DOI: 10.18668/NG.2016.11.03

Piotr Łętkowski, Andrzej Gołąbek, Paweł Budak, Tadeusz Szpunar

Instytut Nafty i Gazu – Państwowy Instytut Badawczy

Robert Nowak, Jarosław Arabas

Politechnika Warszawska

Determination of the statistical similarity of the physicochemical measurement data of shale formations based on the methods of cluster analysis

The paper presents the application of the methods of statistical analysis to the determination of the similarity of measurement data, using boreholes providing access to shale formations as an example. The proposed methodology is based on two statistical techniques: the factor analysis and the cluster analysis. The first method allows the reduction of the number of measurements variables in order to eliminate the redundancy of the data. The second one allows grouping the wells on the grounds of factor variables defining the similarity features of the analysed wells. The available results of geochemical measurements for nine wells providing access to shale structures have been used as measurement data.

Key words: shale formations, factor analysis, cluster analysis.

Określenie podobieństwa statystycznego fizykochemicznych danych pomiarowych formacji łupkowych na podstawie metod analizy klastrowej

W artykule przedstawiono zastosowanie metod analizy statystycznej dla określenia podobieństwa danych pomiarowych na przykładzie odwiertów udostępnionych w formacjach łupkowych. Proponowana metodologia zakłada zastosowanie dwóch technik statystycznych: analizy czynnikowej oraz analizy skupień. Pierwsza z metod pozwala na redukcję ilości zmiennych pomiarowych w celu wyeliminowania redundancji danych, natomiast druga umożliwia pogrupowanie odwiertów w oparciu o zmienne czynnikowe definiujące cechy podobieństwa analizowanych odwiertów. Jako dane pomiarowe wykorzystano dostępne wyniki pomiarów geochemicznych dla dziewięciu odwiertów udostępnionych w strukturach łupkowych.

Słowa kluczowe: formacje łupkowe, analiza czynnikowa, analiza klastrowa.

Introduction

One of the basic elements of the identification and examination of hydrocarbon reservoir are the laboratory tests providing the necessary information on the geological structure, mechanical, physicochemical and petrophysical properties of the reservoir rock, the characteristics of reservoir fluids etc. Due to the amount and complexity of the available data, their statistical analysis constitutes a certain challenge. The problem is that the various types of measurement values (measured in various physical units) are correlated in a varying degree, which combined with the number of single calculations for each measurement value generates problems when interpreting and attempting to identify their internal structure. The use of the characteristics of descriptive statistics, i.e. the measure of placement, diversification, concentration or asymmetry seems to be insufficient, since it gives no detailed information about the variability of data following the depth, which constitutes significant information for the construction of geological and simulation models.

Detecting the internal structure of the measurement data acquired for the individual boreholes would allow an attempt at the determination of the features of similarity between the wells. Searching for groups of similar wells has, e.g. certain significance in reservoir simulations, where the possibility to define a structure of the alternating layers of a simulative model allows the determination of the possible directions of the flow of reservoir fluids. Another possible application is the use of similarity features to predict the extraction from wells which provide access to shale formations in a situation of the lack of production test data.

The goal of the conducted analysis was an attempt to isolate groups of similar wells based on the available measurement data. The proposed method of determining the degree of the similarity of wells based on the measurement data is based on two data exploration techniques: the factor analysis and the cluster analysis. The first one allows the reduction of the number of measurement variables in order to remove the redundancy of data; the second one allows the grouping of wells with respect to the variables resulting from the factor analysis.

The cluster analysis is a method allowing the detection of the structure of data by grouping then into clusters, meaning datasets with similar values. Depending on the used version of the method, it is possible to select the number of groups into which the data is to be divided (using the *k*-means method), or the so-called agglomeration, meaning the gradual merging of the grouped values according to the selected method of merging and the manner of measuring the distance. This method is usually used to group a set of measurements. There are however no obstacles against using it to group measurement vectors in a multi-dimensional space. However, this requires the preceding determination of these vectors in such a way that the resulting vector would contain as much information about the measurement values as possible.

The use of the proposed method of the categorisation of wells based on statistical similarity requires the implementation of the following stages:

- data "screening",
- the elimination of correlated variables,
- the analysis of factor variables,
- grouping the factor variables into equinumerous datasets,
- cluster analysis.

The paper uses the geochemical data available for nine boreholes providing access to the shale formation as the original measurement data.

The cluster analysis method

The term cluster analysis covers several different algorithms of the classification of datasets. In general, it could be said that its purpose is to group the elements of a dataset into reasonable, relatively uniform groups (classes, clusters). The basis for the grouping in the used algorithms is the similarity between the elements of a dataset, expressed as the function of similarity (the metric of distance). The methods used in the paper to assess the similarity of data (wells), which are the hierarchical agglomeration method and grouping using the *k*-means method, are presented below.

The hierarchical agglomeration method

The hierarchical agglomeration method allows the determination of the so-called hierarchical tree plot of the elements of the analysed dataset. The tree of connections is obtained as a result of the use of the algorithm of the progressive agglomeration (meaning merging into subsets) of the subsequent datasets. In the first step of the procedure it is assumed that each case (measurement) constitutes a separate subgroup. Subsequently, a distance matrix is calculated (with a dimension of N, where N is the number of measurements), in which the smallest element is sought outside of the main diagonal. This distance (called the agglomeration distance) is locally minimal, since in each step of iteration it has a different value. After merging the subgroups for which the agglomeration distance is minimal, the distance matrix is prepared again, its maximum dimension now amounting to N-1. The presented procedure is repeated until obtaining a single cluster comprising all N measurement points.

A characteristic feature of the method is that it does not require the a-priori assessment of the number of groups, but only the criterion of stopping the agglomeration procedure. However, it is necessary to select the manner of measuring the distance between groups (clusters) and the methods of their merger, meaning the agglomeration. The most frequently used measures of distance (metrics, e.g. euclidean distance, squared euclidean distance, city block distance, exponential distance) and agglomeration methods (e.g. single linkage method, ward's method, complete linkage method) are presented below.

The measures of distance (metrics)

Euclidean distance. Probably the most frequently selected measure of distance. Determined as a geometric distance in a multidimensional space. It should be noted that if Euclidean distances are calculated based on raw (not normalised) data, then the distance between any two measurements is not affected by adding new measurements to the measurement variable. However, because the differences in measurements between the measurement variables considerably affect Euclidean distances, it is recommended to conduct preliminary standardisation of variables in order to obtain variables of comparable scales.

Squared Euclidean distance. A modification of the Euclidean distance which involves squaring it in order to assign greater weight to more distant objects.

City block (Manhattan) distance. A measure of distance calculated as the sum of differences measured along the dimensions. In most cases it provides similar results to the Euclidean distance. Let us note, however, that in the case of this measure, the impact of individual great differences (e.g. outliers) is suppressed, since they are not squared.

Chebyshev distance. The measure of distance used in a situation when we wish to define two measurements as "different" when they differ in one dimension.

Exponential distance. A measure being the generalisation of the Euclidean distance. It allows the control of the weights assigned to measurement variables or measurements by using two real parametres.

The methods of agglomeration

The single linkage method (nearest neighbourhood). In this method, the distance between two clusters is defined by the distance between two closest measurements (in the sense of the adopted metric) belonging to different clusters. According to this rule, the measurements form clusters by merging into series, and the resulting clusters create long "chains".

The complete linkage method (farthest neighbourhood). The method defines the distance between clusters as the greatest distance between any two measurements belonging to various clusters (i.e. "the farthest neighbours"). The method works in situations where objects form naturally separated clusters.

Unweighted pair-group average. In this method the distance between two clusters is calculated as an average distance between all pairs of objects belonging to two different clusters. The method is effective both when the measurements form natural clusters as well as in a situation where they exhibit the nature of "chains".

The weighted par-group average. This method is a modification of the group average method, involving the introduction of weights in the form of the number of measurements. It is used in situations when we want to avoid the extensive diversification of the number of measurements in the individual clusters.

The unweighted par-group centroid. A method based on the definition of the centre of gravity of a cluster in a multidimensional space, defined by measurement variables. In this method, the distance between two clusters is defined as the difference between the centres of gravity.

The weighed par-group centroid (median). This method is the equivalent of the weighted average method for the centre of gravity method. It involves weighing the distance by the number of measurements belonging to the corresponding clusters.

The Ward's method. A method using a variational approach to assess the distance between clusters, involving the minimisation of the sum of squared deviations of the sum of two clusters, which may be formed at each stage. Although deemed very effective, this method tends to create small-sized clusters.

A characteristic feature of hierarchical methods is that once a decision to merge two clusters (subgroups) is made, it cannot be changed until the end of the agglomeration procedure being in effect. It seems that in a case where we are dealing with organised datasets (the measurements of hydrocarbon reservoir parameters as a function of depth), such solution seems natural.

Grouping using the *k*-means method

One independent dataset grouping method is the *k*-means method, whose characteristic feature is the a priori assumption of the number of clusters. Let us assume that their number has been assessed in some manner. We now want to divide the dataset into a predetermined number of groups (clusters) which will be as different as possible. The action of the algorithm begins by picking two random *k* clusters, followed by moving the grouped objects between these clusters in order to (1) minimise the variability inside the clusters and (2) maximise the variability between the clusters. In other words, our goal is to obtain the maximum similarity in the group, accompanied by a maximum diversification between the groups.

Data similarity analysis

Due to the results of laboratory examinations, the proposed procedure of grouping borehole data is based on statistical data exploration techniques and consists of the following elements:

- data "screening",
- the elimination of dependent variables (correlated),
- factor analysis,
- agglomerative analysis and grouping by means of the *k*-means method.

The first stage is the screening of input data, involving the selection of wells for which the same set of data is available, and the rejection of incomplete or uncertain data. As a result of the analysis of the available data, a set of geochemical data was prepared for nine boreholes providing access to shale structures (Silurian–Llandovery). The analysis was conducted based on the following measurements (measurement variables):

TOC – total organic carbon content [wt%],
T_{max} – the temperature at which the maximum amount of hydrocarbons is created during cracking of kerogen [C],
S1 – the amount of free hydrocarbons content present in a rock sample [mg HC/g of rock],
S2 – the amount of hydrocarbons released during the original cracking of kerogen [mg HC/g of rock],
S3 – the amount of carbon dioxide released during the destruction of organic substance (mg CO_2/g of rock),
PI – the so-called 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].

The wells for which the above-mentioned data have been compiled are: L-1, O-2, O-3, B-1, W-1, K-1, T-1, G-1 and Z-1.

For the purposes of further analysis, correlations between the individual variables were examined for each well, which allowed the elimination of dependent (correlated) variables from the dataset. To this end, TOC was adopted as a reference variable, and subsequently those variables for which the coefficient of correlation with TOC was higher than 0.85 were removed from further analysis. A sample Pearson correlation matrix is presented in Table 1 (variables eliminated from further analysis are marked red).

The presented operation allowed the reduction of the number of measurement variables from 10 to 6, with a relatively small loss of information regarding their variability. Ultimately, further analysis was based on six measurement variables: TOC, T_{max}, S3, PI, HI, OI.

At this stage of the analysis it is necessary to perform the standardisation of data, involving the introduction of all measurement data into a space of variables in which the distance between measurements does not depend on the coordinates.

The standardisation was performed in accordance with the following formula:

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

where:

x_i' – standardised measurement variable,
$\overline{x}$ – the average value of all data for the given measurement variable,
S_x – the standard deviation of all data for the given measurement variable.

Further reduction of the number of measurement variables was conducted based on the so-called factor analysis, involving the replacement of measurement variables with the so-called factor variables. The idea of factor analysis is based on the assumption that the variables measured directly may be expressed as linear combinations of unobservable variables

Table 1. The matrix of correlation coefficients for measurement variables – the O-2 borehole

Variable	T_{max}	S1	S2	S3	PI	PC	RC	TOC	HI	OI
T_{max}	1.00	0.50	0.44	0.10	–0.47	0.46	0.43	0.43	0.16	–0.47
S1	0.50	1.00	0.94	–0.02	–0.70	0.96	0.87	0.88	0.14	–0.67
S2	0.44	0.94	1.00	0.01	–0.79	1.00	0.97	0.98	0.04	–0.56
S3	0.10	–0.02	0.01	1.00	0.10	0.03	0.02	0.02	–0.18	0.20
PI	–0.47	–0.70	–0.79	0.10	1.00	–0.78	–0.78	–0.78	–0.02	0.51
PC	0.46	0.96	1.00	0.03	–0.78	1.00	0.96	0.97	0.05	–0.58
RC	0.43	0.87	0.97	0.02	–0.78	0.96	1.00	1.00	–0.13	–0.52
TOC	0.43	0.88	0.98	0.02	–0.78	0.97	1.00	1.00	–0.11	–0.53
HI	0.16	0.14	0.04	–0.18	–0.02	0.05	–0.13	–0.11	1.00	–0.24
OI	–0.47	–0.67	–0.56	0.20	0.51	–0.58	–0.52	–0.53	–0.24	1.00

known as factor variables, the number of unobservable variables not having to be equal to the number of measurement variables. The factor analysis involves finding the factor variables based on the introduced measurement variables. Each factor variable combines within itself the information concerning several measurement variables, which considering the lack of correlation between them allows the significant reduction of the number of variables.

Depending on the predetermined number of factor variables, it is possible to retain the input amount of information on the variability of the input data. We have at our disposal three criteria assessing the number of factor variables taken into account: The Kaiser criterion, the Cattel criterion and the percentage criterion. According to the Kaiser criterion, one should take into account those principal components (factor variables) which have their eigenvalues higher than 1. On the other hand, the Cattel criterion (the so-called scree test) recommends finding such a point on the plot of eigenvalues, to the right of which a gentle drop in the eigenvsalues takes place. The use of the percentage criterion means leaving as many factor variables as it is necessary to explain the arbitrarily selected percentage of the variances of primary variables. The problem is that when using the first two criteria, it is not uncommon to obtain various numbers of factor variables, or for their number to not guarantee retaining a sufficient amount of information about data variability. Moreover, in the case of the Cattel criterion we might get a chart for which it is impossible to unambiguously determine the threshold point.

For the reasons stated above, the selection of the number of factor variables was conducted based on the Kaiser criterion and the percentage criterion. Table 2 presents the eigenvalues as well as the cumulated and individual (values in brackets) percentages of the variance taken into account for the subsequent factor variables. Adopting the Kaiser criterion, two factor variables would be taken into account for the B-1, O-2, T-1, W-1 and Z-1 wells. For the remaining wells (L-1, G-1, K-1, O-3) the criterion indicates three factor variables. Because for the needs of further analysis it is necessary to assess the same number of variables for each well, the decisive role will be played by the percentage criterion. Adopting two factor variables means taking into account between 72% (the T-1 well) and 80% (B-1, Z-1) of the information on the variability of the data, while adopting three variables retains between 74% (the K-1 well) and over 92% (Z-1) of information. Because including the third variable causes a considerable increase in the amount of information taken into account (an average of 69% to 85%), three factor variables were adopted for further analysis. Such a solution is also supported by the fact that for the O-2, T-1 and W-1 wells the third eigenvalue is very close to one, which results in adopting three factor variables according to the Kaiser criterion.

For further analysis, it is necessary to ensure the equinumerosity of datasets for the analysed factor variables, which would allow the examination of the similarity of data variability (factor variables) over the whole measurement range. To this end, it is necessary to normalise the measurement

Table 2. Eigenvalues, the number of factor variables, the cumulated and individual percentage of the explained variance

Borehole	Eigenvalues	The number of factor variables			
		1	2	3	4
B-1	3.83/1.02/0.65	63.885%	80.81% (16.93%)	91.63% (10.82%)	95.91% (4.28%)
L-1	2.52/1.43/1.02	42.00%	65.84% (23.84%)	82.90% (17.06%)	91.59% (8.69%)
G-1	3.07/1.34/1.00	51.19%	73.61% (22.42%)	90.25% (16.64%)	96.19% (5.94%)
K-1	2.09/1.30/1.08	34.89%	56.52% (21.63%)	74.51% (17.99%)	88.69% (14.18%)
O-2	2.87/1.18/0.94	47.86%	67.55% (19.69%)	83.28% (15.73%)	91.08% (7.8%)
O-3	2.16/1.28/1.08	36.05%	57.45% (21.40%)	75.38% (17.93%)	88.59% (13.21%)
T-1	3.1/1.24/0.96	52.66%	72.32% (19.66%)	88.38% (16.06%)	94.18% (5.8%)
W-1	2.91/1.21.0.98	48.51%	68.61% (20.10%)	84.89% (16.28%)	94.81% (9.92%)
Z-1	3.40/1.45/0.68	56.67%	80.77% (24.10%)	92.15% (11.38%)	97.50% (5.35%)

depths, adopt the same number of measurement points for each well and interpolate the factor variables for new normalised measurement depths. In a situation where the boreholes differ considerably in the number of measurement points this may distort the results for wells which have considerably lower numbers of original measurements.

The next stage is conducting a cluster analysis for the examined wells. To this end, the *k*-means method and the hierarchical agglomeration method were used, adopting the determined factor variables as the grouped values. The analysis was conducted independently for each factor variable. Following the analysis conducted using the *k*-means method, the results which are presented in Table 3 were obtained for 3 clusters.

Table 3. The grouping of wells using the *k*-means method

	Cluster 1	Cluster 2*	Cluster 3*
Variable 1	O-3 Z-1	B-1 L-1 G-1 O-2 W-1	K-1 T-1
Variable 2	B-1 L-1 O-3	G-1 K-1 O-2 T-1 Z-1	W-1
Variable 3	O-3 T-1 W-1 Z-1	L-1 G-1 K-1	B-1 O-2

* PLEASE NOTE: Assuming 2 clusters in the method resulted in merging clusters 2 and 3 into one group

Considering the definition of factor variables and the percentage of variance explained by each factor variable (Table 2), the obtained results may be interpreted in the following manner. The clusters obtained for the first variable result from taking into account the largest part of information involving the variability of data in the wells. Depending on the well, in the analysed case this means taking into account between over 34% (the K-1 well) and almost 64% (the B-1 well) of information about the variability of data. The information not taken into account by the first factor variable is taken into account by the second one within a range from almost 17% (the B-1 well) to 24% (the Z-1 well). The remaining factor variables form clusters depicting similarities of higher orders not taken into account so far. However, this means that for the subsequent variables we would get an independent division, i.e. there are no identical clusters for any two factor variables. This is a consequence of the fact that these variables are linearly independent and each of them, being a linear combination of the original measurement variables, reflects their variability to a varying extent.

In order to verify the obtained results, an agglomeration analysis was also conducted for various combinations of distance metrics and agglomerations. Sample results are presented in Figures 1–6. These figures present graphically the manner of merging wells depending on the distance between them in terms of the selected method of agglomeration and metrics (linkage distance). For example, in Figure 1 for a linkage distance equaling 6.5 each well constitutes a separate group, which means that the distance between any two wells is greater than 6.5. If we increase this distance to 7.0, the B-1 and O-2 wells will be merged into one group. This means that the distance between them ranges between 6.5 and 7.0 (the precise value of the distance amounts to 6.6). By gradually increasing the threshold value of distance we get decreasing numbers of well groups. Therefore, by adopting various threshold values of distance we may perform divisions into 2, 3, 4 and a higher number of groups, e.g. for a linkage distance of 9.5 we get two well groups. One of them consists of the O-3 and Z-1 wells and the other one of all the remaining wells.

The presented manner of merging similar wells allows the assessment of the sequence of the merging (agglomeration) of wells with respect to the similarity of the variability of data. Figure 7 presents the sequence for the second factor variable using the Euclidean norm and the centre of gravity method. Following the adopted assumptions, G-1 and Z-1 turn out to be the most similar wells. Subsequent mergers were performed for the following wells: K-1, O-2, T-1, etc. However, it should be noted that the sequence of agglomeration may depend on the adopted methods of measuring the distance between the variables and their merger.

Figures 1–2 present the results for the first factor variable and the Euclidean norm after adopting the group average method and the Ward method, respectively. Although while dividing into two clusters in both cases we get a division identical to the *k*-clusters method (Table 3), we do notice certain differences when dividing into 3 groups. In each case the O-3 and Z-1 wells constitute one group. However, the division of the remaining wells differs depending on the method of the analysis, with varying results also obtained during the agglomeration analysis. The difference involves the L-1 well, which in the case of the group average method constitutes one group with the B-1, O-2 and W-1 wells, and one group with G-1, K-1 and T-1 in the Ward method. Such differences are missing in the case of the second (Figures 3–4) and third factor variable (Figures 5–6), where each method results in an identical division for both two and three clusters.

The analysis of the similarity of wells was conducted based on the results of agglomeration for the Ward method and the Euclidean norm (see Figures 2, 4, 6). For the first factor vari-

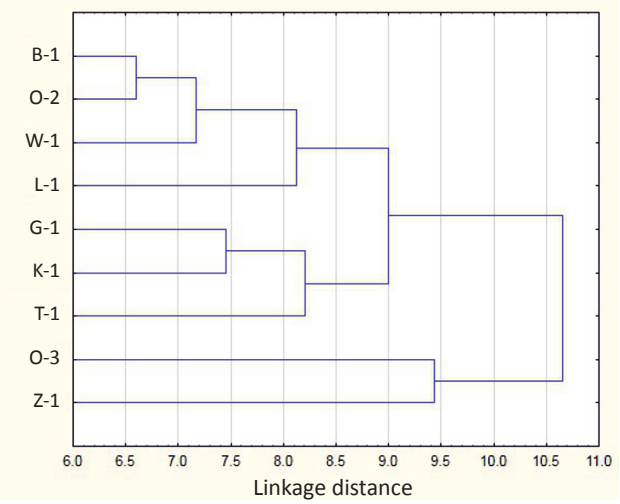

Fig. 1. The result of agglomeration: first factor variable, group average method, Euclidean norm

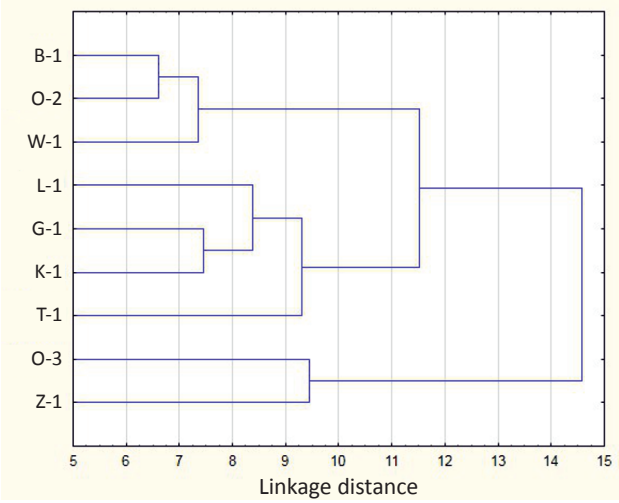

Fig. 2. The result of agglomeration: first factor variable, Ward method, Euclidean norm

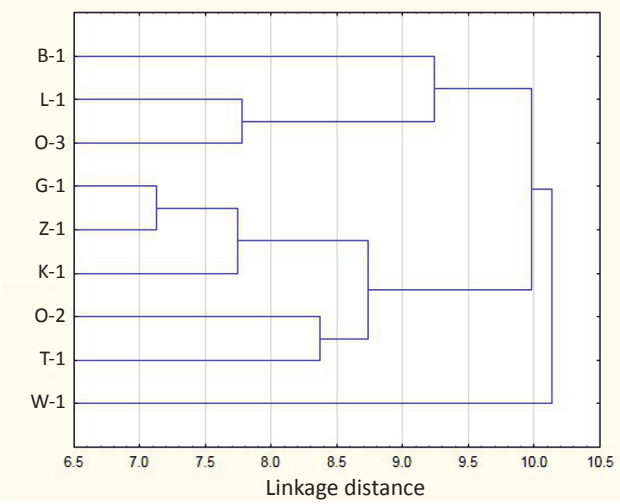

Fig. 3. The result of agglomeration: second factor variable, group average method, Euclidean norm

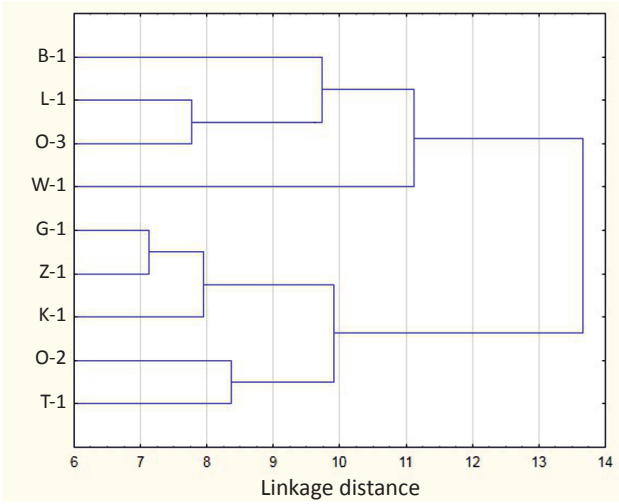

Fig. 4. The result of agglomeration: second factor variable, Ward method, Euclidean norm

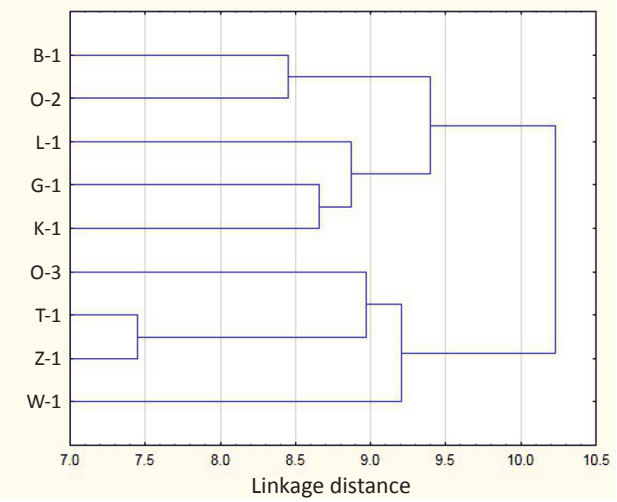

Fig. 5. The result of agglomeration: third factor variable, group average method, Euclidean norm

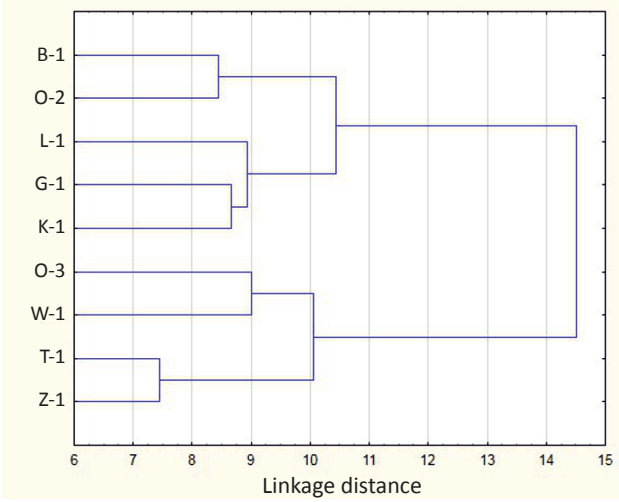

Fig. 6. The result of agglomeration: third factor variable, Ward method, Euclidean norm

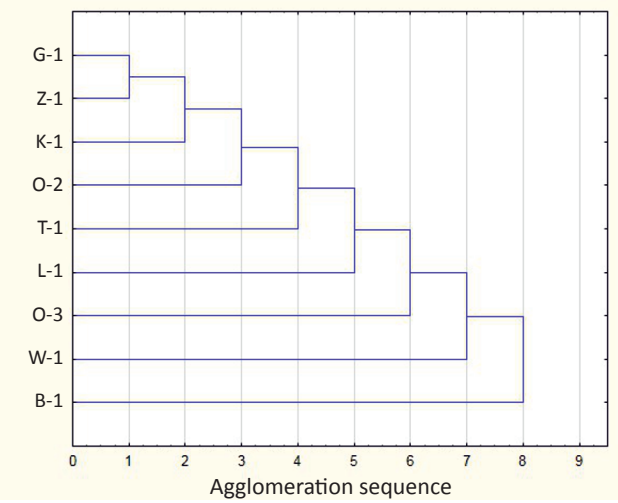

Fig. 7. The sequence of agglomeration: second factor variable, centre of gravity method, Euclidean norm

able (Figure 2), we get an irregular division into two groups. In the first group there are the following wells: B-1, O-2, W-1, L-1, G-1, K-1 and T-1, while the O-3 and Z-1 wells are in the second group. Although the O-3 and Z-1 wells are in one group, the linkage distance between them is relatively large and is only slightly greater than the linkage distance for the T-1 well from the first group, which evidences a considerable difference between these wells. For the second factor variable (Figure 4), the wells form more uniform groups: the first group consists of 4 wells (B-2, L-1, O-3, W-1), while the second group consists of 5 (G-1, Z-1, K-1, O-2, T-1). As can be seen, taking into account the second factor variable caused the disclosure of differences between wells both for the first and the second group for the first factor variable. A division resulting from taking the third variable into account exhibits a certain similarity to the division obtained for the first factor variable. The difference involves the fact that the group which, for the first factor variable, consisted of the O-3 and Z-1 wells, has now been complemented by the W-1 and T-1 wells.

Limiting the similarity analysis to the first factor variable, taking into account the majority of information about the variability of parameters, for splitting into three groups of wells we get a division presented in Table 3. According to Figure 2, the most similar wells are B-1 and O-2 (linage distance 6.6), which form a group with the W-1 well. The G-1 and K-1 wells also exhibit a great similarity (linkage distance 7.5), forming a group with the L-1 and T-1 wells. However, T-1 visibly stands out from its group (the linkage distance from the group equals 9.45).

Summary

The paper presents an attempt at the categorisation of wells which provide access to shale formations (Silurian–Llandovery) with respect to the statistical similarity of laboratory data. The suggested method is based on two data exploration techniques: the factor analysis and the cluster analysis, for which two methods were used: the *k*-means method and the hierarchical agglomeration method. The conducted statistical analysis authorises us to formulate the following comments and conclusions:

- The methods of cluster analysis may be used in order to categorise the wells based on the measurement data. However, in order to guarantee the unambiguity of results, it is necessary to use methods for the reduction of the number of measurement variables.
- When using the presented methods, the selection of measurement data constituting a basis for the conducted analysis may considerably affect the results. The paper uses biochemical data due to its availability over a relatively wide range of measurement variables and of the numbers of individual measurements.
- The effective reduction of measurement variables may be conducted using a two-stage process utilising the correlation analysis and the factor analysis. The used combination of the above-mentioned methods enabled an over threefold reduction of the number of variables, while retaining control over the amount of lost information.
- Due to their definition, the cluster analysis based on factor variables makes it possible to perform independent divisions with respect to the given factor variable. However, it does not enable the gradual particularisation of the categories when taking into account the subsequent variables. However, as part of the analysis of one variable, it is possible to assess the degree of similarity between the wells.
- The conducted comparative analysis indicated sufficient coherence of the obtained results depending on the used method of cluster analysis (the *k*-means method, the hierarchical agglomeration method) and the number of groups.
- The results of agglomeration may depend on the adopted merging method and the manner of measuring the distance, however for the Euclidean norm and the two most effective merging methods, i.e. the group average method and the Ward method, the obtained results turned out to be almost identical to the results of the *k*-means method.
- A certain inconvenience of using the methods of cluster analysis to categorise the wells in a manner presented in the paper is the necessity to operate on equinumerous datasets for various wells. In a situation where the numbers

of cases (measurements in the available interval) differ considerably, a distortion of results may take place for the wells whose numbers of measurements are significantly smaller than the average.

Please city as: Nafta-Gaz 2016, no. 11, pp. 910–918, DOI: 10.18668/NG.2016.11.03

Article contributed to the Editor 1.04.2016. Approved for publication 21.07.2016.

The paper is the result of research conducted as part of the project: *Methodology of determining sweet spots on the basis of geochemical, petrophysical and geomechanical properties, based on the correlation of the results of laboratory examinations with the geophysical measurements and the 3D generational model. Subject 8: the construction of a computer system for storing and processing information involving shale formations as the potential deposits of unconventional hydrocarbons* – co-funded by the National Centre for Research and Development as part of the BLUE GAS – POLISH SHALE GAS programme. Contract No. BG1/MWSSSG/13.

Literature

[1] Budak P., Łętkowski P., Szpunar T., Nowak R., Radzikowski K., Arabas J.: *SweetSpot – a computer system for storing and sharing data on the rock properties of shale formations*. Nafta-Gaz 2015, no. 12, pp. 944–952.

[2] Filar B., Kwilosz T., Miziołek M., Piesik-Buś W., Zamojcin J.: *The use of cluster analysis for the segmentation of the physicochemical properties of shale gas deposits*. Nafta-Gaz 2015, no. 11, pp. 898–909, DOI: 10.18668/NG2015.11.13.

[3] Internetowy podręcznik statystyki; www.statsoft.pl/textbook/stathome.html (acces: 28.03.2016).

[4] Jain A. K., Murty M. N., Flynn P. J.: *Data clustering: a review*. ACM Computing Surveys 1999, vol. 31(3), pp. 264–323.

[5] Klaja J., Łykowska G.: *Wyznaczenie typów petrofizycznych skał czerwonego spągowca z rejonu południowo-zachodniej części niecki poznańskiej na podstawie analizy statystycznej wyników pomiarów laboratoryjnych*. Nafta-Gaz 2014, no. 11, pp. 757–764.

Dr. Eng. Piotr ŁĘTKOWSKI PhD.
Assistant Professor
Department of Hydrocarbon Deposits and UGS Facilities Simulation
Oil and Gas Institute – National Research Institute
ul. Lubicz 25 A, 31-503 Kraków
E-mail: *piotr.letkowski@inig.pl*

Andrzej GOŁĄBEK M.Sc. Eng.
Junior Scientist
Department of Hydrocarbon Deposits and UGS Facilities Simulation
Oil and Gas Institute – National Research Institute
ul. Lubicz 25 A, 31-503 Kraków
E-mail: *andrzej.golabek@inig.pl*

Paweł BUDAK M.Sc. Eng.
Senior Science and Research Specialist
Head of the Department of Petroleum Engineering
Oil and Gas Institute – National Research Institute
ul. Lubicz 25 A
31-503 Kraków
E-mail: *pawel.budak@inig.pl*

Dr. Eng. Tadeusz SZPUNAR PhD.
Assistant Professor
Department of Petroleum Engineering
Oil and Gas Institute – National Research Institute
ul. Lubicz 25 A
31-503 Kraków
E-mail: *tadeusz.szpunar@inig.pl*

Dr hab. Eng. Robert NOWAK
Lecturer/Assistant Professor
Warsaw University of Technology
Department of Electronics and Information Technology
Institute of Electronic Systems
ul. Nowowiejska 15/19, 000-665 Warszawa
E-mail: *R.M.Nowak@elka.pw.edu.pl*

Prof. dr hab. Eng. Jarosław ARABAS
Associate Professor
Warsaw University of Technology
Department of Electronics and Information Technology
Institute of Electronic Systems
ul. Nowowiejska 15/19, 00-665 Warszawa
E-mail: *jarabas@ise.pw.edu.pl*