Cent. Eur. J. Chem. • 11(2) • 2013 • 151-159
DOI: 10.2478/s11532-012-0145-8

VERSITA

Central European Journal of **Chemistry**

Application of chemometrics to identifying solid fuels and their origin

Research Article

Marcin Sajdak*

Institute for Chemical Processing of Coal,
41-803 Zabrze, Poland

Received 29 June 2012; Accepted 21 August 2012

Abstract: The aim of this work was to implement a chemometric analysis to detect the relationships between the analysed variables in samples of solid fuels. Efforts are being made to apply chemometrics methods in environmental issues by developing methods for the rapid assessment of solid fuels and their compliance with the required emission characteristics regulations. In the present investigation, two clustering techniques—hierarchical clustering analysis (HCA) and principal components analysis (PCA)—are used to obtain the linkage between solid fuel properties and the type of sample. These analyses allowed us to detect the relationships between the studied parameters of the investigated solid fuels. Furthermore, the usefulness of chemometrics methods for identification of the origin of biofuels is shown. These methods will enable control of the degree of contamination.

Keywords: *Hierarchical clustering analysis • Data mining • Solid fuels • Biomass • Principal components analysis*

1. Introduction

Chemometric analysis is a powerful tool for data analysis that increases the amount of information that can be obtained in the same amount of research time. This technique makes it possible to resolve problems and answer questions about the nature of tested objects and the relationships between them. Chemometrics is a branch of chemistry that uses mathematics, probability, statistics and informatics in research experiments to obtain the maximum amount of useful information based on an analysis of the data. Chemometrics also makes it possible to find mutual correlations between fuel properties and within a group of tested objects. The following are additional benefits of implementing a chemometric analysis:

- presentation of measurement data in a form allowing their effective use,
- cost optimisation using a smaller amount of expensive reagents,
- reduction in the time needed to obtain the necessary data,
- environmental protection (reduction in the amount of reagents released into the environment) [1].

Due to the many advantages of this method, it is frequently used in other branches of science, for example, to assess the quality of the various fuels used in industry. This problem is particularly important for environmental protection because the characteristics and quality of fuels significantly affect the emission of pollutants into the environment. The aim of the present research was to identify and describe variables that might be used as markers of solid fuels origin. At this step it can be helpful to implement a chemometric analysis to detect relationships between the contents of the analysed parameters (elements) present in the tested materials. Other project targets included applying these techniques to evaluate the degree of biofuel contamination by undesirable substances and reducing the number of necessary laboratory tests and the time of analysis, thus leading to a decrease in the costs.

1.1. Clustering method

The clustering method is a tool of exploration and data mining, and thus, is an excellent method of acquiring new knowledge and information about the examined

* *E-mail: msajdak@ichpw.zabrze.pl*

object. The results of two methods are presented in the investigation: hierarchical clustering analysis (HCA) and principal components analysis (PCA) of solid fuels, including biomass and recyclable solid fuels. Particular emphasis was placed on those aspects of the analysis that affect the usefulness of these methods in the process of data mining. Attention was also drawn to the data preparation techniques used for different chemometric methods.

2. Experimental procedure

2.1. Collected data

The data used for the chemometric analysis, the results of which are presented in this article, were collected from the generally available PHYLLIS database. For the analysis, the values of oxides (Al – aluminium, Ca - calcium, Fe - iron, K - potassium, Mg - magnesium, Na - sodium, P - phosphorus, Si - silicon, Ti - titanium) were calculated based on the content of elements in the dry weight material.

The chemometric analysis was performed on the data matrix containing 17 variables (the analysed parameters) for each of the 104 solid fuels.

In the present research, data was used describing 104 solid fuels belonging to 6 groups: grass and plant biomass, untreated and treated wood, biochar (char), peat and coal. The choice of these materials was required to properly perform research, further analysis and interpretation of results. Above mentioned groups are characterized by different energy content, mineralogical composition and the degree of coalification. Degree of coalification rises, starting from biomass to coal. These sources of energy, with the exception of biochar, belong to previously described groups, and are commonly used as fossil and renewable fuels. Biochar has been placed among them because of its potential use as a energy source.

Before examining the results of the tests, all results have been standardised according to Eq. 1. Standardisation of data was necessary to conduct each phase of the chemometric analysis properly, and to interpret the results correctly [2].

Standardisation is designed to ensure equal influence (validity) of each variable, regardless of its size range and type of used units. With the standardisation of the variables, obtained values are characterised by an average equal to 0 and a variance equal to 1.

$$x_{ij} = \frac{a_j - b_j}{s_j} \quad (1)$$

where: x_{ij}- standardised parameter value; a_{ij} - initial value of the parameter; b_j-average value of the parameter; s_j - standard deviation of the j-th parameter.

3. Results and discussions

3.1. Cluster analysis

Cluster analysis is an exploratory multivariate method that can be used to describe the relationships among variables. Some mathematical rules can be used to examine the similarity between variables cases. For the hierarchical clustering analysis (HCA), Single linkages (nearest neighbor) method was used to obtain a cleaner plot of clusters. This method is believed to be as a very efficient one. The joining of the tree clustering method uses the dissimilarity distance between objects when using the clusters. Similarities are collection of rules that serves as criterion for grouping or separating items. The Euclidean distance was chosen to carry analysis [3].

The purpose of cluster analysis is to determine the mutual similarities between the analysed objects and attributes that describe them. In this analysis, there is no information about the homogeneity of the analysed data set. The first step in clustering objects is evaluating their similarity (or dissimilarity): the distance or correlation coefficient can be used as a measure of (dis)similarity. One way of measuring the distance between two objects *i* and *j* in HCA is to use the Euclidean distance [4,5].

$$d_{ij} = \sqrt{\sum_{k=1}^{m} \left(x_{ik} - x_{jk}\right)^2} \quad (2)$$

where: m – the number of variables; d_{ij}-Euclidean distance between objects i and j; x_{ik}-value of the variable I; x_{jk}-value of the variable j.

Using vector notation, Eq. 2 becomes:

$$d_{ij}^{\ 2} = \left(x_i - x_j\right)^T \cdot \left(x_i - x_j\right) \quad (3)$$

where x_i and x_j are the column vectors of the two objects and *T* stands for “transpose”.

The smallest value is the Euclidean distance, and the largest is the similarity between the objects. The Euclidean distance can be graphically interpreted as the length of the vector starting at i and ending at j.

This method does not directly apply the values of the variables. Rather, the variables are subjected to a previous standardisation. An analysis of the agglomeration of the distance matrix indicating the degree of similarity between the studied variables was performed. The results are presented in Table 1.

Table 1. Euclidean distance matrix for study variables.

	Ash	HHV	C	H	O	N	S	Cl	Al	Ca	Fe	K	Mg	Na	P	Si	Ti
Ash	0.0																
HHV	15.0	0.0															
C	14.9	1.6	0.0														
H	17.5	16.0	16.5	0.0													
O	17.0	18.8	18.9	6.3	0.0												
N	11.9	12.3	12.4	14.4	15.7	0.0											
S	13.2	9.3	9.6	15.6	17.7	12.2	0.0										
Cl	9.5	15.5	15.3	15.6	14.8	12.9	14.2	0.0									
Al	8.8	11.2	11.2	17.1	18.2	11.7	11.8	13.4	0.0								
Ca	8.0	15.5	15.2	16.3	15.2	12.7	14.5	11.2	12.6	0.0							
Fe	7.5	10.6	10.8	15.6	16.9	11.0	10.0	11.9	6.6	10.9	0.0						
K	9.4	16.5	16.4	15.1	14.1	12.7	14.7	5.3	13.9	10.8	12.7	0.0					
Mg	7.6	14.9	14.8	16.3	16.0	11.5	13.7	7.9	11.8	8.5	10.0	7.7	0.0				
Na	6.6	15.9	15.5	17.0	15.8	12.7	14.1	8.3	11.7	8.2	10.3	9.0	7.4	0.0			
P	10.2	14.7	14.6	13.7	13.4	10.2	13.8	10.3	13.3	9.3	12.1	8.2	8.6	11.4	0.0		
Si	7.8	14.8	14.7	16.0	15.8	12.8	13.6	11.4	10.5	9.9	9.6	12.0	10.9	9.9	10.7	0.0	
Ti	11.4	10.9	10.4	17.1	17.6	10.8	12.6	14.0	9.2	12.4	9.2	14.9	13.1	12.2	13.9	12.3	0.0

Direct interpretation of the data set because of its large size is relatively difficult; therefore, a Horizontal Hierarchical Tree Plot was used for easier data interpretation.

The combination of the horizontal hierarchical tree plot features with the chart of the stages of binding (as presented in Figs. 1a and 1b) facilitates correct distribution of the tree plot for each subgroup and makes it possible to avoid grouping mistakes.

Considering the entire data set without distribution into groups by type of material (Fig. 1a), the following conclusions can be drawn:

- The concentrations of carbon, sulphur, nitrogen, hydrogen and oxygen in solid fuels have the greatest influence on the determination of combustion heat of material. These are evident conclusions from the viewpoint of chemistry, but in the initial stage of the analysis, they confirm the correctness of the calculation (Chemometric analysis) and the inference and correctness of the initial data preparation.

- A similarity relation between oxygen and hydrogen concentrations indicates the existence of structures of organic-rich oxygen, such as cellulose, hemicellulose and lignin, which occur in biomass.

A hierarchical clustering analysis was performed for the divided data set, including the type of analysed material: coal, char, and several types of biomass. The results of this analysis are shown in Fig. 2.

By analysing the characteristics of the horizontal hierarchical tree plot shown in Fig. 2, it is possible to find variables indicating the origin (type) of the sample and determine whether the sample is of plant origin (biomass or solid fuel is secondary) or fossil origin by assessing only the research results. As presented in Fig. 2, one should pay attention to the relevant groups of co-occurring elements: hydrogen and oxygen, aluminium and iron, and the ash content in combination with magnesium, sodium or potassium. These groups are present only in samples of plant origin. Char from biomass is similar in chemical composition to carbon. In fact, thanks to the tools used, we can easily qualify char as material originating from the thermal processing of biomass due to the correlations between oxygen and hydrogen and iron and aluminium.

The hierarchical clustering analysis can be used to obtain the linkage between the parameters and the types of investigated material. This will help reduce the time required for future analyses and allow for appropriate conduct with material and analysis.

3.2. Principal component analysis PCA

Multivariate statistical tools, such as the principal component analysis (PCA) technique, have been widely applied in the treatment of high-complexity data sets [6,7]. The PCA technique extracts the Eigenvalues and Eigenvectors from the covariance matrix of the original

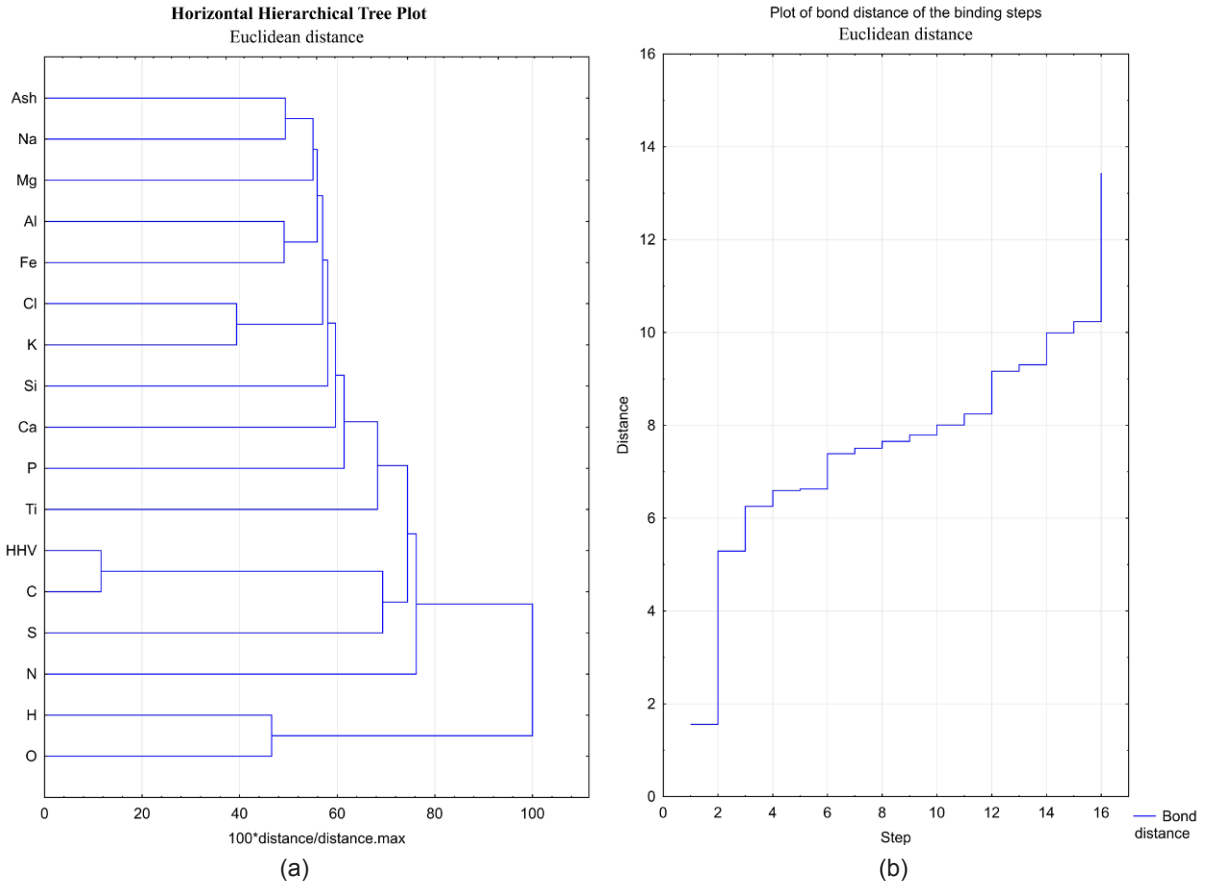

Figure 1. a) Horizontal Hierarchical Tree Plot characteristics analysed in the material; b) Graph HCA bond distances (Ash - ash, HHV - enthalpy of combustion).

variables. It makes it possible to find the association between variables, thus reducing the dimensionality of the data set. The principal components (PCs) are the uncorrelated (orthogonal) variables obtained by multiplying the original correlated variables with the Eigenvector (loading or weighing). The Eigenvalues of PCs are the measure of their associated variance. The participation of the original variables in the PCs is given by the loading, and the individual transformed observations are called scores [8]. The use of correlated variables in this analysis gives the best results. The main task is detecting the internal data structure and describing it with the other parameters resulting from the above-mentioned structure [3,4]. A characteristic feature of principal components is the ability to determine the main factors (PCs) in order of decreasing volatility of the stock. The measure of this stock is the eigenvalues. In some cases, it is possible to assign the main components and groups of objects to a certain chemical or physical interpretation [4].

The principal component analysis technique is used primarily for:

1. The reduction of the data space,
2. The transformation of correlated input variables in the output main components, and
3. The graphical presentation of the structure of a multidimensional data set in the plane with minimal distortion of information.

These techniques were used to obtain orthogonal factors from the seventeen variables remained in a partial correlation. Seven PCs were selected, which is approximately 91% of the studied variables. The individual variance values of the principal components are shown in Table 2.

As shown in Fig. 3, the first three PCs already describe 71% of the variability of the data. This makes a simple (three-dimensional) expression of the relationship between PCs possible. Fig. 3 shows how to change the scree test graph for the total variance coordinates.

The total number of PCs in this case is 16 components. All variables were compensated more than twice. As mentioned earlier, the main observation can be attributed to a physical interpretation or chemical means.

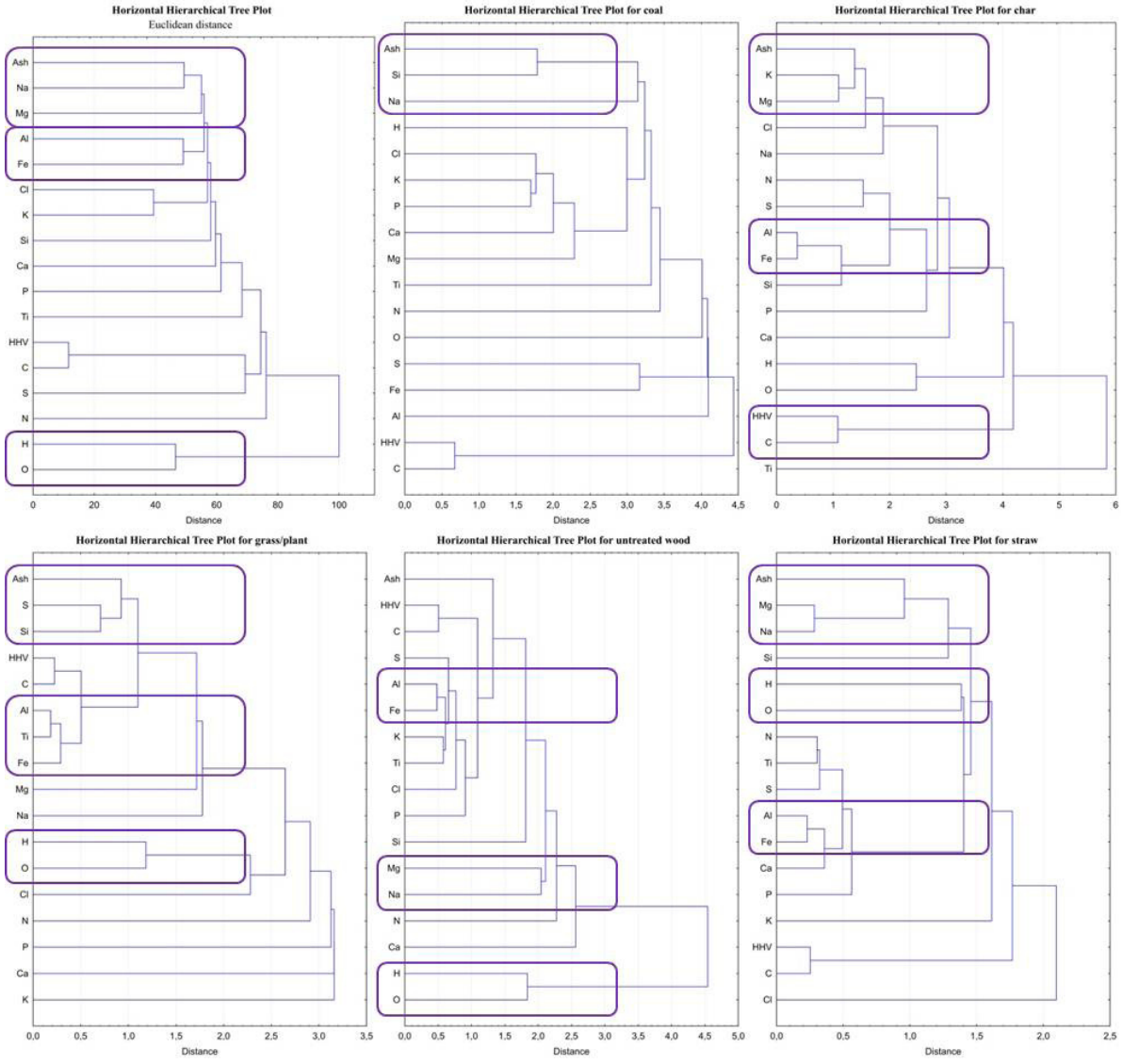

Figure 2. Horizontal Hierarchical Tree Plot for each type of test material (Ash - ash, HHV - enthalpy of combustion).

With this method, two primary principal components (2 and 3) from the analysis of data covering coal, biomass, and the products of their thermal treatment of the sample can confirm the affiliation to a particular group of fuels.

Application of principle component analysis has confirmed the results obtained by cluster analysis. As presented in Fig. 4a, agglomeration determined using principal component correspond to those obtained by the cluster analysis, and presented on Fig. 1a.

A chemical interpretation of the second and third PC's was also possible, combining the results of the analyzed variables and analyzed materials (Fig. 4B).

The second PC allows you to describe the materials studied in terms of:

- degree of coalification sample - the higher the value of the sample PC 2, the greater the degree of coalification in the samples
- enthalpy of combustion - the higher the value, the greater the energy content of the sample (inverse relationship between carbon content and the enthalpy of combustion of oxygen and hydrogen in the samples)

The thirtd PC allows you to describe the materials studied in terms of the content of micro and macro elements. Thanks to the second and thirtd PC's, it is possible to classify the test materials to the respective groups, for reasons of their kind. This classification is presented in graphical form in Fig. 5.

The relationships among the variables can be represented in the form of predictive equations, as presented in Table 3.

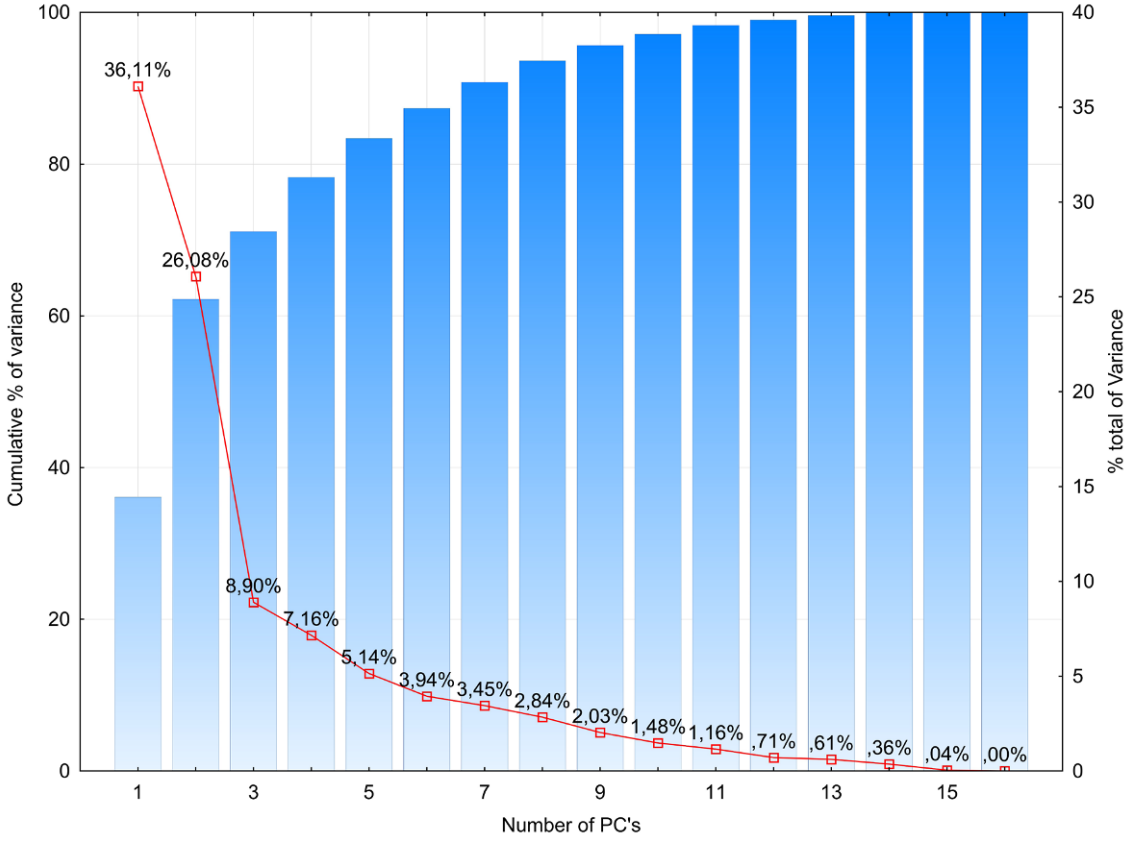

Figure 3. The scree test plot.

Table 2. The values characterising the principal components designated for the data set.

Value	Eigenval	% total Variance	Cumul. Eigenval	Cumul. %
1	5.78	36.11	5.78	36.11
2	4.17	26.08	9.95	62.19
3	1.42	8.90	11.37	71.09
4	1.15	7.16	12.52	78.25
5	0.82	5.14	13.34	83.39
6	0.63	3.94	13.97	87.33
7	0.55	3.45	14.52	90.78
8	0.45	2.84	14.98	93.62
9	0.32	2.03	15.30	95.65
10	0.24	1.48	15.54	97.13
11	0.19	1.16	15.73	98.29
12	0.11	0.71	15.84	99.00
13	0.10	0.61	15.94	99.60
14	0.06	0.36	15.99	99.96
15	0.01	0.04	16.00	100.00
16	0.00	0.00	16.00	100.00

Table 3. Summary of the correlation function for the studied char variables.

Sample types	Function	The correlation coefficient
Char	C=0.5023+0.227·HHV	0.997
	O=-6.008+6.551·H	0.879
	Al=-947.800+1.912·Fe	0.991
	K=-2289+1121.9·Ash	0.991
	Mg=460.18+175.44·Ash	0.991
	Mg=837.370+0.154·K	0.992

As showed in Table 3, the coefficients of determination for the analyzed variables in these char samples is high and ranges from 0.88 to 0.99. Such high values of the determination coefficient allows the use of equations derived within the considered types of materials.

It is possible to calculate the correlation of each variable contained in the subgroup using principle component analysis, which shortens the time needed to perform the analysis because it is not necessary to perform all 17 tests to obtain the maximum viewable information about the examined object.

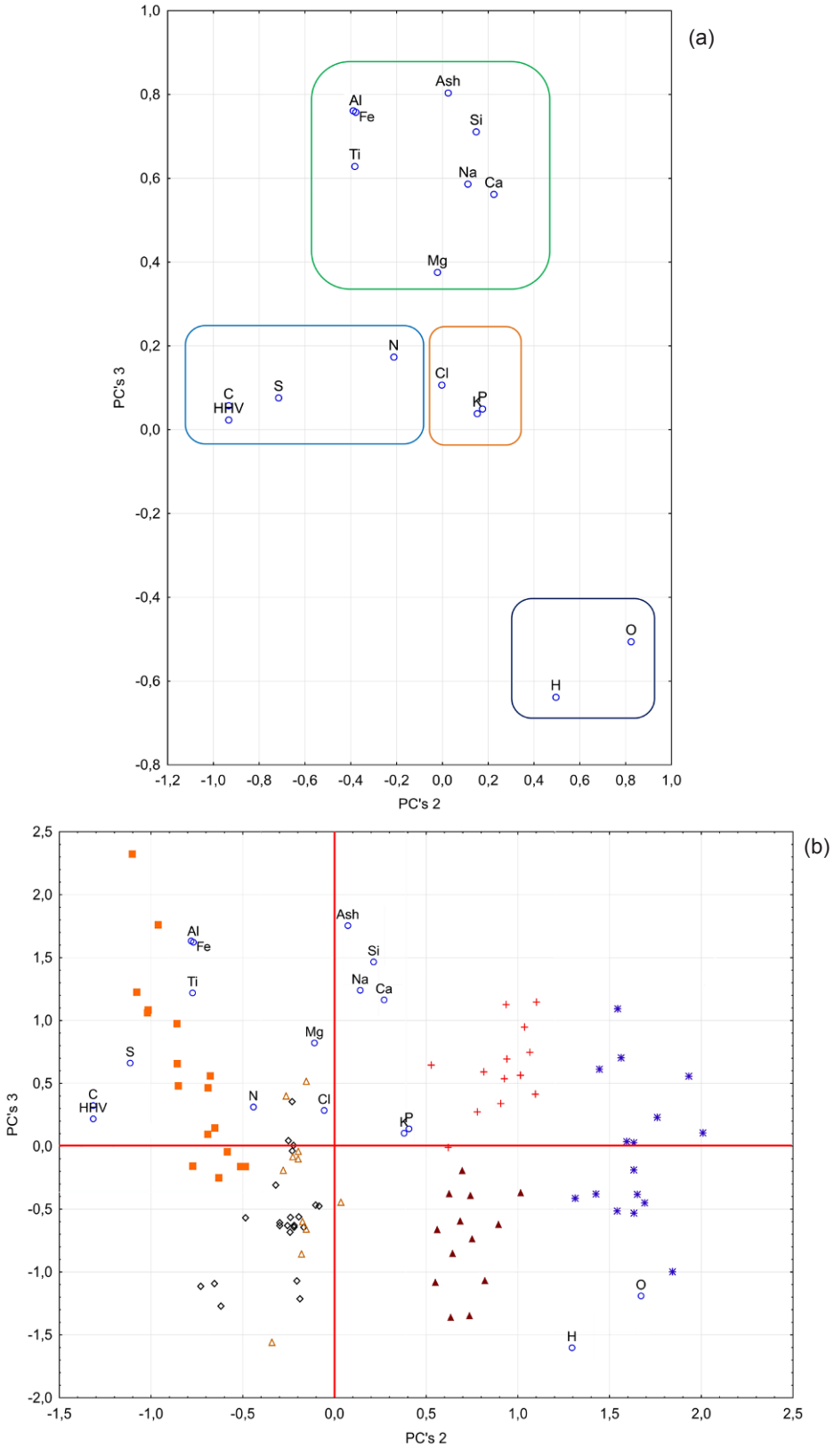

Figure 4. Plot of Principal Component Analysis of a) analysed variables, b) analysed materials.

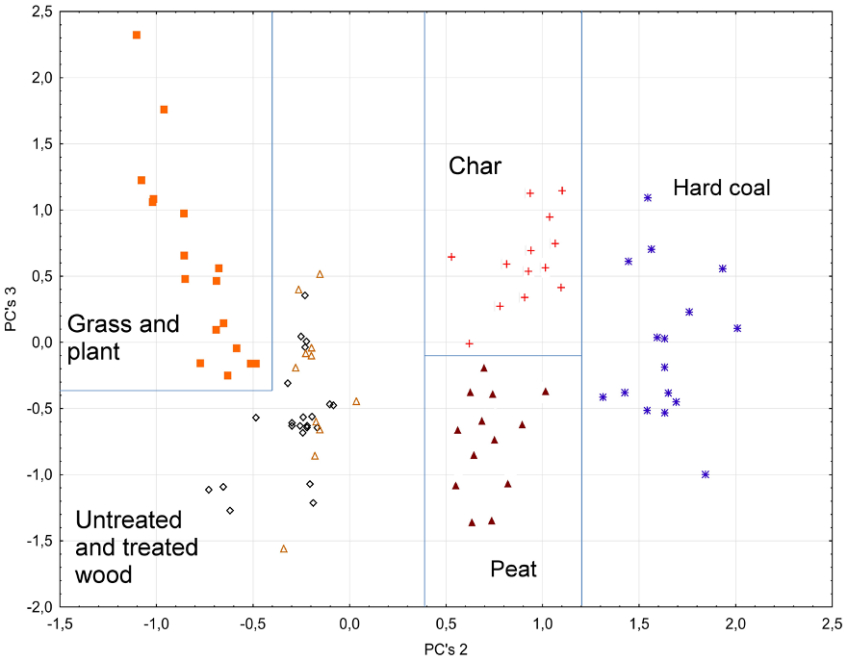

Figure 5. Plot of Principal Component Analysis of analysed materials.

Analysing the results of correlation analysis for the rest of types these materials, it can be seen that in the case of alternative fuels such as grass and untreated wood, as well as in the case of the thermal processing of biochar (char), there is a strong linear correlation between the content of iron and aluminum. However, this correlation does not occur in fossil fuels like coal or chemically modified biomass (by impregnation). The results of determination coefficients for the relationship between iron and aluminum are presented in Table 4.

Table 4. Coefficient of determination and function of samples analysed.

Sample types	Function	Coefficient of determination
Coal	Al = 0.015+0.7501·Fe	0.279
Biochar (char)	Al = -0.0054+1.0144·Fe	0.990
Grass	Al = -0.0249+0.9954·Fe	0.864
Husk	Al = $-5.8157 \cdot 10^{-5}$+1·Fe	1.000
Untreated wood	Al = -0.0014+0.9345·Fe	0.942
Treated wood	Al = 0.0876+0.5443·Fe	0.407
Straw	Al = 0.0047+1.0058·Fe	0.994

4. Conclusion

In the present investigation, the possibility of using chemometrics tools in search of markers (variables) that might indicate the origin of different solid fuels was investigated, thus confirming the reliability of the certificate of origin. The study used a variety of materials classified as biomass (straw, grasses, energy crops, wood) and coal and biomass char.

By applying chemometric analysis (grouping methods), it was possible to detect the interaction between the contents of the analysed elements occurring in the investigated materials. The study showed the usefulness of the chemometrics methods for identifying the origin and the type of biofuels.

Studies using chemometric techniques are going to make it possible to implement these methods to assess the degree of contamination with undesirable substances biofuels.

In the future, these studies may bring very tangible benefits by reducing the number of necessary laboratory tests and the analysis time, thereby lowering costs.

Acknowledgments

These studies were funded from the budget of research task No. 4 “Development of integrated technology of fuels and energy from biomass, agricultural waste, and others as part of the strategic program of research and development: Advanced technology of obtaining energy” provided by the NCBiR.

References

[1] A. Astel, Chemometria jako wydajne narzędzie analizy danych środowiskowych, LAB Laboratoria, Aparatura, Badania 1, 44 (2008) (In Polish)

[2] A. Astel, J. Mazerski, J. Namieśnik, In: J. Namieśnik, W. Chrzanowski, P. Szpinek (Eds.), Nowe horyzonty i wyzwania w analityce i monitoringu środowiska (CEEAM, Gdańsk, 2003) 131-162 (In Polish)

[3] J. Mazerski, Chemometria praktyczna (Malamut, Warszawa, 2009) (In Polish)

[4] O. Abollino, M. Aceto, M. Malandrino, E. Mentasti, C. Sarzanini, F. Petrella, Chemosphere 49, 545 (2002)

[5] P. Khare, B. P. Barah, P.G. Rao, Fuel 90, 3299 (2011)

[6] A. Giacomino, O. Abollino, M. Malandrino, E. Mentasti, Analytica Chimica Acta 688, 122 (2011)

[7] E. Marengo, M.C. Genaro, E. Robotti, P. Rossanigo, C. Rinaude, M. Roz-Gastaldi, Analytica Chimica Acta 560, 172 (2006)

[8] K.P. Singh, A. Malik, V.K. Singh, D. Mohan, Analytical Chemistry Acta 550, 82 (2005)