DOI: 10.18668/NG2015.11.13

Bogdan Filar, Tadeusz Kwilosz, Mariusz Miziołek, Wacława Piesik-Buś, Jadwiga Zamojcin

Oil and Gas Institute – National Research Institute

The use of cluster analysis for the segmentation of the physicochemical properties of shale gas deposits

Cluster analysis methods have been adapted for the segmentation of data describing the generative and reservoir properties of shale gas-type formations. Tests were carried out for the segmentation of data describing the geochemical properties of core samples collected from eight wells within the stratigraphic unit of Llandovery (Silurian). The study was conducted in two stages. During the first stage, data segmentation was performed for three wells described by the largest number of measurements. The second stage of data segmentation involved a set of samples originating from all wells.

Key words: geochemistry, cluster analysis, gas-bearing shale, Silurian.

Zastosowanie analizy skupień do segmentacji właściwości fizykochemicznych złóż typu *shale gas*

Zaadaptowano metody analizy skupień do segmentacji danych opisujących właściwości generacyjne i zbiornikowe formacji typu *shale gas*. Przeprowadzono próby segmentacji danych opisujących parametry geochemiczne próbek rdzeniowych pobranych z ośmiu odwiertów w obrębie piętra landoweru. Badania przeprowadzono w dwóch etapach. W pierwszym etapie dokonano segmentacji danych w obrębie trzech odwiertów opisanych największą liczbą pomiarów. W drugim etapie dokonano segmentacji danych na zbiorze próbek pochodzących z wszystkich odwiertów.

Słowa kluczowe: geochemia, analiza skupień, łupek gazonośny, sylur.

Introduction

One of the most important issues associated with the analysis of data originating from the study of geological structures – prospective for the exploration, accessing and extraction of unconventional gas deposits – is the evaluation and classification of obtained information in relation to the possible occurrence of economically viable gas reserves, the selection of optimal and safe strategies for accessing the deposits, and the methods of their extraction. On the one hand, enterprises owning exploratory concessions will receive enormous streams of data from the completed research: seismic, geophysical, petrographic, geochemical and others, and on the other hand, these enterprises will experience the pressure of time and economics in the face of the necessity to make choices and subsequent decisions related to the continuation of exploration and accessing preliminarily identified deposits. One helpful tool may prove to be a system (set of procedures and methods) designed to seek analogies to well identified and documented (regarding measurements and research results, and the used methods of access and effects of extraction) geological structures with unconventional natural gas deposits in the USA, Canada and (with time) domestic deposits. The developed methods (algorithms), operating on enormous datasets originating from domestic and foreign deposits, will be used in the search, as well as in a qualitative and quantitative analysis of multidimensional datasets and construction of similitude models, utilising the methods used in artificial intelligence, such as: cluster analyses, multidimensional comparative analysis (WOP),

neural networks, fuzzy sets and fractal analysis. The use of these types of tools will allow the preliminary evaluation and classification of analysed geological structures with respect to the possible occurrence of economically viable gas reserves.

In order to determine the characteristics of a deposit formation, in an exhaustive way reflecting the generative and reservoir capabilities of the rock, as well as the possibilities of accessing the gas reserves, a carefully selected set of attributes is required, involving a wide range of geological, geochemical, geophysical or geomechanical data. A review of all available data obtained during the stage of work related to the present paper allowed for the conclusion that the only complete and coherent set of data describing more than two wells and originating from the same stratigraphic unit consists of the results of geochemical examinations performed on cores collected from the following boreholes: Lubocino-1, Opalino-2, Kochanowo-1, Gdańsk-IG1, Opalino-3, Tępcz-1, Wysin-1 and Żarnowiec-IG1 – from within the Llandovery series. Since the research work at this stage was an experiment related to testing the data analysis methods and evaluation of their usefulness in searching for analogies, it has been decided to use a geochemical dataset. Another datasets will be used in subsequent papers in a more complex manner describing the generative and reservoir properties of shale gas-type formations.

Regarding this part of the paper, a review has been made concerning the available artificial intelligence methods, and the cluster analysis method (data clustering) was the first to be selected as the most prospective (within the scope of identifying analogies of conditions allowing the construction of similitude models of attributes describing geological structures). This does not mean discontinuation of research on the use of other abovementioned methods during further stages of work on the project.

Data review and evaluation

Commencing execution of this part of the task, the authors of the paper had at their disposal the data being the results of measurements, laboratory research, as well as interpretation results obtained from wells located within the following concessions: Wejherowo and Stara-Kiszewa, as well as the Żarnowiec drilling area.

According to the report prepared by Daniel M. Jarvie, Worldwide Geochemistry, LLC for the Lubocino-1 borehole, one of the most prospective (with respect to hydrocarbon generation properties) stratigraphic units within the Wejherowo concession is the Llandovery strata (including the Jantar member of bituminous claystones). It is also the best documented stratigraphic unit inside the Wejherowo concession with respect to examination results. For these reasons the authors of the paper have decided to choose the data from this unit to conduct tests of selected data analysis methods.

An assumption has been made that data analysis methods will be tested on datasets originating from at least three boreholes from the same stratigraphic unit. A review has been conducted for the results of measurements and geochemical examinations performed for 25 wells. It has been assumed that a given borehole will be an object of further analysis if geochemical analyses are performed involving more than two parameters: TOC (total organic carbon content) and R_o (vitrinite reflectance index) for at least ten samples collected from the selected stratigraphic unit. Based on the conducted analyses it has been concluded that the only complete and coherent set of data fulfilling the described objectives are the results of geochemical examinations performed on cores collected from the following boreholes: Lubocino-1, Opalino-2, Kochanowo-1, Gdańsk-IG1, Opalino-3, Tępcz-1, Wysin-1 and Żarnowiec-IG1 from within the Llandovery series. The data describing geochemical parameters (the results of a pyrolytic analysis for core samples) determined for cores collected from selected boreholes have been qualified for further analyses.

The use of cluster analysis for data segmentation

Grouping (segmentation) of data is a process involving the division of a set of data (observations, examination results) into subsets (classes) comprising "similar" elements (according to the predetermined rule of similitude). Such division is handled via the identification of natural groups, in which objects similar to each other are to be placed in one group, and the objects which differ considerably in different groups [2, 3]. In contrast to the pattern classification characterised by attributing objects to groups of predetermined properties, in this case the properties (or the number) of created groups are not known a priori.

The primary goal of developing effective methods of data segmentation as part of the implementation of the present paper is obtaining tools intended to examine whether the studied geological structures are similar with respect to their generative properties and hydrocarbon reserves. It is assumed that if the objects – samples (defined by a combination of results from measurements of various types, but having to

do with the generative properties and hydrocarbon reserves) originating from various geological structures are placed in the same groups (subsets), then these structures are "similar" and may be considered analogical.

One of the most efficient methods of data segmentation is cluster analysis, intended to compare and categorise objects described by numerous attributes (variables). The cluster analysis procedures allow the formation of groups (clusters) of objects "the least distant from each other" or "the most similar to each other", considered as points in multidimensional space, where the spatial dimension is determined by the number of variables describing the given objects.

Hierarchical agglomerate grouping is among the most frequently used, and is considered to be one of the more efficient grouping methods [1, 6]. In this method new clusters (aggregations) are formed by merging existing clusters. One condition of their merging is their adequate distance (or other used measure of proximity).

Such grouping involves the use of the following algorithm:

1) select the initial set of clusters,
2) find the closest pair of clusters and merge them into one,
3) repeat step 2 until fulfilling the rule of completion.

The rule of completion is (usually):

- the lack of cluster pairs located less than the given threshold distance apart (d_{max}),
- merging of all clusters into one set.

The measure of remoteness in a multidimensional space is the properly defined distance. This kind of different variants of the cluster analysis method will be used in the present paper.

The distance may be defined in multiple ways depending on the type of attributes describing the individual data (quantitative, qualitative data, ranks). Among the most frequently used are:

- Euclidean distance

$$d_{ik} = \sqrt{\frac{1}{m}\sum_{j=1}^{m}\left(x'_{ij} - x'_{kj}\right)^2}$$

- city block (Manhattan) distance

$$d_{ik} = \frac{1}{m}\sum_{j=1}^{m}\left|x'_{ij} - x'_{kj}\right|$$

- Chebyshev distance

$$d_{ik} = \max_j\left|x'_{ij} - x'_{kj}\right|$$

The individual variants of cluster analysis differ in the manner of determining the distance between clusters.

- the nearest neighbour method (single linkage) – the distance between clusters is the distance between the two closest objects,
- the farthest neighbour method (complete linkage) – the distance between clusters is the distance between the two most distant objects,
- the median method – the distance between two clusters is the median of the distance between the units of the first and the second cluster,
- the group average method – the distance between two clusters is the average distance between the units of the first and the second cluster,
- the centre of gravity method – the distance between two clusters is the distance between the centres of gravity of the first and the second clusters,
- the Ward method – sampling of merging all cluster pairs and selection of such merging where the variance of distance inside a formed cluster is the smallest.

All of the abovementioned methods have been tested with respect to efficiency for selected datasets.

The individual attributes – data describing the examined object (sample, core etc.) are usually of various types (TOC, permeability, porosity etc.), with a varying range of values. In order for objects defined in such a manner to be considered points in multidimensional space, in which distance does not depend on coordinates, data transformation must be conducted. This may be accomplished in two ways:

- data standardisation

$$x'_i = \left(\frac{x_i - \bar{x}}{S_x}\right)$$

- data unification

$$x'_i = \left(\frac{x_i}{x_{\max} - x_{\min}}\right)$$

where:

$\bar{x}$ – the average value of all data for the given attribute,
S_x – standard deviation of all data for the given attribute,
x_{max}, x_{min} – maximum and minimum value in the dataset for the given attribute.

The use of cluster analysis for the segmentation of core samples described by the results of geochemical measurements – validation of data model

Having analysed the sets of data describing the results of geochemical examinations (of samples representing the Llandovery strata) from the 8 qualified boreholes, it may be concluded that three of them: Lubocino-1 (41 samples), Opalino-2 (88 samples) and Kochanowo-1 (45 samples) are considerably more extensive than the others. The next one

with respect to the amount of examined samples is Tępcz-1 (23 samples). Due to this, the experiment associated with data segmentation was conducted in two stages:

- during stage I data segmentation was conducted within the three indicated wells, for each one of them separately,
- during stage II the segmentation was conducted for data originating from all wells collectively.

Data segmentation within a single well – stage I

According to what has been indicated in the introduction, segmentation has been conducted for data originating from the results of geochemical examinations performed for core samples collected from three boreholes located in the Wejherowo concession:

- Lubocino-1 (2850÷2907 m) – 41 samples,
- Opalino-2 (2803÷2884 m) – 88 samples,
- Kochanowo-1 (3150÷3212 m) – 45 samples.

The following were selected as attributes (geochemical parameters) describing the elements of sets (of examined samples) [6]:

- TOC – total organic carbon content [wt%],
- T_{max} – temperature at which the maximum amount of hydrocarbons is created during cracking of kerogen [°C],
- S_1 – free hydrocarbon content [mg HC/g of rock],
- S_2 – the amount of hydrocarbons released during cracking of kerogen [mg HC/g of rock],
- S_3 – the amount of CO_2 created from the destruction of the organic substance [mg CO_2/g of rock],
- PI = $S_1/(S_1 + S_2)$ – generation index,
- PC – pyrolytic carbon content [wt%],
- RC – residual carbon content [wt%],
- HI – hydrogen index [mg HC/g TOC],
- OI – oxygen index [mg CO_2/g TOC],
- Total MINC – overall mineral carbon content [wt%].

All the used algorithms have been implemented by the authors of the paper as procedures (macros) in MS Excel calculation sheets.

The main selected goal of the conducted analysis was the examination of the internal structure of datasets describing each borehole individually with respect to distinguishing subsets of samples with "similar" characteristic properties. Checking the usefulness of various variants of the cluster analysis method for the similitude analysis used for this type of data.

Before commencing segmentation, validation of data model was executed. Variables (attributes; parameters describing the studied object) which did not contribute significant information were eliminated from the datasets. This was accomplished based on the following assumptions.

Assumptions of data model validation:

1) A leading variable (attribute) which will not be removed from the model is to be selected. To this variable (explained variable) will be referenced the remaining variables of the model (explanatory variables).
2) The variables should be characterised by adequate dispersion (variation) of the values. The variation coefficient V_x is assumed to be the measure of variation; its value should not be lower than 0.2 (20%). This condition is not considered to be categorical. If there are other premises, the variable remains.
3) Variables should not be correlated to each other. The Pearson coefficient r_{ij}, which should not be higher than 0.9, is assumed as the measure of correlation. If this value is exceeded, the variable for which the Pearson correlation coefficient with the explained variable is higher is removed from the model.

TOC has been selected as the leading variable. The variation coefficients for all variables of the model have been calculated. The results are shown in Table 1.

Table 1. Coefficients of variation for model attributes

	Lubocino-1	Opalino-2	Kochnowo-1
X_i	Vx_i	Vx_i	Vx_i
T_{max}	2.22%	11.84%	16.84%
S_1	113.36%	132.98%	116.76%
S_2	123.87%	162.18%	139.26%
S_3	46.90%	62.54%	114.71%
PI	24.34%	27.47%	22.85%
PC	115.37%	147.53%	128.21%
RC	111.02%	137.60%	129.45%
HI	38.05%	54.63%	113.93%
OI	113.48%	116.91%	161.57%
Total MINC	87.30%	93.82%	86.58%
TOC	111.27%	138.80%	129.08%

The T_{max} variable does not fulfil the condition $V_x > 20\%$ for the data from all three boreholes. In spite of this, it has been decided not to remove it from the model. In the next step of model validation it was checked to what degree the individual variables are correlated to each other. The Pearson linear correlation coefficient R has been used. The results are presented in Tables: 2, 3, 4.

The following may be concluded from the analysis of the obtained results:

- The highest (exceeding 0.9) correlation coefficients (for all boreholes) have pairs of variables: (S_1, S_2), (S_1, PC), (S_1, RC), (S_2, RC), (S_2, PC) and (PC, RC).
- The following variables have the highest coefficients of correlation with the TOC variable: S_2, PC and RC.

On this basis, using the previously described assumptions

Table 2. Multiple correlation coefficients *R* for the data from the Lubocino-1 borehole

R	T_{max}	S_1	S_2	S_3	PI	PC	RC	HI	OI	Total MINC	TOC
T_{max}	1.000	–0.594	–0.424	0.063	–0.327	–0.469	–0.364	–0.652	0.638	0.106	–0.385
S_1	–0.594	1.000	0.942	–0.109	0.160	0.966	0.903	0.676	–0.599	–0.015	0.919
S_2	–0.424	0.942	1.000	–0.104	–0.072	0.996	0.975	0.609	–0.550	0.026	0.984
S_3	0.063	–0.109	–0.104	1.000	–0.205	–0.099	–0.091	–0.240	0.241	0.248	–0.093
PI	–0.327	0.160	–0.072	–0.205	1.000	–0.018	–0.094	–0.049	–0.277	–0.115	–0.080
PC	–0.469	0.966	0.996	–0.099	–0.018	1.000	0.968	0.630	–0.566	0.018	0.979
RC	–0.364	0.903	0.975	–0.091	–0.094	0.968	1.000	0.476	–0.560	0.067	0.999
HI	–0.652	0.676	0.609	–0.240	–0.049	0.630	0.476	1.000	–0.368	–0.283	0.508
OI	0.638	–0.599	–0.550	0.241	–0.277	–0.566	–0.560	–0.368	1.000	0.208	–0.564
Total MINC	0.106	–0.015	0.026	0.248	–0.115	0.018	0.067	–0.283	0.208	1.000	0.058

Table 3. Multiple correlation coefficients *R* for the data from the Opalino-2 borehole

R	T_{max}	S_1	S_2	S_3	PI	PC	RC	HI	OI	Total MINC	TOC
T_{max}	1.000	0.200	0.196	–0.197	–0.580	0.197	0.217	0.285	–0.456	–0.120	0.215
S_1	0.200	1.000	0.947	–0.192	–0.460	0.966	0.912	0.792	–0.504	0.029	0.925
S_2	0.196	0.947	1.000	–0.141	–0.518	0.998	0.985	0.671	–0.441	–0.016	0.990
S_3	–0.197	–0.192	–0.141	1.000	0.097	–0.141	–0.137	–0.177	0.641	0.238	–0.138
PI	–0.580	–0.460	–0.518	0.097	1.000	–0.512	–0.549	–0.486	0.504	0.095	–0.545
PC	0.197	0.966	0.998	–0.141	–0.512	1.000	0.978	0.702	–0.454	–0.002	0.985
RC	0.217	0.912	0.985	–0.137	–0.549	0.978	1.000	0.613	–0.456	0.002	0.999
HI	0.285	0.792	0.671	–0.177	–0.486	0.702	0.613	1.000	–0.537	0.174	0.631
OI	–0.456	–0.504	–0.441	0.641	0.504	–0.454	–0.456	–0.537	1.000	0.089	–0.457
Total MINC	–0.120	0.029	–0.016	0.238	0.095	–0.002	0.002	0.174	0.089	1.000	0.001

Table 4. Multiple correlation coefficients *R* for the data from the Kochnowo-1 borehole

R	T_{max}	S_1	S_2	S_3	PI	PC	RC	HI	OI	Total MINC	TOC
T_{max}	1.000	0.441	0.393	0.016	0.083	0.404	0.417	0.154	–0.529	0.126	0.416
S_1	0.441	1.000	0.968	–0.026	–0.362	0.982	0.961	0.205	–0.449	0.241	0.965
S_2	0.393	0.968	1.000	–0.016	–0.453	0.998	0.978	0.169	–0.389	0.164	0.982
S_3	0.016	–0.026	–0.016	1.000	0.188	–0.010	0.034	–0.154	0.379	0.429	0.029
PI	0.083	–0.362	–0.453	0.188	1.000	–0.425	–0.388	–0.158	0.016	0.068	–0.393
PC	0.404	0.982	0.998	–0.010	–0.425	1.000	0.980	0.184	–0.402	0.188	0.984
RC	0.417	0.961	0.978	0.034	–0.388	0.980	1.000	0.040	–0.397	0.229	1.000
HI	0.154	0.205	0.169	–0.154	–0.158	0.184	0.040	1.000	–0.172	–0.047	0.055
OI	–0.529	–0.449	–0.389	0.379	0.016	–0.402	–0.397	–0.172	1.000	–0.030	–0.398
Total MINC	0.126	0.241	0.164	0.429	0.068	0.188	0.229	–0.047	–0.030	1.000	0.225

for validation, it has been determined that variables S_2, PC and RC are to be removed from the model.

Upon completion of validation, a model described by 8 variables (attributes) X = (TOC, T_{max}, S_1, S_3, PI, HI, OI, Total MINC) was obtained. The next step was the transformation of data in order to obtain, for all attributes, the values from the same range. It has been decided to choose data standardisation. For the data converted in this manner (and for the source data), matrices D' (and D) were constructed for the distances between all pairs of elements (samples) from within the space under consideration. The Euclidean distance was selected, being natural for the considered type of data.

A very important element of conducting data segmentation is the determination of the threshold distance in the data space, d_{max}, below which mutually distant objects would be qualified as part of the same cluster [5]. Selection of too short a distance will result in too high a number of unnaturally generated clusters, and selection of too long a distance will result in qualifying distant and mutually "dissimilar" objects as parts of the same clusters, which will contradict the idea of segmentation. As can be seen, this is a critical element of the whole procedure. Because we are dealing with a space of elements with standardised values, it has been assumed, using an analogy to the statistical data analysis, that the threshold distance would be associated with standard deviation of the sample, calculated from the set of distances generated for all pairs of samples from the space in question. It was decided to conduct segmentation tests for the maximum distance equalling: one ($S(d)$), two ($2S(d)$) and three ($3S(d)$) standard deviations. The results of the calculations are presented in Table 5.

Table 5. Average values and standard deviations for a set of distances for pairs of points in the space of source (d) and standardised (d') data

	Lubocino-1	Opalino-2	Kochanowo-1
d_{sr}	83.800	107.900	115.700
$S(d)$	51.800	64.200	122.300
d'_{sr}	3.812	3.669	3.703
$S(d')$	1.494	1.694	1.713
$2S(d')$	2.988	3.388	3.426
$3S(d')$	4.482	5.082	5.139

d – actual data, d' – standardised data.

Tests of data segmentation were conducted using various variants of the cluster analysis method for the data from each borehole separately, analysing the obtained results.

Results of using cluster analysis for the segmentation of core samples

Results of using cluster analysis for the maximum distance equalling one standard deviation

The following maximum distances have been used:

- for Lubocin-1, d_{max} = 1.494,
- for Opalino-2, d_{max} = 1.696,
- for Kochanowo-1, d_{max} = 1.713.

All the used methods allowed the creation of multiple individual clusters not exceeding three elements. Individual larger clusters have been generated using the following methods: the nearest neighbour method (11÷66 elements), the median method (9÷43 elements), the centre of gravity method (9÷46 elements), the group average method (5÷11 elements). The farthest neighbourhood method, and the Ward method – numerous small clusters (5÷9 elements).

Conclusions:

Too small a threshold distance – numerous individual clusters.

1. The nearest neighbourhood method, the median method and the centre of gravity method worked similarly – quickly generating numerous individual clusters.
2. The Ward method instantly creates numerous small clusters.
3. The threshold distance is to be increased to $2S(d')$, repeating the whole process.

The results of using cluster analysis for the maximum distance equalling two standard deviations

The following maximum distances have been used:

- for Lubocin-1, d_{max} = 2.988,
- for Opalino-2, d_{max} = 3.392,
- for Kochanowo-1, d_{max} = 3.426.

The nearest neighbourhood and centre of gravity methods do not prove valid. They generate one large cluster in the form of a chain. The remaining methods generate groups of larger clusters. The remaining methods work quite well – especially the farthest neighbourhood method and the Ward method. Isolated points are clearly visible, the same, generated by all methods.

Conclusions:

1. The nearest neighbourhood and centre of gravity methods (used for all boreholes) have generated one large cluster comprising almost all the elements plus individual clusters with isolated points.
2. The median and group average methods generated 2÷3 large clusters comprising almost all the elements plus several clusters with isolated points.
3. The farthest neighbourhood and Ward methods began generating several (4÷8) more extensive (4÷20-element) clusters plus numerous clusters with individual isolated points.

4. The threshold distance is to be increased to $3S(d')$, repeating the whole process.

The results of using cluster analysis for the maximum distance equalling three standard deviations

The following maximum distances have been used:

- for Lubocin-1, $d_{max} = 4.482$,
- for Opalino-2, $d_{max} = 5.088$,
- for Kochanowo-1, $d_{max} = 5.138$.

The nearest neighbourhood and centre of gravity methods did not considerably change the results for the increased distance. The farthest neighbourhood method merged two medium-sized clusters into one large cluster and merged several small clusters with larger ones. The median and group average methods generated one large cluster and an isolated point. The Ward method increased the number of small and medium-sized clusters at the expense of the smallest ones. All methods kept the same isolated points.

Conclusions upon tests conducted for all threshold distances:

1. Increasing the distance from two to three standard deviations did not change the effects of using the nearest neighbourhood and centre of gravity methods. The remaining methods allowed the further consolidation of clusters involving the merging of medium-sized clusters into large ones (the median and group average methods) and merging small ones (two-element clusters and isolated points) with larger ones (the nearest neighbourhood and Ward methods).
2. Increasing the distance did not result in the absorption of some isolated points. Due to this, the removal of samples represented by these points from further work on the model is to be considered.
3. The following methods have been qualified as the most useful for the model in question: the farthest neighbourhood method and the Ward method.
4. The distance equalling three standard deviations was chosen as the most advantageous from the standpoint of conducting the segmentation process (for the farthest neighbourhood and Ward methods).
5. The median method, which clearly segments the sets into two large clusters, is worth noting.

The results of cluster analysis for two variants: the farthest neighbourhood method and the Ward method are presented below. Segmentation has been conducted for data from each borehole separately, assuming the maximum distance, $d_{max} = 3S(d')$. The average values for a given attribute within a cluster are presented in Tables 6–8. The red colour indicates high values, blue – low values, violet – close to the average value from the borehole. The maximum and minimum average values are indicated in bold text.

Table 6. Results of cluster analysis for the data from the Lubocino-1 borehole

Farthest neighbourhood method									
Cluster no.	Elementary number	TOC [wt%]	T_{max} [°C]	S_1 [mg HC/g of rock]	S_3 [mg CO_2/g of rock]	PI [-]	HI [mg HC/g TOC]	OI [mg CO_2/g TOC]	Total MINC [wt%]
1	15	0.699	459.200	0.184	0.257	0.206	111.222	43.110	0.441
2	6	**0.378**	467.500	**0.125**	0.153	0.302	**75.253**	**50.397**	0.520
3	6	**5.853**	454.000	2.558	**0.118**	0.211	165.516	**2.065**	**0.735**
4	2	5.850	453.500	**2.905**	**0.400**	0.210	188.402	6.815	0.445
5	8	2.185	440.250	1.780	0.185	0.311	**184.978**	8.410	0.449
Ward method									
Cluster no.	Elementary number	TOC [wt%]	T_{max} [°C]	S_1 [mg HC/g of rock]	S_3 [mg CO_2/g of rock]	PI [-]	HI [mg HC/g TOC]	OI [mg CO_2/g TOC]	Total MINC [wt%]
1	4	1.010	457.000	0.200	0.360	0.218	73.000	36.250	**0.795**
2	2	**0.200**	471.500	**0.055**	0.180	0.250	82.331	**90.476**	**1.115**
3	7	0.681	458.857	0.209	0.231	0.229	102.148	40.550	0.377
4	4	0.420	462.000	0.125	0.200	**0.156**	165.322	54.449	0.200
5	4	0.468	465.500	0.160	0.140	**0.327**	**71.714**	30.357	0.223
6	7	**5.376**	453.571	2.423	**0.120**	0.220	165.946	**2.510**	0.727
7	3	5.157	450.667	**2.973**	**0.400**	0.257	176.935	8.210	0.443
8	6	1.867	437.667	1.587	0.158	0.310	**192.883**	8.517	0.412

Table 7. Results of cluster analysis for the data from the Opalino-2 borehole

Farthest neighbourhood method									
Cluster no.	Elementary number	TOC [wt%]	T_{max} [°C]	S_1 [mg HC/g of rock]	S_3 [mg CO_2/g of rock]	PI [-]	HI [mg HC/g TOC]	OI [mg CO_2/g TOC]	Total MINC [wt%]
1	43	0.547	457.256	0.207	**0.095**	0.297	82.558	28.000	**0.473**
2	10	0.510	460.300	0.197	**0.244**	0.297	90.500	74.900	0.928
3	14	**0.169**	**327.071**	**0.050**	0.182	0.432	**42.214**	**109.214**	0.721
4	18	**3.871**	448.833	**1.902**	0.096	**0.249**	**168.778**	**3.056**	0.504
Ward method									
Cluster no.	Elementary number	TOC [wt%]	T_{max} [°C]	S_1 [mg HC/g of rock]	S_3 [mg CO_2/g of rock]	PI [-]	HI [mg HC/g TOC]	OI [mg CO_2/g TOC]	Total MINC [wt%]
1	16	0.857	450.250	0.425	0.134	0.283	130.000	16.688	0.720
2	12	0.377	459.333	0.091	0.124	0.263	69.667	43.000	0.610
3	6	0.388	465.333	0.133	**0.285**	0.297	85.667	104.333	**0.875**
4	6	0.187	**326.500**	0.052	0.255	0.402	45.333	**141.000**	0.838
5	8	**0.156**	327.500	0.049	0.128	**0.455**	39.875	85.375	0.633
6	4	0.165	503.500	**0.038**	0.150	0.368	38.500	92.000	0.260
7	9	0.658	461.000	0.184	0.039	0.261	73.667	10.444	**0.239**
8	6	0.243	432.333	0.062	0.035	0.408	**34.667**	13.667	0.387
9	7	**6.099**	455.429	**2.281**	0.113	**0.187**	164.857	**1.857**	0.577
10	11	2.454	444.636	1.661	**0.085**	0.289	**171.273**	3.818	0.457

Table 8. Results of cluster analysis for the data from the Kochnowo-1 borehole

Farthest neighbourhood method									
Cluster no.	Elementary number	TOC [wt%]	T_{max} [°C]	S_1 [mg HC/g of rock]	S_3 [mg CO_2/g of rock]	PI [-]	HI [mg HC/g TOC]	OI [mg CO_2/g TOC]	Total MINC [wt%]
1	25	0.858	427.800	0.358	**0.026**	0.344	64.807	8.104	0.390
2	7	**0.113**	**314.714**	**0.027**	0.069	0.372	**45.741**	**64.391**	0.554
3	5	**6.340**	466.400	**2.008**	0.036	**0.258**	**91.903**	**0.503**	**0.388**
4	5	3.376	459.600	1.234	0.068	0.324	85.851	4.740	**1.830**
Ward method									
Cluster no.	Elementary number	TOC [wt%]	T_{max} [°C]	S_1 [mg HC/g of rock]	S_3 [mg CO_2/g of rock]	PI [-]	HI [mg HC/g TOC]	OI [mg CO_2/g TOC]	Total MINC [wt%]
1	13	0.518	475.846	0.185	0.034	0.363	58.829	10.862	0.378
2	5	0.122	**313.200**	0.030	0.058	0.414	**39.037**	48.647	0.572
3	2	**0.090**	318.500	**0.020**	0.095	0.268	62.500	**103.750**	0.510
4	6	0.115	315.667	0.023	**0.008**	0.290	57.209	9.005	**0.338**
5	6	2.340	435.833	1.068	0.025	0.356	85.358	1.225	0.467
6	5	**6.340**	466.400	**2.008**	0.036	**0.258**	**91.903**	**0.503**	0.388
7	5	3.376	459.600	1.234	0.068	0.324	85.851	4.740	**1.830**

Data segmentation for examination results obtained from all boreholes – Stage II

In the next step of using data segmentation, data from eight boreholes have been merged into one set, using the cluster analysis for two variants: the farthest neighbourhood method and the Ward method. Due to the lack of calculation of the

Total MINC parameter for most samples from five new added boreholes, this parameter has been dropped and the analysis has been conducted using seven parameters. Upon completion of numerous tests it was concluded that the best results, due to the optimal number of clusters (10÷20 clusters), have been obtained for the maximum distance of d_{max} = 3.2 (two standard deviations for the set of standardised distances) in the case of the farthest neighbourhood method, and d_{max} = 7 (an order of four standard deviations for a set of standardised distances). Both methods generated four clusters comprising isolated points, which were skipped during further analysis. Clusters comprising less than five elements have also been skipped (three such clusters generated by both methods). The results are presented in the Tables below. In Tables (9–10) the average values have been presented for the given attribute within the cluster. The red colour marks values above average, blue marks values below average, and bold marks maximum values. In Tables 11 and 12 the percentage (weighted by the number of samples from the given borehole) of samples from each borehole in the given cluster has been presented.

Table 9. Results of cluster analysis for the data from eight boreholes in the farthest neighbourhood method version

Cluster no.	Elementary number	TOC [wt%]	T_{max}	S_1	S_3	PI	HI	OI
1	10	0.881	447.600	0.382	0.315	0.276	115.249	40.633
2	17	0.462	464.000	0.121	0.212	**0.214**	103.381	60.183
3	45	0.720	457.667	0.263	0.109	0.273	92.301	19.675
4	26	0.877	466.942	0.379	0.039	0.383	62.331	8.342
5	34	5.341	456.838	2.143	0.104	0.227	143.537	**2.102**
6	8	**6.056**	452.875	**2.678**	**0.353**	0.228	157.232	6.186
7	18	2.087	442.667	1.556	0.114	0.302	**175.174**	5.763
8	8	0.189	**477.875**	0.064	0.181	0.409	52.250	96.000
9	9	0.172	327.111	0.050	0.222	0.402	47.667	**127.333**
10	6	**0.147**	**309.833**	0.042	0.090	**0.499**	**31.202**	59.726
11	10	0.298	329.400	0.131	**0.017**	0.323	58.797	9.631
12	9	0.169	331.889	**0.034**	0.107	0.346	45.774	70.931
13	15	3.740	464.933	1.280	0.113	0.336	69.672	2.952
14	13	2.147	462.615	0.649	0.213	0.397	47.308	11.846

Table 10. Results of cluster analysis for the data from eight boreholes in the Ward method version

Cluster no.	Elementary number	TOC [wt%]	T_{max}	S_1	S_3	PI	HI	OI
1	12	0.596	463.500	0.143	**0.286**	0.244	75.670	61.850
2	9	0.610	460.444	0.161	0.202	**0.177**	133.494	44.214
3	19	0.378	472.316	0.105	0.064	0.356	48.306	21.526
4	19	0.827	451.632	0.358	0.205	0.292	106.240	30.584
5	27	2.657	446.444	1.640	0.120	0.281	**167.161**	4.971
6	20	**6.480**	454.850	**2.345**	0.152	0.192	157.297	2.399
7	10	0.820	447.400	0.460	0.095	0.285	144.832	12.494
8	9	0.189	**474.333**	0.066	0.218	0.383	61.667	118.111
9	9	**0.172**	327.111	0.050	0.222	0.402	47.667	**127.333**
10	16	0.660	460.563	0.172	0.068	0.245	75.625	18.854
11	12	0.176	**324.750**	**0.040**	0.102	**0.439**	**33.432**	58.936
12	10	0.298	329.400	0.131	**0.017**	0.323	58.797	9.631
13	14	1.266	463.536	0.593	0.026	0.390	73.353	2.850
14	8	6.028	462.438	2.158	0.025	0.248	110.978	**0.417**
15	14	4.602	465.071	1.389	0.091	0.325	78.055	2.490
16	18	2.420	462.111	0.786	0.200	0.388	52.000	9.944

Table 11. Results of cluster analysis – percentage of samples from each borehole in the given cluster for the data from eight boreholes in the farthest neighbourhood method version

Cluster no.	Elementary number	L-1	O-2	K-1	G-IG1	O-3	T-1	W-1	Ż-IG1
1	10	28	6	0	0	27	0	0	39
2	17	63	10	0	0	11	0	0	16
3	45	14	41	5	19	11	0	10	0
4	26	3	6	44	22	0	16	0	9
5	34	12	8	9	14	20	3	0	35
6	8	20	0	0	22	39	0	0	19
7	18	51	29	6	0	14	0	0	0
8	8	0	30	0	0	0	0	38	32
9	9	0	68	0	0	0	32	0	0
10	6	0	17	52	0	0	31	0	0
11	10	0	10	71	0	0	19	0	0
12	9	0	13	36	0	0	16	34	0
13	15	0	0	15	0	0	70	15	0
14	13	0	0	0	19	0	23	58	0

Table 12. Results of cluster analysis – percentage of samples from each borehole in the given cluster for the data from eight boreholes in the Ward method version

Cluster no.	Elementary number	L-1	O-2	K-1	G-IG1	O-3	T-1	W-1	Ż-IG1
1	12	65	16	0	0	18	0	0	0
2	9	74	0	0	0	0	0	0	26
3	19	12	18	41	0	0	9	20	0
4	19	17	16	0	28	16	0	0	23
5	27	27	16	3	0	26	0	0	28
6	20	18	9	3	12	28	0	0	30
7	10	0	67	0	0	33	0	0	0
8	9	0	33	0	0	0	0	36	31
9	9	0	68	0	0	0	32	0	0
10	16	0	68	10	0	22	0	0	0
11	12	0	22	45	0	0	33	0	0
12	10	0	10	71	0	0	19	0	0
13	14	0	2	35	33	0	16	0	14
14	8	0	3	28	53	15	0	0	0
15	14	0	0	19	14	0	41	14	12
16	18	0	0	0	16	0	37	47	0

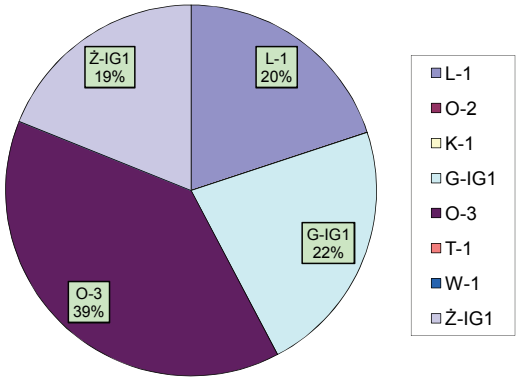
Having analysed the percentages of samples in the individual clusters and the characteristic features of clusters with respect to the average values of parameters, it may be concluded that four cluster groups with the following properties can be distinguished:

1. Clusters with numbers: 10, 11 (from the farthest neighbourhood method) and 11, 12 (from the Ward method) characterised by the maximum and high values of: TOC, T_{max}, S_1, S_3, HI. The following boreholes predominate in those clusters: Lubocino-1, Opalino-2, Żarnowiec-IG1 and Opalino-3.
2. Clusters with numbers: 4, 12, 13 (from the farthest neighbourhood method) and 3, 13, 15 (from the Ward method) characterised by the minimum and low values of: S_3 and HI. The presence of most boreholes is observed in those clusters, with the exception of Opalino-3 and Lubocino-1 (minor contribution).
3. Clusters with numbers: 10, 11 (from the farthest neigh-

bourhood method) and 11, 12 (from the Ward method) characterised by the minimum values of: TOC, T_{max}, HI, S_1, S_3 as well as the maximum and high values of: PI. The following boreholes are predominant in those clusters: Opalino-2, Tępcz-1 and Kochanowo-1.

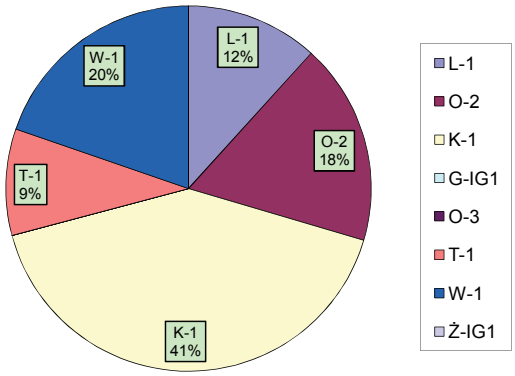
4. Clusters with numbers: 13, 14 (from the farthest neighbourhood method) and 16 (from the Ward method) characterised by high values of: TOC, T_{max} and PI. The following boreholes are predominant in those clusters: Tępcz-1, Wysin-1, Kochanowo-1 and Żarnowiec-IG1.

The percentage of samples from boreholes in selected clusters has been shown in Figures 1–4.



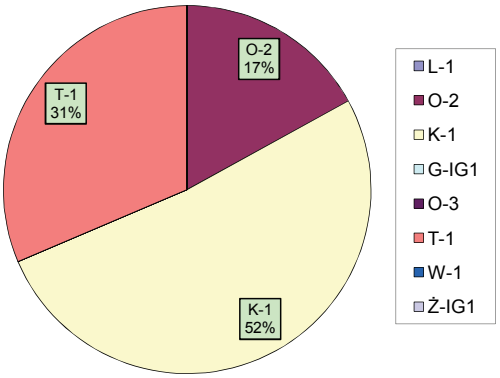
Characteristic features:
– max. value of S_1 = 2.345 [mg HC/g of rock]
– max. value of TOC = 6.480 [wt%]
– high value of HI = 157.297 [mg CO_2/g TOC]

Fig. 1. Percentage of samples in cluster no. 6; number of samples: 8; the farthest neighbourhood method



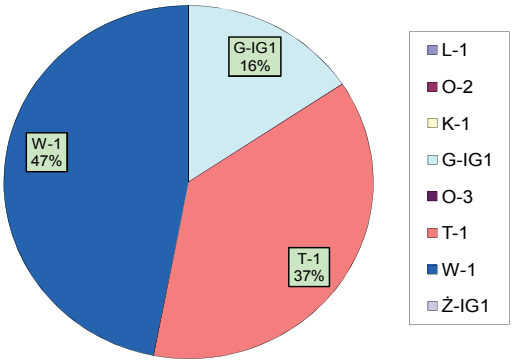
Characteristic features:
– low value of S_3 = 0.064 [mg CO_2/g of rock]
– low value of TOC = 0.378 [wt%]
– low value of HI = 48.306 [mg CO_2/g TOC]
– high value of PI = 0.356 [-]

Fig. 2. Percentage of samples in cluster no. 3; number of samples: 19; the Ward method



Characteristic features:
– min. value of PI = 0.499 [-]
– min. value of T_{max} = 309.833 [°C]
– min. value of TOC = 0.147 [wt%]
– min. value of HI = 31.302 [mg CO_2/g TOC]

Fig. 3. Percentage of samples in cluster no. 10; number of samples: 6; the farthest neighbourhood method



Characteristic features:
– high value of S_1 = 0.064 [mg HC/g of rock]
– high value of TOC = 2.420 [wt%]
– low value of HI = 52.000 [mg CO_2/g TOC]
– high value of PI = 0.388 [-]

Fig. 4. Percentage of samples in cluster no. 16; number of samples: 18; the Ward method

Conclusions

1. The cluster analysis method is an effective tool used to examine the differences and similarities in sets of data describing generative features of the creation of hydrocarbons in shale gas-type formations.
2. The variants of cluster analysis methods: the farthest neighbourhood method and the Ward method are the most efficient with respect to the analysis of datasets describing the geochemical properties of rocks.
3. The conducted analysis indicates that due to the geochemical properties, the rocks penetrated by the Tępcz-1 and Wysin-1 boreholes considerably differ from the remaining examined formations.

Please cite as: Nafta-Gaz 2015, no. 11, pp. 898–909, DOI: 10.18668/NG2015.11.13

Article contributed to the Editor 2.09.2015. Approved for publication 16.10.2015.

The article is the result of research conducted in connection with the project: *Selection of optimal methods for estimation of resources and (geological and commercial) risk at prospecting stage in relation to unconventional "shale gas", "shale oil" and "tight gas" deposits in Poland, and development of methods for documentation of unconventional deposits*, co-funded by the National Centre for Research and Development as part of the programme BLUE GAS – POLISH SHALE GAS. Contract No. BG1/ŁUPZAS/13.

Literature

[1] Duda R. O., Hart P. E.: *Pattern Classification and Scene Analysis*. Wiley, New York, 1973.
[2] Jain A. K., Dubes R.: *Algorithms for Clustering Data*. Prentice Hall, New Jersey, 1988.
[3] Jain A. K., Murty M. N., Flynn P. J.: *Data clustering: a review*. ACM Computing Surveys 1999, 31(3), pp. 264–323.
[4] Klaja J., Lykowska G.: *Wyznaczenie typow petrofizycznych skal czerwonego spagowca z rejonu poludniowo-zachodniej czesci niecki poznanskiej na podstawie analizy statycznej wynikow pomiarow laboratoryjnych*. Nafta-Gaz 2014, no. 11, pp. 757–764.
[5] Mroczkowska-Szerszen M., Ziemianin K., Brzuszek P., Matyasik I., Jankowski L.: *The organic matter type in the shale rock samples assessed by FTIR-ATA analyses*. Nafta-Gaz 2015, no. 6, pp. 361–369.
[6] Yager R. R.: *Intelligent control of the hierarchical agglomerative clustering process*. IEEE Transactions on Systems, Man, Cybernetics – Part B: Cybernetics 2000, 30(6), pp. 835–845.

Bogdan FILAR M.Sc. Eng.
Senior Technical Research Specialist
Head of the Department of Underground Gas Storage
Oil and Gas Institute – National Research Institute
ul. Lubicz 25 A
31-503 Kraków
E-mail: *bogdan.filar@inig.pl*

Dr. Tadeusz KWILOSZ PhD.
Assistant Professor
Department of Underground Gas Storage
Oil and Gas Institute – National Research Institute
ul. Lubicz 25 A
31-503 Kraków
E-mail: *tadeusz.kwilosz@inig.pl*

Mariusz MIZIOŁEK M.Sc.
Senior Technical Research Specialist
Department of Underground Gas Storage
Oil and Gas Institute – National Research Institute
ul. Lubicz 25 A
31-503 Kraków
E-mail: *mariusz.miziolek@inig.pl*

Wacława PIESIK-BUŚ M.Sc. Eng.
Senior Technical Research Specialist
Department of Underground Gas Storage
Oil and Gas Institute – National Research Institute
ul. Lubicz 25 A
31-503 Kraków
E-mail: *piesik@inig.pl*

Jadwiga ZAMOJCIN M.Sc. Eng.
Senior Technical Research Specialist
Department of Underground Gas Storage
Oil and Gas Institute – National Research Institute
ul. Lubicz 25 A
31-503 Kraków
E-mail: *jadwiga.zamojcin@inig.pl*