Strona czasopisma: http://analit.agh.edu.pl/

Application of chemometric methods in analysis of environmental data

Zastosowanie metod chemometrycznych w analizie danych środowiskowych

Anthime Chevallier[a], Małgorzata Jakubowska[b]

[a] IUT University Institute of Technology, Biological Engineering, Environmental Engineering, 2 avenue Adolphe Chauvin 95302 Cergy-Pontoise, France
[b] AGH Akademia Górniczo-Hutnicza, Wydział Inżynierii Materiałowej i Ceramiki, al. Mickiewicza 30, 30-059 Kraków, Polska

1. Introduction

Chemometrics [1-3]is a science based on data analysis and aimed for extraction of hidden information by using a dedicated software. This method uses skills in multivariate statistics, applied mathematics and computer science. This numerical methods can be used in chemistry, biochemistry, medicine, biology and chemical engineering but also psychometrics and econometrics. For descriptive applications the goal is to identify structure of the system and hidden relationships. For predictive applications the goal is model definition and finding class memberships for new objects. However, this science can only be applied to a large set of digital data to ensure reliable results. In this article we shall focus on multivariate data analysis, this corresponds to factor analysis technique which allows us to uncover the latent structure of a data set. It also allows to obtain a smaller number of factors for data modeling with the reduction of number and attribute. In analytical chemistry chemometrics is often used to compare different samples when the data set is big. For example, it can be used to analyze water in lakes or in rivers because with chemometrics it's possible to compare concentration of heavy metals in the area of sampling. It's can be also use in soil or blood analysis but it's the same principle every time, a lot of data and variables for similar samples. Sometimes during analysis errors may appear, it's call outlier. It is a point very distant of the other measures, it corresponds to an experimental error and for this reason it should not be taken into account in interpretation.

2. Experimental data and software

2.1. Experimental data

For our study we shall use 2 different data set issued from a work [4,5] in which there were studied two types of forests: “Sugi” forest and “Hinoki forest” with geographically weighted variables. In these data sets we have 4 classes of data:

- S: ('Sugi' forest)
- H: ('Hinoki' forest)
- D: ('Mixed deciduous' forest)
- O: ('Other' non-forest land)

And we have 9 columns which correspond to geographically weighted variables for each area. The first data set is named “Training” and contains 198 areas identified columns B1 to B9. The second data set is named “Testing” and contains 325 areas identified also by columns B1 to B9.

2.2. Software

For the Forest analysis we have used two computational tools. Statgraphic64 software developed by Francestat, we use to obtain spider plot, matrix plot, cluster analysis and principal component analysis. The second software is Matlab and we have used it for LDA and CART method.

3. Methods

3.1. Spider Plot

The first method used in this study is a radar plot. It is a picture which displays multivariate data in the form of a two-dimensional diagram. With radar plot we can compare at least three variables which are represented on the axes starting at the same point in the center of the plot. The relative position and the axes angle is typically uninformative. This allows a detailed analysis of several objects, and a comparison between the analyzed samples.

3.2. Cluster analysis

The second method used in this article is cluster analysis. It is an exploratory analysis which classifies data into several groups to identify hidden structures. It tries to identify homogenous groups of data using experimental results, while often the groups can hardly be distinguished before such analysis. This technique is often used in medicine, marketing, education, biology or a lot of study with a big data set.

3.3. Matrix plot

The third method used in this study is a matrix plot, this analysis shows us the relationships between pairs of variables. The results appear on a matrix with X/Y plots for each comparison of variables. Here we will use a scatterplot matrix to obtain the information about correlations between variables. The difficulty of this technique is the density of information issued from the different diagrams of the matrix plot.

3.4. Principal Component Analysis

The next method used in this article is the Principal Component Analysis (PCA). It’s a procedure which converts experimental variables into linear and uncorrelated variables, called principal components. During the transformation the software tries to present the highest percent of the variability using the lowest number of principal components.

3.5. LDA

The Linear Discriminant Analysis is a method for explaining and predicting an individual's membership to a predefined class based on its measured characteristics. Here we will use this technique to classify the samples referring to the forests types. Data set "Training" will be used to define classification model. Dataset "Testing" will be applied to check the membership of the unknown objects.

3.6. CART

Classification And Regression Trees is a method that builds a decision tree using “Training” dataset. This tree provides a template for classifying new samples.

4. Results and discussion

4.1. Unsupervised learning methods

4.1.1. Spider plot

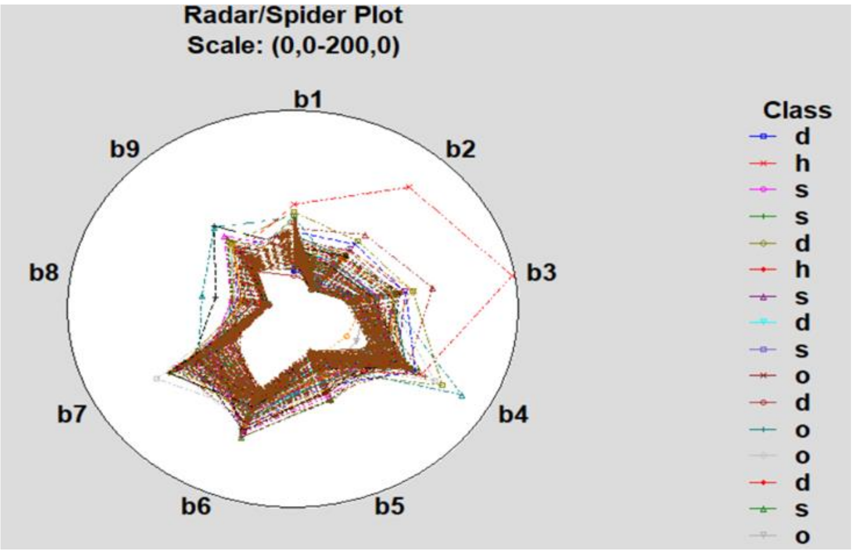

Figure 12. Spider plot – comparison of samples of different forest types (Set 1).

On this radar plot (**Fig. 1**), we can see that the majority of points have the same appearance, we can see only differences in the positions of certain classes. This spider plot is not very effective for getting good interpretations.

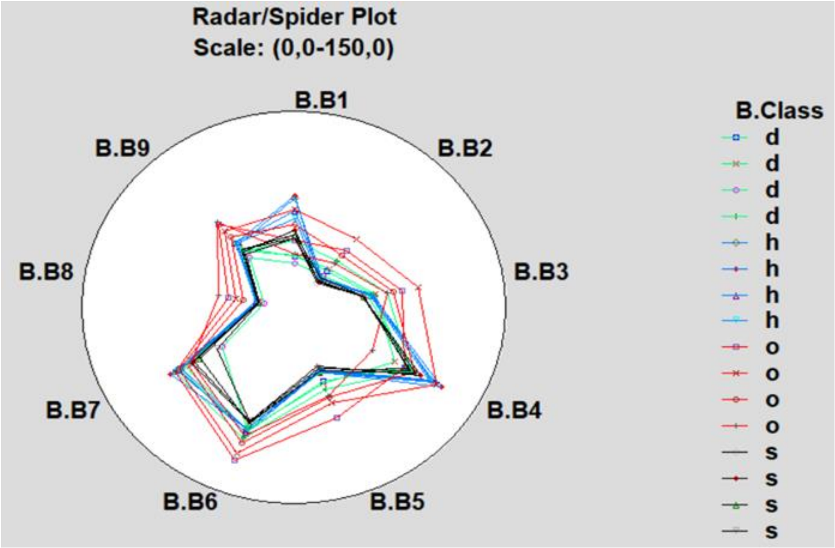

Figure 2. Spider plot – comparison of samples of different forest types, randomly selected 16 objects (Set 1).

For this second radar plot (**Fig. 2**) we use only 4 objects of each class because on the first spider plot we had too much values. Now we can see a homogeneity for the class S. The class H is characterized by a little higher values than the class S and the results are heterogeneous. The class D has also higher values than S but the values are homogeneous. And for the class O we can see a big variability, the values are very heterogeneous and higher than in all other classes.

We can conclude that the forest areas are not very differentiated, even within the same forest-type class. But in the non-forest area, there are large variations compared to other areas but also between the measured values. The mixed zone is not very far from the two types of the forest zones.

4.1.2. Matrix plot

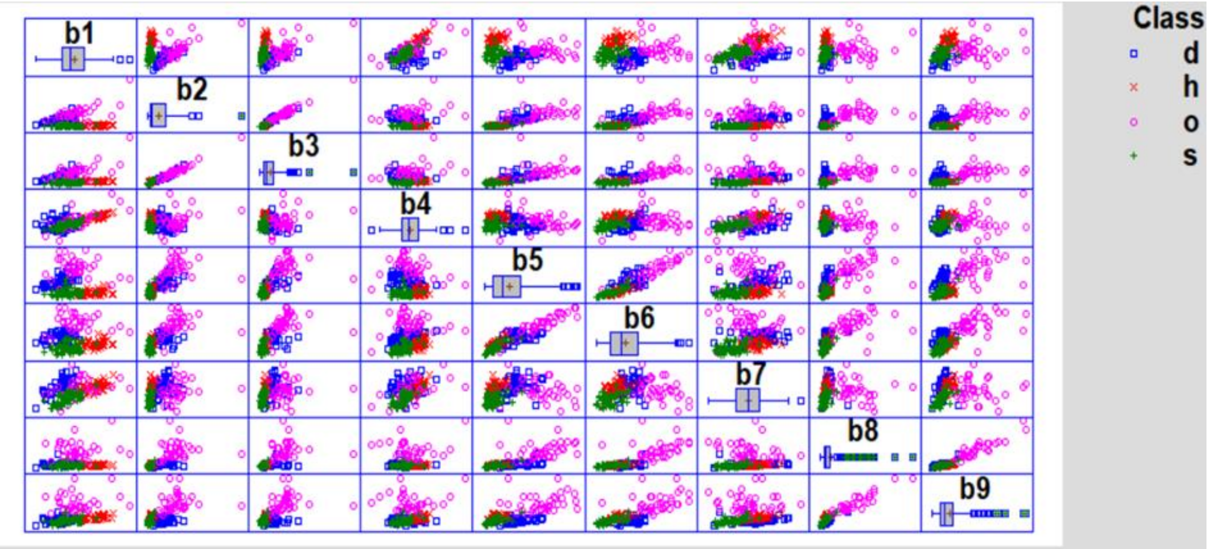

Figure 3. Matrix plot – comparison of variables pars describing the forest types (Set 1).

With that matrix plot (**Fig. 3**) we compare position of the different classes in plot. We can see a linear relation between variables B2/B3, B5/B6 and B8/B9. We can also observe a large variability of the class O. The class S, as in the spider plot, have low variability. The characteristics of class D, H and S are close but variability of S is different from the other classes. We can also see an outlier in the top/right corner from the class O.

4.1.3. Cluster analysis

We shall analyze the link between data with a cluster analysis.

At the beginning it is necessary to optimize parameters of cluster analysis that is distance measure and agglomeration algorithm.

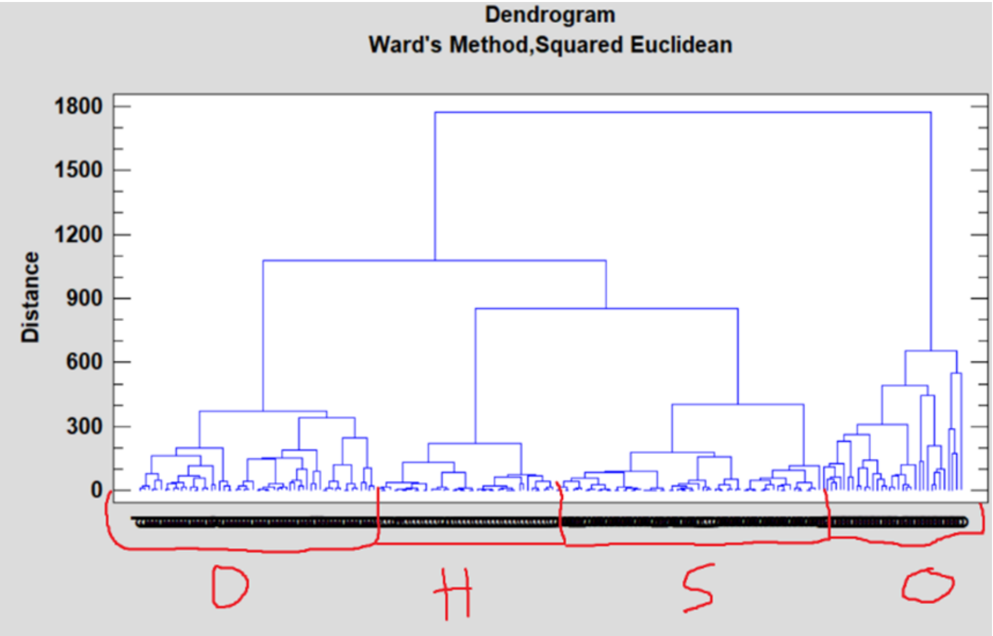

Figure 4. Second dendrogram for data set: Forests types (Set 1).

By using the "Ward's" option, we obtain a dendrogram which can be easily interpreted. We can see two main clusters **(Fig. 4)**, the first one containing the classes D, H and S and the other containing the class O. The first cluster is divided into three clusters D, H and S. We can also observe for each different classes several clusters. In all of this clusters we have points who are not in the proper position.

Now we will analyze the relations between variables using cluster analysis.

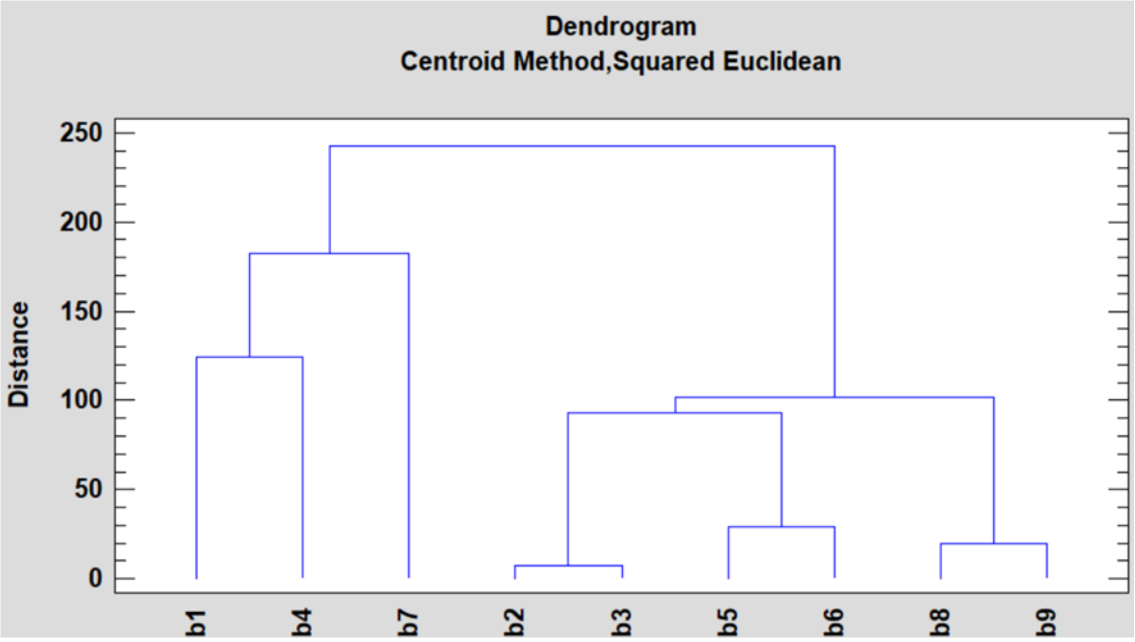

Figure 5. Dendrogram for variables describing forest types (Set 1)

This dendrogram **(Fig. 5)** shows us 2 clusters of variables, the first one with variables B1, B4 and B7. We can notice a correlation between B1 and B4 but they are however not so similar. In the second cluster we can see variables B2, B3, B4, B5, B6, B8 and B9. We can see 3 groups of variables: B2 and B3 who are very similar because the distance is very low, B5 and B6 who are quite similar and B8 and B9 are also correlated. This information confirms the interpretations of the results based on the matrix plot.

4.1.4. Principal Component Analysis

Then we perform the Principal Component Analysis to complete our analysis and to confirm if it is correct.

Table 1. Variance of components Principal Components Analysis.

Component Number	Eigenvalue	Percent of Variance	Cumulative Percentage
1	4,80492	53,388	53,388
2	2,04805	22,756	76,144
3	0,91693	10,188	86,332
4	0,571191	6,347	92,679
5	0,347292	3,859	96,538
6	0,249559	2,773	99,310
7	0,038215	0,425	99,735
8	0,015644	0,174	99,909
9	0,0081961	0,091	100,000

The purpose of the analysis is to obtain a small number of linear combinations of the experimental variables which account for most of the variability in the data. In this case, 2 principal components have been extracted, since 2 components had eigenvalues greater than or equal to 1.0. Together they explain 76.1442% of the variability in the original data set.

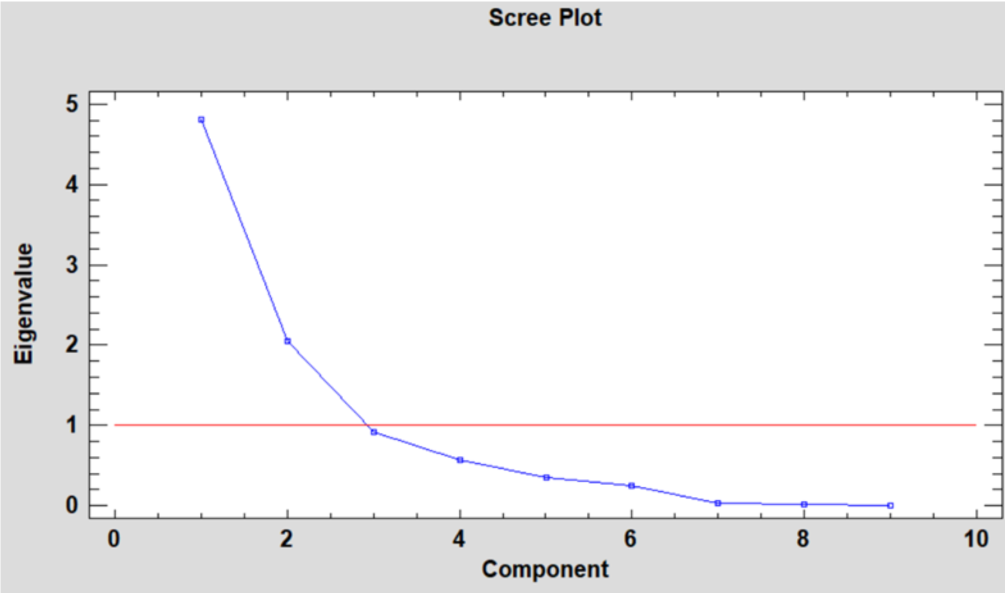

Figure 6. Scree plot for data set: Forests types (Set 1).

With Scree Plot (**Fig. 6**), we can see that only two principal components are important. This results are taken from the **Tab. 1**.

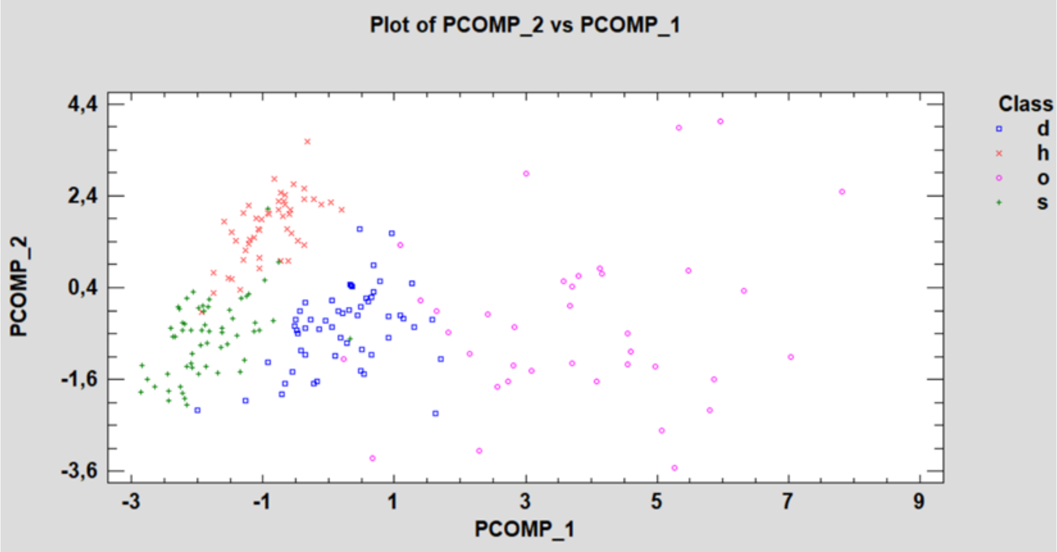

Figure 7. Projection of the objects on PC1/PC2 plane for data set: Forests types (Set 1).

Using projection of the objects (**Fig. 7**) on principal components 1 and 2 we can analyze the results. As we have seen before, the class D, H and S are very similar and homogenous but class O are very different and heterogeneous. This graph confirms the previous analysis.

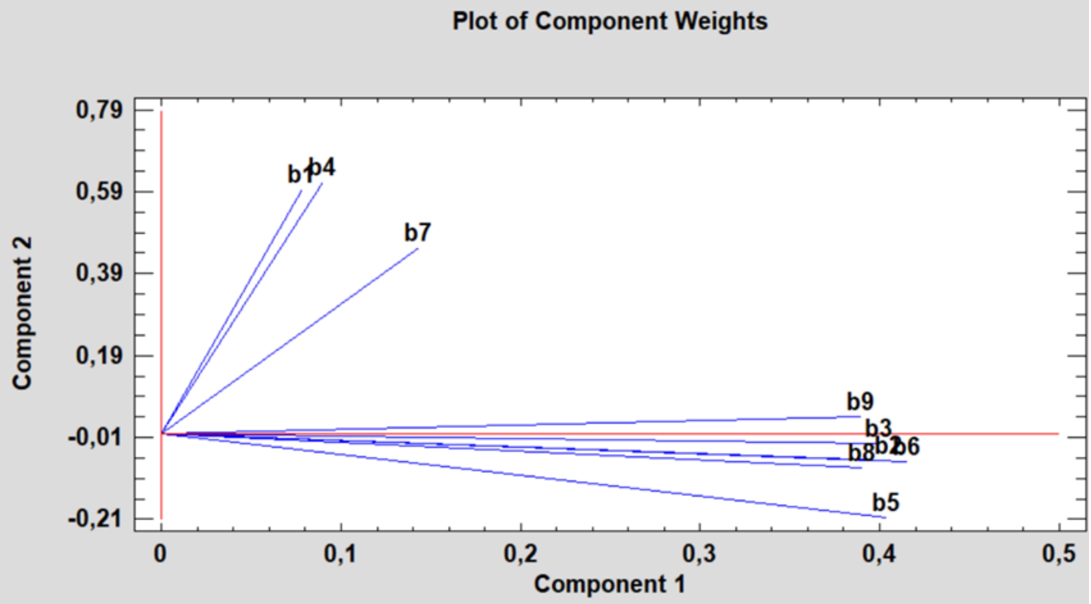

Figure 8. Plot of component weights for data set: Forests types (Set 1).

This plot (**Fig. 8**) gives information about relations between the variables. B1/B4/B7 have influence on the principal component 2, B1 and B4 are correlated. The other variables have influence to the principal component 1, they are correlated, except B5.

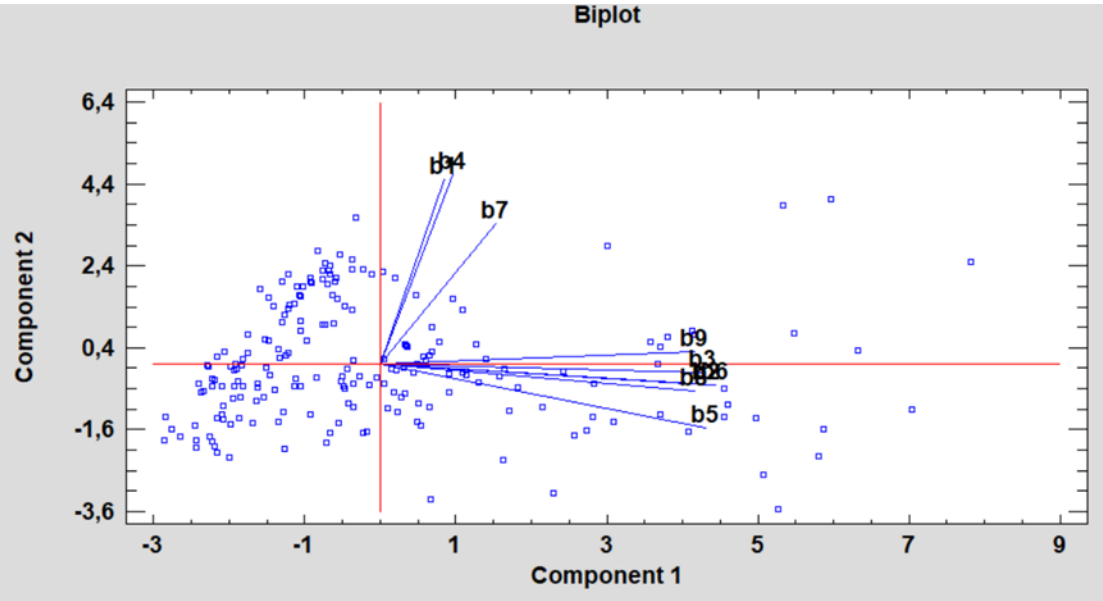

Figure 9. Projection of the objects and variables on PC1/PC2 for data set (biplot): Forests types (Set 1).

Fig. 9 shows the information of the two graphs, presented in **Fig.** 7 and **8**. We can see points from class S which have large values of variables B2/ B3 / B5 / B6 / B8 / B9 and points from class D, H and O which have large values of variables B1/ B4 / B7.

4.2. Supervised learning methods - classification

4.2.1. Linear discriminant analysis

First we will use the LDA method. With that method we define classification model using data set “Training”. Next, with the data set “Testing”, we control if defined model works correctly. For this part of our study we use another software: Matlab. With the function confusionmat we obtain a confusion matrix (**Tab.2**).

Table 2. Confusion matrix for data set 1 after LDA.

Reference data	S	H	D	O	Total	Correctly classified (%)
S	123	12	1	0	136	90%
H	6	32	0	0	38	84%
D	17	1	83	4	105	79%
O	2	1	12	31	46	67%

This classification method is rather effective because the percentage of classification are quite high. For example 90% objects from class S are correctly classified. The least favorable result was obtained for class D. 33% of objects were incorrectly included to other classes.

4.2.2. Classification and Regression Trees

Then we use CART. This method (Classification and Regression Trees) builds a decision tree (**Fig. 10**) using the values of the data “Training” set and creates a model to classify new samples.

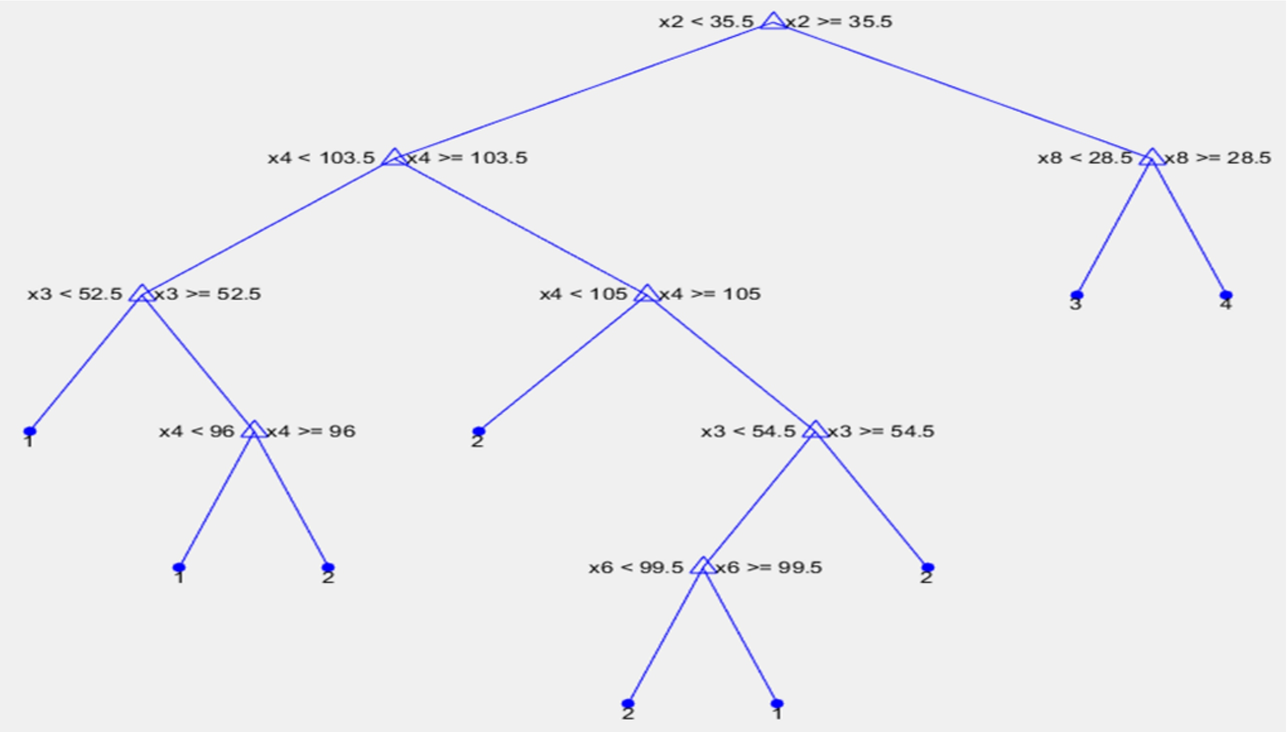

Figure 10. CART analysis for data set: Forests types – Training.

After this classification and definition of model, we decided to validate it using a new data set, ”Testing” to judge the reliability of our classification method. To see if our results are viable we realized a confusion matrix (**Tab.3**).

Table 3. Confusion matrix for data set 1 after CART.

Reference data	S	H	D	O	Total	PA (%)
S	102	29	5	0	136	75%
H	7	31	0	0	38	84%
D	18	2	70	15	105	67%
O	2	0	6	38	46	83%

Unlike the first confusion matrix, here the percentages are rather weak and a large number of errors are observed. So, our classification method is not efficient enough to get reliable results to be used correctly.

4. Conclusions

In this study we analyzed data set which included photographs of forest areas of various types, as well as areas without a forest. We wanted to shows the utility of chemometrics in analysis of big data set. We use different aspect of this science to compare the data and extract hidden data, it was why we use different technique, to discover some of the tools used in this field.

For the Forest analysis we used two computational tools. Statgraphic64 software developed by Francestat, we use to obtain spider plot, matrix plot, cluster analysis and principal component analysis. The second software is Matlab and we used it for LDA and CART method.

First we applied spider plot and matrix plot to visualize multivariate data. We can see a homogeneity for the class S. The class H is characterized by a little higher values than the class S and the results are heterogeneous. The class D has also higher values than S but the values are homogeneous. And for the class O we can see a big variability, the values are very heterogeneous and higher than in all other classes.

Next cluster analysis with "Ward's" option was used. We can see two main clusters, the first one containing the classes D, H and S and the other containing the class O. Using PCA we first obtained information that two principal components are important which describe ca. 78% of variability included in the dataset. As before we have seen, that the class D, H and S are very similar and homogenous but class O are very different and heterogeneous.

Classification models were built using LDA and CART. Next they were validated with external data set. The correctness of classification was on the level 67-90%, depending on the class. LDA more often it provided reliable results.

Acknowledgments

I would like to thank the university AGH of Krakow for its welcome that allowed me to do my internship but more particularly my tutors, professors Jakubowska and Reczyński for their help and support throughout the course.

References

[1] http://www.statisticssolutions.com/cluster-analysis-2/
[2] J. Miller , Jane C. Miller, Statistics and Chemometrics for Analytical Chemistry, Pearsons 2018.
[3] Richard G. Brereton, Chemometrics: Data Analysis for the Laboratory and Chemical Plant, Wiley 2004.
[4] https://archive.ics.uci.edu/ml/datasets/Forest+type+mapping
[5] B. Johnson, R. Tateishi, Z. Xie, Using geographically weighted variables for image classification, Remote Sensing Letters, 3 (2012) 491-499.