

THE USE OF CHOSEN METHODS OF STATISTICAL AND CHEMOMETRIC ANALYSIS IN FORENSIC EXAMINATIONS OF GLASS

Grzegorz ZADORA

Faculty of Chemistry, The Jagiellonian University, Cracow

Zuzanna BROŻEK-MUCHA

The Institute of Forensic Research, Cracow

ABSTRACT: Selected statistical and chemometric methods were applied in the interpretation of analytical data obtained during examinations of glass fragments by means of refractive index measurement and elemental analysis. Results of refractive index measurements with the use of a GRIM system were analysed using the „3 sigma” rule and Student’s t-test. The elemental composition of the studied glass samples was determined by the SEM-EDX method and cluster analysis. An attempt was made to utilise selected statistical and chemometric methods both in order to classify an unknown glass sample into a use-type group of glass objects, and to discriminate between an evidential glass sample and a comparative glass sample.

KEY WORDS: SEM-EDX; GRIM; Forensic glass identification; Student’s t-test; “3 sigma” rule; Cluster analysis.

Z Zagadnień Nauk Sądowych, z. XL, 1999, 33–71

Received 15 June 1999; accepted 15 October 1999

INTRODUCTION

There are two main aims of forensic examinations of evidence materials. In the case of lack of a comparative material, the evidence material is studied in order to classify it into a use-type of glass objects, taking into account their specific chemical and physical properties. When both evidence and comparative materials are available the task of a forensic chemist is to answer, whether they could have come from the same object, i.e. to carry out a comparative analysis, also called a discrimination [10].

Only in the case of detection of significant differences in physical properties or chemical contents of the materials being compared, is it possible to ascertain that they did not originate from the same object. To discriminate between objects of different classes is one of the easiest tasks. However, to perform a reliable differentiation between objects of the same class, which by definition have similar features, is more difficult.

With the advent of modern analytical methods characterised by high sensitivity and precision, the short duration of the process of analysis and “user-friendly” operating systems resulting from equipment automation, one can obtain a large amount of good quality analytical data. The following analytical methods, frequently used for forensic glass examinations, possess the above mentioned attributes: glass refractive index measurement technique – GRIM [4, 5, 6, 7, 8, 12, 13] as well as methods of elemental analysis, e.g. X-ray spectrometry realised either with scanning electron microscopy (SEM-EDX) [1, 15], or as X-ray fluorescence (XRF) [11]. Nevertheless, the appropriate interpretation of the analytical results obtained remains an important challenge for the forensic chemist.

In the presented work selected methods of statistical analysis were described, which according to the authors could be helpful in an assessment of variability of the studied material and in revealing discrete differences between samples of objects belonging to the same class. With the use of an appropriate statistical method (especially in the case of multivariable description of objects) group identification can be limited to a group of smaller population.

The choice of the particular method of statistical analysis to be applied in order to solve a problem of classification or discrimination, depends mainly on the number of features taken into account when describing the studied objects.

CLASSIFICATION OF OBJECTS DESCRIBED BY A SINGLE VARIABLE

In order to classify an object described by one feature only, a data base of ranges of values which the feature can attain in various groups is required. Let us consider the following example: five glass fragments D1, D2, D3, D4 and D5 were recovered on the clothes of a car accident victim. Comparative material (P) taken from the damaged headlamp of a car belonging to a person suspected to have caused the fatal accident, was also provided for examination.

Refractive index, RI, was measured for all the samples [13]. Results of the measurement are collected in Table I. A fragment of the data base, including values of refractive indices of glass samples taken from car windows (c) and car headlamps (r) is shown in Table II.

A classification of the evidence samples should be carried out first. If any of the evidence samples were classified as coming from car headlamps, the next step of the examinations would be discrimination, i.e. answering the question, whether the evidence and the comparative samples could have originated from the same object.

Before the classification had been completed, mean values of refractive index for each of the evidence and the comparative samples were calculated from the following formula:

$$\overline{RI}_i = \frac{\sum_{j=1}^n RI_{ij}}{n} \quad \{1\}$$

where: i – index of the sample ($i = D1, \dots, D5, P$); n – number of measurements performed for each sample ($n = 9$).

TABLE I. RESULTS OF THE MEASUREMENT OF REFRACTIVE INDEX FOR EVIDENCE SAMPLES D1, ..., D5 AND COMPARATIVE SAMPLE P.

Measurement	Refractive index					
	Evidence samples					Comparative sample
	D1	D2	D3	D4	D5	P
1	1.5143	1.5173	1.5252	1.5124	1.5218	1.5122
2	1.5146	1.5166	1.5254	1.5126	1.5219	1.5122
3	1.5144	1.5168	1.5256	1.5124	1.5218	1.5123
4	1.5148	1.5168	1.5258	1.5128	1.5216	1.5128
5	1.5149	1.5171	1.5253	1.5124	1.5211	1.5124
6	1.5142	1.5169	1.5255	1.5123	1.5214	1.5122
7	1.5146	1.5165	1.5251	1.5123	1.5216	1.5127
8	1.5145	1.5166	1.5255	1.5123	1.5217	1.5129
9	1.5146	1.5163	1.5255	1.5124	1.5218	1.5126
Mean value of RI	1.51454	1.51677	1.52543	1.51243	1.52163	1.51248

TABLE II. RANGES OF VALUES OF REFRACTIVE INDEX FOR GLASS SAMPLES OF CAR WINDOWS AND CAR HEADLAMPS (OUR DATA).

Group	Refractive index
Car headlamps (r)	1.5111 – 1.5169
Car windscreens (c)	1.5157 – 1.5244

In order to classify sample D1 the mean value of refractive index $\overline{RI}_{D1} = 1.51454$ is compared with the data base (Table II). A group of glass is searched for that would contain a refractive index of sample D1. One can observe that a value of 1.51454 is included in the characteristic range for the

car headlamps glass. This means that evidence sample D1 could originate from a car headlamp.

The mean value of the refractive index measured for sample D2 ($\overline{RI}_{D2} = 1.514547$) allows us to classify it both into the car window glass group and into the car headlamps group; the mean refractive index of sample D3 ($\overline{RI}_{D3} = 1.52543$) does not fall within any of the ranges of the presented data bases. In this case classification is not possible. The mean values of refractive indices for samples D4 and D5 ($\overline{RI}_{D4} = 1.514547$; $\overline{RI}_{D5} = 1.514547$) show that sample D4 could originate from a car headlamp, and sample D5 from a car window.

Summarising this stage of the study, two evidence samples – D1 and D4 – were classified in the car headlamp glass group. Now, we should answer the question as to whether any of these evidence samples show similarity to the comparative sample P.

Discrimination of the glass samples on the basis of their refractive indices can be realised using, e.g. “3 sigma” rule or Student’s t-test.

DISCRIMINATION OF OBJECTS DESCRIBED BY A SINGLE VARIABLE

The “3 sigma” rule

In order to use the “3 sigma” rule [2, 9], the confidence ranges based upon the mean values of refractive indices are determined for the compared evidence samples, D1 and D4, and the comparative sample P. Direct comparison of mean values of refractive indices determined for the analysed samples is not allowed because the mean value is only an estimation of the real value of the refractive index for each sample (μ_i , $i = D1, D4, P$). That is why ranges of values of refractive indices (confidence ranges) – within which the real value of the refractive index (μ_i) is most probably contained – ought to be calculated. This probability is equal to 99.73% in the case of the “3 sigma” range.

The confidence range is built in such a manner that the low limit is equal to the difference between RI_i value and triple value of the standard deviation and the upper limit is equal to the sum of RI_i value and triple value of the standard deviation. The value of the standard deviation σ_i can be calculated from the following formula:

$$s_i = \sqrt{\frac{\sum_{j=1}^n (RI_{ij} - \overline{RI}_i)^2}{n-1}} \quad \{2\}$$

where: $i = D1, D4, P$; n – number of measurements ($n = 9$).

The results are shown in Table III. The confidence ranges of refractive indices determined for D1 and P do not overlap, so evidence sample D1 did not originate from the same glass object as comparative sample P. On the contrary, for D4 and P samples the confidence ranges of refractive indices overlap. This means that evidence sample D4 and comparative sample P could have originated from the same car headlamp.

TABLE III. THE CONFIDENCE RANGES OF RI_i DETERMINED FOR SAMPLES D1, D4, AND P.

Sample	Mean value	s_i	Confidence range
D1	1.51454	$2.2 \cdot 10^{-4}$	1.51387 – 1.51522
D4	1.51243	$1.6 \cdot 10^{-4}$	1.51194 – 1.51293
P	1.51248	$2.7 \cdot 10^{-4}$	1.51165 – 1.51331

Student's t-test

Another way of performing discrimination is Student's t-test [2, 9] for the mean value of independent trials. The test allows us to check the hypothesis of equality of two mean values which are being compared to one another. This is the so-called null hypothesis (H_0). In this case we test the null hypothesis, that mean values of refractive indices, determined for the evidence and the comparative samples, are statistically equal, $H_0: \mu_D = \mu_P$. If the hypothesis is confirmed, one can state that the fragments of the evidence and the comparative glass are similar and could originate from the same glass object. Rejecting the null hypothesis, one could say that the fragments of the evidence and the comparative glass were different and most probably originated from different objects.

The Student's t-test can be used when two assumptions are fulfilled. Firstly, it is assumed that the considered results have a normal distribution [2, 9]. Usually, this can be considered true. If the hypothesis has to be tested, the Kolmogorov-Smirnov or chi-squared (χ^2) tests can be performed [9]. Taking into account the above remarks, in the presented example a normal distribution of results of refractive indices determined for examples D1, D4 and P was assumed.

The second assumption concerns variances of the results obtained for both analysed samples (σ_1^2 and σ_2^2).

In this case the null hypothesis assumed equality of these variances ($H_0: \sigma_1^2 = \sigma_2^2$). It is to be tested with the use of the F test [2, 9] in which the value F is calculated from the following formula:

$$F = \frac{s_1^2}{s_2^2}, \quad \text{gdzie } s_1^2 > s_2^2; \quad \{3\}$$

$$s_1^2 = \frac{\sum_{j=1}^n (x_{ji} - \bar{x}_i)^2}{n-1} \quad (i=1, 2). \quad \{4\}$$

Then the obtained value F is compared with the value of $F_{r_1, r_2, \alpha}$ taken from tables for r_1 and r_2 degrees of freedom and for the assumed confidence level α . The values of degrees of freedom were calculated from the following formulae:

$$r_1 = n_1 - 1, \quad \{5\}$$

$$r_2 = n_2 - 1, \quad \{6\}$$

where n_1 and n_2 – numbers of measurements performed for samples of variances s_1^2 and s_2^2 .

As the number of measurements of refractive index was 9 ($n_1 = n_2 = 9$) for each of the analysed samples, thus $r_1 = r_2 = 8$.

The value of α is most frequently taken as 0.05. Thus, in the case of accepting the null hypothesis there is a 5% risk of committing a fault of rejecting the opposite hypothesis that could have been true.

When for a pair of compared objects $F_{r_1, r_2, \alpha}$ is greater than F calculated from formula (3), the null hypothesis can be accepted and Student's t-test for mean values of the independent trials can be applied. The result from test F performed for pairs D1 – P and D4 – P was that the null hypothesis could be accepted in both cases as the values of F calculated were smaller than $F_{8,8,0.05} = 3.44$ (Table IV).

TABLE IV. VALUES OF VARIABLE F FOR PAIRS OF SAMPLES D1 – P AND D4 – P.

Pair of samples	Value of variable F
D1 – P	1.53
D4 – P	2.80

Prior to an analysis of the null hypothesis on the equality of the mean values of refractive indices measured for the evidence and the comparative samples, i.e. for pairs: D1 – P and D4 – P, one should determine the value of variable t . As the number of measurements of RI for each of the considered samples is the same ($n_1 = n_2 = n = 9$), the variable t can be expressed according to the following formula:

$$t = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\bar{s}^2}} \sqrt{\frac{n}{2}}, \quad \{7\}$$

where: \bar{x}_1, \bar{x}_2 – the mean values of the established features in the compared samples. The value of \bar{s}^2 can be calculated from the following formula:

$$\bar{s}^2 = \frac{s_1^2 + s_2^2}{2}. \quad \{8\}$$

For the considered pairs the following values were obtained:

$$\begin{aligned} \bar{x}_1 &= \overline{RI}_{D1} = 1.51454, \\ \bar{x}_2 &= \overline{RI}_P = 1.51248, \\ t_{D1-P} &= 17.382 \text{ (from \{7\})}, \\ \bar{x}_1 &= \overline{RI}_{D1} = 1.51454, \\ \bar{x}_2 &= \overline{RI}_P = 1.51248, \\ t_{D4-P} &= 0.413 \text{ (from \{7\})}. \end{aligned}$$

Then, the calculated values of t_{D1-P} and t_{D4-P} are compared to $t_{r,\alpha}$ taken from tables for $r = n-1$ ($r = 8$) and $\alpha = 0.05$.

The null hypothesis should be accepted when the value of $t_{r,\alpha}$ is greater than that calculated, t_{cal} . In the opposite case, the null hypothesis should be rejected.

Variable $t_{r,\alpha}$ taken from tables for $r = 8$ and $\alpha = 0.05$ equals 2.306. It is greater than $t_{D4-P} = 0.413$. In this case one can accept the hypothesis and say that evidence sample D4 and comparative sample P were similar to each other and so they could have come from the same glass object. Conversely, for pair D1 – P, $t_{D1-P} = 17.382$ being greater than $t_{r,\alpha}$ was obtained and so this required rejection of the null hypothesis.

CLASSIFICATION OF OBJECTS OF MULTIVARIABLE DESCRIPTION

An example

Problems of classification and discrimination of objects described by more than one feature can be illustrated in the following example. For five evidence glass samples D1, ..., D5 and a comparative glass sample, P, the content of six elements (Na, Mg, Al, S, K, Ca) was established by means of SEM-EDX. Each sample was analysed five times. The obtained results together with the calculated mean values are collected in Table V.

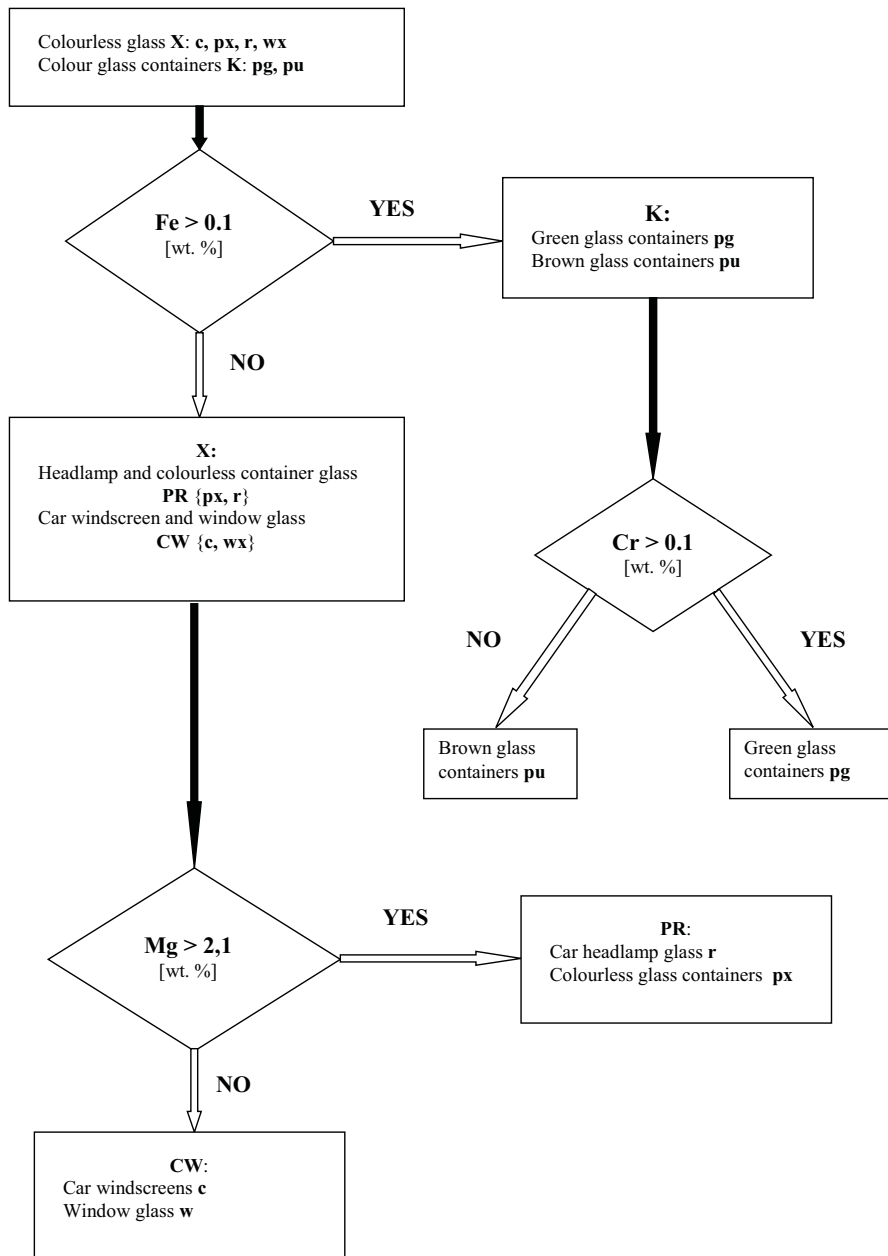


Fig. 1. A glass classification scheme – a non-statistical approach.

Six objects ($m = 6$) described by six variables ($n = 6$) will be considered. Included in Table V are mean values of element contents; they can also be presented as a matrix of size $m \times n$ {9}.

	Na	Mg	Al	S	K	Ca	
D1	9.36	0	1.18	0.12	1.58	4.78	} {9}
D2	9.46	2.78	0.16	0.12	0.18	5.88	
D3	10.46	1.48	0.50	0.08	0	6.84	
D4	9.56	0	1.06	0.14	1.32	5.78	
D5	9.58	2.26	0.34	0.08	0.18	5.86	
P	9.66	0	1.12	0.14	1.20	5.76	

Moreover, a data base was created that contained results of elemental analysis performed using SEM-EDX method for glass samples belonging to different use-type groups of glass, i.e. colourless container glass (px), green container glass (pg), brown container glass (pu), car headlamp glass (r), car windscreen glass (c) and window glass (w).

Non-statistical approach

Firstly, the non-statistical classification approach should be described. The data base was analysed in order to choose the elements whose concentration ranges do not overlap from one particular glass group to another. The concentrations of these chosen elements became the criteria of separation in a scheme of glass classification (Figure 1).

It is worth mentioning that the scheme only allows unambiguous classification of green (pg) and brown (pu) container glass and the differentiation of car window and windscreen glass groups (CW) from the car headlamp and colourless container glass groups (PR). The method does not allow more detailed classification within the CW and PR groups.

Using the above scheme, one can perform classification of sample D1 on the grounds of its elemental content (Table V). The sample did not contain iron (Fe) and magnesium (Mg) and could be classified into the PR set. Thus, the sample could have originated from either a car headlamp (r), or a colourless glass container (px). The remaining samples were classified in the same way. The obtained results are collected in Table VI.

Evidence samples D1, D3 and D4 as well as comparative sample P were classified into set PR. However, taking into account the elemental content (Table V) sample D3 can be eliminated from the group of samples similar to the comparative one. Sample D3 contained magnesium and did not contain potassium. Samples D1 and D4 revealed the same elemental content as sample P and so these samples were subjected to the discrimination procedure.

TABLE V. THE ELEMENTAL CONTENT [wt. %] OF THE STUDIED EVIDENCE SAMPLES D1, ..., D5 AND COMPARATIVE SAMPLE P.

Sample	Measurement	Element					
		Na	Mg	Al	S	K	Ca
D1	1	9.4	0.0	1.2	0.1	1.5	4.8
	2	9.3	0.0	1.2	0.1	1.6	4.8
	3	9.3	0.0	1.1	0.1	1.6	4.8
	4	9.4	0.0	1.2	0.1	1.6	4.7
	5	9.4	0.0	1.2	0.2	1.6	4.8
Mean value		9.36	0.00	1.18	0.12	1.58	4.78
D2	1	9.6	2.8	0.1	0.1	0.2	5.6
	2	9.3	2.9	0.1	0.2	0.2	6.0
	3	9.5	2.6	0.2	0.1	0.1	5.9
	4	9.4	2.9	0.2	0.1	0.2	6.0
	5	9.5	2.7	0.2	0.1	0.2	5.9
Mean value		9.46	2.78	0.16	0.12	0.18	5.88
D3	1	10.4	1.4	0.6	0.0	0.0	6.9
	2	10.5	1.5	0.5	0.1	0.0	6.9
	3	10.5	1.4	0.4	0.1	0.0	6.8
	4	10.4	1.6	0.5	0.1	0.0	6.8
	5	10.5	1.5	0.5	0.1	0.0	6.8
Mean value		10.46	1.48	0.50	0.08	0.00	6.84
D4	1	9.4	0.0	1.0	0.1	1.6	6.00
	2	9.8	0.0	1.0	0.1	1.3	5.7
	3	9.6	0.0	1.1	0.1	1.2	5.7
	4	9.5	0.0	1.0	0.2	1.3	5.8
	5	9.5	0.0	1.2	0.2	1.2	5.7
Mean value		9.56	0.00	1.06	0.14	1.32	5.78
D5	1	9.7	2.3	0.3	0.1	0.2	5.8
	2	9.4	2.2	0.4	0.1	0.2	6.0
	3	9.6	2.2	0.3	0.0	0.2	5.9
	4	9.7	2.4	0.3	0.1	0.2	5.8
	5	9.5	2.2	0.4	0.1	0.1	5.8
Mean value		9.58	2.26	0.34	0.08	0.18	5.86
P	1	9.6	0.0	1.0	0.2	1.2	5.9
	2	9.6	0.0	1.2	0.1	1.2	5.9
	3	9.7	0.0	1.2	0.1	1.3	5.7
	4	9.7	0.0	1.2	0.2	1.2	5.7
	5	9.7	0.0	1.0	0.1	1.1	5.6
Mean value		9.66	0.00	1.12	0.14	1.2	5.76

TABLE VI. RESULTS OF CLASSIFICATION OF THE STUDIED SAMPLES USING THE NON-STATISTICAL APPROACH.

Sample	Sets
D1	PR
D2	CW
D3	PR
D4	PR
D5	CW
P	PR

PR – set of car headlamp glass and colourless container glass.

CW – set of car windscreen glass and window sheet glass.

TABLE VII. COMPARISON OF THE CONTENT OF PARTICULAR ELEMENTS IN SAMPLES D1, D4 AND P USING THE “3 SIGMA” RULE.

Element	Confidence ranges for samples			Results of comparison	
	D1	P	D4	D1 – P	D4 – P
Na	9.20 – 9.52	9.50 – 9.82	9.11 – 10.02	+	+
Al	1.05 – 1.31	0.79 – 1.45	0.79 – 1.33	+	+
S	0.00 – 0.25	0.00 – 0.30	0.00 – 0.30	+	+
K	1.45 – 1.71	0.99 – 1.41	0.83 – 1.81	+	+
Ca	4.65 – 4.91	5.36 – 6.16	5.39 – 6.17	–	+

(+) – overlapping ranges of confidence, (–) – separate ranges of confidence.

Student’s t-test and the “3 sigma” rule

In order to utilise the “3 sigma” rule and Student’s t-test in an analogous way to the refractive index measurements, the appropriate calculation should be performed for each element content separately. The results of these calculations are presented in Tables VII and VIII. As shown in Tables VII and VIII, sample D4 reveals similarity to sample P for each of the analysed elements.

TABLE VIII. RESULTS OF STUDENT'S t -TEST FOR SIMILARITY OF PARTICULAR ELEMENT CONTENTS FOR PAIRS OF SAMPLES D1 – P AND D4 – P.

Element	Value of t_{cal} for pairs of samples			
	D1 – P		D4 – P	
Na	8.660	(–)	1.387	(+)
Al	1.134	(+)	0.949	(+)
S	0.632	(+)	0.000	(+)
K	10.156	(–)	1.500	(+)
Ca	15.495	(–)	0.239	(+)

(+) – stays for $t_{cal} < t_{r,\alpha}$; (–) – stays for $t_{cal} > t_{r,\alpha}$; $t_{r,\alpha} = t_{4,0.05} = 2.776$.

Some difficulties occurred during interpretation of results obtained for pair of samples, D1 – P. It is only in the case of calcium that the range of concentration for these samples did not overlap. The ranges of concentrations of the remaining elements could suggest some similarity between samples D1 and P. However, Student's t -test performed for this pair of samples (Table VII) resulted in dissimilarity. Coefficients t_{cal} for sodium, potassium and calcium were greater than $t_{r,\alpha} = 2.776$ ($r = 4$; $\alpha = 0.05$), meaning that the null hypothesis of equality of the mean values of element contents in samples D1 and P should be rejected, whereas values of t_{cal} for aluminium and sulphur allow us to accept this hypothesis. As the two ways of discrimination provided opposite results, the question arose as to which of these methods is more suitable. In the case of choosing Student's t -test another question should be answered; whether the ratio 3:2 (for 3 elements – negative results of the test and for 2 elements – positive results) would be sufficient to draw the conclusion of the samples being different. One could find it helpful to check whether there are correlations between the analysed element composition of the compared samples [2, 9] and the nature of the correlation, i.e. does an increase in one element cause any increase or decrease in the content of another element? Important correlations found for a pair of the elements should be taken into account during interpretation of the results by both the “3 sigma” rule and Student's t -test. In the absence of correlations the interpretation of the results for the D1 – P pair still remains ambiguous.

One can conclude from the considerations above that in the case of examinations of objects described by several features a non-statistical approach did not provide final solutions to classification and also results of the “3 sigma” rule or Student's t -test applied to each feature separately proved

to be rather ambiguous in the process of discrimination. Thus methods of statistical analysis that consider many features simultaneously must be applied.

Cluster analysis

Cluster analysis [3, 14] will be more broadly described in this paper. This method allows us to create clusters out of similar objects. The measure of similarity of objects studied, e.g. glass samples, is the distance between them. A simple relation is used in cluster analysis: the smaller the distance between objects the more similar to each other they are.

Calculations of the distances between objects described by several quantitative features, will be presented using the example of the square of the Euclidean distance. Let us consider objects D1 and D2 described by two initial rows in matrix {9} and put them in the plane defined by axes of concentration of sodium and potassium, C_{Mg} and C_{Na} (Figure 2).

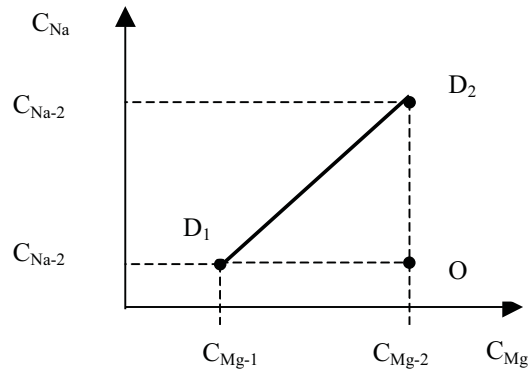


Fig. 2. Representation of objects D1– D2 in an orthogonal two-dimensional space.

From Pythagoras' theorem:

$$D_{D1,D2}^2 = (OD_1)^2 + (OD_2)^2. \quad \{10\}$$

From Figure 2:

$$OD_1 = c_{Mg-1} - c_{Mg-2}, \quad \{11\}$$

$$OD_2 = c_{Na-1} - c_{Na-2}. \quad \{12\}$$

From equations {10}, {11} and {12} the following formulae can be obtained:

$$D_{D1,D2}^2 = (c_{Mg-1} - c_{Mg-2})^2 + (c_{Na-1} - c_{Na-2})^2, \quad \{13\}$$

In an n-dimensional case this relation can be expressed as:

$$D_{D1,D2} = \sum_{i=1}^n (c_{D1-i} - c_{D2-i})^2 . \quad \{14\}$$

Including mean concentrations of sodium, magnesium, aluminium, sulphur, potassium and calcium, determined for samples D1 and D2 in equation {14}, one can obtain the value of the squared Euclidean distance between these samples, i.e. 12.042.

Mutual distances between the remaining objects, obtained in the same manner were presented in the form of a symmetric matrix {15}.

$$\begin{array}{c} \begin{array}{cccccc} & D1 & D2 & D3 & D4 & D5 & P \\ D1 & \left(\begin{array}{cccccc} 0 & & & & & & \\ 11.949 & 0 & & & & & \\ 10.604 & 3.761 & 0 & & & & \\ 1.122 & 9.858 & 6.184 & 0 & & & \\ 8.990 & 0.319 & 2.401 & 6.936 & 0 & & \\ 1.199 & 9.754 & 5.825 & 0.028 & 6.776 & 0 & \end{array} \right) & \\ D2 & & & & & & \\ D3 & & & & & & \\ D4 & & & & & & \\ D5 & & & & & & \\ P & & & & & & \end{array} \end{array} \quad \{15\}$$

It is sufficient to note down only half of the matrix as the distance measured from object D1 to object D2 is the same as the distance from object D2 to object D1.

Having calculated the matrix of mutual distances of objects D1, ..., D5 and P., one can start building clusters. There is a great number of clustering methods. They are generally divided into non-hierarchical and hierarchical ones, and the latter are divided into agglomerative and divisive ones. Hierarchical methods, most frequently used in empirical research, are based on the rule that smaller clusters are parts of the larger cluster. In the case of hierarchic-agglomerative methods at the beginning there are as many clusters as the considered objects, i.e. every cluster consists of one object. The clustering process is finished when one cluster is obtained that consists of all the described objects. In the case of divisive hierarchical methods the clustering process is opposite to that above.

The clustering process in agglomerative-hierarchical methods consists of several stages. The first stage is selection of the pair of objects from the similarity matrix, which reveals the smallest mutual distance, i.e. objects D4 and P ($d_{D4-P} = 0.028$). The next stage is determination of distances between the new cluster S (D4; P) and other objects D1, D2, D3 and D5. Clustering methods differ in the method of definition of distances between the new cluster and other objects (single objects or/and clusters). Among the methods are, e.g. the farthest neighbourhood and the nearest neighbourhood method. In the case of the farthest neighbourhood method the distance between ob-

ject S (D4, P) and another object, e.g. D1(d_{S-D1}) equals the largest distance among distances calculated for pairs D4 – D1 and P – D1, i.e. $d_{S-D1} = \max(d_i)$ (Table IX):

$$d_{S-D1} = d_{P-D1} = 1.199 .$$

Similarly, in the further neighbourhood method the distance between cluster S (D4, P) and object D1 is the smallest distance among these distances ($d_{S-D1} = \min(d_i)$):

$$d_{S-D1} = d_{D4-D1} = 1.122 .$$

The distance from cluster S (D4, P) to other objects can be obtained in the same manner and placed in new similarity matrices, separately for farthest neighbourhood {16} and for nearest neighbourhood {17}. Dimensions of these matrixes are diminished by one row and one column compared to the initial similarity matrixes {15}.

	D1	D2	D3	S	D5	
D1	0					{16}
D2	12.042	0				
D3	10.681	3.601	0			
S	1.199	9.858	6.184	0		
D5	9.091	0.319	2.374	6.936	0	
	D1	D2	D3	S	D5	
D1	0					{17}
D2	12.042	0				
D3	10.681	3.601	0			
S	1.122	9.745	5.825	0		
D5	9.091	0.319	2.374	6.776	0	

TABLE IX. THE SQUARE EUCLIDEAN DISTANCE FROM OBJECTS D4 AND P TO OBJECT D1.

Pair of objects	Distance d_i
D4 – D1	1.122
P – D1	1.199

Proceeding with the clustering, again we are searching for objects of the smallest distance and create new distance matrixes using either the farthest neighbourhood or the nearest neighbourhood method. Results of the cluster

analysis can be presented in the form of dendrograms. Dendrograms obtained for the presented examples are shown in Figures 3 and 4.

Fig. 3. A dendrogram resulting from cluster analysis using the nearest neighbourhood method.

Fig. 4. A dendrogram resulting from cluster analysis using the furthest neighbourhood method.

The cluster analysis performed showed that samples D1 and D4 were similar to sample P. They created a three-element cluster. Samples D2, D3 and D5 formed another three-element cluster, indicating their mutual simi-

larity and the difference from comparative sample P. Distances between objects inside these separated clusters are smaller than distances between cluster D1 – D4 – P and cluster D2 – D3 – D5. Especially well expressed differences have been shown in the case of the farthest neighbourhood method (Figure 4).

Which method of calculating distances between objects and which of the clustering methods are most suitable, depends on the particular task to be solved. From the literature [10] one can conclude that the squared Euclidean distance and the farthest neighbourhood method are most frequently used for glass sample classification.

The presented analysis did not enable us to classify samples D1, ..., D5. It is not possible to state that samples D2 and D1 could have originated from certain groups of glass, e.g. from car headlamps or car windscreens. This analysis showed only which objects are similar and which are not. In order to carry out classification, it is necessary to have information about the elemental composition of various use-type groups of glass, obtained by the same analytical method as used for the evidence and comparative samples.

Let us classify the evidence samples, having at our disposal data concerning elemental composition of the following glass samples: car headlamps (r), colourless containers (px) and windscreen glass (c) obtained by the SEM/EDX method. Prior to classification of evidence samples using cluster analysis, a method of classifying the data base itself must be worked out; this requires selection of variables (elemental composition) as well as the distance and the clustering methods. The chosen parameters should allow us to obtain a dendrogram built of three clusters, each cluster containing objects of the same group. In the ideal case (Figure 5) we will obtain three such clusters: one for the headlamps glass (r), one for the colourless containers (px) and one for windscreen glass (c). An adequate solution to this problem could be the dendrogram presented in Figure 6. The distance between clusters should be considerably greater than distances within the cluster.

The choice of variables concerns selection of chemical elements, whose mean concentrations significantly differ in particular groups. In Figure 7 results of cluster analysis are presented, obtained for samples included in the data base, using concentration of all determined elements (Al, Ba, Ca, Cr, Fe, K, Mg, Na, O, S, Si, Ti). Figure 8 shows the results of cluster analysis for the same samples, using concentrations of elements, chosen using the Tukey HSD method [2, 9]. This method allows us to choose variables (elemental contents), of mean values which differ significantly in the studied groups. The analysis of the data base with this method enables us to find significant differences in mean concentration of Al, Ca, K, Mg and Na among the considered groups.

Fig. 5. A dendrogram representing the ideal solution to classification.

Fig. 6. A dendrogram representing an adequate solution to classification.

Dendrograms in Figure 7 and 8 were examined in order to check whether their constructions were consistent with the models assumed in Figures 5 and 6. In the dendrogram in Figure 7 the following clusters could be distinguished: px1, ..., px9 for the colourless glass containers group; two clusters for car windscreen glass group: c1, ..., c4 and c8, ..., c12; and a cluster for the headlamp glass group r1, ..., r5. The remaining 15 objects are located beyond the clusters. Objects px12, ..., px22, are located close to each other, but do not create one cluster. One can distinguished among them two groups: px12, ..., px14 and px16, ..., px19 and single objects px15 and px20 of mutual distances greater than within clusters. This evaluation of distances has only got a qualitative character. For the purpose of quantitatively establishing what value of distance would be sufficient to deviate sets of objects into two separate clusters it would be necessary to analyse distances present on a dendrogram similar to that in Figure 7, but it should have been performed for a greater number of samples in every group. The pair px21 and px22 showed similarity to objects c5 – c6. In the remaining part of the dendrogram one can also distinguish single objects c7 and px 23. The dendrogram strongly deviates from models in Figure 5 and 6.

Fig. 7. A dendrogram resulting from cluster analysis for data base samples using concentrations of all the determined elements.

Fig. 8. A dendrogram resulting from cluster analysis for data base samples, using concentrations of all selected elements

In Figure 8 three large clusters can be observed: for colourless glass containers (px1, ..., px17), for car windscreen glass (c1, ..., c11) and for car headlamp glass (r1, ..., r6). However, they do not contain all of the examined objects, as colourless container glass samples create an additional three-element cluster (px19, ..., px20) and objects px21, px22, c12, px21 – px17 remain beyond the large clusters. The structure of this dendrogram, in spite of 5 objects present beyond the main clusters, seems to resemble the model in Figure 6 more than the structure of the dendrogram obtained when all element concentrations were used as variables (Figure 7). The advantage does not seem convincing, therefore classification of objects D1, ..., D5 was carried out with both methods. For this aim results obtained for the evidence and the comparative samples were added to the data base and the cluster analysis was carried out. Concentration of all analysed elements included in the data base were used as variables. Thus, for the examined samples not only concentration of elements noted at Table V should be taken into consideration, but also: oxygen (O), silicon (Si), iron (Fe), chromium (Cr), barium (Ba), and titanium (Ti) – Table X. Dendrograms in Figures 9 and 10 show results of the classification worked out.

TABLE X. MEAN VALUES OF ELEMENT CONCENTRATIONS IN SAMPLES D1, ..., D5 AND P – A SUPPLEMENT TO TABLE V.

Sample	Element content [wt. %]					
	O	Si	Fe	Cr	Ba	Ti
D1	53.84	29.08	0.00	0.00	0.06	0.00
D2	46.74	34.68	0.00	0.00	0.00	0.00
D3	44.20	36.44	0.00	0.00	0.00	0.00
D4	53.18	28.84	0.00	0.00	0.12	0.00
D5	44.90	36.80	0.00	0.00	0.00	0.00
P	42.28	39.84	0.00	0.00	0.00	0.00

Comparative material P, which is a sample of car headlamp glass (r), was incorrectly classified into the colourless glass containers group (px) when all variables were used in the method of classification (Figure 9). In the case of objects D1, ..., D5 the classification could also be incorrectly performed. The correct classification of sample P was achieved after a preliminary selection of variables, i.e. element concentrations, had been performed with Tukey HSD (honestly significant differences) method (Figure 10). On the grounds of the same dendrogram one can find that samples D1 and D4 most probably came from headlamps (are included in cluster (r)). Pieces D2 and D5 were classified into the car windscreen glass group and sample D3 could have come from a colourless glass container. It was classified into the group of colourless glass containers (px).

As was mentioned before, the structure of in Figure 7 and Figure 8 was not entirely consistent with that assumed in Figures 5 and 6. The reason for this could be the fact that element concentration in the analysed objects differed, even to three decimal places (e.g. for Si 28.84–40.85 wt. % and for S 0.1–0.2 wt. %). In such a situation a scaling of variables can be performed. Depending on the type of variables describing features of the analysed objects the primordial score, also called the raw score can be transformed in various ways. In the case of quantitative data, the most commonly used transformations are presented in Table XI.

The decision concerning which transformation to use can be made after examining relations between the variance and the experimental average. An additional frequently used “scaling” method is dividing the raw score by the standard deviation, calculated for the mean value of the established variable. One can transform not only quantitative, but also qualitative data.

Moreover, there are methods of “scaling” of both qualitative and quantitative variables, describing objects simultaneously.

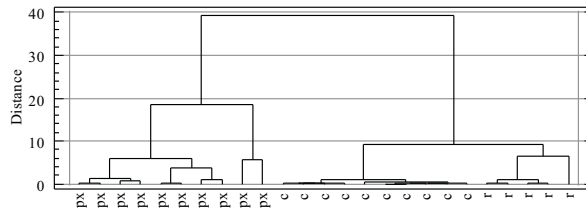


Fig. 9. As dendrogram resulting from the classification of objects from the data base and the examined samples, using all of the determined elements as variables.

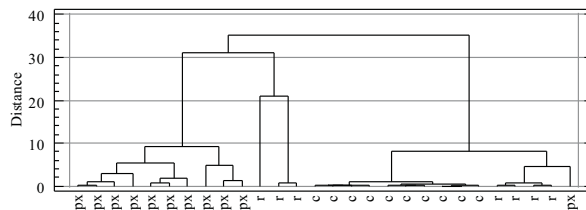


Fig. 10. A dendrogram resulting from the classification of objects from the data base and the examined samples, using all of the selected elements as variables.

TABLE XI. MOST POPULAR TRANSFORMATION OF PRIMORDIAL SCORE OF ANALYTICAL RESULTS.

Transformations	Expression
Square root	\sqrt{X}
Logarithm	$\log X$
Quotient	$1/X$
Arcus sinus	$\arcsin X$

The decision concerning which transformation to use can be made after examining relations between the variance and the experimental average. An additional frequently used “scaling” method is dividing the raw score by

the standard deviation, calculated for the mean value of the established variable. One can transform not only quantitative, but also qualitative data. Moreover, there are methods of “scaling” of both qualitative and quantitative variables, describing objects simultaneously.

In the considered example, the problem of differences (even within three decimal places) in element concentrations could be solved by use of the logarithmic transformation. However, some element concentrations equal 0, making it impossible to use this transformation as $\log 0$ does not exist. The problem can be overcome by replacing null values of some element concentrations with very small values. It is also possible to use the following formula:

$$Z_i = \frac{X_i - \min X}{\max X - \min X}, \quad \{18\}$$

where: X_i – the value obtained at measurement no. “ i ”, $\min X$ and $\max X$ – minimum and maximum value of x -th element in the considered model.

After having completed this transformation, the values for the established element concentration were enclosed in the interval of $<0; 1>$. The dendrogram obtained for the transformation of values of all analysed elements is presented in Figure 11. Conversely, the dendrogram in Figure 12 was created using (as variables) the transformed values of element concentrations, selected by Tukey HSD method (Al, Ca, K, Mg and Na) [2, 9]. During cluster analysis the square Euclidean distance and the farther neighbourhood method were used, as earlier.

Comparing the dendrogram obtained using transformed data (Figure 11) to that obtained using raw score (Figure 7) one can conclude that no significant improvement occurred in this case. The structure of the dendrogram obtained for the transformed data still deviated from the model ones in Figures 5 and 6; furthermore, the classification of samples D1, ..., D5 and P was incorrect. In Figure 13 one can observe that sample P was incorrectly classified to cluster (px), instead of to cluster (r). However, in the case of the variables transformed and selected by the Tukey HSD method (Figure 12) a significant improvement in comparison to dendrogram in Figure 8 (for raw score) was obtained. Only one cluster for each considered glass group (px1, ..., px20; c1, ..., c11; r1, ..., r7) and only 4 objects beyond them (px21, c12, px22 and px23) can be found in this dendrogram. Figure 14 presents the results of the classification of the analysed glass microtraces. The results are consistent with the classification presented in the dendrogram (Figure 8).

One can conclude that in order to solve the problem of classification of objects that belong to a number of groups, it is necessary to analyse various combinations of raw score transformation methods, distance expressions and clustering methods, so that optimal results can be obtained.

Fig. 11. A dendrogram resulting from cluster analysis with the use of the transformed values of concentrations of all the examined elements.

Fig. 12. A dendrogram resulting from cluster analysis with the use of the transformed values of concentrations of all selected elements.

DISCRIMINATION OF OBJECTS OF MULTIVARIABLE DESCRIPTION

Now we are faced with the problem as to whether samples D1 and D4 originated from the broken car headlamp P. It will be solved by means of cluster analysis.

As was mentioned before, in the course of cluster analysis, clusters which include similar objects are created. For the discrimination of two samples: the evidence and the comparative sample all of the single measurements performed were utilised. In order to carry out discrimination of glass microfragments, i.e. to establish – whether they could have come from the same objects or from two different objects, concentration of all the determined elements will be utilised. In the ideal case, samples of the same objects would be similar taking into account each element content. If two separate clusters are obtained, each containing results obtained for one object (e.g. one cluster of results for D1 and for P), one can conclude that the fragments are not similar and did not come from the same objects. However, if

Fig. 13. A dendrogram resulting from cluster analysis of elements of the data base and the examined samples with the use of transformed element concentrations.

Fig. 14. A dendrogram resulting from cluster analysis of elements of the data base and the examined samples with the use of selected and transformed element concentrations.

the analysed results are gathered in a cluster containing analytical results originating from both objects, one could state that the evidence and the comparative fragments were similar and could have come from the same objects.

Let us carry out discrimination of the evidence samples, D1 and D4, and the comparative sample P. Figures 15 and 16 represent results of the cluster analysis performed.

One can observe in Figure 15 that fragment D1 most probably did not come from car headlamp glass, as two separated clusters occurred, each containing measurements of one object only. On the other hand, the structure of Figure 16, obtained for samples D4 and P enables one to conclude that they could originate from the same car headlamp.

CONCLUDING REMARKS

The presented article provided a review of methods of statistical and analysis that were utilised for interpretation of results obtained for glass

Fig. 15. A dendrogram resulting from cluster analysis of the analytical results obtained for samples D1 and P.

Fig. 16. A dendrogram resulting from cluster analysis of all the analytical results obtained for samples D4 and P.

microtraces during both basic studies and routine examination for forensic purposes. Selected statistical and chemometric methods, used for classification and discrimination of glass fragments, may be successfully applied in examinations of other types of materials. In particular, the methods of cluster analysis are extensively used nowadays not only in sciences such as chemistry or physics, but also in economics, psychology, sociology, etc.

References:

1. Brożek-Mucha Z., Zadora G., Differentiation between various types of glass using SEM-EDX elemental analysis. A preliminary study, *Z Zagadnień Nauk Sądowych* 1998, z. XXXVII, s. 68–89.
2. Czermiński J. B., Iwaszkiewicz A., Paszek Z. [i in.], *Metody statystyczne dla chemików*, Wydawnictwo Naukowe PWN, Warszawa 1992.
3. Everitt B. S., *Cluster analysis*, Arnold, Oxford 1993.
4. Evett I. W., The interpretation of refractive index measurement, *Forensic Sciences International* 1977, vol. 9, pp. 209–217.
5. Evett I. W., The interpretation of refractive index measurement II, *Forensic Sciences International* 1978, vol. 12, pp. 37–47.
6. Evett I. W., Lambert J. A., The interpretation of refractive index measurement III, *Forensic Sciences International* 1982, vol. 20, pp. 237–245.
7. Evett I. W., Lambert J. A., The interpretation of refractive index measurement IV, *Forensic Sciences International* 1984, vol. 24, pp. 149–163.
8. Evett I. W., Lambert J. A., The interpretation of refractive index measurement V, *Forensic Sciences International* 1985, vol. 27, pp. 97–110.
9. Ferguson G. A., Takane Y., *Analiza statystyczna w psychologii i pedagogice*, Wydawnictwo Naukowe PWN, Warszawa 1997.
10. Hickman D. A., Harbottle G., Sayer E. V., The selection of the best elemental variables for the classification of glass samples, *Forensic Science International* 1983, vol. 23, pp. 189–212.
11. Howden C. R., Dudley R. J., Smalldon K. W., The analysis of small glass fragments using energy dispersive X-ray fluorescence spectrometry, *Journal of Forensic Sciences* 1978, vol. 18, pp. 99–112.
12. Lambert J. A., Evett I. W., The refractive index distribution of control glass samples examined by the forensic sciences laboratories in the United Kingdom, *Forensic Sciences International* 1984, vol. 26, pp. 1–23.
13. Locke J., Underhill M., Automatic refractive index measurement of glass particles, *Forensic Sciences International* 1985, vol. 27, pp. 247–260.
14. Massart D. L., Kaufman L., *The interpretation of analytical chemical data by the use of cluster analysis*, John Wiley & Sons, New York 1983.
15. Terry K. W., van Riessen A., Lynch B. F. [et al.], Quantitative analysis of glasses used within Australia, *Forensic Sciences International* 1985, vol. 25, pp. 19–34.

ZASTOSOWANIE WYBRANYCH METOD STATYSTYCZNYCH I CHEMOMETRYCZNYCH W KRYMINALISTYCZNYCH BADANIACH SZKŁA

Grzegorz ZADORA, Zuzanna BROŻEK-MUCHA

WSTĘP

W kryminalistycznych badaniach dowodów rzeczowych wyróżnia się dwa podstawowe cele. Przy braku materiału porównawczego ujawnione mikroślady dowodowe badane są w celu dokonania ich klasyfikacji, tj. zaliczenia do określonej grupy użytkowej przedmiotów na podstawie charakterystycznych dla nich właściwości fizykochemicznych. Gdy prócz materiału dowodowego do badań dostarczony jest materiał porównawczy, zadaniem eksperta jest udzielenie odpowiedzi na pytanie, czy mogły one pochodzić z tego samego przedmiotu, a więc przeprowadzenie analizy porównawczej zwanej też dyskryminacją [10].

Tylko w przypadku wykazania zdecydowanych różnic we właściwościach fizycznych bądź w składzie chemicznym porównywanych materiałów możliwe jest kateryczne stwierdzenie, że nie pochodzą one z tego samego przedmiotu. Najłatwiej jest rozróżnić próbki materiałów należących do odrębnych klas. Jednak dokonanie rzetelnej dyskryminacji próbek przedmiotów należących do tej samej klasy, a więc z definicji wykazujących podobne cechy, jest zadaniem znacznie trudniejszym.

Wraz z wprowadzeniem do nauk sądowych nowoczesnych metod analitycznych, charakteryzujących się wysoką czułością i precyzją, krótkim czasem analizy, prostą obsługą aparatury badawczej wynikającą z jej zautomatyzowania, możliwe stało się uzyskanie dobrej jakości wyników analitycznych. Tego rodzaju zalety wykazują, często stosowane w badaniach mikrośladów szkła, metody oparte zarówno na pomiarach współczynnika załamania światła, np. technika GRIM (ang. glass refractive index measurement) [4, 5, 6, 7, 8, 12, 13], jak i metody analizy pierwiastkowej, np. spektrometria promieniowania rentgenowskiego realizowana bądź w sprzężeniu ze skaningową mikroskopią elektronową (SEM-EDX – ang. scanning electron microscopy with energy dispersive X-ray spectrometry) [1, 15], bądź jako fluorescencja rentgenowska (XRF – ang. X-ray fluorescence) [11]. Istotnym problemem eksperta z dziedziny badań fizykochemicznych pozostaje jednak interpretacja wyników analitycznych.

W prezentowanej pracy przedstawione zostały wybrane metody analizy statystycznej, które – w przekonaniu autorów – mogą być pomocne w ocenie zmienności badanego materiału oraz w określaniu niewielkich różnic pomiędzy próbkami różnych przedmiotów należących do tej samej klasy. Dzięki wykorzystaniu odpowiedniej metody statystycznej (w szczególności w przypadkach wielowymiarowego opisu badanych obiektów) identyfikacja grupowa może być zawężona do grup o mniejszej liczebności.

Wybór metody analizy statystycznej stosowanej w celu rozwiązania problemu klasyfikacji lub dyskryminacji zależy m.in. od ilości cech, którymi opisywane są badane obiekty.

KLASYFIKACJA OBIEKTÓW OPISANYCH PRZEZ JEDNĄ ZMIENNĄ

W przypadku, gdy badany obiekt scharakteryzowany jest tylko jedną cechą, w celu jego klasyfikacji wymagana jest baza danych o zakresie (lub zakresach) wartości, jakie ta cecha przyjmuje w różnych grupach. Rozpatrzmy następujący przykład: na odzieży ofiary wypadku drogowego ujawniono pięć okruchów szklanych D1, D2, D3, D4 i D5. Do badań dostarczono również materiał porównawczy P w postaci próbki szkła zabezpieczonego z rozbitego reflektora samochodu należącego do osoby podejrzanej o spowodowanie potrącenia ze skutkiem śmiertelnym.

Wyznaczono współczynniki załamania światła (ang. refractive index – RI) [13], wykonując po 9 pomiarów w różnych miejscach tych próbek. Wyniki zebrano w tabeli I. Fragment bazy danych, zawierający wartości współczynników załamania światła dla szkła pochodzącego z szyb samochodowych (c) i reflektorów samochodowych (r) prezentuje tabela II.

W pierwszej kolejności należy dokonać klasyfikacji próbek dowodowych. W przypadku zaklasyfikowania którejkolwiek z próbek dowodowych do grupy szkieł reflektorowych (r), kolejnym etapem będzie dyskryminacja, czyli odpowiedź na pytanie, czy próbka dowodowa i porównawcza mogły pochodzić z tego samego obiektu.

Przed przystąpieniem do klasyfikacji obliczono wartość średnią współczynnika załamania światła (tabela I) dla próbek dowodowych i próbki porównawczej z następującego wzoru:

$$\overline{RI}_i = \frac{\sum_{j=1}^n RI_{ij}}{n} \quad \{1\}$$

gdzie: i – indeks analizowanej próbki ($i = D1, \dots, D5, P$); n – liczba wykonanych pomiarów RI dla każdej z próbek i ($n = 9$).

W celu klasyfikacji próbki D1 wartość średnią współczynnika załamania światła $\overline{RI}_{D1} = 1,51454$ porównuje się z bazą danych zawartą w tabeli II. Poszukuje się w niej grupy szkła, w zakresie której zawiera się współczynnik załamania światła próbki D1. Wartość 1,51454 mieści się w przedziale wartości współczynnika załamania światła wyznaczonych dla szkła reflektorów samochodowych. Próbka dowodowa D1 może zatem pochodzić z reflektora samochodowego.

Średnia wartość współczynnika załamania światła wyznaczona dla próbki D2 ($\overline{RI}_{D2} = 1,514547$) klasyfikuje ją zarówno do grupy szyb samochodowych, jak i do grupy reflektorów, a w przypadku próbki D3 wartość średnią współczynnika załamania światła ($\overline{RI}_{D3} = 1,52543$) nie zawiera się w żadnym z przedziałów w tej bazie danych. W takich przypadkach klasyfikacja nie jest możliwa. Na podstawie średnich wartości współczynników załamania światła dla próbek D4 i D5 ($\overline{RI}_{D4} = 1,51243$; $\overline{RI}_{D5} = 1,52163$) wywnioskować można, że próbka D4 pochodzi z reflektora samochodowego, a próbka D5 z szyby samochodowej.

Podsumowując ten etap badań, dwie próbki dowodowe D1 i D4 zakwalifikowano do grupy szkieł reflektorów samochodowych. Kolejnym celem staje się teraz odpowiedź na pytanie, czy którakolwiek z tych dwóch próbek dowodowych wykazuje podobieństwo do próbki porównawczej P.

Dyskryminacji tych próbek szkła, na podstawie wyznaczonych dla nich współczynników załamania światła, można dokonać stosując np. metodę „3 sigma” lub test t-Studenta.

DYSKRYMINACJA OBIEKTÓW OPISANYCH PRZEZ JEDNĄ ZMIENNĄ

Reguła „3 sigma”

W celu zastosowania reguły „3 sigma” [2, 9] najpierw wyznacza się przedziały ufności oparte o wartości średnich współczynnika załamania światła dla każdej z porównywanych próbek dowodowych D1 i D4 oraz porównawczej P. Nie można bowiem bezpośrednio porównywać średnich wartości współczynnika załamania światła wyznaczonych dla analizowanych próbek, ponieważ średnie te są jedynie oszacowaniem prawdziwej wartości współczynnika załamania światła każdej z próbek (μ_i ; $i = D1, D4, P$). Dlatego też wyznacza się przedziały wartości współczynnika załamania światła (przedziały ufności), w których najprawdopodobniej zawiera się wartość współczynnika załamania światła (μ_i). W przypadku reguły „3 sigma” prawdopodobieństwo to wynosi 99,73 %.

Przedział ufności buduje się w ten sposób, że jego dolną granicę stanowi różnica średniej wartości \overline{RI}_i i trójkrotnej wartości odchylenia standardowego, a górną suma średniej wartości \overline{RI}_i i trójkrotnej wartości odchylenia standardowego. Wartość odchylenia standardowego σ_i oblicza się według wzoru:

$$s_i = \sqrt{\frac{\sum_{j=1}^n (RI_{ij} - \overline{RI}_i)^2}{n-1}} \quad \{2\}$$

gdzie: $i = D1, D4, P$; n – liczba pomiarów ($n = 9$).

Wyniki powyższych obliczeń zebrano w tabeli III. Przedziały ufności współczynników załamania światła wyznaczone dla próbek D1 i P nie pokrywają się ze sobą. Można więc stwierdzić, że próbka szkła dowodowego D1 nie pochodzi z tego samego obiektu, co próbka porównawcza P. Natomiast w przypadku próbek D4 i P przedziały ufności współczynników załamania światła pokrywają się. Na tej podstawie należy wnioskować, że próbka szkła dowodowego D4 może pochodzić z tego samego reflektora, co próbka porównawcza P.

Test t-Studenta

Innym sposobem przeprowadzenia dyskryminacji jest użycie testu t-Studenta [2, 9] dla średnich z prób niezależnych. Za jego pomocą testuje się hipotezę o równości dwóch porównywanych ze sobą wartości średnich. Jest to tzw. hipoteza zerowa (H_0). W tym przypadku testujemy hipotezę zerową, która zakłada, że wartości średnie współczynników załamania światła wyznaczonych dla próbki dowodowej i próbki porównawczej są statystycznie równe, $H_0 : \mu_D = \mu_P$. W przypadku potwierdzenia tej hipotezy będzie można stwierdzić, że okruszy szkła porównawczego i dowodowego są do siebie podobne i mogą pochodzić z tego samego obiektu. Odrzucając hipotezę zerową, powiemy, że okrusz szkła dowodowego i porównawczego różnią się od siebie i najprawdopodobniej pochodzą z dwóch różnych obiektów.

Test t-Studenta powinno się stosować, gdy spełnione są dwa założenia. Pierwsze z nich wymaga, aby rozkład analizowanych wyników pomiarowych był rozkładem normalnym [2, 9]. Zwykle założenie to uważa się za słuszne. Gdyby zaistniała konieczność potwierdzenia tak postawionej hipotezy, to można zastosować test Kolmogorowa-Smirnowa lub chi kwadrat (χ^2) [9]. Biorąc powyższe stwierdzenia pod uwagę, w rozpatrywanym przykładzie przyjęto, że rozkłady wyników współczynnika załamania światła wyznaczone dla próbek D1, D4 i P mają rozkład normalny.

Drugie założenie dotyczy wariancji wyników obu analizowanych próbek (σ_1^2 i σ_2^2).

Hipoteza zerowa w tym przypadku zakłada równość wariancji ($H_0: \sigma_1^2 = \sigma_2^2$). Jest ona sprawdzana poprzez zastosowanie testu F [2, 9]. W teście tym wyznacza się wartość zmiennej F z zależności {3}:

$$F = \frac{s_1^2}{s_2^2}, \quad \text{gdzie } s_1^2 > s_2^2; \quad \{3\}$$

$$s_i^2 = \frac{\sum_{j=1}^n (x_{ji} - \bar{x}_i)^2}{n-1}, \quad (i = 1, 2). \quad \{4\}$$

Następnie porównuje się uzyskaną wartość zmiennej F z wartością $F_{r1, r2, \alpha}$ odczytaną z tabeli dla odpowiedniej liczby stopni swobody $r1$ i $r2$ oraz założonego współczynnika istotności α . Liczbę stopni swobody oblicza się z zależności {5} i {6}:

$$r1 = n_1 - 1, \quad \{5\}$$

$$r1 = n_2 - 1, \quad \{6\}$$

gdzie: n_1, n_2 – liczba pomiarów wykonanych dla próbek o wariancji s_1^2 i s_2^2 .

Ponieważ liczba pomiarów RI wykonanych dla każdej z analizowanych próbek wynosiła 9 ($n_1 = n_2 = 9$), dlatego też $r1 = r2 = 8$.

Wartość α przyjmuje się najczęściej na poziomie 0,05. Taki poziom istotności oznacza, że w przypadku odrzucenia hipotezy zerowej istnieje pięcioprocentowe ryzyko popełnienia błędu polegającego na uznaniu jej za fałszywą, gdyby w rzeczywistości była ona prawdziwa.

W przypadku, gdy dla porównywanej pary obiektów $F_{r1, r2, \alpha}$ jest większe od F obliczonego z zależności {3}, przyjmuje się hipotezę zerową i można zastosować dla niej proponowany test t-Studenta dla średnich z prób niezależnych. Test F przeprowadzony dla par D1 – P i D4 – P wykazał, że w obu przypadkach można przyjąć hipotezę zerową (tabela IV), bowiem wyznaczone wartości zmiennej F dla powyższych par próbek są mniejsze niż $F_{8,8,0.05} = 3,44$.

Przystępując do analizy hipotezy zerowej mówiącej o równości średnich wartości współczynnika załamania w próbce dowodowej i porównawczej (dla par D1 – P i D4 – P), należy w pierwszej kolejności wyznaczyć wartość zmiennej t . Ponieważ liczba pomiarów RI wykonanych dla każdego z obiektów jest taka sama ($n_1 = n_2 = n = 9$), to można zapisać wyrażenie na zmienną t w postaci następującego wzoru:

$$t = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{s^2}} \sqrt{n}, \quad \{7\}$$

gdzie: \bar{x}_1, \bar{x}_2 – wartość średnia oznaczanych wielkości w porównywanych próbkach. Wartość \bar{s}^2 oblicza się ze wzoru:

$$\bar{s}^2 = \frac{s_1^2 + s_2^2}{2}. \quad \{8\}$$

W rozpatrywanych przypadkach dla par D1–P i D4–P uzyskano odpowiednio:

$$\begin{aligned} \bar{x}_1 &= \overline{RI_{D1}} = 1,51454, \\ \bar{x}_2 &= \overline{RI_P} = 1,51248, \\ t_{D1-P} &= 17,382 \text{ (ze wzoru \{7\})}, \\ \bar{x}_1 &= \overline{RI_{D1}} = 1,51454, \\ \bar{x}_2 &= \overline{RI_P} = 1,51248, \\ t_{D4-P} &= 0,413 \text{ (ze wzoru \{7\})}. \end{aligned}$$

Następnie należy porównać obliczone wartości t_{D1-P} i t_{D4-P} z wartością $t_{r,\alpha}$ odczytaną z tablic dla liczby stopni swobody $r = n - 1$ ($r = 8$) i poziomu istotności $\alpha = 0,05$.

Hipotezę zerową przyjmuje się za prawdziwą, gdy wartość $t_{r,\alpha}$ jest większa od wartości obliczonej t_{obl} . W przypadku przeciwnym hipotezę zerową należy odrzucić.

Zmienna $t_{r,\alpha}$ odczytana z tabeli dla $r = 8$ i $\alpha = 0,05$ wynosi 2,306. Wartość ta jest większa od $t_{D4-P} = 0,413$. Możemy więc przyjąć hipotezę zerową, a tym samym stwierdzić, że próbka dowodowa D4 i porównawcza P są do siebie podobne i mogą pochodzić z tego samego obiektu. Natomiast zmienna t obliczona dla pary D1–P ($t_{D1-P} = 17,382$) jest większa od $t_{r,\alpha}$, co w tym przypadku nakazuje odrzucenie hipotezy zerowej.

KLASYFIKACJA OBIEKTÓW OPISANYCH PRZEZ KILKA ZMIENNYCH

Przykładowy problem

Problem klasyfikacji i dyskryminacji, gdy badane obiekty opisane są przez wiele cech, zilustrowano na przykładzie. Oznaczono zawartości sześciu pierwiastków (Na, Mg, Al, S, K, Ca) metodą SEM-EDX w pięciu próbkach szkła dowodowego D1, ..., D5 i próbce porównawczej P. Każdą z próbek analizowano pięciokrotnie. Wyniki pomiarów i obliczonych wartości średnich zebrano w tabeli V.

Należy więc rozpatrywać 6 obiektów ($m = 6$), z których każdy jest opisany przez 6 zmiennych ($n = 6$). Zawarte w tabeli V dane o średnich zawartościach pierwiastków w badanych próbkach można przedstawić w postaci macierzy o wymiarach $m \times n$ {9}.

$$\begin{array}{l} \text{Na} \quad \text{Mg} \quad \text{Al} \quad \text{S} \quad \text{K} \quad \text{Ca} \\ \text{D1} \left(\begin{array}{cccccc} 9,36 & 0 & 1,18 & 0,12 & 1,58 & 4,78 \\ 9,46 & 2,78 & 0,16 & 0,12 & 0,18 & 5,88 \\ 10,46 & 1,48 & 0,50 & 0,08 & 0 & 6,84 \\ 9,56 & 0 & 1,06 & 0,14 & 1,32 & 5,78 \\ 9,58 & 2,26 & 0,34 & 0,08 & 0,18 & 5,86 \\ 9,66 & 0 & 1,12 & 0,14 & 1,20 & 5,76 \end{array} \right) \end{array} \quad \{9\}$$

Stworzono również bazę danych składającą się z rezultatów analiz składu pierwiastkowego wykonanych metodą SEM-EDX dla próbek szkła pochodzących z różnych grup użytkowych (szkła opakowaniowego bezbarwnego – px, szkła opakowaniowego zielonego – pg, szkła opakowaniowego brązowego – pu, szkła z reflektorów samochodowych – r, szkła z szyb samochodowych – c oraz szkła okiennego – w).

Podejście niestatystyczne

Zastosujemy najpierw niestatystyczne podejście do klasyfikacji. Analizujemy wówczas bazę danych starając się wybrać pierwiastki, których zakresy stężeń w poszczególnych grupach szkła nie pokrywają się. Zawartości tak wytypowanych pierwiastków stają się kryterium podziału w schemacie klasyfikacji szkła (rycina 1).

Należy podkreślić, że schemat ten pozwala jedynie na jednoznaczną klasyfikację szkła opakowaniowego zielonego (pg) i brązowego (pu) oraz na odróżnienie zbioru szkieł okiennych i szyb samochodowych (CW) od zbioru szkieł reflektorowych i opakowaniowych bezbarwnych (PR). Nie pozwala on jednak na jednoznaczną klasyfikację w obrębie zbiorów CW i PR.

Według powyższego schematu dokonajmy klasyfikacji próbki D1 na podstawie jej składu pierwiastkowego (tabela V). Próbka ta nie zawiera żelaza (Fe), ani magnezu (Mg), dlatego klasyfikujemy ją do zbioru PR. Próbka ta może zatem pochodzić z klosza reflektora samochodowego (r) lub z białych opakowań szklanych (px). Tym samym sposobem można przeprowadzić klasyfikację pozostałych próbek. Jej rezultaty przedstawia tabela VI.

Próbki D1, D3 i D4 zostały zakwalifikowane do zbioru PR, podobnie jak próbka porównawcza P. Analiza składu pierwiastkowego tych próbek (tabela V) pozwala wyeliminować próbkę D3 ze zbioru próbek wykazujących podobieństwo do szkła porównawczego. W przeciwieństwie do próbki P, próbka D3 zawiera magnez (Mg), a nie zawiera potasu (K). Próbki D1 i D4 mają natomiast skład podobny do próbki porównawczej i wobec tego to one będą przedmiotem przeprowadzanej dyskryminacji.

Test t-Studenta i metoda „3 sigma”

W celu wykorzystania metody „3 sigma” lub testu t-Studenta w taki sam sposób, jak w przykładzie ze współczynnikiem załamania światła, należy wykonać odpowiednie obliczenia dla każdego z pierwiastków osobno. Wyniki takiego postępowania zamieszczono w tabelach VII i VIII. Z tabel tych wynika, że analiza przeprowadzona metodą „3 sigma” i testem t-Studenta dla próbki D4 wykazuje jej podobieństwo z próbką P w przypadku każdego z oznaczanych pierwiastków.

Trudności występują przy interpretacji wyników uzyskanych dla pary D1 – P. Analiza metodą „3 sigma” wykazuje, że tylko w przypadku wapnia wyznaczone dla tych próbek przedziały ufności nie pokrywają się. Pozostałe wyniki mogłyby więc sugerować, że próbka D1 jest podobna do próbki P. Jednakże na podstawie testu t-Studenta dla średnich z prób niezależnych można wnioskować, że próbki szkła dowodowego (D1) i porównawczego (P) nie są do siebie podobne. Na takie stwierdzenie pozwala analiza rezultatów zawartych w tabeli VII. Wartości t_{obl} dla sodu, potasu i wapnia są większe od $t_{r,\alpha}$ równego 2,776 ($r = 4$; $\alpha = 0,05$), co oznacza, że nie można przyjąć hipotezy zerowej mówiącej o równości średnich zawartości tych pierwiastków w próbkach D1 i P. Wartość t_{obl} dla glinu (Al) i siarki (S) pozwala natomiast na przyję-

cie hipotezy zerowej. Metoda „3 sigma” wskazuje więc na podobieństwo pary D – P, a test t-Studenta daje rezultat przeciwny. Chcąc rozwiązać ten problem, należałoby odpowiedzieć na pytanie, która z metod jest bardziej miarodajna. Jeżeli test t-Studenta, to nasuwa się kolejne pytanie, czy stosunek 3:2, tj. dla trzech pierwiastków rezultat testu t-Studenta negatywny, a dla dwóch pozytywny, jest wystarczający do wnioskowania o braku podobieństwa próbki D1 do próbki P? Odpowiedź twierdząca na ostatnie pytanie wydaje się subiektywna. Pomocne mogłoby być sprawdzenie, czy występują korelacje pomiędzy zawartością oznaczanych pierwiastków w porównywanych próbkach [2, 9] oraz jaki jest ich kierunek, tzn. czy wzrost zawartości jednego pierwiastka powoduje wzrost, czy też spadek zawartości innego pierwiastka. Wykrycie istotnych korelacji między parą lub parami pierwiastków powinno być wówczas uwzględnione w trakcie interpretacji uzyskanych wyników metodą „3 sigma” i testem t-Studenta. Gdy jednak brakuje korelacji pomiędzy wynikami składu pierwiastkowego analizowanych mikrookruszków, wówczas interpretacja wyników pary D1 – P jest nadal problematyczna.

Z powyższych rozważań wynika, że w przypadku analizowania obiektów opisanych przez kilka zmiennych, podejście niestatystyczne do celów klasyfikacji nie daje rozstrzygających rezultatów, a użycie w celu dyskryminacji metod „3 sigma” i (lub) testu t-Studenta dla każdej z cech oddzielnie może stwarzać problemy interpretacyjne. Właściwe wydaje się więc zastosowanie takich metod statystycznych, które jednocześnie rozpatrywałyby kilka cech, jak np. analiza dyskryminacyjna i analiza skupień.

Analiza skupień

W niniejszej pracy szerzej opisana zostanie analiza skupień [3, 14]. Metoda ta pozwala grupować obiekty w zbiory zwane skupieniami lub klastrami (z ang. cluster), które są zbudowane z obiektów podobnych do siebie. Miarą podobieństwa jest wzajemna odległość pomiędzy rozpatrywanymi obiektami (np. próbkami szkła). W analizie skupień wykorzystuje się też prostą zależność, że im mniejsza odległość między obiektami, tym są bardziej do siebie podobne.

Obliczanie odległości pomiędzy obiektami, gdy są one opisane przez kilka cech ilościowych, zostanie przedstawione na przykładzie kwadratu odległości Euklidesowej. Weźmy obiekty D1 i D2, opisane przez dwa pierwsze wiersze w macierzy {9} i umieścmy je na płaszczyźnie określonej przez osie stężeń sodu i magnezu C_{Mg} i C_{Na} (rycina 2).

Na podstawie twierdzenia Pitagorasa można zapisać:

$$D_{D1,D2}^2 = (OD_1)^2 + (OD_2)^2 . \quad \{10\}$$

Z rysunku 2 odczytujemy:

$$OD_1 = c_{Mg-1} - c_{Mg-2} , \quad \{11\}$$

$$OD_2 = c_{Na-1} - c_{Na-2} . \quad \{12\}$$

Podstawiając wzory {11} i {12} do równania {10} uzyskuje się:

$$D_{D1,D2}^2 = (c_{Mg-1} - c_{Mg-2})^2 + (c_{Na-1} - c_{Na-2})^2 . \quad \{13\}$$

Dla przypadku n-wymiarowego otrzymujemy:

$$D_{D_1, D_2} = \sum_{i=1}^n (c_{D_1-i} - c_{D_2-i})^2. \quad \{14\}$$

Podstawiając do wzoru {14} średnie zawartości sodu, magnezu, glinu, siarki, potasu i wapnia oznaczone w próbkach D1 i D2, uzyskujemy wartość kwadratu odległości Euklidesowej pomiędzy tymi próbkami wynoszącą 12,042.

Po obliczeniu wzajemnych odległości pozostałych obiektów uzyskuje się symetryczną macierz podobieństwa tych obiektów {15}.

$$\begin{array}{c} \begin{array}{cccccc} & D1 & D2 & D3 & D4 & D5 & P \\ D1 & \left(\begin{array}{cccccc} 0 & & & & & \\ 11,949 & 0 & & & & \\ 10,604 & 3,761 & 0 & & & \\ 1,122 & 9,858 & 6,184 & 0 & & \\ 8,990 & 0,319 & 2,401 & 6,936 & 0 & \\ 1,199 & 9,754 & 5,825 & 0,028 & 6,776 & 0 \end{array} \right) & & & & & \\ D2 & & & & & & \\ D3 & & & & & & \\ D4 & & & & & & \\ D5 & & & & & & \\ P & & & & & & \end{array} \end{array} \quad \{15\}$$

Wystarczy zapisać tylko połowę tej macierzy, ponieważ odległość mierzona od obiektu D1 do obiektu D2 jest identyczna, jak odległość mierzona od obiektu D2 do obiektu D1.

Mając do dyspozycji macierz wzajemnych odległości analizowanych obiektów (D1, ..., D5 i P), można przystąpić do budowy skupień. Istnieje wiele metod ich konstruowania. Ogólnie dzieli się je na niehierarchiczne i hierarchiczne, a te ostatnie z kolei na aglomeracyjne (łącające) i dzielące. Metody hierarchiczne, które są najczęściej używane w badaniach empirycznych, opierają się na zasadzie, że mniejsze skupienia wchodzi w skład większych. W przypadku metod hierarchiczno-aglomeracyjnych na początku mamy tyle samo skupień, co rozpatrywanych obiektów, a to oznacza, że każde skupienie składa się z jednego obiektu. Proces skupiania kończy się w momencie zbudowania jednego klastra, w skład którego wchodzi wszystkie rozpatrywane obiekty. W metodach hierarchicznych dzielących postępuje się odwrotnie.

Proces konstruowania skupień w metodach hierarchiczno-aglomeracyjnych składa się z kilku etapów. W pierwszym wybiera się z macierzy podobieństwa {15} tę parę obiektów, pomiędzy którymi odległość jest najmniejsza. W macierzy odległości są to obiekty D4 i P ($d_{D_4-P} = 0,028$). Następnie określa się odległość nowego skupienia S (D4, P) od pozostałych obiektów D1, D2, D3 i D5.

Poszczególne metody konstruowania skupień różnią się sposobem określania odległości pomiędzy nowo powstałym skupieniem a pozostałymi obiektami (pojedyncze obiekty lub/i skupienia). Wśród nich znajdują się m.in. metody najdalszego i najbliższego sąsiedztwa. W przypadku pierwszej z nich odległość obiektu S (D4, P) od innego obiektu, np. D1 (d_{S-D_1}) jest równa największej odległości spośród odległości wyznaczonych dla par D4 – D1 i P – D1, tj. $d_{S-D_1} = \max(d_i)$ (tabela IX):

$$d_{S-D_1} = d_{P-D_1} = 1,199.$$

Analogicznie w metodzie najbliższego sąsiedztwa odległość skupienia S (D4, P) od obiektu D1 będzie najmniejszą z tych odległości ($d_{S-D_1} = \min(d_i)$). Na podstawie danych z tabeli IX uzyskujemy:

$$d_{S-D_1} = d_{D_4-D_1} = 1,122.$$

Na tych samych zasadach wyznacza się odległość skupienia S (D4, P) od pozostałych obiektów (D2, D3, D5) i tworzy nowe macierze podobieństwa – osobno dla metody najdalszego {16} i najbliższego sąsiedztwa {17}. Wymiar tych macierzy jest pomniejszony o jeden rząd i jedną kolumnę w stosunku do macierzy podobieństwa sprzed tej operacji {15}.

$$\begin{array}{c}
 \begin{array}{ccccc}
 & D1 & D2 & D3 & S & D5 \\
 D1 & \left(\begin{array}{c} 0 \\ 12,042 \\ 10,681 \\ 1,199 \\ 9,091 \end{array} \right. & & & & \\
 D2 & & \left(\begin{array}{c} 0 \\ 3,601 \\ 9,858 \end{array} \right. & & & \\
 D3 & & & \left(\begin{array}{c} 0 \\ 6,184 \\ 2,374 \end{array} \right. & & \\
 S & & & & \left(\begin{array}{c} 0 \\ 6,936 \end{array} \right. & \\
 D5 & & & & & \left. \begin{array}{c} 0 \\ 0 \end{array} \right) \\
 \end{array} & & & & & \{16\}
 \end{array}$$

$$\begin{array}{c}
 \begin{array}{ccccc}
 & D1 & D2 & D3 & S & D5 \\
 D1 & \left(\begin{array}{c} 0 \\ 12,042 \\ 10,681 \\ 1,122 \\ 9,091 \end{array} \right. & & & & \\
 D2 & & \left(\begin{array}{c} 0 \\ 3,601 \\ 9,745 \end{array} \right. & & & \\
 D3 & & & \left(\begin{array}{c} 0 \\ 5,825 \\ 2,374 \end{array} \right. & & \\
 S & & & & \left(\begin{array}{c} 0 \\ 6,776 \end{array} \right. & \\
 D5 & & & & & \left. \begin{array}{c} 0 \\ 0 \end{array} \right) \\
 \end{array} & & & & & \{17\}
 \end{array}$$

Następnie postępuje się jak poprzednio, tzn. z macierzy {16} i {17} zostają wyszukane obiekty, pomiędzy którymi odległość jest najmniejsza i tworzy się nową macierz odległości, stosując przy jej konstruowaniu metodę najdalszego bądź najbliższego sąsiedztwa. Rezultaty analizy skupień mogą być zaprezentowane graficznie w postaci dendrogramów. Dendrogramy uzyskane dla powyższego przykładu są przedstawione na rycinach 3 i 4.

Z przeprowadzonej analizy skupień wynika, że próbki D1 i D4 wykazują podobieństwo do próbki P. Tworzą one trójelementowe skupienie. Próbki D2, D3 i D5 tworzą również trójelementowe skupienie, wykazując podobieństwo względem siebie, ale nie względem próbki porównawczej P. Odległości pomiędzy obiektami wewnątrz tak wyodrębnionych skupień są mniejsze niż odległość pomiędzy skupieniem D1 – D4 – P i skupieniem D2 – D3 – D5. Szczególnie wyraźnie różnice te są zarysowane w przypadku zastosowania metody najdalszego sąsiedztwa (rycina 4).

Odpowiedź na pytanie, która z metod obliczania odległości między obiektami i konstruowania skupień jest najlepsza, zależy od konkretnego zadania, które należy rozwiązać. Na podstawie danych literaturowych [10] można wnioskować, że dla celów klasyfikacji próbek szkła stosuje się najczęściej kwadrat odległości Euklidesowej oraz metodę najdalszego sąsiedztwa.

Wykonana powyżej analiza nie pozwala na jednoznaczną klasyfikację (identyfikację grupową) próbek dowodowych D1, ..., D5. Nie możemy bowiem na jej podstawie powiedzieć, że próbka D4 jest np. szkłem reflektorowym, a próbka D2 np. szkłem samochodowym. Analiza ta mówi jedynie o tym, które obiekty są do siebie podobne, a które nie. Aby przeprowadzić klasyfikację, konieczne jest użycie informacji o składzie pierwiastkowym szkieł, pochodzących z różnych grup użytkowych, która została uzyskana tą samą metodą, jak dla badanej próbki dowodowej i porównawczej.

Dokonajmy klasyfikacji próbek dowodowych dysponując bazą danych o składzie pierwiastkowym szkieł reflektorowych (r), szkła opakowaniowego bezbarwnego (px)

i szyb samochodowych (c) uzyskanych metodą SEM-EDX. Przed przystąpieniem do właściwej klasyfikacji próbek dowodowych metodą analizy skupień należy opracować metodę, która pozwoli na klasyfikację grupową próbek szkła. Opracowanie takiej metody polega na wyborze zmiennych (stężeń pierwiastków) używanych do obliczania odległości pomiędzy obiektami, sposobu obliczania odległości oraz metody skupiania. Wybrane parametry powinny pozwolić na uzyskanie dendrogramu złożonego z trzech jednorodnych skupisk zbudowanych tak, aby w skład każdego wchodziły próbki tylko jednego rodzaju szkła. W idealnym przypadku (rycina 5) uzyskamy trzy takie skupienia: jedno skupienie dla szkielek reflektorowych (r), jedno dla szkielek opakowaniowych bezbarwnych (px) i jedno dla szkła z szyb samochodowych (c). Również poprawnym rozwiązaniem tego problemu byłby dendrogram pokazany na rycinie 6. Odległości pomiędzy grupami powinny znacznie przewyższać odległości występujące wewnątrz skupień.

Wybór zmiennych polega na wytypowaniu tych pierwiastków, których średnie zawartości istotnie różnią się w poszczególnych zbiorach. Rycina 7 przedstawia rezultaty analizy skupień przeprowadzonej dla obiektów zawartych w bazie danych, w której jako zmiennych użyto zawartości wszystkich pierwiastków oznaczonych w badanych próbkach (Al, Ba, Ca, Cr, Fe, K, Mg, Na, O, S, Si, Ti). Rycina 8 przedstawia rezultat analizy skupień tego samego zbioru próbek, w której jako zmiennych użyto zawartości pierwiastków wybranych metodą Tukeya [2, 9]. Metoda ta pozwala na wybór tych zmiennych (zawartości pierwiastków), których średnie wartości w sposób statystycznie istotny różnią się w badanych zbiorach. Analiza wyników zawartych w bazie danych wspomnianą metodą wykazała istotne różnice średnich zawartości Al, Ca, K, Mg i Na w rozpatrywanych grupach.

Należy prześledzić, czy budowa powyższych dendrogramów jest zbieżna z budową zakładaną na rycinach 5 i 6. Na rycinie 7 można wyróżnić następujące skupienia: dla grupy szkła opakowaniowego bezbarwnego skupienie px1, ..., px9, dla szkła z szyb samochodowych dwa skupienia: c1, ..., c4 oraz c8, ..., c12, a dla szkła z kloszy reflektorów samochodowych dwa skupienia: r1, ..., r5. Pozostałych 15 obiektów leży poza tymi skupieniami. Występujące obok siebie na dendrogramie obiekty px12, ..., px22 nie tworzą jednego skupienia. Wyróżnia się w nim dwie grupy: px12, ..., px14 i px16, ..., px19 oraz pojedyncze obiekty: px15 i px20, bowiem wzajemne odległości pomiędzy nimi są znacznie większe niż wewnątrz skupień. Ta ocena odległości ma charakter jakościowy. W celu ilościowego określenia, jaką odległość można uznać za wystarczającą, aby mówić o dwóch oddzielnych skupieniach, należałoby przeanalizować odległości występujące na dendrogramie podobnym do uwidocznionego na rycinie 7, ale uzyskanego dla większej liczby obiektów w poszczególnych grupach. Para px 21 i px 22 wykazuje natomiast podobieństwo do grupy obiektów c5 – c6. W pozostałej części dendrogramu można ponadto wyróżnić pojedyncze obiekty c7 i px23. Dendrogram ten odbiega więc w znacznym stopniu od dendrogramów modelowych przedstawionych na rycinach 5 i 6.

Na rycinie 8 daje się zaobserwować 3 duże skupienia po jednym dla każdej z rozważanych grup szkła: dla szkła z opakowań białych (px1, ..., px17), dla szkła z szyb samochodowych (c1, ..., c11) oraz dla szkła z kloszy reflektorów samochodowych (r1, ..., r6). Jednak i one nie zawierają wszystkich analizowanych obiektów, bowiem próbki szkła opakowaniowego białego tworzą dodatkowo trójelementowe skupienie (px19, px20), a obiekty px21, px22, c12, px23 i r7 znajdują się poza wyżej wymienio-

nymi skupieniami. Budowa tego dendrogramu, pomimo występowania 5 obiektów poza głównymi skupieniami, jest bardziej zbliżona do przedstawionego na rycinie 6, niż budowa dendrogramu uzyskanego wówczas, gdy jako zmiennych użyto zawartości wszystkich oznaczanych pierwiastków (rycina 7). Przewaga ta nie wydaje się jednak przekonująca, dlatego poddano klasyfikacji obiekty D1, ..., D5 obydwoma metodami. W celu dokonania klasyfikacji próbek dowodowych D1, ..., D5 oraz próbki porównawczej P wyniki ich analizy pierwiastkowej dołączono do bazy danych i przeprowadzono analizę skupień. Jako zmienne zastosowano zawartości wszystkich pierwiastków znajdujących się w bazie danych. Dla rozpatrywanych próbek należało więc uwzględnić zawartości pierwiastków nie tylko wymienionych w tabeli V, lecz także: tlenu (O), krzemu (Si), żelaza (Fe), chromu (Cr), baru (Ba) oraz tytanu (Ti) – por. tabela X. Dendrogramy na rycinach 9 i 10 obrazują rezultaty tak przeprowadzonych klasyfikacji.

Próbka porównawcza P będąca szkłem z klosza reflektora samochodowego (r) została błędnie zakwalifikowana do grupy szkła opakowaniowego bezbarwnego (px), gdy poddano ją klasyfikacji metodą opartą na analizie wszystkich zmiennych (rycina 9). W tym przypadku klasyfikacja obiektów D1, ..., D5 również mogła być nieprawidłowa. Dopiero po zastosowaniu wstępnej selekcji zmiennych, tj. zawartości pierwiastków wytypowanych metodą Tukeya, uzyskano poprawną klasyfikację próbki P (rycina 10). Na podstawie tego samego dendrogramu można stwierdzić, że próbki D1 i D4 najprawdopodobniej pochodzą ze szkła reflektorów samochodowych, ponieważ wchodzi w skład skupienia (r). Okruchy D2 i D5 zostały zakwalifikowane do grupy szkła szyb samochodowych, a próbka D3 może pochodzić np. z bezbarwnej butelki lub słoika. Została ona bowiem zakwalifikowana do grupy szkła bezbarwnego opakowaniowego (px).

Jak wspomniano wcześniej, budowa dendrogramów na rycinach 7 i 8 nie jest w pełni zbieżna z założoną jak na rycinach 5 i 6. Powodem tego może być fakt, że zawartości pierwiastków w analizowanych próbkach różnią się nawet o trzy rzędy wielkości (np. dla Si od 28,84% wag. do 40,85% wag., dla S od 0,1% wag. do 0,2% wag.). W takiej sytuacji można przeprowadzić tzw. „skalowanie” zmiennych. W zależności od charakteru danych, które opisują analizowane obiekty, można stosować różne formy transformacji pierwotnych wyników zwanych też wynikami surowymi. W przypadku danych ilościowych można wybrać jedno z powszechnie stosowanych przekształceń zamieszczonych w tabeli XI.

Decyzja, które przekształcenie zastosować, podejmuje się po zbadaniu związku pomiędzy wariancją i średnimi eksperymentalnymi [9]. Jeszcze inną, często stosowaną metodą, jest dzielenie wyników pierwotnych przez wartość odchylenia standardowego obliczonego dla średniej wartości oznaczanej zmiennej. Niekoniecznie muszą to być dane tylko ilościowe czy tylko jakościowe. Istnieją też metody „skalowania” dla obiektów opisanych przez obydwa rodzaje zmiennych.

W rozważanym przykładzie problem różnicy trzech rzędów wielkości w zawartościach poszczególnych pierwiastków można byłoby rozwiązać stosując przekształcenie logarytmiczne. Jednak zawartość niektórych pierwiastków wynosi 0, co uniemożliwia bezpośrednie zastosowanie tego typu transformacji ($\log 0$ nie istnieje). Problem ten omija się w ten sposób, że zerową zawartość jakiegoś pierwiastka zastępuje się niewielką liczbą. Istnieje też możliwość zastosowania zależności {18}:

$$Z_i = \frac{X_i - \min X}{\max X - \min X}, \quad \{18\}$$

gdzie: X_i – wyznaczona zawartość pierwiastka uzyskana w i -tym pomiarze; $\min X$, $\max X$ – minimalna i maksymalna zawartość x -tego pierwiastka w rozpatrywanym modelu.

Po przeprowadzeniu tej operacji zawartości oznaczanych pierwiastków zostają zawarte w przedziale $<0; 1>$. Dendrogram na rycinie 11 uzyskany został dla przypadku, gdy jako zmiennych użyto przekształconych zawartości wszystkich oznaczanych pierwiastków w badanych próbkach szkła. Natomiast dendrogram na rycinie 12 powstał, gdy jako zmiennych użyto przeskalowanych zawartości pierwiastków wytypowanych metodą Tukeya (Al, Ca, K, Mg i Na) [2, 9]. W trakcie analizy skupień zastosowano, jak uprzednio, kwadrat odległości Euklidesowej oraz metodę najdalszego sąsiedztwa.

Porównując dendrogramy na rycinie 7 przed „skalowaniem” i rycinie 11 po „skalowaniu” stwierdzono, że w tym przypadku nie nastąpiła istotna poprawa. Budowa dendrogramu uzyskanego po transformacji danych surowych w dalszym ciągu odbiega od tej zakładanej na rycinach 5 lub 6, a klasyfikacja obiektów D1, ..., D5 i P także nie jest poprawna. Na rycinie 13 można zaobserwować, że próbka P z grupy (r) została błędnie zakwalifikowana jako próbka z grupy (px). Natomiast w przypadku dendrogramu uzyskanego dla zmiennych przekształconych i wyselekcjonowanych metodą Tukeya (rycina 12) uzyskano znaczną poprawę w porównaniu z dendrogramem z ryciny 8 (przed „skalowaniem”). Można bowiem wyróżnić po jednym skupieniu dla każdej z rozpatrywanych grup szkła (px1, ..., px20; c1, ..., c11; r1, ..., r7) i tylko 4 obiekty poza nimi (px21, c12, px22 i px23). Rycina 14 przedstawia rezultat klasyfikacji analizowanych okruszków szklanych. Jej wyniki są zgodne z rezultatami klasyfikacji przedstawionej na dendrogramie na rycinie 8.

Podsumowując, należy stwierdzić, że w celu rozwiązania problemu klasyfikacji obiektów należących do kilku grup konieczne jest przeanalizowanie różnych kombinacji metod transformacji danych pierwotnych, sposobów wyrażania odległości między obiektami oraz metod konstruowania skupień tak, aby znaleźć optymalne rezultaty.

DYSKRYMINACJA OBIEKTÓW OPISANYCH PRZEZ KILKA ZMIENNYCH

Teraz rodzi się pytanie, czy próbki D1 i D4 mogą pochodzić z rozbitego reflektora P. Rozwiążemy ten problem stosując analizę skupień.

Jak wspomniano wcześniej, w trakcie analizy skupień powstają skupienia, w skład których wchodzi obiekty wykazujące wzajemne podobieństwo. Przyjmijmy, że w celu dyskryminacji dwóch próbek metodą analizy skupień będziemy rozpatrywać pojedyncze wyniki uzyskane w trakcie oznaczania składu pierwiastkowego próbki dowodowej i porównawczej. Gdy chcemy dokonać dyskryminacji dwóch mikrookruszków szklanych, czyli stwierdzić, czy mogą one pochodzić z tego samego obiektu, czy też z dwóch różnych, jako zmienne stosujemy zawartości wszystkich oznaczanych pierwiastków. W idealnym przypadku próbki pochodzące z tego samego obiektu powinny być podobne w obrębie każdego z pierwiastków. W przypadku uzyskania oddzielnych skupień, w skład których wchodzi wyniki tylko jednego z obiek-

tów (np. jedno skupienie dla wyników D1 i jedno dla wyników P), można wnioskować, że okruchy te nie są do siebie podobne, czyli nie pochodzą z tego samego obiektu. Natomiast w przypadku, gdyby analizowane wyniki utworzyły skupienie lub skupienia złożone z wyników pochodzących z obu okruchów, uzasadnione jest stwierdzenie, że okruch dowodowy i porównawczy są do siebie podobne i mogą pochodzić z tego samego obiektu.

Dokonajmy więc dyskryminacji pomiędzy okruchami dowodowymi D1, D4 i okruchem porównawczym P. Ryciny 15 i 16 prezentują wyniki przeprowadzonej analizy skupień.

Z rysunku 15 wynika, że okruch D1 najprawdopodobniej nie pochodzi ze szkła reflektorowego P. Obserwujemy bowiem powstanie dwóch oddzielnych skupień, w skład których wchodzi wyniki (obiekty) uzyskane tylko dla jednego z okruchów, w przeciwieństwie do sytuacji na dendrogramie przedstawionym na rycinie 16, a uzyskanym dla pary D4 – P. Budowa tego dendrogramu pozwala wnioskować, że okruch dowodowy D4 i porównawczy P mogą pochodzić z tego samego reflektora samochodowego.

UWAGI KOŃCOWE

Prezentowana praca stanowi przegląd metod analizy statystycznej i chemometrycznej, jakie wykorzystano w trakcie opracowywania wyników badania mikroskładów szkła, uzyskanych zarówno w ramach badań podstawowych, jak i podczas wykonywania ekspertyz dla potrzeb wymiaru sprawiedliwości. Elementy chemometrii i analizy statystycznej użyte do klasyfikacji i dyskryminacji mikrookruchów szkła mogą być z powodzeniem zastosowane w badaniach innego rodzaju materiałów. W szczególności metody analizy skupień znajdują coraz szersze zastosowanie nie tylko w naukach ścisłych, lecz także w takich dziedzinach, jak ekonomia, psychologia, pedagogika czy socjologia.