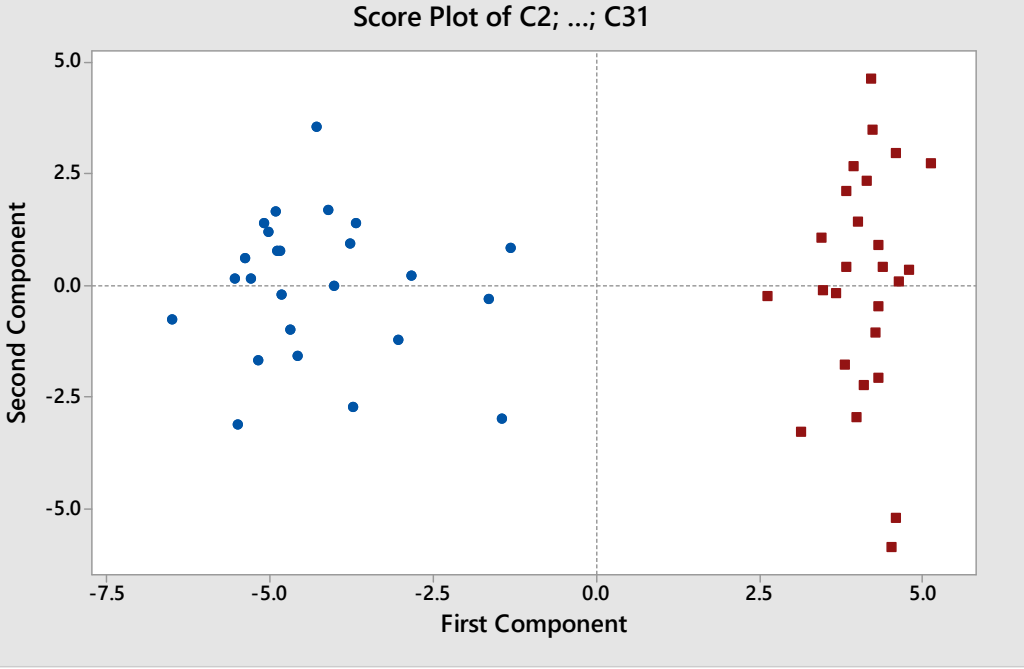
As a result, the 26 wild salmon samples were classified in Cluster 1 and the 25 farmed ones were classified in Cluster 2.

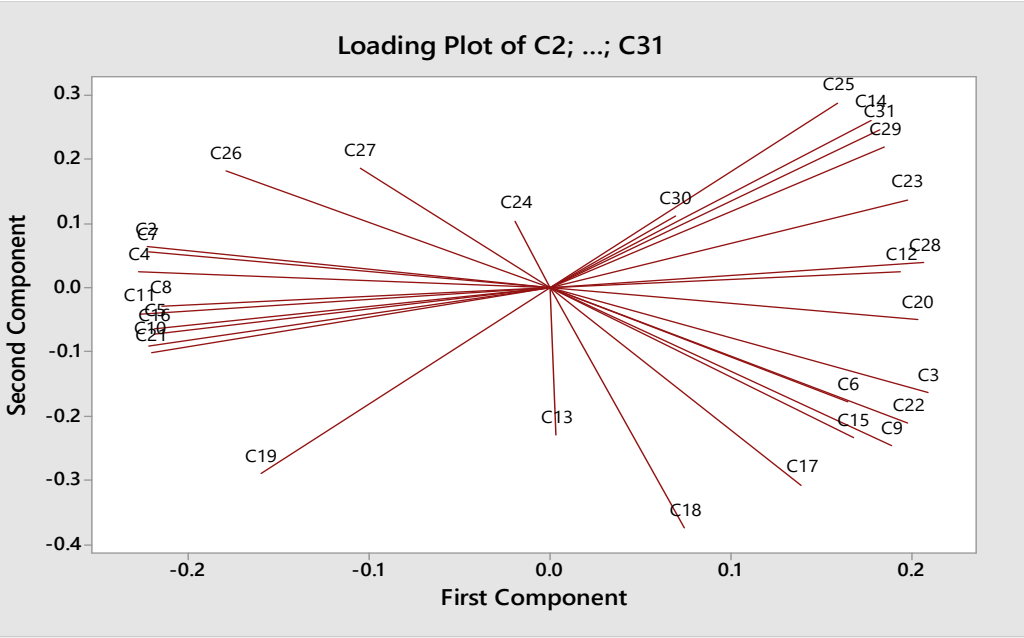
Final Partition

	Number of observations	Within cluster sum of squares	Average distance from centroid	Maximum distance from centroid
Cluster1	26	865.200	5.338	10.522
Cluster2	25	1160.798	4.910	26.129

Interestingly, the application of Principal Component Analysis to the same dataset confirmed the clear distinction between the two types of salmon (blue- farmed, red – wild).

When the identity of fatty acids mainly responsible for the distinction was evaluated from the loadings plot, it turned out that polyunsaturated FA with 20 or 22 carbon atoms, including ω-3 FA 20:5 and 22:6, were prevailing in wild salmon, whereas FA 18:1, 18:2 and 18:3, along with oxidized derivatives, were more abundant in farmed salmon.





Dendrograms with «heatmaps»

Several programs for chemometrics elaborations are able to provide special dendrograms after Cluster Analysis.

In this case, two dendrograms are actually drawn, one of which indicating the clustering of samples and the other the clustering of variables.

Moreover, the values assumed by variables (usually after autoscaling) are represented by small boxes colored according to a color scale, resembling a heatmap.

This representation, shown in the figure on the right for the clustering of durum wheat samples according to the contents of six elements, is very effective, since it emphasizes the similarities/differences of variable values between clustered samples.