PCA – principal component analysis

PCA

PCA is often used to reduce the number of variables in a statistical dataset by determining new variables (principal components) that correlate with the source variables while preserving as much information as possible.

We can reduce the number, for example, to two or three variables, which enables a simplified multidimensional ordering of objects and a visual assessment of object diversity in a two or three-dimensional coordinate system.

Each of the principal components (i.e. new variables) explains a certain% of the total variation of all variables. If the% of the explained variation by the first and second principal components is large (close to 100%), then we can use the two-dimensional coordinate system to properly assess the multidimensional differentiation of the examined objects.

Such a reduction in the number of dimensions is possible only in the case of fairly strong correlations between individual features

PCA – example

The division of apple varieties is based on several characteristics

cultivar	trunk cross-sectional area (cm^2)	average tree productivity (kg/cm^2)	average yield (t/ha)	the average weight of the fruit (g)
Arlet	35,2	0,48	29,3	144
Red Boskoop	34	0,44	24,3	179
Fiesta	36,3	0,51	33,2	148
Gala Must	24,5	0,52	23,6	141
Golden Delicious Reinders	33,5	0,52	29	147
Jonagored	31,2	0,32	20,2	164
Jonica	28,6	0,34	18,2	167
Wilmuta	34,5	0,5	25,8	187
Jonagold	33,1	0,45	26,8	186
Elstar	62	0,3	33	135
Red Elstar	66,2	0,27	27	138
Elshof	60,2	0,3	32	139
Gloster	35,2	0,45	34,2	177

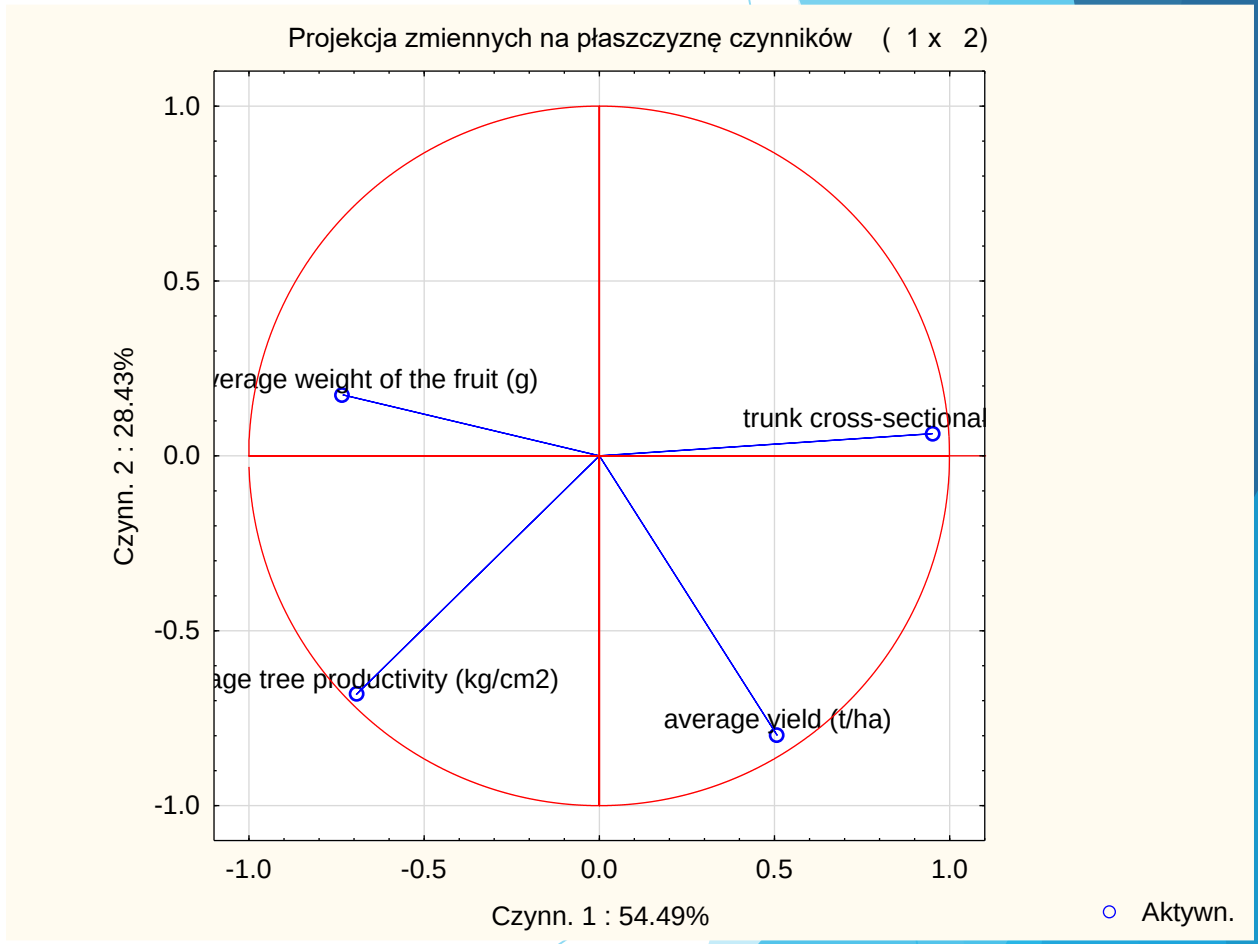
Based on:
http://www.up.poznan.pl/ogrodnictwo/Ogrodnictwo%2041/67%20Wocir.pdf

PCA – example

https://datatab.net/statistics-calculator/factor-analysis

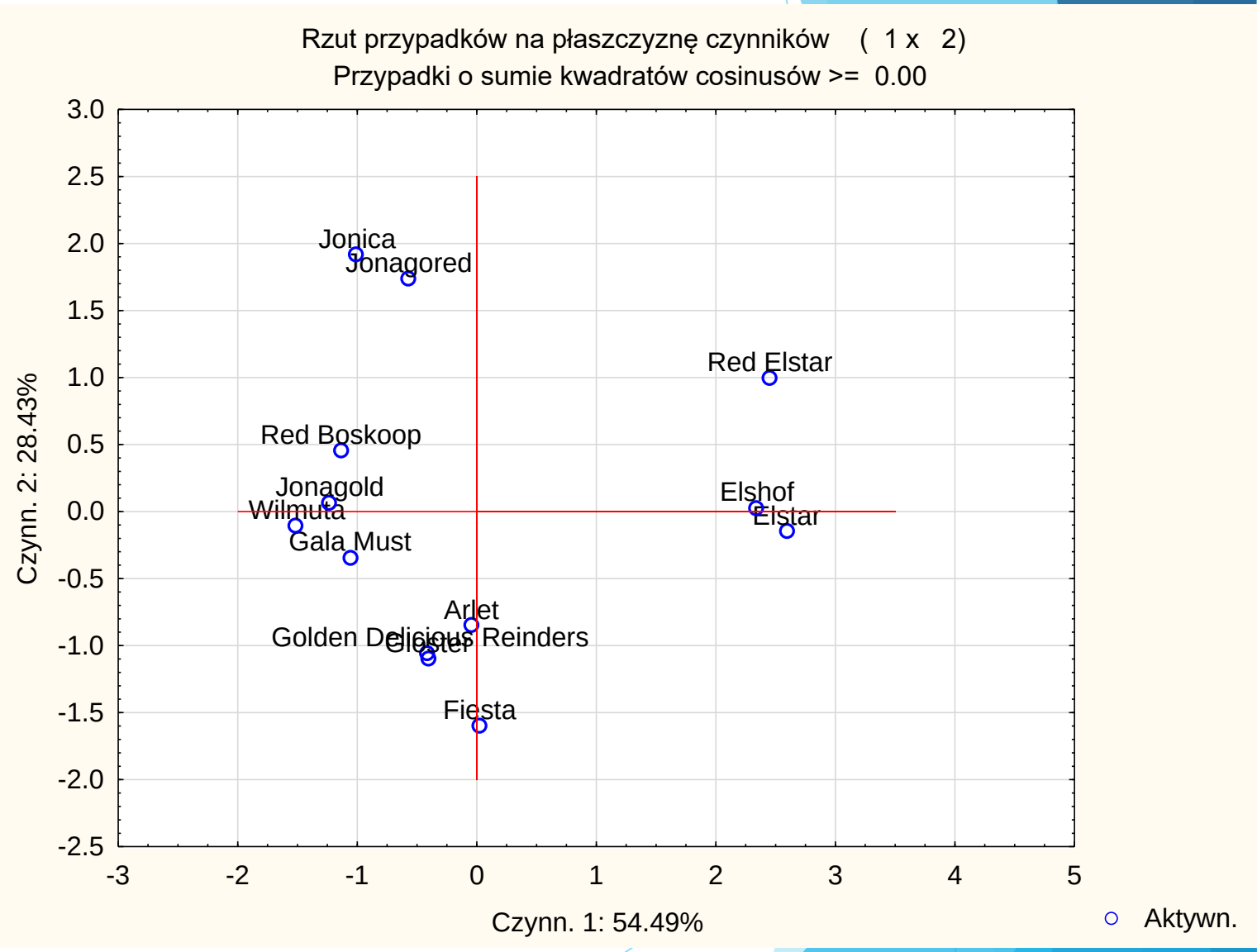
PCA – example

	PC1	PC2
trunk cross-sectional area (cm2)	0.950580	0.063465
average tree productivity (kg/cm2)	-0.693142	-0.682087
average yield (t/ha)	0.507113	-0.798387
the average weight of the fruit (g)	-0.733825	0.174754



PCA – example

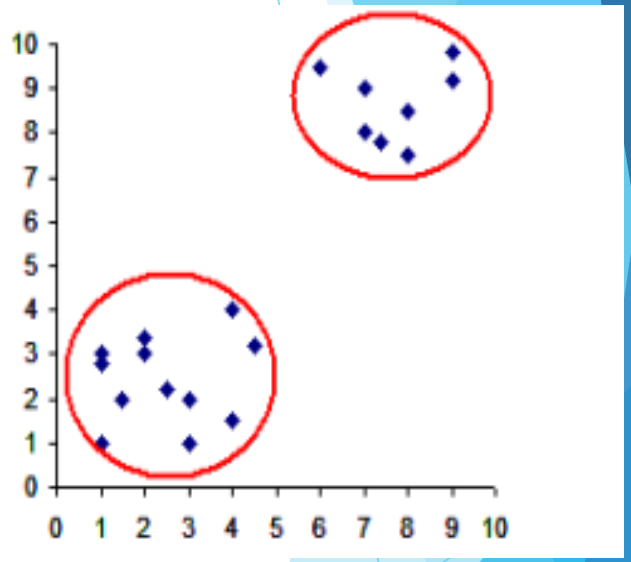
	PC1	PC2
Arlet	-0.04468	-0.85081
Red Boskoop	-1.13969	0.45486
Fiesta	0.02324	-1.59670
Gala Must	-1.06066	-0.34231
Golden Delicious Reinders	-0.42020	-1.05821
Jonagored	-0.57435	1.73807
Jonica	-1.00955	1.91540
Wilmuta	-1.51520	-0.10440
Jonagold	-1.23941	0.06942
Elstar	2.59356	-0.14678
Red Elstar	2.45280	0.99449
Elshof	2.33806	0.02828
Gloster	-0.40392	-1.10131



Cluster analysis

Cluster analysis - multidimensional classification of objects

A method that allows objects to be grouped in terms of many features at the same time. In the case of two or three features, it is possible to select objects similar to each other on the basis of the value of these features (X, Y or Z) on a scatter plot. In the case of grouping features in terms of more than 3 features (variables), it is not possible to graphically present the values of all features on the chart. However, it is possible to determine the distance between objects in a multidimensional space



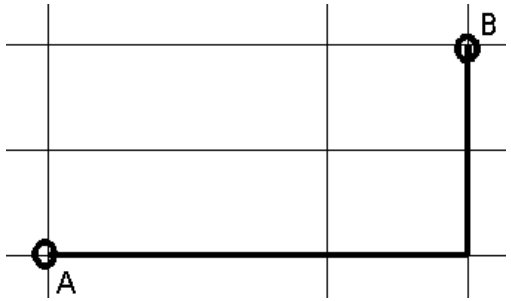
Cluster analysis - multidimensional classification of objects

Euclidean distance in a space with p dimensions between two objects:

$X_{ij;}\, X_{kj}$ – values of the j-th feature for objects i and k
p - number of features / variables

$$d(x_i, x_k) = d_{ik} = \sqrt{\sum_{j=1}^{p} (x_{ij} - x_{kj})^2}$$

There are also other measures for determining distance, e.g. city distance (city block or Manhattan type)



Cluster analysis - multidimensional classification of objects

Due to the use of different units of individual features and different value scales, usually the distance between objects is determined on the basis of standardized variables

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{S_j}$$

z_{ij} – *value after standardization*

x_{ij} – *value before standardization*

x_j – *average value of the feature*

S_j – *standard deviation*

The standardization of variables allows, in the analysis of clusters, to maintain a similar weight of each variable in the classification of objects.

Cluster analysis - grouping methods

Hierarchical

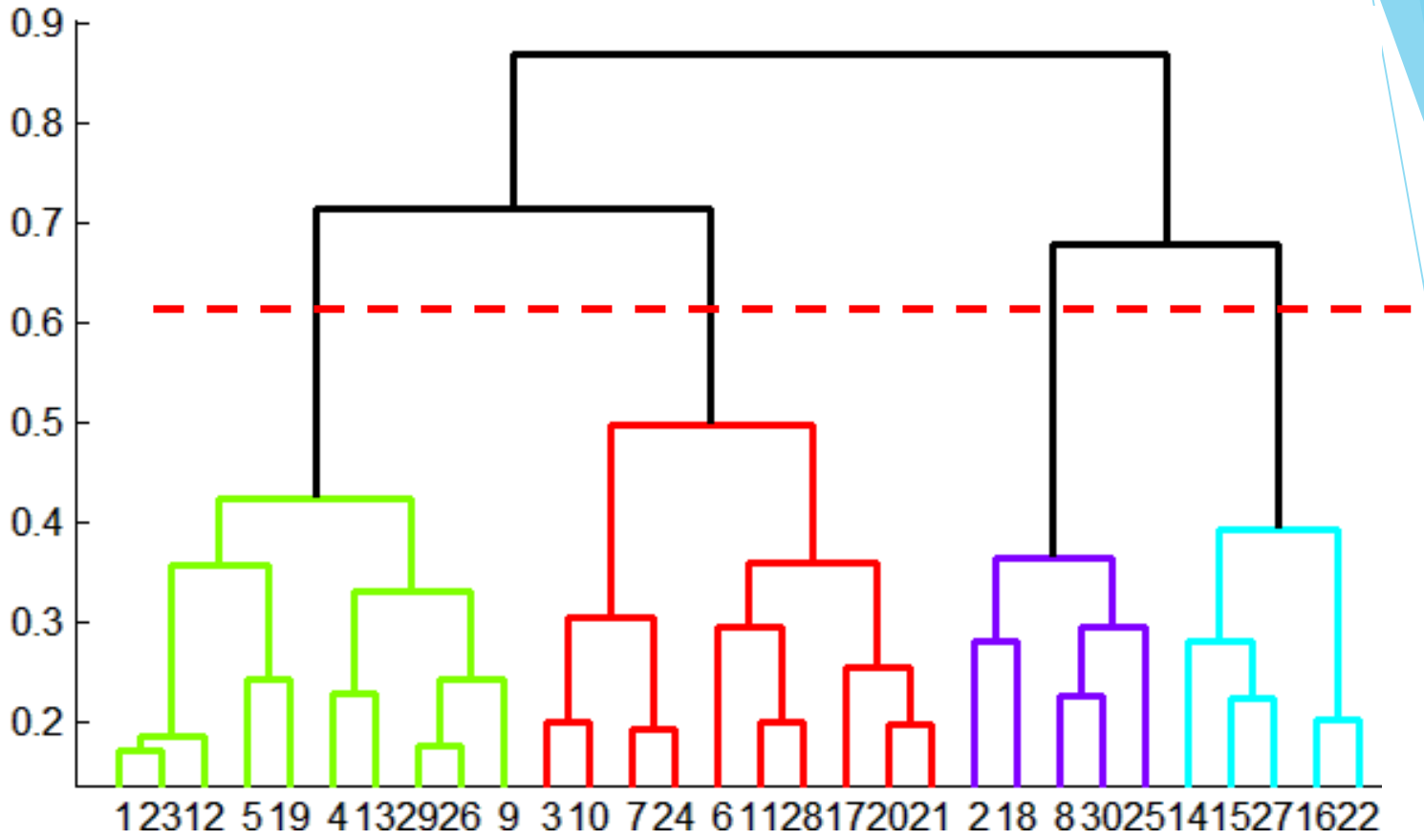
allow to combine objects into groups maintaining the hierarchy, i.e. we can determine on the basis of the dendrogram which objects within the selected groups are similar and which are more distant

Selected methods of agglomeration (combining objects) in hierarchical grouping:
- nearest neighbor method
- farthest neighbor method
- Ward's method
- centroid method

Non-hierarchical

after assigning an object to a given group, we cannot say which of the objects and which of the objects within one group are more similar.

The non-hierarchical method is the k-means method.



Sample dendrogram, created as a result of cluster analysis. The dashed red line shows the division of objects into 4 groups. There is freedom in determining the number of groups, so objects can be divided into a larger or smaller number of groups depending on the objectives of the analysis

Analiza skupień – przykłady zastosowań

1) Identifying groups of apple varieties similar in terms of many characteristics, e.g.

fruit color (color must be quantified, i.e. in the form of a number, e.g.

on a 5-point scale 1- green…. 5- red)

the size of the fruit

growth rate

susceptibility to disease

e.t.c.

2) Separation of groups of communes similar in terms of many features, e.g.

the number of residents

per capita income

unemployment rate

share of agricultural land, forests and orchards

e.t.c.

Cluster analysis - application examples

To characterize individual selected groups of objects, one can use the mean values of individual features (variables) on the basis of which the classification was made.

The average values of the features for objects from particular groups are called centroids.

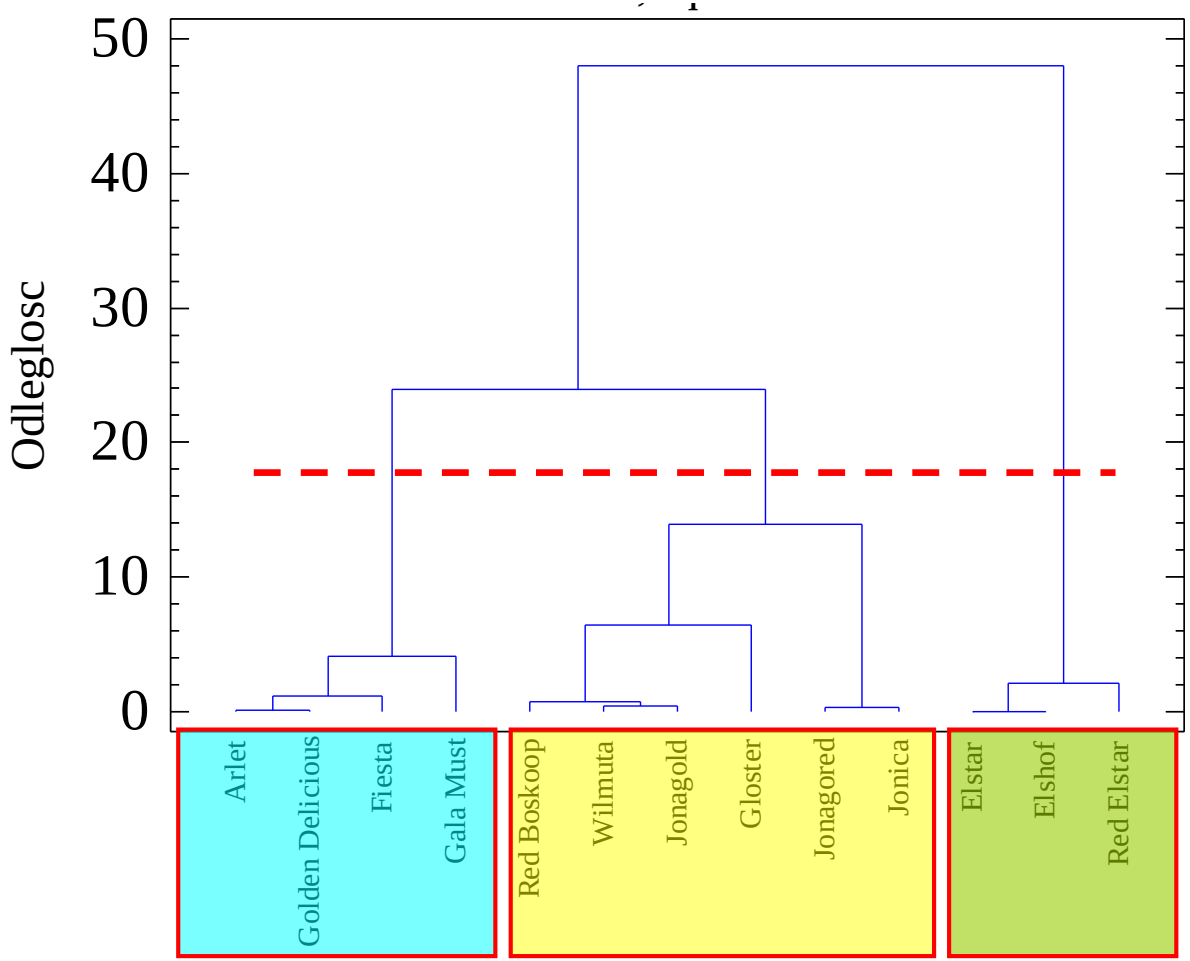
There is no single objective method for determining the number of separate groups of objects. This number may be determined either before performing the analyzes or after performing the analyzes, e.g. on the basis of a dendrogram.

Cluster analysis - example

The division of apple varieties is based on several characteristics

cultivar	trunk cross-sectional area (cm^2)	average tree productivity (kg/cm^2)	average yield (t/ha)	the average weight of the fruit (g)
Arlet	35.2	0.48	29.3	144
Red Boskoop	34	0.44	24.3	179
Fiesta	36.3	0.51	33.2	148
Gala Must	24.5	0.52	23.6	141
Golden Delicious Reinders	33.5	0.52	29	147
Jonagored	31.2	0.32	20.2	164
Jonica	28.6	0.34	18.2	167
Wilmuta	34.5	0.5	25.8	187
Jonagold	33.1	0.45	26.8	186
Elstar	62	0.3	33	135
Red Elstar	66.2	0.27	27	138
Elshof	60.2	0.3	32	139
Gloster	35.2	0.45	34.2	177

Based on:
http://www.up.poznan.pl/ogrodnictwo/Ogrodnictwo%2041/67%20Wocir.pdf



Mean values (centroids) of the assessed features for separate groups in the cluster analysis

Grupa	trunk cross-sectional area (cm^2)	average tree productivity (kg/cm^2)	average yield (t/ha)	the average weight of the fruit (g)
1	32,4	0,51	28,8	145,0
2	32,8	0,42	24,9	176,7
3	62,8	0,29	30,7	137,3

Cluster analysis - an example

Statistica

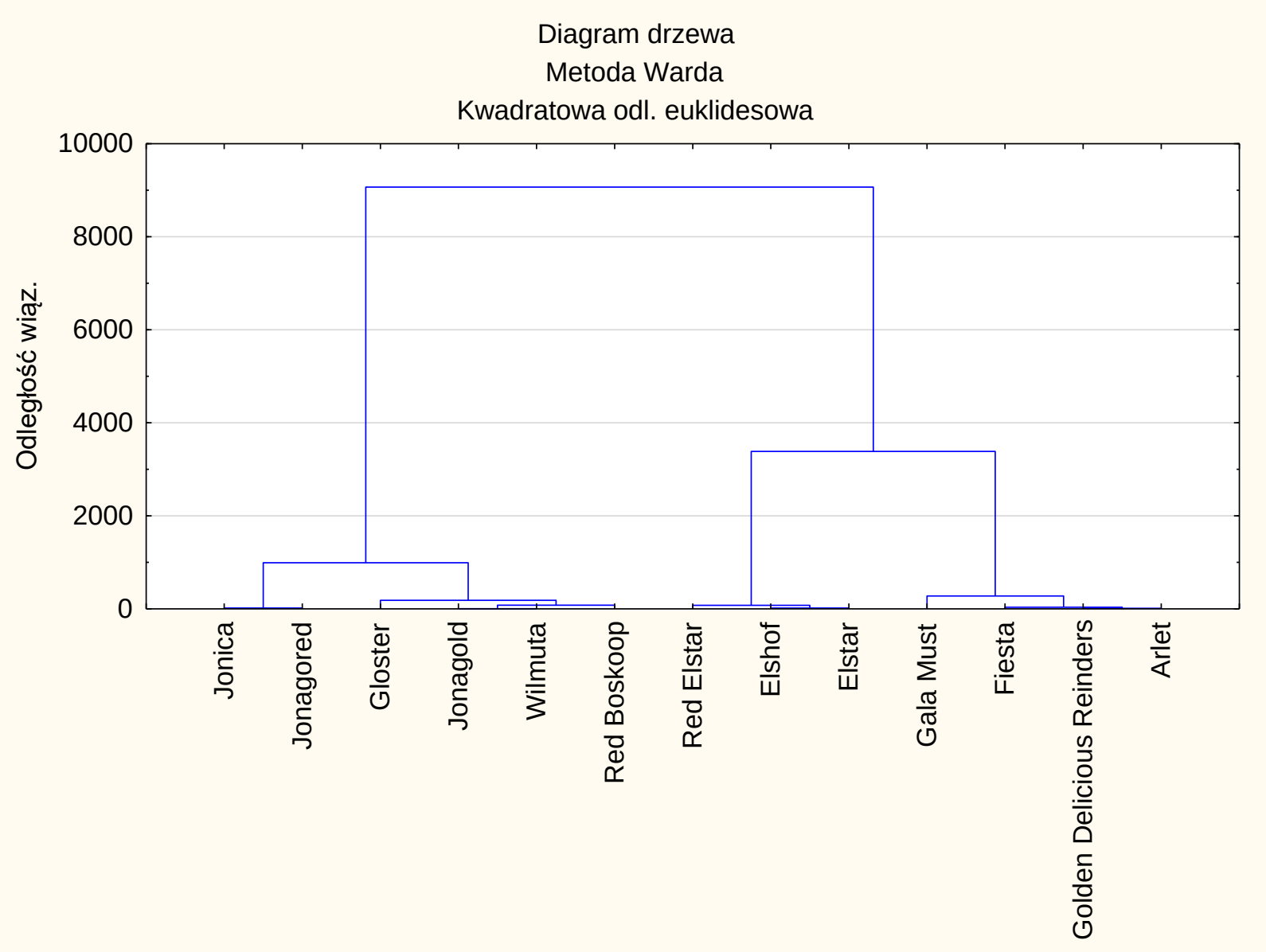
Score:

The chart shows which objects, i.e. variants, are similar in terms of the assessed features.

Red Elstar, Elshof and Elstar are similar to each other.

Gloster, Jonagold, Wilmuta and Red Boskoop are similar to each other.

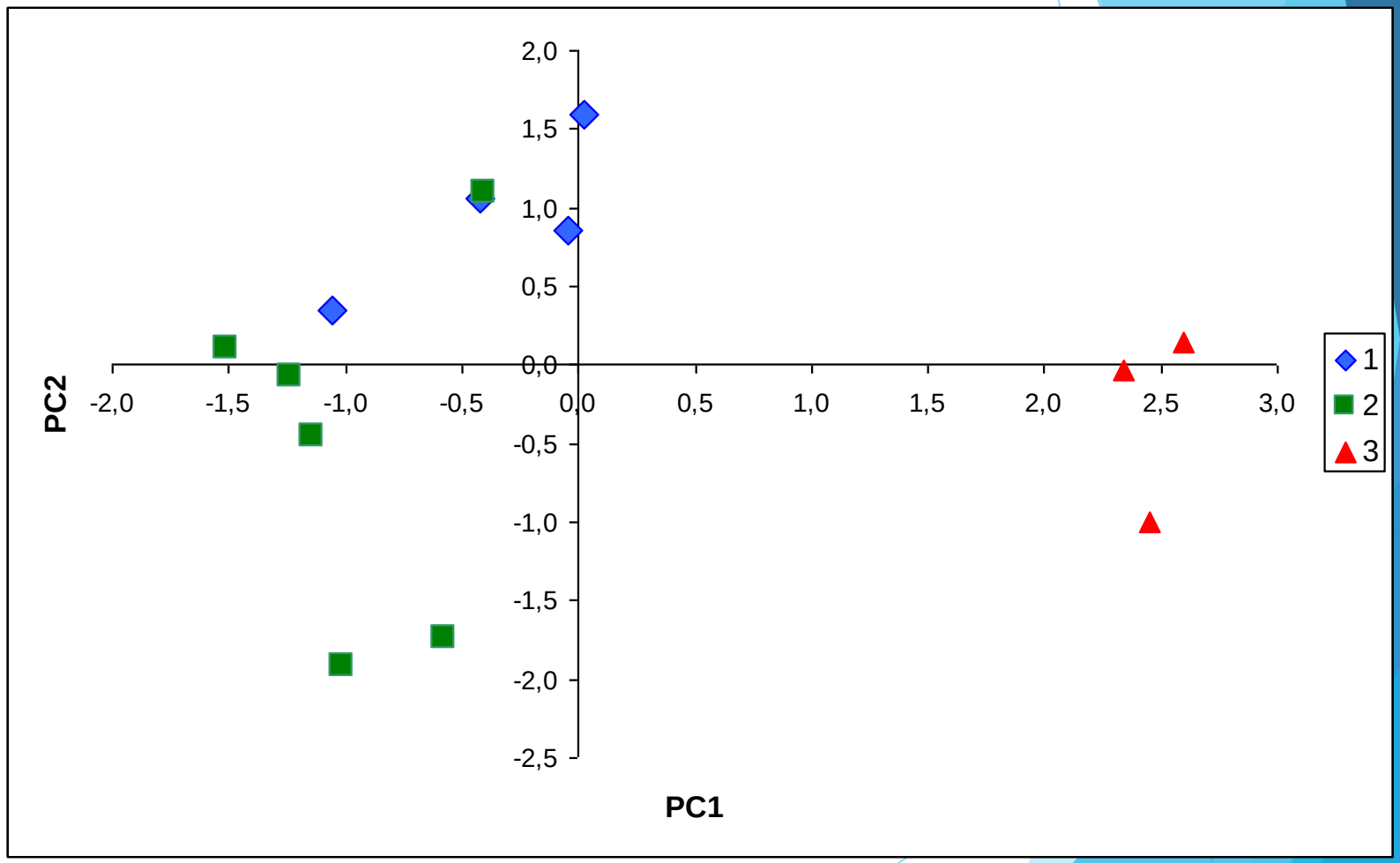
Gala Must, Fiesta, Golden Delicious Reinders and Arlet are similar to each other.



Comparison of PCA with cluster analysis

Values of the first two principal components for individual varieties (the numbers represent the variation groups identified in the cluster analysis).

Principal component analysis facilitates the characterization of groups of objects distinguished in the cluster analysis, allows for a graphic representation of the diversity of these objects in a two-dimensional system.



https://datatab.net/statistics-calculator/cluster