Cluster analysis of core measurements using heterogeneous data sources: an application to complex Miocene reservoirs

N. P. Szabó[1,2*], K. Nehéz[1,3], O. Hornyák[1,3], I. Piller[1,3], Cs. Deák[1], P. P. Hanzelik[4], Cs. Kutasi[4], K. Ott[4]

[1]University of Miskolc, 3515 Miskolc-Egyetemváros, Hungary

[2]MTA-ME Geoengineering Research Group, University of Miskolc, 3515 Miskolc-Egyetemváros, Hungary

[3]HEIC, Higher Education Industrial Cooperation Centre, University of Miskolc, 3515 Miskolc-Egyetemváros, Hungary

[4]MOL Group, 1117 Október huszonharmadika Street 18, Budapest, Hungary

[*]corresponding author, Department of Geophysics, University of Miskolc, 3515 Miskolc-Egyetemváros, Hungary, e-mail: norbert.szabo.phd@gmail.com

Abstract

An integrated clustering approach is suggested for the interpretation of petrophysical properties measured on core samples. Porosity, carbonate content, bulk and grain density, permeability, irreducible water saturation and capillary pressure curves measured from different sources are processed simultaneously for a more reliable evaluation of hydrocarbon reservoirs. The specialty of the problem is that the input dataset is composed of observations at different boreholes and depth intervals, where the number of rock specimens and measurement types strongly vary. Several statistical methods such as cluster and factor analyses are traditionally used for rock typing, which are either highly limited or not feasible at all for the processing of such incomplete datasets. To overcome this difficulty, the sparse matrix of multivariate observations is fully filled with reliable estimates of petrophysical parameters before cluster analysis. In the first phase, a correlation-based interpolation method is proposed for replacing the missing data with synthetic ones estimated from the available petrophysical information. Then, principal component analysis of the filled data matrix is performed to investigate the relative contribution of different variables to the solution and reduce the large number of measured parameters

into fewer variables. In the last step, a non-hierarchical cluster analysis of the principal components is made to separate the lithological units and reservoir zones. The suggested statistical workflow is tested in a Hungarian oilfield, where Miocene reservoirs of different lithologies are evaluated using a large amount of laboratory data collected for three decades. When clustering all petrophysical variables in a joint procedure, not just the lithological properties but also the fluid and other reservoir characteristics are taken into account to differentiate the pay zones from unproductive intervals. In addition to the current application, the statistical method may serve as useful tool for improved well log analysis, well-to-well correlation and reservoir modeling on larger scales.

Keywords K-means clustering; petrophysical properties; sparse matrix; core measurement; Miocene reservoir; Hungarian oilfield

1. Introduction

The role of multivariate statistical approaches has been continuously increasing with the growing amount of observations and highly developing computational resources in petroleum geosciences. Hempkins (1978) suggests the practical use of multiple regression, cluster and principal component analysis as powerful tools in formation evaluation. For today, these techniques became as routinely applied methods in petrophysics, used primarily for the identification of hydrocarbon reservoirs, facies analysis, determination of rock physical relations, trends and correlation analyses, and for the replacement of missing data along the boreholes. Principal component analysis and its extended variant called factor analysis were originally developed to reduce the dimension of the data space defined by the measurement variables (Lawley and Maxwell, 1962), the geological application of which makes it possible to emphasize the common geological/geophysical information in measured dataset and explore hidden variables dependent on lithological and rock physical characteristics that cannot be measured directly with geophysical instruments. A modern application of factor analysis can be found in Jarzyna et al. (2017), where the heterogeneity of Paleozoic shale gas formations and the complex geophysical responses of organic-rich reservoirs were studied. The robust forms of the above exploratory statistical

methods make a significant improvement in the quality of parameter estimation. Szabó and Dobróka (2017) processed oilfield well logs by iteratively reweighted factor analysis to give an outlier-free estimate to the shale volume of clastic formations. Permeability as a related quantity was also predicted by the same methodology using an evolutionary computation-based optimization approach (Szabó and Dobróka, 2018).

Cluster analysis being an effective rock typing tool has a wide literature record. As an advanced example, Skalinsky et al. (2006) suggests the clustering of mercury injection and well-logging data to predict the types of carbonate rocks. Beyond its capabilities with stratigraphic evaluation of rock formations, cluster analysis can provide the inversion of well logging data with quantitative information such as matrix and fluid parameters of the separated lithological units (Szabó et al., 2013). A clustering-based uncertainty assessment of porosity estimation is proposed by Masoudi et al. (2018), where the accuracy information of the former parameter can be extended to predict irreducible water saturation and permeability by means of fuzzy arithmetic. In the petrophysical characterization of unconventional reservoirs, cluster analysis seems to be also an inescapable tool (Ma and Holditch, 2016). Moreover, soft computing methods including self-organizing maps, neural networks, machine learning techniques and other artificial intelligence approaches will be more and more important in the near future of hydrocarbon exploration (Cranganu et al., 2015).

However, the performance of clustering spectacularly decreases in big and incomplete datasets. Classical non-hierarchical clustering algorithms are known to be very sensitive to the initial setting of cluster centers, which may cause interpretation problems in the lack of a priori geological information. On the other hand, traditional methods would either drop the data columns of the missing data, or fill the missing data with zero values. The former approach dramatically reduces the amount of data involved in clustering, the latter may introduce false data. Thus, neither of the above methods provide feasible solution in petrophysical applications. Considering the aforementioned level of incompleteness, it is expedient to estimate missing data before applying data mining techniques.

Several matrix completion and imputation algorithms exist for handling missing data. One of the simplest approaches replaces the missing entries with the mean (or median) of each column of the input data matrix (Little, 1986). As another alternative, the nearest neighbour algorithm weights the sample elements using the mean squared difference of columns for those two rows where there are observed data. In our application, there are huge blocks of missing sub-matrices because of the separated groups of unknowns that are available in different wells. In this case, the traditional methods do not work properly. Recently, newer soft-impute, nuclear norm minimization and classical direct matrix factorization algorithms have been introduced to carry out the completion of missing values. The soft-impute method which iteratively replaces the missing elements with those obtained from a soft-thresholded singular value decomposition did not yield satisfactory results for our problem (Mazumder et al., 2010). Nuclear norm minimization offers an exact matrix completion via convex optimization (Candès et al., 2009). The direct matrix factorization method transforms the incomplete matrix into an *M*-by-*N* matrix with an L_1 sparsity penalty on the elements of the *M*-th row and an L_2 penalty on the elements of the *N*-th column using gradient descent algorithm (Gemulla et al., 2011). By using the above methods, serious problems were raised in the physical interpretation of completed data matrices. Our preliminary investigations showed the application of the above mentioned interpolation algorithms results in a matrix with wrong data, for example negative or outlying values were frequently obtained being out of the expected physical ranges of petrophysical variables.

In this study, an efficient approach is proposed to eliminate the disadvantages of the clustering procedure. By integrating numerous variables measured on core samples in the laboratory, which are based on different physical principles, we aim to increase the performance of clustering and resolve the problem of ambiguity. When doing this, we transform the sparse matrix of measured variables into a fully filled one by replacing the missing data by quasi measured data using a novel regression-based (multivariate) imputation method. Cluster analysis of the data matrix completed with the synthetic values of petrophysical quantities gives highly reliable results for rock typing and allows an improved identification of reservoir zones, which is demonstrated by a Hungarian oilfield study.

2. Methods

2.1 Imputation algorithm

To perform a reliable imputation of core laboratory data, a correlation-based algorithm is proposed in the paper, which effectively predicts the missing values from the available petrophysical observations. The heterogeneity of data sources is a result of the measurements taken at different times, during which sometimes new measurement types are added or older ones are taken away due to the development of measuring technology. The change of the observed parameters with time may also be the result of the re-interpretation of geological formations, which requires modification of the measurement program. It is expedient to maximize the number of measured properties (i.e., the number of the data columns), because the larger the statistical sample the more reliable the result is, revealing more characteristics of the studied formations.

The correlation-based imputation method assumes a sparse data matrix as input, where places of missing data are filled with NaN (Not-a-Number) values. It is noted that the replacement of all NaN values is not always possible. The proposed algorithm shows a viable solution for the datasets used in practice, but the limitations of the described heuristic approach require further investigation. It is assumed that the columns of the matrices have been sorted by significance, which is not an exact measure in this case. The assumption is based on the fact that the data matrices have been arranged by petrophysical experts. They tend to choose lower column indices for more important properties, while the higher indices remain for the least important ones. The imputation algorithm uses this implicit information by filling the missing data from lower indices to higher indices in multiple steps. As a basic calculation step of the estimation process, one fits a linear regression model for two selected columns. Consider variables x and y as two columns of the input data matrix. First, we collect (x_k, y_k) pairs of available values (where x_k and $y_k \in \mathrm{R}$). It is important to check whether the number of points is enough for linear regression. (The examined core datasets are proper in the sense that we have enough points for linear regression in all cases.) In positive case, we can fit a linear regression model using those rows where we have data in

both columns. Then, one can give an estimate for the missing values by substituting the data into the linear equation, where exactly one of the x_k or the y_k is missing (NaN). The imputation process is shown in Fig.1. The selected columns of variables x and y with four and five measured data, respectively, are represented in Fig.1a. The white cells denote the NaN values, where measurements are not available. In Fig. 1b, the (x_k, y_k) pairs with real values are highlighted. This example has only three matching points for linear regression, but the number of data points is significantly larger in the processed big dataset. Regression analysis finds an optimal linear function, the parameters of which are estimated by minimizing the sum of squared distances of the selected points from the line. The above described algorithm is applied to each row of the input data matrix until it is fully filled with values extracted from the resultant linear equations (Fig. 1c). The pseudo-code of the correlation-based interpolation algorithm is presented in the Appendix.

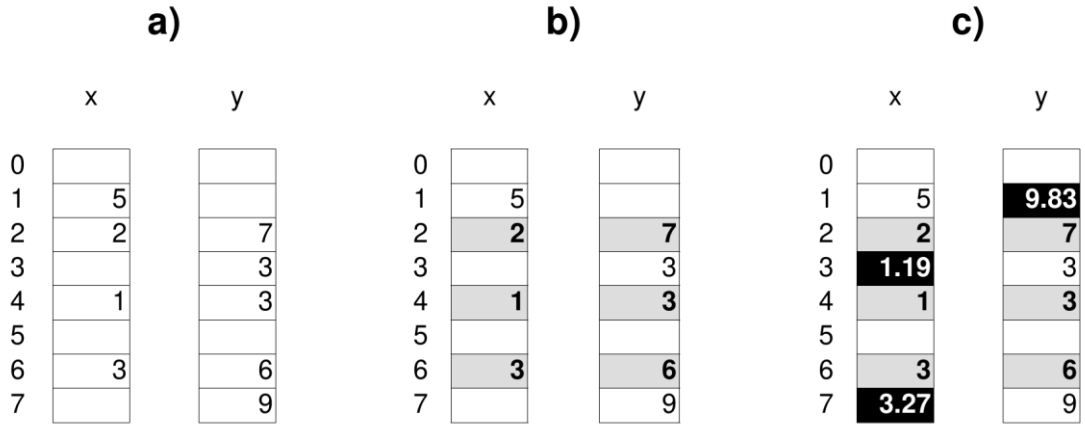


Fig. 1. Scheme of the correlation-based imputation method as a data preparation and filling procedure before cluster analysis

A comparison can be made between the result of matrix factorization as a traditional completion method (see Section 1) and the proposed imputation procedure for a Miocene dataset (Fig. 2). On the axis of abscissae, the number of rows of the input data matrix is plotted representing the depth of core samples collected from neighboring wells. As it is seen in Fig. 2, the matrix factorization algorithm, on the examined data, overestimates the water saturation (~0.9 v/v) in high porosity gas-bearing formations. It estimates too low gas saturation or forecasts water-bearing formations furthermore it gives noisier

solution than the correlation-based imputation algorithm (Fig. 2a–b). In addition to it, non-physical estimates for the petrophysical parameters are also given, e.g., negative values of permeability are indicated in Fig. 2c. The reliability of matrix factorization is questionable, because it overestimates the permeability of calcareous marls and silts (e.g., for the first eight core samples in Fig. 2c) and at the same time, it underestimates the same property in fractured basalt formations (compare the 77th and 80th core samples in Fig. 2c–d). The permeability values in fractured zones do not differ significantly from their environment by means of classical matrix factorization, while they actually do separate using our proposed method (see the two peaks above 60 mD in Fig. 2d). In contrast, the correlation-based imputation method gives more realistic results and physically correct values of petrophysical parameters.

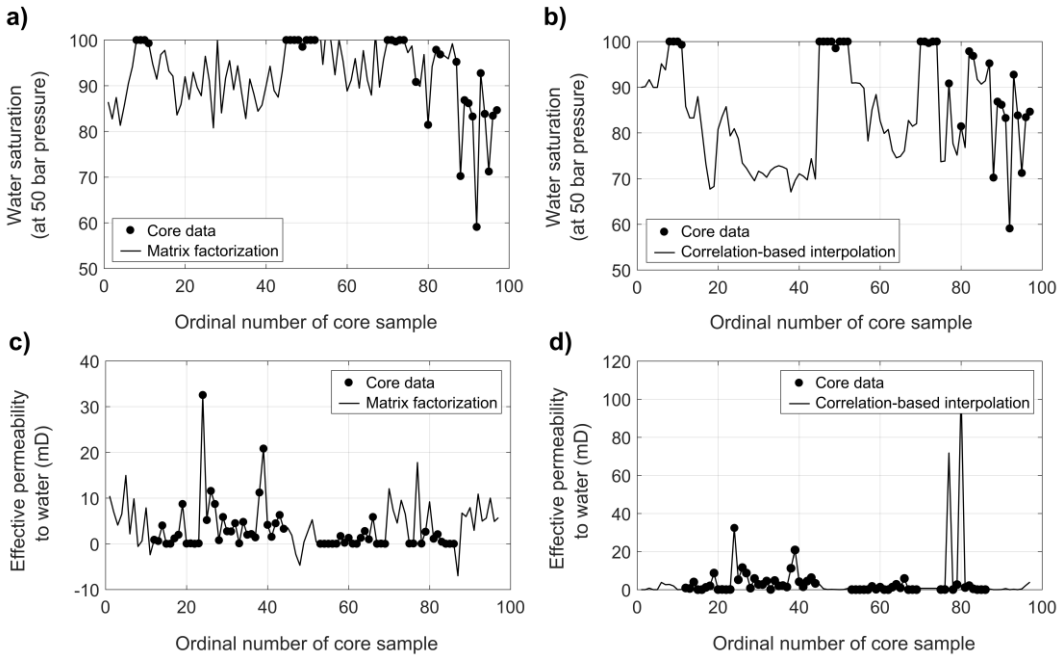


Fig. 2. Interpolation results for the capillary pressure derived water saturation (a–b) and effective permeability (c–d) using a traditional matrix factorization method (a, c) and the suggested correlation-based imputation procedure (b, d) in a Hungarian Miocene formation

2.2 Dimension reduction

In order to make the interpretation of multivariate measurements easier, it is advantageous to reduce the number of observed parameters to fewer statistical variables (Jolliffe, 2002). In this study, principal

component analysis (PCA) is used to decrease the number of observed petrophysical variables by reducing the number of columns of the input data matrix into smaller sizes. In complex oilfield problems, PCA can be beneficially used as a preliminary data processing step to find the main geological characteristics of the reservoir model. As an example, Jung et al. (2018) established a PCA assisted support vector machine approach for setting a proper initial model, which allows an improved history matching and prediction of reservoir performances in heterogeneous channel reservoirs. The PCA transformation is orthogonal, which gives new uncorrelated variables, i.e., principal components (PCs). Consider **D** as an *N*-by-*M* fully filled data matrix, where *N* denotes the number of core samples collected in different wells and *M* is the total number of observed petrophysical variables in all wells. Let us write the data matrix as a product of the *N*-by-*r* matrix of PCs (**P**) and that of their coefficients of *M*-by-*r* size (**W**)

$$\mathbf{D} = \mathbf{P}\mathbf{W}^{\mathrm{T}} \tag{1}$$

where *r* is the number of PCs. With the knowledge of the PCs' coefficients, a unique solution can always be found to Eq. (1). Scores of the *j*-th principal component are given by elements P_{lj}, where l=1,2,...,N and j=1,2,...,r. By multiplying Eq. (1) from the right with an orthonormal matrix $\mathbf{W}^{\mathrm{T}}$, the PCs can be determined as a linear combination of the measured variables by equation $\mathbf{P}=\mathbf{D}\mathbf{W}^{\mathrm{T}}$. The estimation of the weighting coefficients leads to the solution of an eigenvalue problem, in which the covariance matrix of the matrix product $\mathbf{D}^{\mathrm{T}}\mathbf{D}$ for centralized data gives

$$\lambda_k = \sigma_k^2, \tag{2}$$

where λ_k is the *k*-th eigenvalue and σ_k is the variance of the *k*-th petrophysical parameter. As a consequence of Eq. (2), the extracted PCs are sorted in a manner that the first few of them explaining most of the variance of the original variables showing the directions of biggest variances of the original sample. PCA specifies the coordinates of the objects of observed data in a new coordinate system stretched by the PCs, and rotates the original variables into the direction of the principal axes. The

number of PCs are arbitrarily chosen or selected by studying the distribution of eigenvalues in the function of the PCs. By examining the elements of the *M*-by-*M* loading matrix $\mathbf{W}^{\mathrm{T}}$, one can investigate the individual contribution of measurement variables to the PCs. In this study, the relative importance of the *i*-th petrophysical variable is expressed by

$$s_i = \max_j \left| W_{ij} \right|, \tag{3}$$

where **s** gives the vector of maximal absolute values of PCs' weights (i=1, 2, …, M and j=1, 2,…,r). In the reduced coordinate system of PCs, cluster analysis can be made to group the data objects and infer petrophysical characteristics of the available core information.

2.3 Clustering of big core datasets

A non-hierarchical K-means clustering was applied to the interpolated core data matrix, which is a commonly used clustering approach for performing unsupervised learning tasks (Hartigan, 1979). It can be effectively used for the grouping of core data in such a way that the *M*-dimensional objects specified by petrophysical properties measured on given rock samples (collected from given depths and wells) are more similar than others observed on different samples. From the point of view of the proposed method, it is of great importance that data objects connected to the same cluster define approximately the same lithological and petrophysical character, while other clusters represent dissimilar ones. Given an initial set of center groups calculated as the average of cluster elements, the algorithm proceeds by repeating two steps. We assign each observation to the cluster by the nearest mean principle. Then, we calculate the new means to be the centroids of the observations in the new clusters. The solution is obtained when assignments no longer change. The optimal number of clusters (K) to be formed can be selected by the minimization of SSE (Sum of Squared Error), which gives the deviation between the centroid and group elements for all groups. In the preliminary determination of the number of clusters, it is important to consider a priori geological and rock physical information. Too small number of

clusters does not allow proper spatial resolution, but a large number of groups may result in false conclusions, i.e., non-existent lithological categories.

Another factor influencing the result of cluster analysis is the distance between the data objects as the degree of similarity. Let $\mathbf{x}^{(i)}$ and $\mathbf{x}^{(j)}$ two vectors having M number of observed variables in the data space of $X_1,\ldots,X_M$. In our case, X_p denotes the p-th petrophysical variable measured in the laboratory. In a more detailed form, the i-th and j-th objects are given as $\mathbf{x}^{(i)}=[x_1^{(i)},\ldots,x_M^{(i)}]^{\mathrm{T}}$ and $\mathbf{x}^{(j)}=[\,x_1^{(j)},\ldots,x_M^{(j)}]^{\mathrm{T}}$ (where T is the symbol of transpose). The most commonly used distance metric is the Euclidean norm

$$D_E\left(\mathbf{x}^{(i)},\mathbf{x}^{(j)}\right)=\left[\left(\mathbf{x}^{(i)}-\mathbf{x}^{(j)}\right)^{\mathrm{T}}\left(\mathbf{x}^{(i)}-\mathbf{x}^{(j)}\right)\right]^{1/2}, \tag{4}$$

which may be very sensitive to outliers that are extremely far from the center of the group. A more robust solution can be achieved by using other distance metrics such as the L_1-norm based City block (or Manhattan) distance

$$D_C\left(\mathbf{x}^{(i)},\mathbf{x}^{(j)}\right)=\sum_{p=1}^{M}\left|x_p^{(i)}-x_p^{(j)}\right|, \tag{5}$$

or the Mahalanobis distance taking the correlation of the data into account

$$D_M\left(\mathbf{x}^{(i)},\mathbf{x}^{(j)}\right)=\left[\left(\mathbf{x}^{(i)}-\mathbf{x}^{(j)}\right)^{\mathrm{T}}\mathbf{S}^{-1}\left(\mathbf{x}^{(i)}-\mathbf{x}^{(j)}\right)\right]^{1/2}, \tag{6}$$

where $\mathbf{S}^{-1}$ is the inverse of sample covariance matrix being a weighting operator during the cluster analysis of data objects. By combining the process of K-means cluster analysis with PCA, the following simplified workflow shown in Fig. 3 was developed.

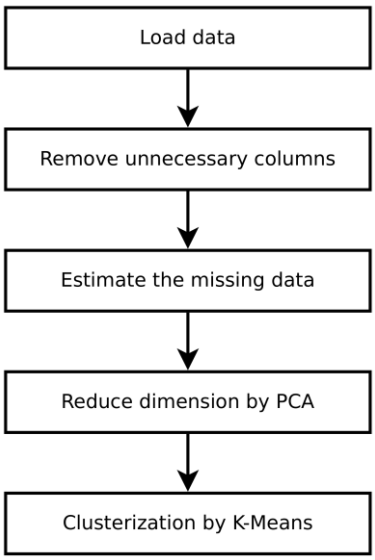


Fig. 3. Flowchart of the correlation-based interpolation- and principal component analysis assisted clustering method

In applying the method, we have considered a column as unusable, when it contains only the same values. One can drop these kind of columns without information loss. It must be also mentioned that clustering can be applied with or without dimension reduction. We prefer the use of PCA method in most cases, because it makes possible to highlight the most important features. Unfortunately, it does not provide the meaning of the resulted new dimensions. However, the physical relations between the PCs and petrophysical characteristics may be explored by partial correlation analysis, which can be added as a new element to the workflow.

3. Case study

The core samples examined in this study originate from the southern part of the Pannonian Basin Province of Central Europe, where several petroleum systems have been discovered. The Pannonian Basin consists of a large extensional basin of Neogene age overlying Paleogene basins and a Mesozoic (or older) basement (Dolton, 2006). A several kilometers thick large-extension Tertiary basin-fill sedimentary sequence contains oil and gas-bearing formations. The main lithological categories are conglomerate, breccia, calcareous marl, clay, aleurolit, and gravellous sandstone of different grain sizes, dolomite and some basalt. The rock specimen was collected in 28 neighboring wells from an interval of

1352 m. The full dataset as input for cluster analysis includes 421 core samples and 59 measured petrophysical properties, i.e. carbonate content, bulk and grain density, helium porosity, Klinkenberg-corrected permeability, oil and water permeability, irreducible water saturation, sample porosity and volume, mercury saturation in the pressure range of 0.1–4,000 bar (and pore-throat radii between $1.88 \cdot 10^{-3}$–75 μm), and water saturation determined by centrifuge method in the pressure range of 0.16 and 6 bar. Most of the observed variables were measured both perpendicular and parallel to the axis of the core drilling forming two separate input parameters from the same petrophysical quantity. For the mercury injection and capillary pressure curves, one column of the data matrix contains those saturation values, which were measured on different core samples under the same pressure. The strength of correlation between the observed variables is found to be moderate. We introduce the following scalar called mean spread for the measure of average correlation:

$$S = \left\{ \frac{1}{M(M-1)} \sum_{i=1}^{M} \sum_{j=1}^{M} \left[R(\mathbf{D})_{ij} - \delta_{ij} \right]^2 \right\}^{1/2}, \tag{7}$$

where $R(\mathbf{D})$ is the Pearson's correlation matrix of the observed variables, M is the number of measurement variables and δ is the Kronecker delta function. In this case study, S resulted around 0.6 for the data matrix of core measurements. The linear regression connection between some of the studied petrophysical quantities is shown in Fig. 4, where the independent and dependent variables were denoted by x and y, respectively.

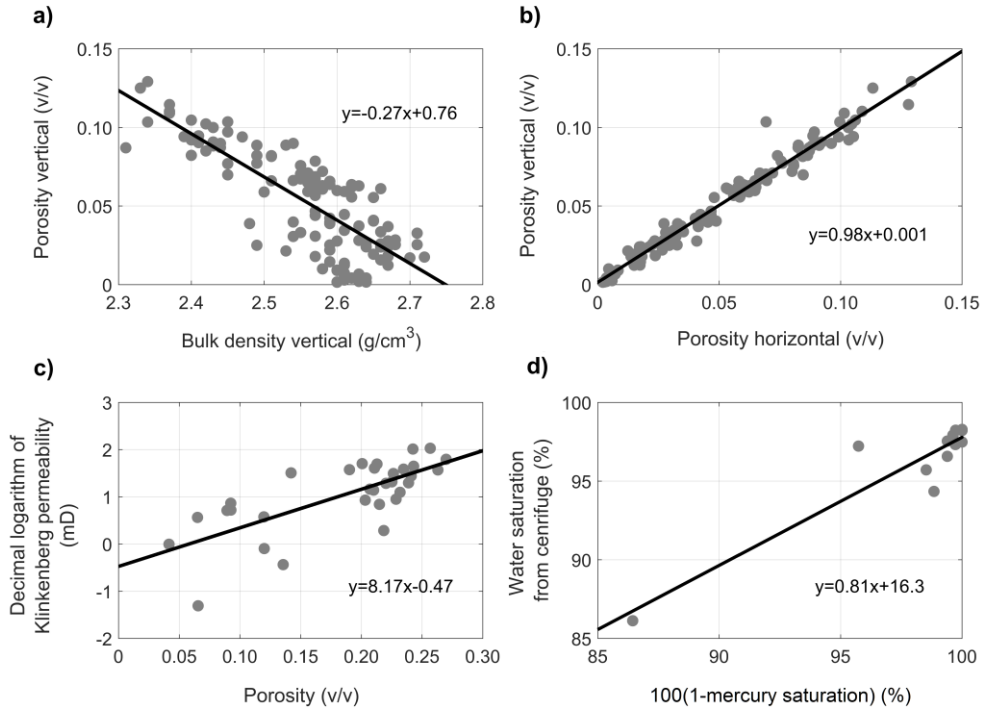
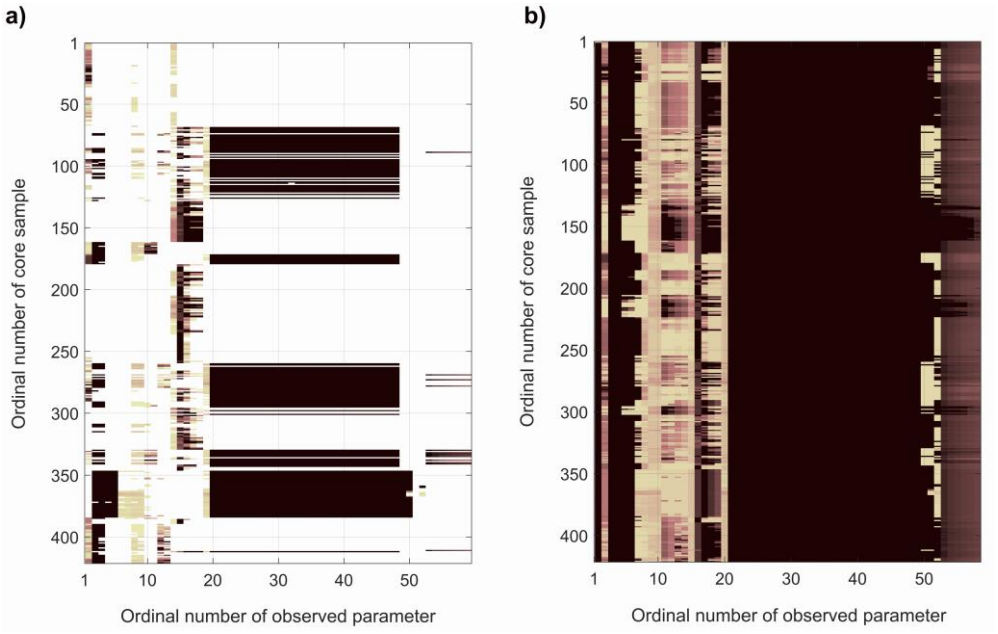


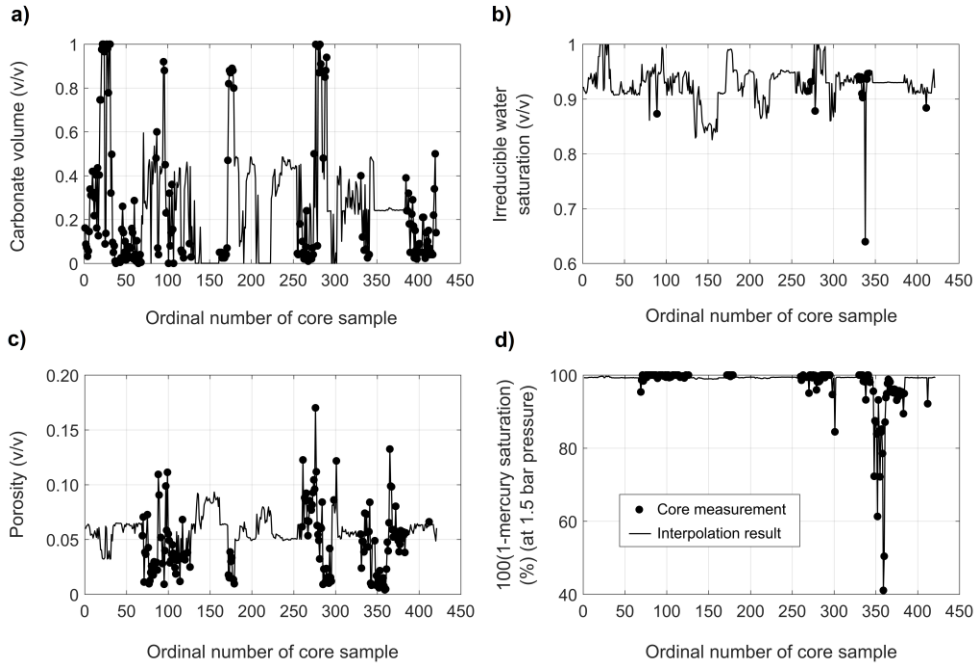
Fig. 4. Linear relationships between the petrophysical parameters in a Hungarian Miocene complex including the regression equations of porosity vs. bulk density (a), porosity measured perpendicular and parallel to the axis of the core drilling (b), Klinkenberg-corrected permeability vs. porosity (c) and water saturation measured by mercury injection and centrifugal methods (d)

Core data referring to different wells and depth intervals are placed in the data matrix in random order. By using the linear regression functions determined between each pair of measured variables, the missing values in the input data matrix is replaced with synthetic data. Our correlation-based imputation method completely fills out the sparse matrix of the observed variables within their physical boundaries. The image of the original data matrix can be seen in Fig. 5a, where the missing data places are indicated by white color and the values of petrophysical parameters are scaled into the range of 0 and 1. The raw data matrix is incomplete at 71 %. One can notice that the data coming from different wells represent distinct groups of observed variables. It is shown in Fig. 5b that how the proposed statistical approach gives an estimate to the missing cells so that creating a complete data matrix of petrophysical quantities. At the end of this process, the large size data matrix is completed at 100 %. The result of interpolation is plotted for four essential petrophysical parameters in Fig. 6, which forms an appropriate input for the subsequent phase of the clustering procedure.



321

322 **Fig. 5.** Input data matrix includes missing values of petrophysical parameters in the investigated
323 Hungarian Miocene complex (a) and fully filled data matrix extracted from the raw dataset using the
324 correlation-based imputation method (b)
325



326

327 **Fig. 6.** Observed data measured on core samples (dots) and interpolation results for petrophysical
328 parameters in a Hungarian Miocene complex (black line): carbonate content (a), irreducible water
329 saturation (b), sample porosity (c), water saturation derived from mercury injection experiment (d)
330

331 The resultant data matrix is decomposed by PCA according to Eq. (1), which concentrates the common
332 observed information in a few variables. By calculating the eigenvalues of the PCs represented in Eq.
333 (2), a decision can be made for the number of extracted variables. For the processed dataset, six PCs

cover the 99.95 % of the total variance of the original data (Fig. 7a), so that the amount of information loss is 0.05 %. Then, the 421 objects in the six-dimensional space of PCs is clustered by a K-means method. The Elbow Method is a commonly used technique in the practice of cluster analysis to select the optimal number of clusters. Consequently, by using the SSE plot shown in Fig. 7b, five clusters were chosen by us for separating the main lithology types in the investigated Miocene complex.

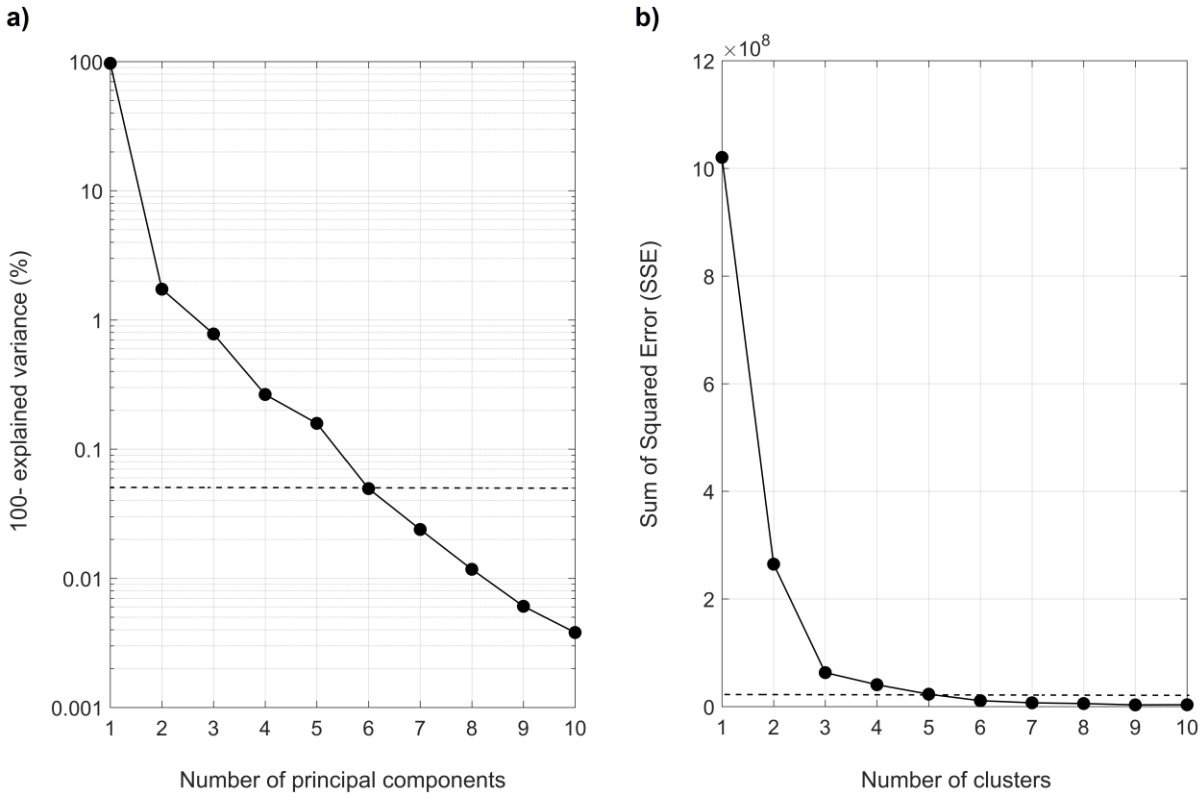


Fig. 7. Variance portions explained by principal components derived from PCA (a), the sum of squared error calculated for different number of clusters given by K-means cluster analysis (b)

The measured variables contribute to the solution to varying degrees. The PCA provides an opportunity for this analysis. The PC loadings given in Eq. (3) are plotted in Fig. 8. Based on the absolute values of the PCs' coefficients obtained in the input Miocene dataset, one can determine the relative weight of each petrophysical parameter. The figure shows the significant role of horizontal and vertical bulk density (RHOBH, RHOBV), horizontal and vertical porosity (PORSH, PORSV), Klinkenberg corrected permeability (KLEKV) and mercury injection data measured on higher pressures (1–SHG20,…,1–SHG28), while it indicates lower weights of the carbonate content (CARB), water

saturation (SW1,…,SW6) and mercury injection data at low pressures (1−SHG1,...,1−SHG17). Maximal impact on the PCs goes to the Klinkenberg corrected permeability and relative permeability to oil and water (KO, KW). In conclusion, PCA emphasizes primarily the pore structure characteristics and the hydraulic conductivity of the given rock types.

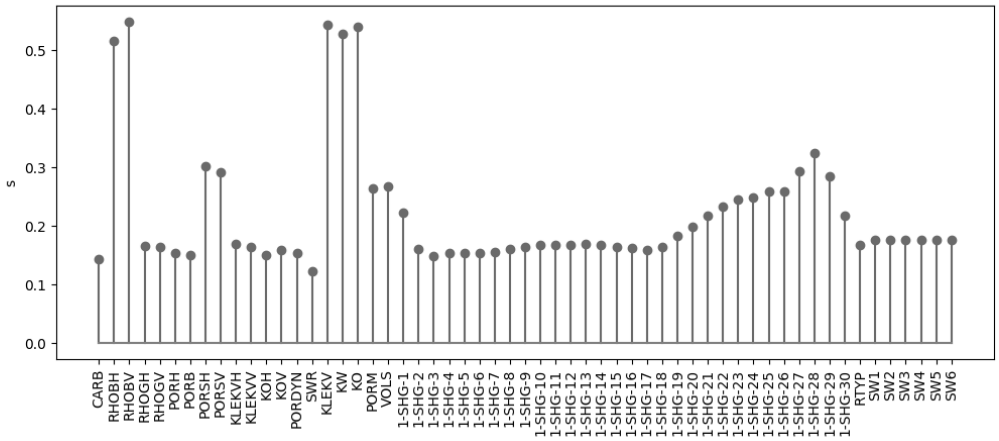


Fig. 8. Coefficients of principal components indicating the relative weights of petrophysical properties on the solution of PCA in the investigated Hungarian Miocene complex

Clustering is made to separate the row vectors of observed petrophysical parameters into K=5 classes. To prevent the data processing from the harmful effect of outliers, the City block distance in Eq. (5) is used for the K-means cluster analysis. The grouped data objects in the space of the first three eigenvalues (eig1−eig3) is given in Fig. 9a, while the same objects are plotted in the coordinate system of three measured quantities (Fig. 9b) such as carbonate volume including calcite and dolomite contents (CARB), porosity obtained by dynamic displacement on full diameter samples (PORDYN) and relative permeability to water (KW). The figures demonstrate that the clusters are well separated (e.g., the high porosity and permeability reservoirs are indicated with dark yellow color) despite the input dataset being compositionally heterogeneous and unevenly distributed in the space.

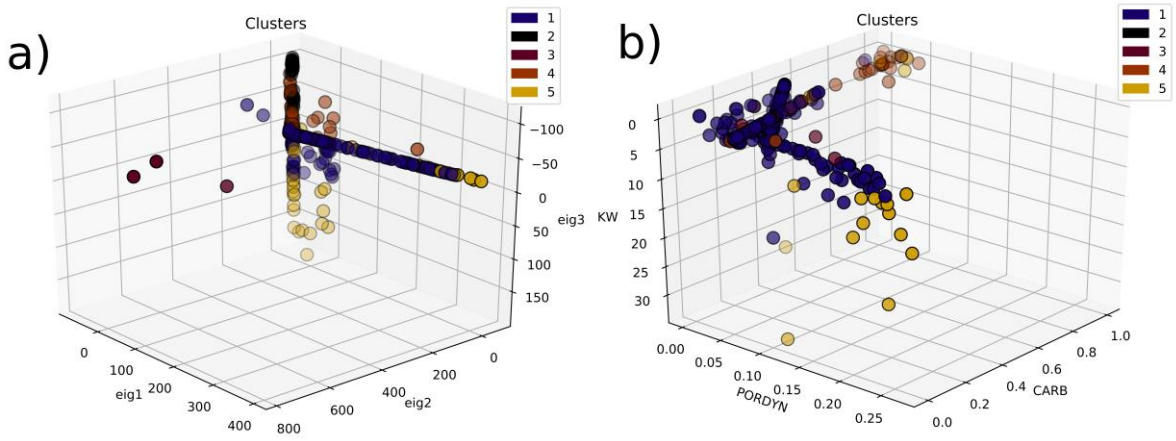


Fig. 9. Result of PCA of the filled data matrix observed in a Hungarian Miocene complex: transformed data in the coordinate system of eigenvectors (a). Result of cluster analysis: clusters vs. petrophysical variables as input parameters (b)

The result of cluster analysis can be seen in Fig. 10, where the section of cluster numbers in the last track shows to which group the rock samples are classified. It is noted that sudden changes in the resultant cluster numbers are also due to the fact that the core data referring to different depth coordinate (and wells) are quasi-randomly selected along the ordinate axis. The hydrocarbon-bearing formations can be clearly separated in the image of cluster numbers. Cluster 5 represented by yellow color is connected to highly porous and permeable zones with small amount of carbonate (e.g., for samples 140–160). Petrophysical properties measured both with horizontal and vertical direction to the axis of sampling practically shows just a little amount of anisotropy. The capillary pressure curves indicates high amount of movable fluids in these intervals. The other four clusters indicate less permeable zones with higher amount of carbonate. Cluster 1 and 2 show impermeable formations with much amount of carbonate and high irreducible water saturation (e.g., around the near vicinity of sample 175), while cluster 3 and 4 form a transition between permeable and impermeable rocks (e.g., for samples 1–70). By comparing the result of cluster analysis to that of the lithology description, cluster 3 and 4 can be identified mostly as conglomerate and clay with some amount of dolomite, respectively. Cluster 5 is mainly composed of sandstone, gravel and fractured breccia with good reservoir storage capacity. Cluster 1 and 2 include calcareous and clayey marls. In the brown intervals, dolomites (e.g., around sample no. 200) and in the blue ones, thin layers of fractured basalt conglomerate are also found (e.g.,

in the range of 290–300). The clustering procedure allows the classification of capillary pressure curves by separating different groups (Fig. 11), which makes the interpretation more reliable. Although there is some overlap between the clusters of different pore geometries, the reservoirs of the highest movable wetting phase saturation can be clearly distinguished (see them by yellow color). The full curves of capillary pressure data gives more information than conventional porosity and permeability data. In addition to pore-size geometry distribution, it holds important information also on other textural characteristics of the clustered rock types, which can be further investigated in future studies.

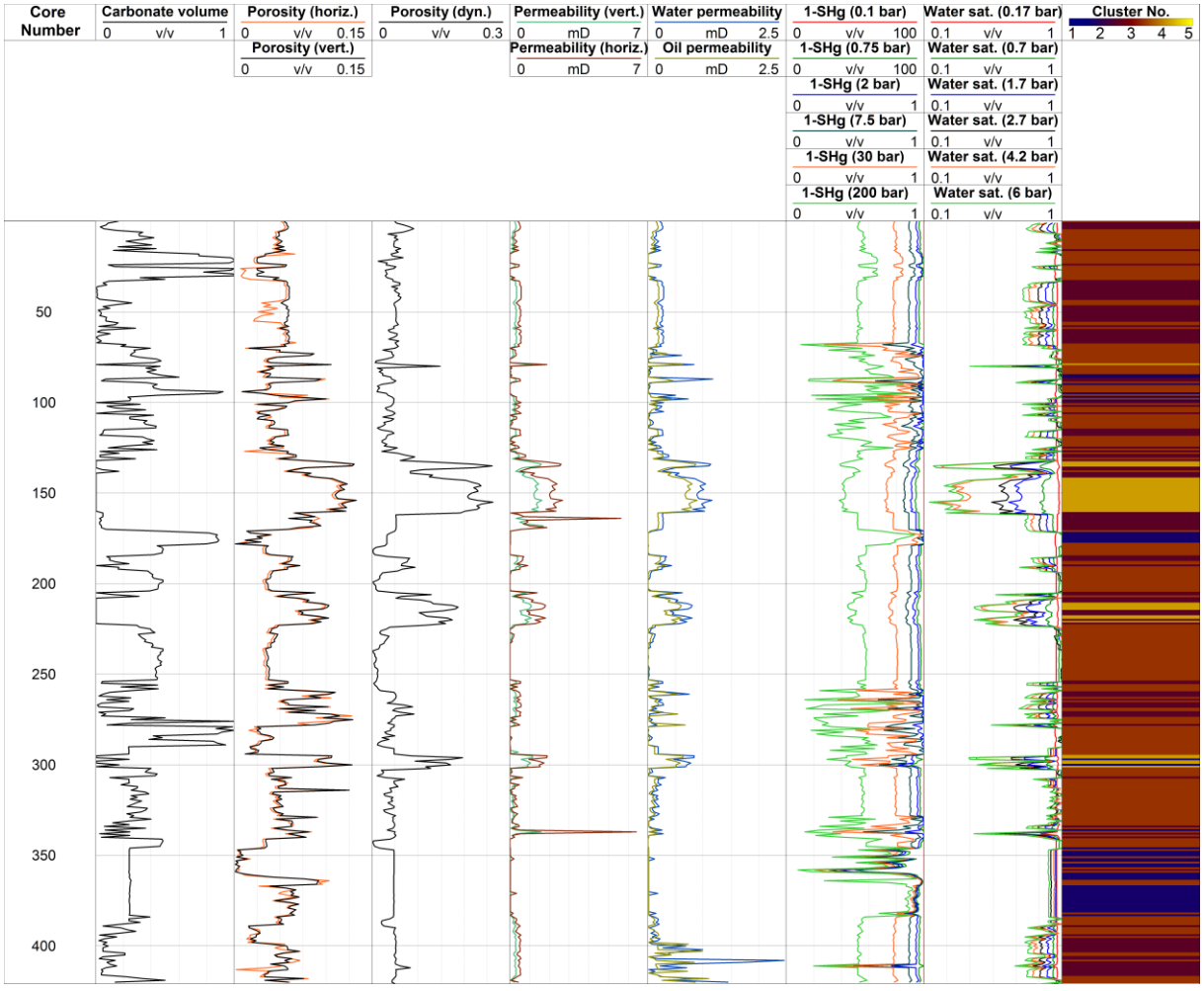


Fig. 10. Cluster analysis of laboratory data measured on core samples collected from a Hungarian Miocene complex: interpolated logs of petrophysical parameters as input for clustering (tracks 1-7), estimated distribution of clusters by using Manhattan distance metric as the result (track 8)

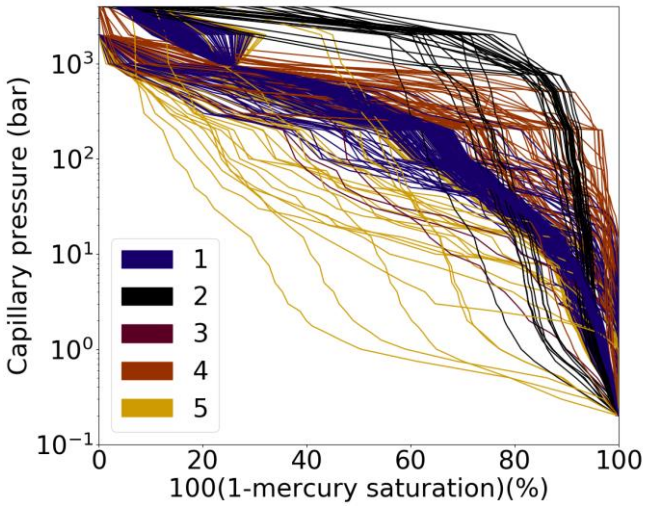


Fig. 11. Results of clustering of mercury injection data measured on core samples collected from the studied Miocene formations

4. Discussion

To test the accuracy of estimation results of the proposed clustering method, a synthetic modeling experiment has been accomplished. An exactly known inhomogeneous petrophysical model is assumed to represent a hydrocarbon formation with varying amount of porosity (Φ), water saturation (S_w), sand volume (V_{sd}), shale content (V_{sh}) and carbonate volume (V_c) along a borehole. Theoretical open-hole wireline logging data are calculated and simultaneously processed by cluster analysis to examine how accurately the exact model is reconstructed. The performance of cluster analysis can be tested using arbitrarily chosen amount of noise added to the input data and the rate of incompleteness of the data matrix formed from the noisy synthetic well logs. The following simplified probe response functions summarized by Alberty and Hashmy (1984) are used to calculate the wireline logs (the effect of invasion is neglected in this approach)

$$\rho_b = \Phi\left[S_w\rho_w + (1-S_w)\rho_h\right] + V_{sh}\rho_{sh} + V_{sd}\rho_{sd} + V_c\rho_c\,, \tag{8}$$

$$GR = \rho_b^{-1}\left(V_{sh}GR_{sh}\rho_{sh} + V_{sd}GR_{sd}\rho_{sd} + V_cGR_c\rho_c\right), \tag{9}$$

$$P_e = \Phi\left[S_w P_{e,w} + (1 - S_w) P_{e,h}\right] + V_{sh} P_{e,sh} + V_{sd} P_{e,sd} + V_c P_{e,c}\,, \tag{10}$$

$$\Delta t = \Phi\left[S_w \Delta t_w + (1 - S_w) \Delta t_h\right] + V_{sh} \Delta t_{sh} + V_{sd} \Delta t_{sd} + V_c \Delta t_c\,, \tag{11}$$

$$\frac{1}{\sqrt{R_d}} = \left[\frac{V_{sh}^{\left(1-\frac{V_{sh}}{2}\right)}}{\sqrt{R_{sh}}} + \frac{\left(\sqrt{\Phi}\right)^m}{\sqrt{a R_w}}\right]\left(\sqrt{S_w}\right)^n, \tag{12}$$

where the quasi-measured parameters are bulk density (ρ_b), natural gamma-ray intensity (GR), photoelectric absorption cross-section index (P_e), acoustic (P-wave) travel-time (Δt) and deep resistivity (R_d). The functional constants included in Eqs. (8)–(12), representing the physical properties of pore fluids, shale and mineral components as well as the textural properties of rocks, are specified in Table 1 (we assume gas phase under hydrocarbon).

Table 1
Zone parameters used for calculating synthetic wireline logs to test the accuracy of the proposed cluster analysis method.

Well log	Zone parameter	Symbol	Selected value	Unit
Natural gamma intensity (GR)	sand	GR_{sd}	10	API
	shale	GR_{sh}	140	
	carbonate	GR_c	5	
Bulk density (ρ_b)	sand	ρ_{sd}	2.65	g/cm^3
	shale	ρ_{sh}	2.55	
	carbonate	ρ_c	2.79	
	pore-water	ρ_w	1.09	
	hydrocarbon	ρ_h	0.016	
Sonic interval-time (Δt)	sand	Δt_{sd}	56	μs/ft
	shale	Δt_{sh}	108	
	carbonate	Δt_c	46	
	pore-water	Δt_w	200	
	hydrocarbon	Δt_h	305	
Photoelectric index (P_e)	sand	$P_{e,sd}$	1.81	barn/e
	shale	$P_{e,sd}$	3.50	
	carbonate	$P_{e,sd}$	4.11	
	pore-water	$P_{e,sd}$	0.81	
	hydrocarbon	$P_{e,sd}$	0.09	
Deep resistivity (R_d)	shale	R_{sh}	1	Ωm
	pore-water	R_w	0.06	
	cementation exponent	m	2.0	–
	saturation exponent	n	2.0	
	tortuosity coefficient	a	1.0	

In this uncertainty analysis, 5 % Gaussian distributed noise is added to the synthetic data calculated in the forward modeling procedure using Eq. (8)-(12). The mean spread (defined in Eq. (7)) calculated for the noisy well logging parameters is 0.44, which represents weak correlation relationship between the five input variables on the average. The full correlation matrix is given in Table 2.

Table 2
Pearson's correlation matrix of synthetic well logs including 5 % Gaussian noise.

	GR	P_e	ρ_b	Δt	R_d
GR	1	0.54	0.44	0.23	–0.50
P_e	0.54	1	0.61	–0.38	–0.49
ρ_b	0.44	0.61	1	–0.34	–0.50
Δt	0.23	–0.38	–0.34	1	0.25
R_d	–0.50	–0.49	–0.50	0.25	1

The exactly known petrophysical model can be seen in tracks 1–2 in Fig. 12, where 6 layers can be distinguished: 5 m thick shale, 10 m thick water-bearing sandstone, 5 m thick shale, 8 m thick hydrocarbon-bearing sandstone, 10 m thick limestone and 6 m thick shale. By assuming 4 main lithological units, the entire well logging dataset (plotted in tracks 3-7) is processed by K-means cluster analysis (in this test PCA is not necessary to be applied). As a result, the depth variation of cluster numbers is shown in the last track.

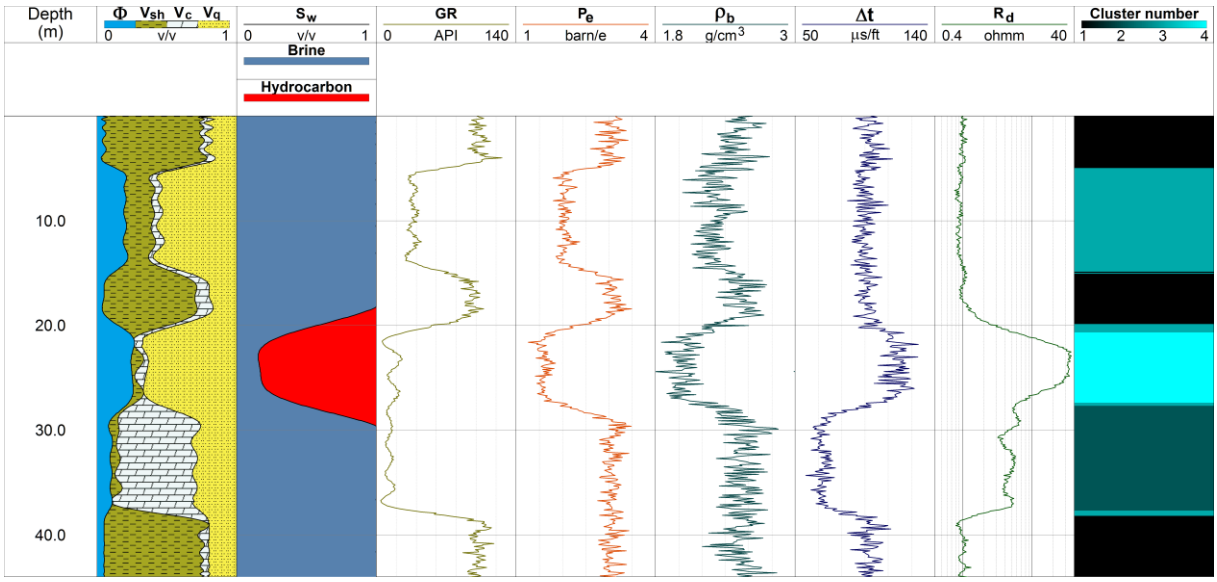


Fig. 12. Cluster analysis of synthetic wireline logs contaminated with 5 % Gaussian distributed random noise

The reliability of the correlation-based imputation method assisted cluster analysis is tested using different datasets derived from the noisy synthetic well logs, where the proportion of missing data is gradually increased. The vertical distribution of the four clusters determined from the noiseless data is illustrated on the third track of Fig. 13, which is considered to be the exact model. For this analysis, a number of data randomly chosen from all logs are substituted with NaN values to form new input datasets and correlation-based imputation is performed before clustering them. To the right from the third track, the result of cluster analysis is shown for the incomplete well logging datasets, where the number of NaN values relative to the total number of data is 5, 10, 20 and 50 %, respectively. The figure shows that even in large incompleteness of the input data matrix the main lithological units can be reliably separated. These results are properly good for a relatively short depth interval (and few number of data) and small number of weakly correlated measured variables. The uncertainty of estimation depending on the accuracy of data and interpolation errors can be reduced by selecting input variables with stronger correlations and removing the outliers from the dataset. The latter can be effectively done by using automated robust clustering methods.

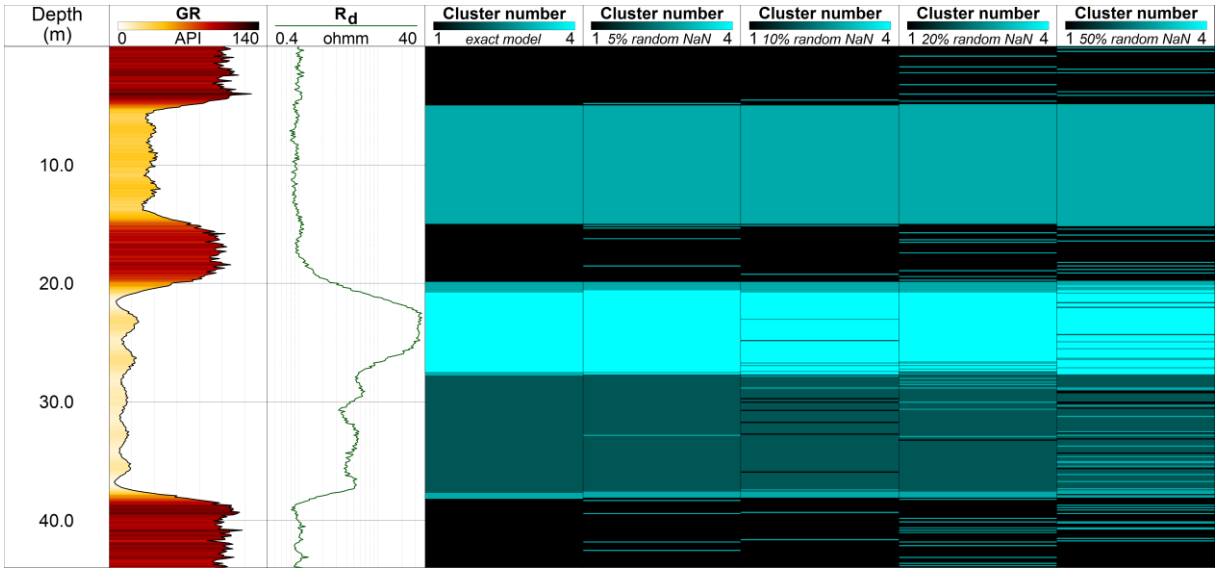


Fig. 13. Effect of the amount of data matrix incompleteness on the result of cluster analysis

The aforementioned synthetic modeling experiment focuses mainly on the effect of the accuracy of original measurements and the interpolation error when replacing the unknown values with estimated ones. Besides these types of noises added to the dataset, the data also include other sources of uncertainty. For a more detailed uncertainty analysis, we have to consider basically three types of errors. At first, one has to quantify the error of the original measurements. The estimation of measurement errors makes it necessary to know the features of the measurement process. Secondly, our model assumes linear connection between all the columns in the data matrix. For a better approximation, one has to consider the underlying physical relations among the columns, which is unknown in most of the cases. At last, the accumulated estimation errors should also be taken into consideration. It is the result of the incremental nature of the imputation algorithm giving an estimate in the actual iteration from the estimated values of earlier iterations. In the ideal case, only the considered value is missing, which means that the estimation is based on the available measurements. In the worst case, all values of the selected columns are estimated. This kind of accumulated estimation error depends on the rate of the counts of measured and predicted values having used for the given estimation.

5. Conclusions

In the paper, an improved clustering algorithm is proposed for the interpretation of core datasets originating from heterogeneous sources. A correlation-based multi-linear regression method is first used for the filling of typically incomplete data matrices, which gives a suitable input for the K-means cluster analysis performed in a subsequent data processing step. After the upload phase, all variables are taken into account during the clustering, but it is also possible to select the parameters individually based on the results of principal component analysis. Those parameters with relative small PC coefficients can be neglected. Since there is no limit for the approach on the number of processed variables forming the input dataset, the clustering method is suitable for the future expansion of observed parameters. Based on this, not only new petrophysical parameters, but also mineralogical (e.g., from X-ray diffraction) and other compositional (e.g., geochemistry) data may be involved. An added advantage of the statistical method is the grouping possibility of the capillary pressure curves, which helps to integrate more

information on the textural and fluid characteristics of the investigated formations. (The analysis of these curves has wide literature, which is beyond the scope of this paper.) The explanation of the behavior of these curves are typically given on an empirical basis in the literature. However, an analytic description may reveal new petrophysical characteristics of the studied formations. In addition to the measured parameters of such analytic models, for a more reliable interpretation of the pore structure (textural properties), pore-fluid types (bound or free water and hydrocarbons) and properties (e.g., viscosity), T2 relaxation time distribution curves of nuclear magnetic resonance measurements can also be added as input in future applications.

Cluster analysis is commonly used also for the processing of wireline logging data. Well log derived petrophysical parameters such as porosity, shale volume, matrix volumes, water and hydrocarbon saturation, permeability etc. as high resolution in situ information can significantly increase the size of the statistical sample and may further improve the performance of cluster analysis. The big amount of data is processed by linear regression and classical K-means clustering algorithm for a quick interpretation. For obtaining a more reliable solution, a robustified cluster analysis approach including a nonlinear interpolation method as imputation phase can be used in the future. As a statistically highly efficient method, the most frequent value-based (automated) weighting procedure introduced by Steiner (1991) gives a robust solution independent from the nature and statistical distribution of the input dataset (e.g., in Zhang, 2017). As a first result, its application to clustering of well logs can be found in Braun et al. (2016), which will be followed by improved clustering algorithms in the near future as intensive research is currently made at the Department of Geophysics, University of Miskolc. The resultant clusters can also be easily correlated between the wells to represent the spatial distribution of clusters for two- and three-dimensional cases.

Acknowledgments

The authors would like to thank the experts and staff at the MOL group for their valuable cooperation and for the measurement and laboratory data they provided. This research was supported by the

European Union and the Hungarian State, co-financed by the European Regional Development Fund in the framework of the GINOP-2.3.4-15-2016-00004 project, aimed to promote the cooperation between the higher education and the industry.

References

Alberty, M., and K. Hashmy, 1984. Application of ULTRA to log analysis: Presented at the SPWLA Symposium Transactions, 1–17.

Braun B. A., Abordán A., Szabó N. P., 2016. Lithology determination in a coal exploration drillhole using Steiner weighted cluster analysis. Geosciences and Engineering 5 (8), 51–64.

Candès E. J., Recht B., 2009. Exact matrix completion via convex optimization. Foundations of Computational mathematics 9(6), 717–772.

Cranganu C., Luchian H., Breaban M. E., 2015. Artificial intelligent approaches in petroleum geosciences. Springer.

Dolton G. L., 2006. Pannonian Basin Province, Central Europe (Province 4808) – Petroleum geology, total petroleum systems, and petroleum resource assessment, *USGS Bull.* 2204–B, 1–47.

Gemulla R., Nijkamp E., Haas P. J., Sismanis Y., 2011. Large-scale matrix factorization with distributed stochastic gradient descent. In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 69–77.

Hartigan J. A., Wong M. A., 1979. Algorithm AS 136: A k-means clustering algorithm. Journal of the Royal Statistical Society. Series C (Applied Statistics), 28.1, 100–108.

Hempkins W. B., 1978. Multivariate statistical analysis in formation evaluation. SPE California Regional Meeting, San Francisco, 7144–MS.

556 Jarzyna J. A., Bała M., Krakowska P. I., Puskarczyk E., Strzępowicz A., Wawrzyniak-Guz K., Więcław
557 D., Ziętek J., 2017. Shale Gas in Poland, Advances in Natural Gas Emerging Technologies. IntechOpen,
558 DOI: 10.5772/67301.

559 Jolliffe I.T., 2002. *Principal component analysis*, 2nd edition, New York, Springer-Verlag.

560 Jung H., Jo H., Kim S., Lee K., Choe J., 2018. Geological model sampling using PCA-assisted support
561 vector machine for reliable channel reservoir characterization. Journal of Petroleum Science and
562 Engineering 167, 396–405.

563 Lawley D. N., Maxwell A. E., 1962. Factor analysis as a statistical method. The Statistician 12, 209–
564 229.

565 Little R., Rubin B., 1986. Statistical Analysis with Missing Data, John Wiley & Sons, Inc., New York.

566 Ma Z., Holditch S., 2015. Unconventional Oil and Gas Resources Handbook: Evaluation and
567 Development. Gulf Professional Publishing.

568 Masoudi P., Aïfa T., Memarian H., Tokhmechi B., 2018. Uncertainty assessment of porosity and
569 permeability by clustering algorithm and fuzzy arithmetic. Journal of Petroleum Science and
570 Engineering 161, 275–290.

571 Mazumder R., Hastie T., Tibshirani R., 2010. Spectral regularization algorithms for learning large
572 incomplete matrices. Journal of Machine Learning Research 11, 2287–2322.

573 Skalinski M., Gottlib-Zeh S., Moss B., 2006. Defining and predicting rock types in carbonates –
574 Preliminary results from an integrated approach using core and log data from the Tengiz Field.
575 Petrophysics 47 (1), 37–52.

576 Steiner F., 1991. The most frequent value: Introduction to a modern conception of statistics. Akadémiai
577 Kiadó, Budapest.

Szabó N. P., Dobróka M., Kavanda R., 2013. Cluster analysis assisted float-encoded genetic algorithm for a more automated characterization of hydrocarbon reservoirs. Intelligent Control and Automation 4 (4), 362–370.

Szabó N. P., Dobróka M., 2017. Robust estimation of reservoir shaliness by iteratively reweighted factor analysis. Geophysics 82 (2), D69–D83.

Szabó N. P., Dobróka M., 2018. Exploratory factor analysis of wireline logs using a Float-Encoded Genetic Algorithm. Mathematical Geosciences 50 (3), 317–335.

Zhang J., 2017. Most frequent value statistics and distribution of 7Li abundance observations. *Monthly Notices of the Royal Astronomical Society* 468 (4), 5014–5019.

Appendix

A simplified pseudo-code of the imputation algorithm applied in the proposed workflow, where *m* denotes the total number of columns of the input data matrix (i.e., measured petrophysical parameters, here variables *x* and *y*) and *n* is the total number of rows of the same matrix (i.e., core samples).

```
for i := 1 to m - 1
      for j := i + 1 to m
      x := i-th column of the data matrix
      y := j-th column of the data matrix
      points = {(x_k, y_k) : x_k != NaN, y_k != NaN, 1 <= k <= n}
      fit linear model to points by linear regression
      for k := 1 to n
            if x_k != NaN and y_k == NaN
                  y_k := fitted value by the linear model from x_k
              end
            if x_k == NaN and y_k != NaN
                  x_k := fitted value by the linear model from y_k
            end
```

```
607         end
608         end
609     end
```