



## Research article

## Robust reservoir identification by multi-well cluster analysis of wireline logging data

N.P. Szabó<sup>\*</sup>, R. Kilik, M. Dobróka

University of Miskolc, Institute of Exploration Geosciences, Department of Geophysics, 3515, Miskolc-Egyetemváros, Hungary



## ARTICLE INFO

## Keywords:

K-MFVs cluster analysis  
Most frequent value  
Well logging  
Rock typing  
Robustness

## ABSTRACT

A novel clustering method is applied to well logs for improved rock type identification in hydrocarbon formations. For grouping the objects in the multi-dimensional data space, we propose a Most Frequent Value (MFV) based clustering technique applied to natural gamma ray, bulk density, sonic, photoelectric index, and resistivity logs. The MFV method is a robust estimator, which assists in finding the cluster centers more reliably than a more noise sensitive K-means clustering approach. The result of K-means cluster analysis highly depends on the choose of the initial centroids. To reduce the risk of inappropriately chosen starting values, we apply a histogram-based selection method to give the best position of the initial cluster centers. We assure the robustness of the solution by calculating the centroid as the MFV of the cluster elements and defining the overall deviation of cluster elements from the center by a weighted Euclidean (Steiner-) distance. The proposed workflow relies on a fully automated weighting of the cluster elements, which does not require a constraint on the statistical distribution of the observed variables. The processing of synthetic data shows high noise rejection capability and efficient cluster recognition even beside considerable amount of outlying and missing data; the accuracy is measured by the difference between the estimated and the exactly known distribution of cluster numbers. The clustering tool is first applied to single borehole data, then the procedure is extended to multi-well logging datasets to reconstruct the multi-dimensional spatial distributions of clusters revealing the lithological and petrophysical characteristics of the studied formations. A large in situ dataset acquired from several boreholes traversing Hungarian gas-bearing clastic reservoirs of Miocene age is analyzed. The accuracy of the field results is confirmed by core permeability measurements, independent well log analysis and a gradient metrics characterizing the noise rejection capability of the clustering method.

## 1. Introduction

Probabilistic data analysis techniques like exploratory statistical methods and inverse modeling have been applied to oilfield well logs since the 1980s when the development of computer-based interpretation software- and modern expert systems was started [1–3]. An early study [4] referred cluster analysis as a promising tool for the recognition of producing hydrocarbon zones. After some early small-scale pattern recognition problems and cross-plot analysis based on a few types of well logs, technological developments of the last decades have made it possible to interpret big multi-dimensional wireline logging datasets jointly and quickly. Today, cluster

<sup>\*</sup> Corresponding author. Institute of Exploration Geosciences, University of Miskolc, 3515, Miskolc-Egyetemváros, Hungary.  
E-mail address: [norbert.szabo@uni-miskolc.hu](mailto:norbert.szabo@uni-miskolc.hu) (N.P. Szabó).

analysis is mentioned as a group of mathematical techniques of unsupervised machine learning and big data analytics [5], which is widely used in rock typing problems in petroleum geosciences [6–8]. Various machine learning algorithms have been applied for lithofacies identification and characterization. A reservoir prediction method was developed based on unsupervised learning and color feature blending, where several seismic attributes were extracted using cluster analysis to highlight oil and gas anomalies [9]. Non-hierarchical cluster analysis was used for assisting permeability prediction with transforming the well logs into electrofacies in dolomite and sandstone intervals in the Ogallah Field, USA [10], specifying the facies for a well in sandstone formation in West Africa before predicting the formation permeability [11], and the identification of heterogeneous carbonate reservoirs in a Southern Iraqi oilfield [12]. Other recent well log applications include improved electrofacies identification and lithology classification [13,14], assisting pseudo-well stochastic seismic inversion [15], automated layer-thickness determination for inversion procedures and estimation of typical log response values of hydrocarbon formations [16], clustering of incomplete core laboratory datasets [17], sweet spot identification and separation of different gas-bearing intervals in unconventional reservoirs [18–20]. As new alternative, machine learning tools can help to solve geophysical inverse problems. A genetic algorithm-based hyperparameter estimation method was applied to estimate the zone parameters (matrix and clay properties) of shallow unconsolidated formations [21]. The special feature of the inversion method was that these parameters were extracted directly from well logs instead of measuring them in the laboratory. Special parameters of objective functions and physical quantities of modeling equations treated previously as constant can be predicted by this technique, too. This approach can be implemented in the future to unconventional hydrocarbon and geothermal reservoirs, where the lithology classification and formation boundary detection made by cluster analysis can significantly help to increase the accuracy and reliability of inversion estimations.

The K-means clustering method performs the partitioning of the data space in such a way that it groups the data observations into a predefined number of clusters based on their similarity [22]. A cluster center is calculated as the mean value of the cluster elements, which is relatively noise sensitive process and gives optimal solution only for Gaussian distributed data. Several attempts have been made to modify the standard algorithms of cluster analysis to give a more robust solution. In practice this may be justified when the assumptions on Gaussian statistics are not fulfilled, or the measured dataset includes a certain number of outlying observations. For example, evolutionary computation as global optimization technique is applied to optimize the fitness function called variance ratio criterion for achieving optimal internal cohesion and external isolation of the clusters [23]. The outlier rejection is found to be an important step before or during cluster analysis. The maximum likelihood estimator was applied to detect outliers and simultaneously partition their complement into some clusters [24]. An adaptive outlier detection technique for exploration geochemical observations was presented in Ref. [25]. A forward search method using the robust Mahalanobis distance to cluster multivariate datasets was published in Ref. [26]. Cluster analysis was performed based on K-nearest neighbor search for improving earthquake mechanism identification and more reliable and stable pattern recognition for active seismicity regions [27]. Integrated petrophysical approaches were suggested in Refs. [28,29], where K-means cluster analysis serves as an important data pre-processing step in rock physical investigations to identify the main lithologies and potential hydrocarbon-bearing zones.

The Most Frequent Value (MFV) method as an efficient statistical estimator was originally developed for geophysical applications [30–32]. In modern statistics, one can use it to calculate the weighted average of observed data, where the weights are automatically optimized during an iterative process. The robust and outlier resistant nature of the MFV method has been proved in several scientific fields, e.g., in nuclear astrophysics and cosmology [33,34], hydrogeology [35] and astronomical geodesy [36]. In applied geophysics, the Steiner weights were employed to assure a reliable and stable joint inversion of direct current resistivity and vertical seismic profiling datasets collected in an underground coal mine [37] and to resolve the problem of ambiguity in 2.5 D inversion of apparent resistivity data collected in a thermal water exploration site [38]. The robust estimator assisted the iteratively re-weighted factor analysis of direct push logs to determine the spatial distribution of water saturation in near surface sediments [39]. Seismic tomography and image processing were successfully robustified by using the MFV technique [40,41]. An inversion-based Fourier transformation was developed for magnetic data processing, where the noise rejection capability of the reduction to pole operator was increased by involving the MFV technique when approximating the frequency spectra [42].

In this paper, the strategy of non-hierarchical cluster analysis (CA) is combined with the MFV estimator to interpret oilfield wireline logs. To give a robust solution, the  $K$  number of centroids are estimated as the MFV of cluster elements, respectively, instead of specifying them by the arithmetic mean. This modification justifies a new name given for the method since the K-MFVs CA method can be considered as an improved variant of the traditional K-means clustering method. The idea of this clustering concept was first applied to wireline logs to identify clayey-shaly Hungarian coal formations [43], and to classify organic-rich shale formations [44]. We further improve the MFV based CA method at some points. Among the others, we extend the CA algorithm to simultaneously process multi-well wireline logging data for non-equidistantly spaced drillings for the determination of 2D/3D subsurface distribution of the clusters. This technique works well for noisy and missing datasets as demonstrated. Then, we reduce the initial model dependence of clustering by setting the centroids by a histogram-based weighted median filtering method suggested originally for noise reduction in modeling digital elevation data [45]. We apply the same technique to narrow the possible range of candidate solutions and find proper starting centroids to get an optimal solution. An MFV-based norm called Steiner-distance is used to measure the similarity between the data objects and give an outlier-resistant solution.

In the conducted workflow an unsupervised machine learning method is used for a more robust identification of hydrocarbon reservoirs. Open-hole wireline logs form the input of non-hierarchical K-MFVs CA procedure, which finds the similarities in the multivariate dataset. The number of clusters are preliminary set by mathematical and geological assumptions. We apply a histogram-based selection process to find optimal values of initial centroids. Unlike the K-means clustering process, we apply an iterative re-weighting to each cluster elements in an inner loop of cluster analysis. In calculating a distance metric, K-means approaches give equal weights to each element, while our MFV estimator gives individual weights to them in an automated weighting procedure. In our

algorithm, both the cluster centers and the inter-cluster distances are specially weighted to reduce the harmful effects of outlying data. The output gives the depth variation of cluster numbers along a well (or several boreholes), effectively cleaning the data from the abrupt changes. The advantage of using the above workflow is the increased noise suppression capability and better vertical resolution, however this robust clustering approach is much more time consuming than the traditional K-means methods. After demonstrating the operation of the MFV clustering method through a permeability estimation problem, we present the results of numerical experiments made to test the performance of the 1D K-MFVs CA and 2D K-MFVs CA methods for rock typing. A real-data case study from a Hungarian hydrocarbon field demonstrates the feasibility of the 2D clustering method.

## 2. Methods

### 2.1. The MFV method

To introduce the concept of most frequent value, consider the following Cauchy-type weight function

$$\varphi(d) = \frac{\varepsilon^2}{\varepsilon^2 + (d - M)^2}, \quad (1)$$

where  $d$  denotes the physical variable observed by a geophysical tool and  $\varepsilon$  is the scale parameter called dihesion. Parameter  $M$  represents the symmetry point of the probability distribution function  $f(d)$

$$M = \frac{\int_{-\infty}^{\infty} \varphi(d) df(d) dd}{\int_{-\infty}^{\infty} \varphi(d) f(d) dd} \quad (2)$$

leading to an implicit expression for estimating  $M$  in Eq. (2). For quantifying the effective number of data playing significant role in calculating  $M$ , we can define quantity  $\xi_{eff}(\varepsilon)$  from the sum of weights in Eq. (1) [30]. The optimal value of dihesion is found at the maximum of the expression  $\xi_{eff}(\varepsilon)/\varepsilon$ , which requires the following objective function to be optimized

$$\Psi = \int_{-\infty}^{\infty} \frac{\varepsilon^{3/2}}{\varepsilon^2 + (d - M)^2} f(d) dd = \max. \quad (3)$$

Technically we simplify Eq. (3) by setting  $M$  to 0, then differentiate function  $\Psi$  with respect to parameter  $\varepsilon$  and make it equal to zero

$$\int_{-\infty}^{\infty} \left\{ \frac{(3/2)\varepsilon^{1/2}(\varepsilon^2 + d^2) - 2\varepsilon^{5/2}}{(\varepsilon^2 + d^2)^2} \right\} f(d) dd = 0. \quad (4)$$

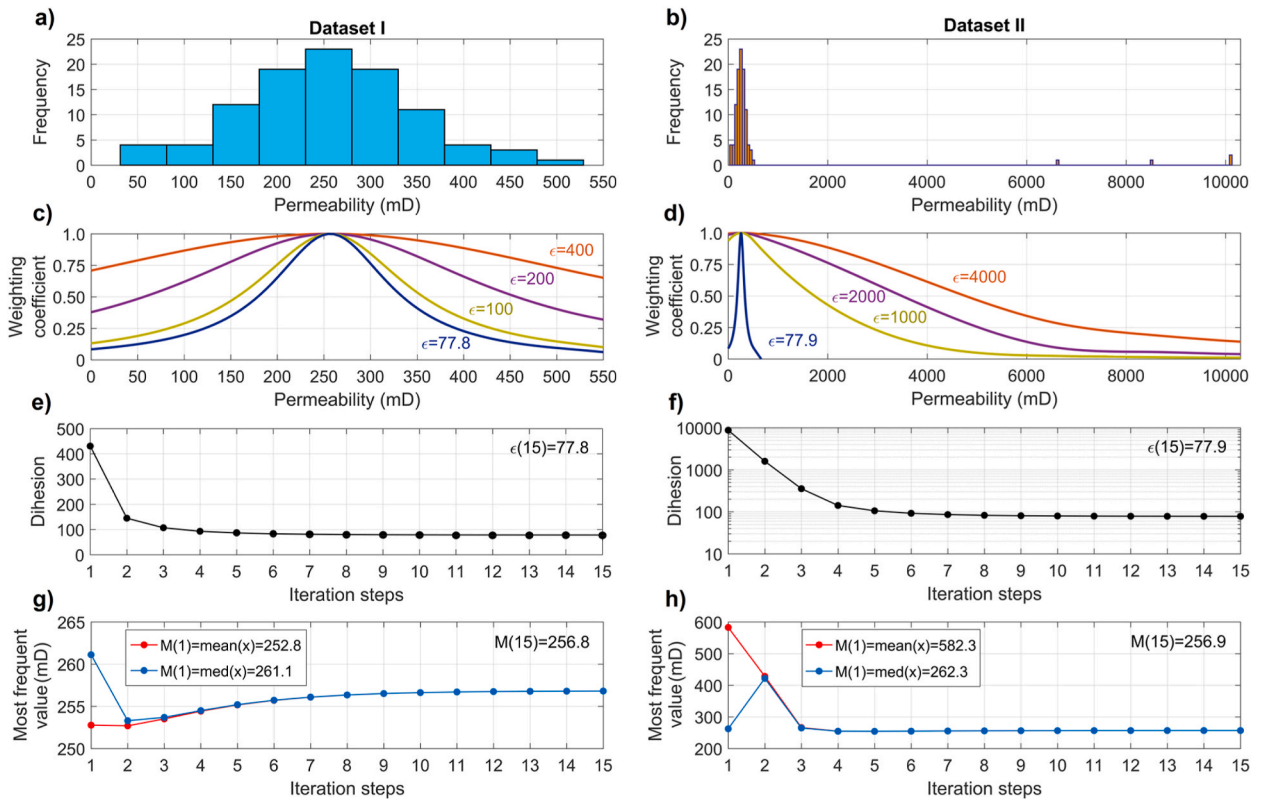
After reorganizing Eq. (4), the square of dihesion can be derived explicitly

$$\varepsilon^2 = \left( \frac{3 \int_{-\infty}^{\infty} \frac{d^2}{[\varepsilon^2 + d^2]^2} f(d) dd}{\int_{-\infty}^{\infty} \frac{1}{[\varepsilon^2 + d^2]^2} f(d) dd} \right). \quad (5)$$

By replacing the expression  $d^2$  with  $(d - M)^2$ , the values of the dihesion and quantity  $M$  as the most frequent value can be derived from Eqs. (2) and (5). For discrete datasets, they are automatically estimated in an iterative process [31]

$$\varepsilon_q = \left( \frac{3 \sum_{j=1}^L \frac{(d_j - M_{q-1})^2}{[\varepsilon_q^2 + (d_j - M_{q-1})^2]^2}}{\sum_{j=1}^L \frac{1}{[\varepsilon_q^2 + (d_j - M_{q-1})^2]^2}} \right)^{1/2}, \quad (6)$$

$$M_q = \frac{\sum_{j=1}^L \left[ \frac{\varepsilon_q^2}{\varepsilon_q^2 + (d_j - M_{q-1})^2} \right] d_j}{\sum_{j=1}^L \left[ \frac{\varepsilon_q^2}{\varepsilon_q^2 + (d_j - M_{q-1})^2} \right]}, \quad (7)$$



**Fig. 1.** Calculating the most frequent value ( $M$ ) of a permeability dataset, (a) frequency plot of Dataset I, (b) that of Dataset II including outliers, (c)–(d) weighting functions at different values of dihesion ( $\epsilon$ ), (e)–(f) dihesion vs. iteration steps, (g)–(h) most frequent value vs. iteration steps. The result is numerically indicated as  $M(15)$  and  $\epsilon(15)$  for Dataset I and II, respectively.

where  $d_j$  denotes the  $j$ -th observed data ( $L$  is the total number of data),  $q$  is the actual iteration step. The added advantage of the MFV method is the automatic calculation of the scale parameter  $\epsilon$  during the iterative procedure, which allows the finding of optimal weights for the actual data set without the knowledge of its statistical distribution.

The MFV estimation is presented through an example involving the processing of absolute permeability data. Dataset I include 100 Gaussian distributed data, the empirical probability density function of which showing the relative occurrence of permeability values can be found in Fig. 1a. To test the noise sensitivity of the MFV method, we expanded the original dataset with 4 outliers to form Dataset II. The histogram of the latter dataset is given in Fig. 1b. The weight functions with different scale parameters in Eq. (1) are plotted in Fig. 1c and d. It is shown that parameter  $\epsilon$  controls the relative importance of the observations in the weighting process. At the beginning of the iterative process, we choose relatively high value of dihesion, when all data get approximately the same weight. With the development of convergence, at smaller values of dihesion, only data being closer to the value of  $M$  affect the estimation considerably. At the end of the procedure, the weight function has a maximum value in the center of gathering, while it quickly decreases to zero for outlying values. We select the initial value of dihesion in proportion with the range of the sample, which is continuously decreased in 15 iterations (Fig. 1e and f). The estimated value of dihesion is  $\epsilon = 78$ . The improvement of  $M$  as symmetry point of the weight function can be seen for Dataset I and II in Fig. 1g and h. Two different procedures run where the initial value of  $M$  is chosen as the mean and the median of the sample in the first iteration, respectively. For Gaussian distributed data, the MFV result is close to that of the mean and median. However, for datasets including some extreme values, it is found that the mean operator gives a poor estimate ( $E \sim 600$  mD) being sensitive to outliers, while MFV act as robust estimator ( $M = 257$  mD) and effectively exclude the harmful effect of inaccurately observed data.

## 2.2. The K-MFVs clustering method

Cluster analysis gathers data objects into different clusters based on their similarity measured by a properly chosen distance metric. Before doing this, all data points are compared, and a large distance matrix is constructed based on which the closest data objects are linked. As a result, a tree-structure is formed which holds detailed information on the hierarchy of data, but unfortunately nothing about the spatial distribution of clusters. For big data analysis, the above dendrogram-based hierarchical clustering is rather time consuming. In well logging applications, we prefer the use of a quicker non-hierarchical clustering approach that groups the data directly with properly changing the cluster centers during the iterations. This partitioning technique constructs a predefined number of



clusters using an evaluation criterion. This often requires the smallest deviation of cluster elements from the center and at the same time creating non-overlapping clusters spaced long distance apart. The greatest difficulty of the non-hierarchical procedure is the selection of the optimal number of clusters and the initial position of the centroids that can highly affect the solution. The former requires all the geological information and mathematical assumptions to be considered, while the latter requires the reduction of the noise sensitivity of the clustering process with the robustification of the K-means CA method.

We collect all the measured data into a common matrix ( $\mathbf{D}$ ), in which the element  $d_{il}$  indicates the observed value of the  $l$ -th physical variable at the  $i$ -th depth coordinate ( $l = 1, 2, \dots, L$ , where  $L$  is the total number of recorded wireline logs). In multi-borehole applications, index  $i$  runs over not only the depth interval of one but also several wells ( $i = 1, 2, \dots, N$ ). In the  $i$ -th row of the data matrix, vector  $\mathbf{d}^{(i)}$  represents an object in the  $L$ -dimensional data space, which may include several data types such as nuclear, electrical, radioactivity, acoustic and of other types, too. The similarity between the data objects  $\mathbf{d}^{(i)}$  and  $\mathbf{d}^{(j)}$  can normally be measured by the Minkowski distance

$$\delta^{(M)} = \left[ \sum_{l=1}^L |d_l^{(i)} - d_l^{(j)}|^p \right]^{1/p}, \quad (8)$$

where  $p = 1$  gives the Manhattan (city-block) distance and  $p = 2$  corresponds to the Euclidean distance. The latter norm gives optimal solution only when the data follow Gaussian distribution. When the observations are not Gaussian distributed, the Manhattan distance performs better as being not so sensitive to outlying values. To give a more robust solution, we define the weighted distance between the data objects and the centroid in a given cluster

$$\delta^{(St)} = \left[ \left( \sum_{l=1}^L \varphi_l \right)^{-1} \sum_{l=1}^L \varphi_l (d_l - c^{(MFV)})^2 \right]^{1/2}, \quad (9)$$

where the weighting coefficients ( $\varphi$ ) are calculated using Eq. (1). (Formally, when the weighting coefficients are equal to 1 and the centroid is calculated as the mean of the cluster elements, the above formula leads to the Euclidean distance-based K-means CA method.) As a further modification of the K-means CA method, we make the cluster center ( $c$ ) equal to  $M$  using Eq. (7), instead of calculating it by the arithmetic mean of the cluster elements. We name the above similarity measure as Steiner-distance. The K-MFVs CA method seeks a pre-defined number of clusters ( $K$ ) including their initial elements and centroids. In each iteration, each object is assigned to the closest centroid forming a new cluster. The actual configuration is iteratively improved by re-calculating the positions of centroids. When they do not change significantly, the clustering procedure is stopped.

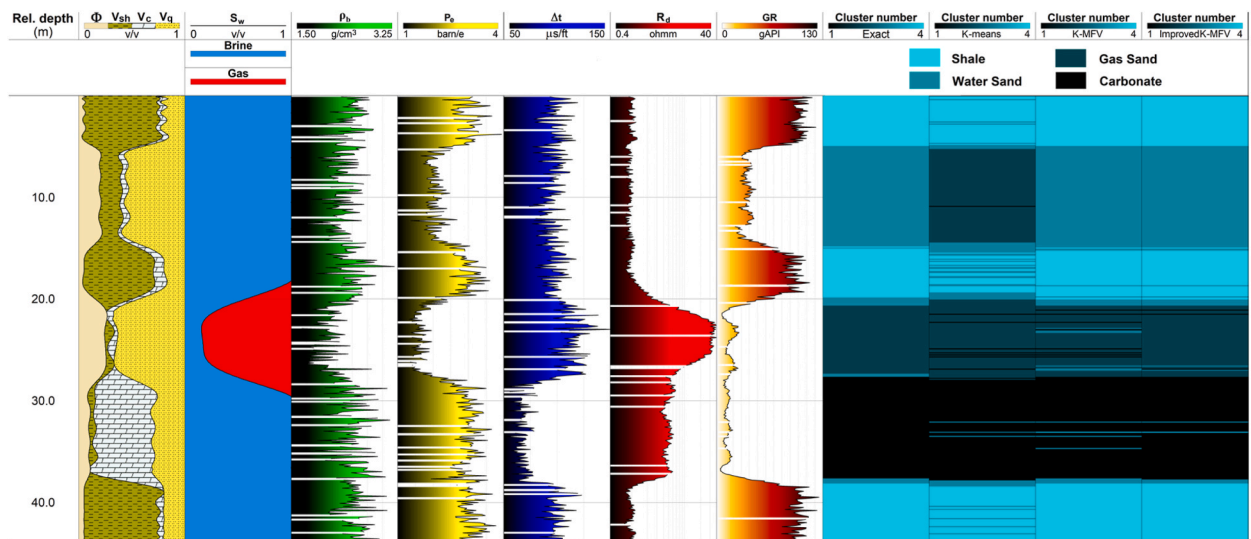
The optimal number of clusters is normally specified by calculating the Sum of Squared Error (SSE). In our algorithm, the SSE is calculated by involving the MFV method to calculate the centroids to measure the robust distance of the data objects to their closest centroids

$$\text{SSE} = \sum_{k=1}^K \sum_{i=1}^{I_k} \delta^{(St)2}(\mathbf{c}^{(MFV)(k)}, \mathbf{d}^{(i)}), \quad (10)$$

where  $\mathbf{c}$  represents the vector of  $K$  number of centroids ( $I_k$  is the total number of elements in the  $k$ -th cluster). To find the optimal solution, one must minimize the sum of the within-cluster variances calculated by Eq. (10). At the beginning of the process, the SSE vs. cluster number curve is usually used to select the optimal number of clusters. At the “elbow” of the SSE curve, one can find the optimal cluster number on the axis of abscissa. While this curve steadily converges to zero, at this selected point, there is no considerable reduction in the value of SSE. It must be mentioned that using smaller cluster numbers than the optimal one gives poor spatial resolution in rock typing, while choosing greater numbers may yield not existing lithological categories.

We optimize the selection of the initial centroids by narrowing their possible ranges using a histogram-based selection method proposed in Ref. [45]. The core of the multi-step process is based on the sorting of the elements and then calculating the weighted mean of elements to correct their central value at each position of a moving window going through the dataset. The possible outliers are effectively eliminated by the weighting process and the median is estimated of the reduced dataset using its histogram. In this paper, two independent window narrowing processes take place to estimate the initial centroids as the weighted mean of the result of the pure MFV technique and that of the histogram approach assisted MFV method. Both methods compute the centroids and partition the data around the cluster centers. However, the difference between MFV and the histogram modified MFV methods is that in the former the initial cluster centers are set by using the Manhattan distance, while for the latter the centroids are computed as the medians of the data partitions generated by the histogram operations. On the other hand, while the MFV method multiplies the elements of the partition vectors by automatically chosen weights when partitioning the data to the current centroid, this weighting does not occur for the modified method. The weights expressing the relative contribution of the two individual MFV methods are set empirically, and the quality of initialization can be checked as described in detail in Ref. [45].

In this study, the performance and applicability of the proposed K-MFVs CA method is tested on synthetic wireline logs. For solving rock typing problems, we suggest the use of the 1D and 2D variants of the robust clustering method. In addition to our detailed synthetic modeling experiments using noisy wireline logs, we demonstrate its practical feasibility to a multi-borehole dataset collected in a Hungarian hydrocarbon field.



**Fig. 2.** Result of cluster analysis of synthetic data, tracks 1–2 show the exactly known petrophysical model, tracks 3–7 include the incomplete input wireline logs contaminated by 10% Gaussian distributed noise, track 8 contains the exact distribution of clusters, track 9–11 show the cluster number logs estimated by K-means CA, K-MFVs CA, histogram-modified K-MFVs CA procedures, respectively.

**Table 1**

Zone parameters of wireline tool response functions chosen for calculating synthetic data to test the accuracy of the K-MFVs CA method.

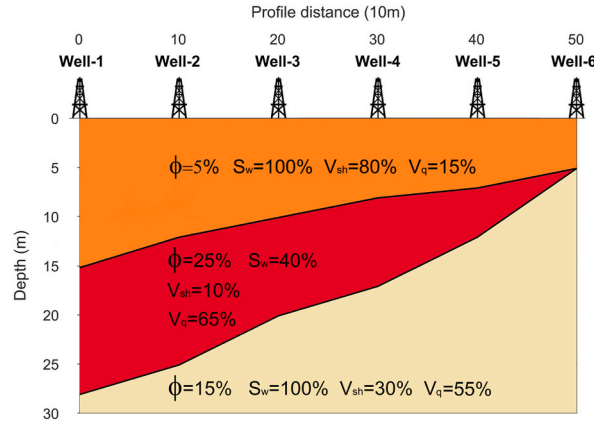
Wireline log	Functional constant	Symbol	Selected value	Unit
Natural gamma intensity (GR)	quartz	$GR_q$	10	API
	shale	$GR_{sh}$	140	
	calcite	$GR_c$	5	
Bulk density ( $\rho_b$ )	quartz	$\rho_q$	2.65	$g/cm^3$
	shale	$\rho_{sh}$	2.55	
	calcite	$\rho_c$	2.79	
	brine	$\rho_w$	1.09	
	gas	$\rho_g$	0.016	
Acoustic transit-time ( $\Delta t$ )	quartz	$\Delta t_q$	56	$\mu s/ft$
	shale	$\Delta t_{sh}$	108	
	calcite	$\Delta t_c$	46	
	brine	$\Delta t_w$	200	
	gas	$\Delta t_g$	305	
Photoelectric index ( $P_e$ )	quartz	$P_{e,q}$	1.81	barn/e
	shale	$P_{e,sh}$	3.50	
	calcite	$P_{e,c}$	4.11	
	brine	$P_{e,w}$	0.81	
	gas	$P_{e,g}$	0.09	
Deep resistivity ( $R_d$ )	shale	$R_{sh}$	1	$\Omega m$
	brine	$R_w$	0.06	
	cementation exponent	$m$	2.0	
	Saturation exponent	$n$	2.0	
	Tortuosity coefficient	$a$	1.0	

### 3. Test results

#### 3.1. 1D synthetic modeling experiments

In formation evaluation, cluster analysis can be used to separate different lithologies based on the similarity of well log responses along a borehole. To test the performance of the K-MFVs CA method a synthetic modeling experiment is made. We assume a fictive (exactly known) petrophysical model, which is used to calculate synthetic well logs. The cluster analysis of these noiseless data gives a solution for the depth distribution of clusters forming the target (known) cluster model. To simulate real geophysical measurements, one can add any amount of noise from known statistical distribution to the synthetic (noiseless) data to test how accurately and reliably the distribution of clusters is reconstructed by the K-MFVs CA method.

A six-layered inhomogeneous model is built representing a shaly calcareous sand formation (Fig. 2 tracks 1–2). The unit volume of rock is composed of pore space ( $\phi$ ) including water ( $S_w$ ) and gas ( $S_g$ ), quartz ( $V_q$ ), shale ( $V_{sh}$ ) and calcite ( $V_c$ ). (For the sake of simplicity,



**Fig. 3.** Two-dimensional three-layered (exactly known) petrophysical model for generating synthetic wireline logs for cluster analysis,  $\phi$  is porosity,  $S_w$  is water saturation,  $V_{sh}$  is shale volume,  $V_q$  is quartz content.

the invasion of drilling fluid into the permeable formations is neglected in this case.) The following tool response functions are used to compute the synthetic well logs [2].

$$\rho_b = \Phi [S_w \rho_w + (1 - S_w) \rho_g] + V_{sh} \rho_{sh} + V_q \rho_q + V_c \rho_c, \quad (11)$$

$$GR = \rho_b^{-1} (V_{sh} GR_{sh} \rho_{sh} + V_q GR_q \rho_q + V_c GR_c \rho_c), \quad (12)$$

$$P_e = \Phi [S_w P_{e,w} + (1 - S_w) P_{e,g}] + V_{sh} P_{e,sh} + V_q P_{e,q} + V_c P_{e,c}, \quad (13)$$

$$\Delta t = \Phi [S_w \Delta t_w + (1 - S_w) \Delta t_g] + V_{sh} \Delta t_{sh} + V_q \Delta t_q + V_c \Delta t_c, \quad (14)$$

$$R_d^{-1/2} = [R_{sh}^{-1/2} V_{sh}^{(1-V_{sh}/2)} + (a R_w)^{-1/2} \Phi^{m/2}] S_w^{n/2}, \quad (15)$$

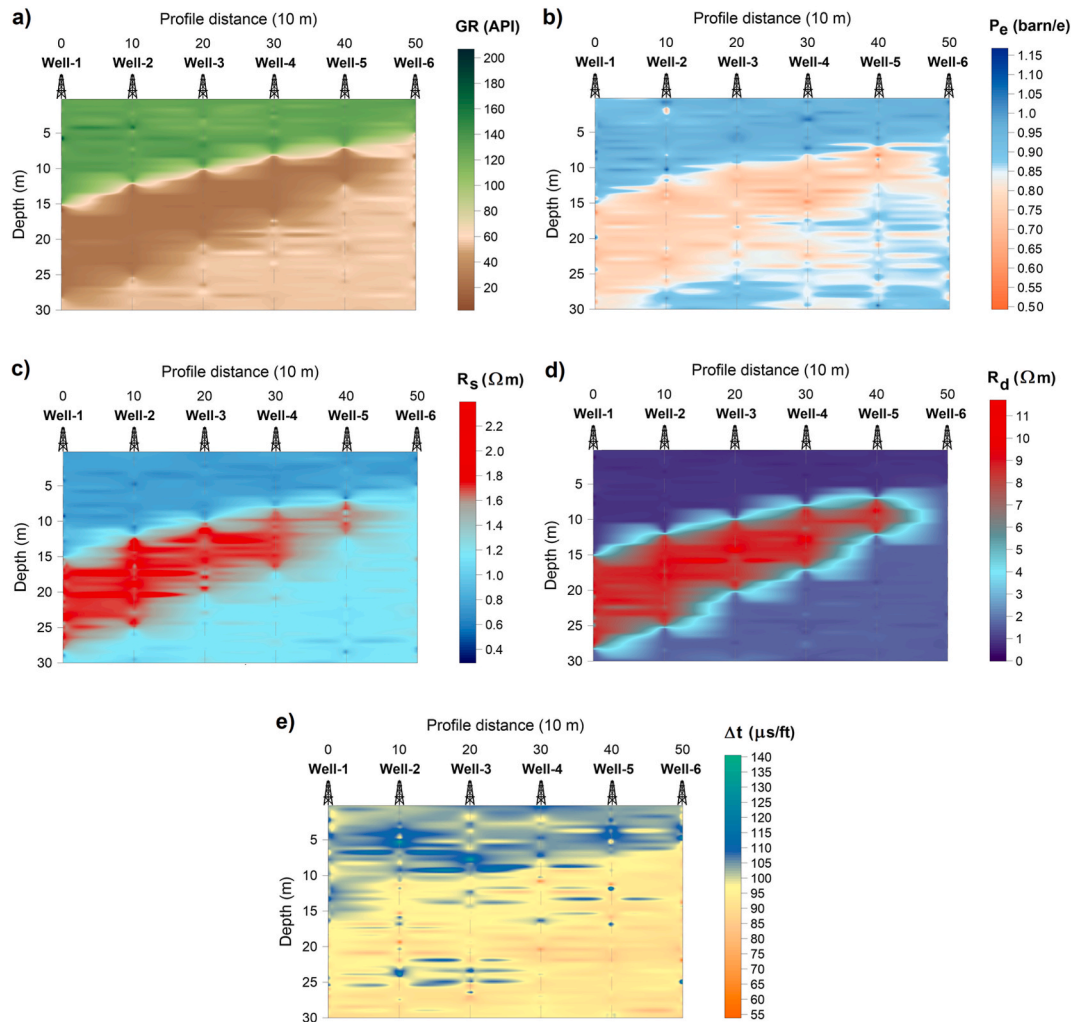
where the calculated data are formation (bulk) density ( $\rho_b$ ), natural gamma-ray intensity ( $GR$ ), photoelectric absorption cross-section index ( $P_e$ ), P-wave sonic travel-time ( $\Delta t$ ) and deep resistivity ( $R_d$ ). The function constants in Eqs. 11–15 are listed in Table 1, which were selected based on the result of well log inversion made in a similar geological formation [46] than studied in section 4.

The known values of the petrophysical parameters (tracks 1–2 in Fig. 2) are used to calculate synthetic well logs using Eqs. 11–15. The input logs of cluster analysis are generated by adding 10% Gaussian distributed noise to the synthetic data. In addition to it, we randomly select 1/20 part of the total data and delete them from the dataset shown by white stripes in tracks 3–7. (We treat them as NaN values in MATLAB.) The K-means CA method requires a complete data matrix, which was built by a correlation-based imputation method suggested originally to core data in Ref. [17]. The multivariate statistical regression-based approach effectively replaces the missing data with synthetic ones estimated from the available observed data up to 50–70% incompleteness and having poor a priori information to the correlated variables. After this phase, we run the K-means, K-MFVs and the improved (weighted) K-MFVs CA methods, separately.

The cluster analysis of noiseless data gives the vertical distribution of the four clusters (i.e., the depth variation of cluster numbers) on track 8, which is identical to the exact solution. The rock types along the processed interval from the shallow depths are: 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. The result of cluster analysis is included in tracks 8–11, where four clusters are assumed given on a lithological basis. The limestone layer is associated to black color, the calcareous sediments are lighter colors, while the hydrocarbon-bearing sand is illustrated by dark blue (track 8). By comparing the results of the K-means CA and the two MFV-based CA solutions, we can see that the traditional K-means method is more sensitive to data noises, confuses the water reservoir with the hydrocarbon reservoir in many cases, delays layer boundaries, and its estimation for the effective layer thickness are not so accurate. The latter is conspicuous in the case of the high porosity (2nd and 4th) layers. The improved K-MFVs CA method seems to reconstruct the cluster model the best (see tracks 8 and 11), its resistance against the peaky noises is the highest and can define the reservoir zones with the smallest error.

### 3.2. 2D synthetic modeling test

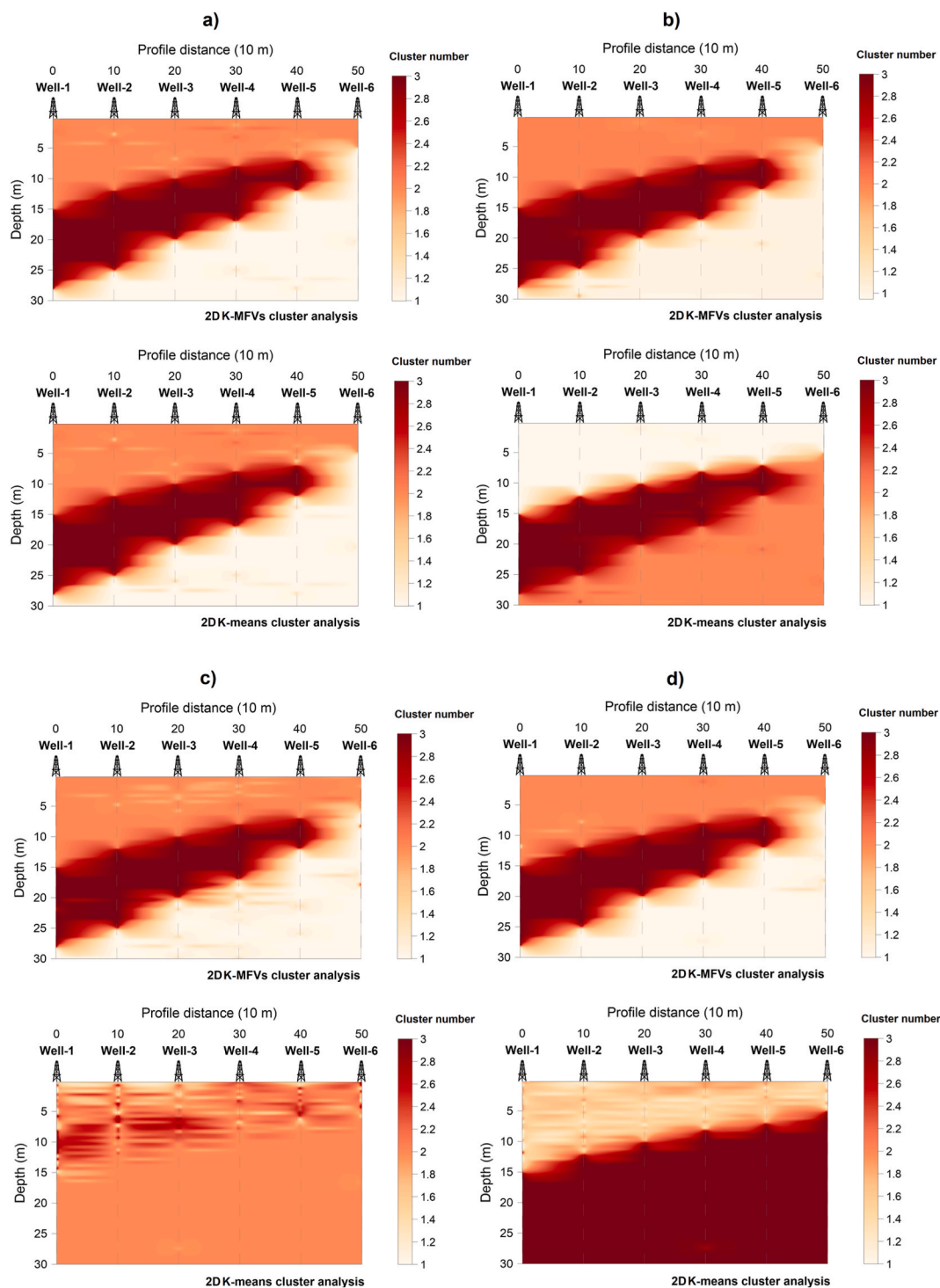
We extend the robust clustering method to multi-dimensional applications by integrating all data of several wells drilled along a line on the earth's surface. A 2D three-layered petrophysical model is constructed with the volumetric parameters quantitatively given in Fig. 3. By starting from the top, the sequence includes a shale, a hydrocarbon sand, and a tight water-bearing sand formation, respectively. By using these model parameters, we generate the synthetic logs  $\rho_b$ ,  $GR$ ,  $P_e$ ,  $\Delta t$ ,  $R_d$  by applying Eqs. 11–15. The zone



**Fig. 4.** Synthetic wireline logs calculated over the two-dimensional three-layered petrophysical model contaminated with 5% Gaussian distributed noise, (a) natural gamma-ray intensity (GR), (b) photoelectric absorption index ( $P_e$ ), (c) shallow resistivity ( $R_s$ ), (d) deep resistivity ( $R_d$ ), (e) sonic travel-time ( $\Delta t$ ).

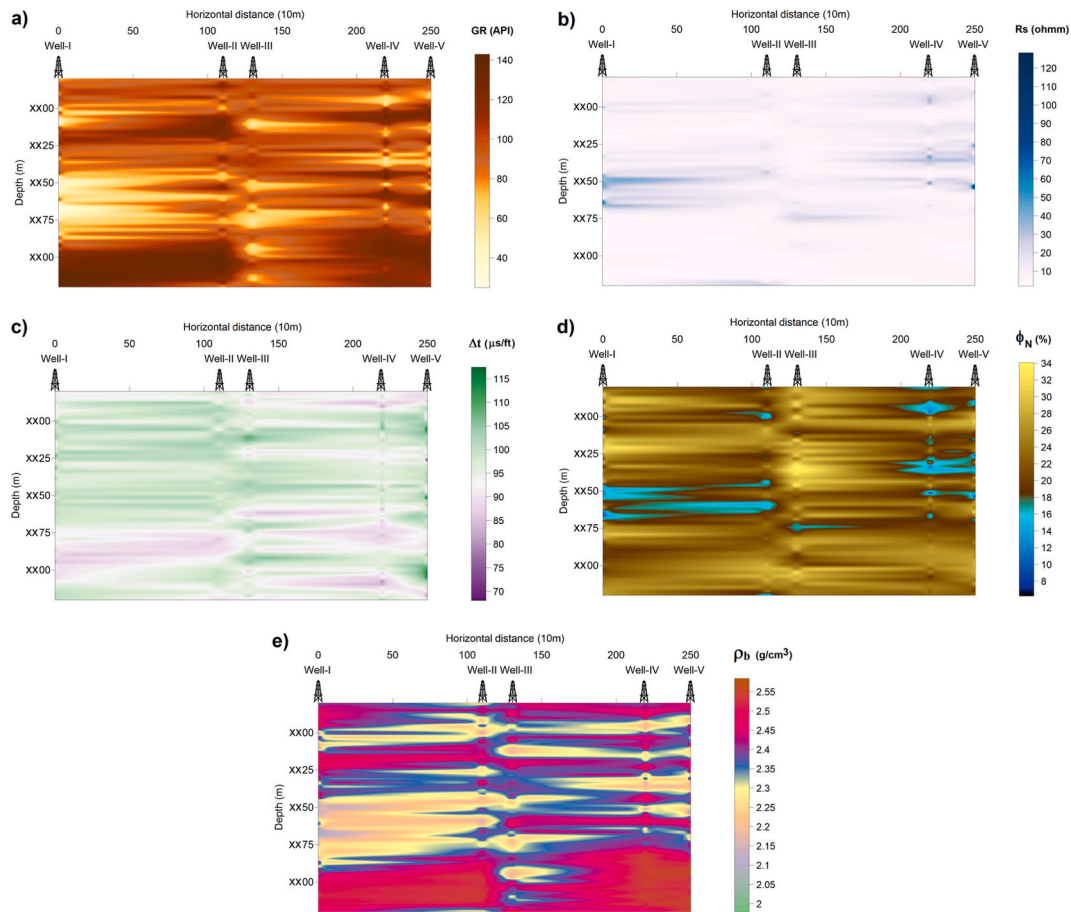
parameters of response equations are listed in Table 1. Data samples of the same type of wireline logs are collected into the same column of a data matrix forming the input of the 2D cluster analysis procedure. By selecting a sampling distance for vertical direction 0.1 m, and 100 m separation between the drill-holes (Well-1–Well-6), we process a total number of 8970 data. A wireline logging dataset contaminated by 5% Gaussian noise is illustrated as 2D sections in Fig. 4.

In addition to basic rock typing, it is particularly important how accurately cluster analysis can separate the hydrocarbon-bearing layer with a saturation of 60% ( $1 - S_w$ ). The sensitivity of the traditional K-means CA procedure to the initial centroids is also tested. For this reason, we generate four different datasets by adding 5% Gaussian distributed noise to the synthetic data, respectively. To simulate the real oilfield conditions, these data matrices were 90% complete including 10% missing data. To enrich the input dataset, we use the correlation-based imputation method suggested in Ref. [17]. Then, four different computer runs were made using the K-means CA and the improved K-MFVs CA methods to compute the 2D spatial distribution of clusters (Fig. 5a–d). According to the consistent results of K-MFVs CA method the clusters are water sand (Cluster 1), shale (Cluster 2), hydrocarbon reservoir (Cluster 3). The clustering results included in Fig. 5a show almost equally good estimations for both methods (K-MFVs CA is slightly less noisy). The layers are identified correctly by both. In that case, the random positions of centroids for the K-means CA method are properly (luckily) chosen. The same output is given in Fig. 5b; however, the shale and water sand layers are mixed by the traditional clustering method. In the following two plots, the K-MFVs CA method correctly reconstructs the layers, while the K-means CA method cannot make it. In the bottom panel of Fig. 5c, the shale dominates, and the hydrocarbon layer is mixed with the groundwater formation appearing on the top of the processed depth interval. In Fig. 5d, K-MFVs CA still gives reliable results independently from the initial centroids, but the K-means CA incorrectly identifies two hydrocarbon reservoir formations instead of one. It can be stated that the MFV-based clustering method behaves more robust against the random noises and identifies the different lithologies more reliably than the traditional K-means CA



**Fig. 5.** Results of two-dimensional cluster analysis, (a)–(d) MFVs (top panel) and K-means (bottom panel) CA procedure with different synthetic datasets including 5% Gaussian distributed noise and 10% missing input data.





**Fig. 6.** Interpolated cross-sections of observed well logging parameters, (a) natural gamma-ray intensity (GR), (b) shallow resistivity ( $R_s$ ), (c) acoustic travel-time ( $\Delta t$ ), (d) neutron porosity ( $\phi_N$ ), (e) bulk density ( $\rho_b$ ).

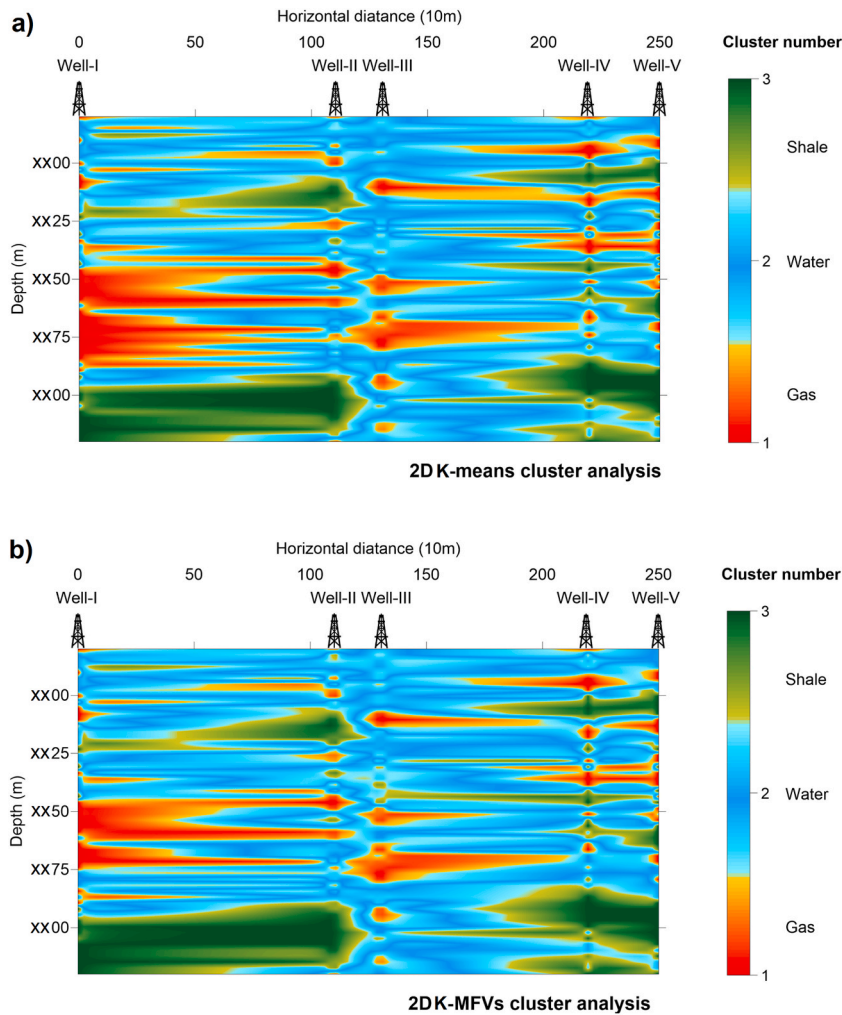
method.

#### 4. Field results

We test the suggested clustering approach using in situ wireline logs collected from an East-Hungarian gas field. Along a 140 m long depth interval of Miocene shaly (silty) sand sequences, we process simultaneously open-hole well-logging data measured in five boreholes (Well-I–Well-V). The horizontal coordinates of the drill-holes along a line are 0 m, 1100 m, 1300 m, 2200 m, 2500 m. The input logs are natural gamma-ray intensity (GR), shallow resistivity ( $R_s$ ), acoustic travel-time ( $\Delta t$ ), neutron porosity ( $\phi_N$ ) and bulk density ( $\rho_b$ ). The basis of selecting the well logs is that these data types are complete in all boreholes. The 2D distribution of observed variables is obtained by interpolating the same type of well logs between the boreholes using ordinary kriging (Fig. 6). Based on the GR plot, the permeable and impermeable zones can be clearly separated, we use the assumption that there are no other radioactive minerals other than clay. The possible gas reservoirs are shown with dark blue colors in the  $R_s$  section. On the  $\Delta t$  map, rocks with higher primary porosity are shown with purple color, while those with lower porosity and higher fluid content appear with green color. The  $\phi_N$  log is highly sensitive to hydrogen content; thus, the gas zones (bright blue color) can be well separated from the formations with higher total porosity formations including crystalline and pore-water (brown color). On the  $\rho_b$  profile, the clays appear with higher density, while the rocks with lower bulk density values are the fluid-saturated permeable rocks. Gas indications were previously confirmed by quantitative formation evaluation [46]. We run the clustering process using the K-means and improved K-MFVs CA methods, separately. We assume three clusters to differentiate the gas reservoirs (Cluster 1) from water saturated sands (Cluster 2) and shale beds (Cluster 3). Fig. 7 shows that both methods give consistent results. Small differences can be seen in the delineation of gas saturated zones.

For validating the results of cluster analysis, we use the permeability logs estimated earlier by factor analysis of raw wireline logs [47]. It was shown that the first statistical factor correlates highly to formation permeability calculated by Timur's equation [48] and sidewall core data in the same area. In this research, we remove the GR data from the input data matrix and repeat the 2D CA calculations. By reducing the lithological effect, one can find a good correlation between the resultant cluster numbers and the magnitude





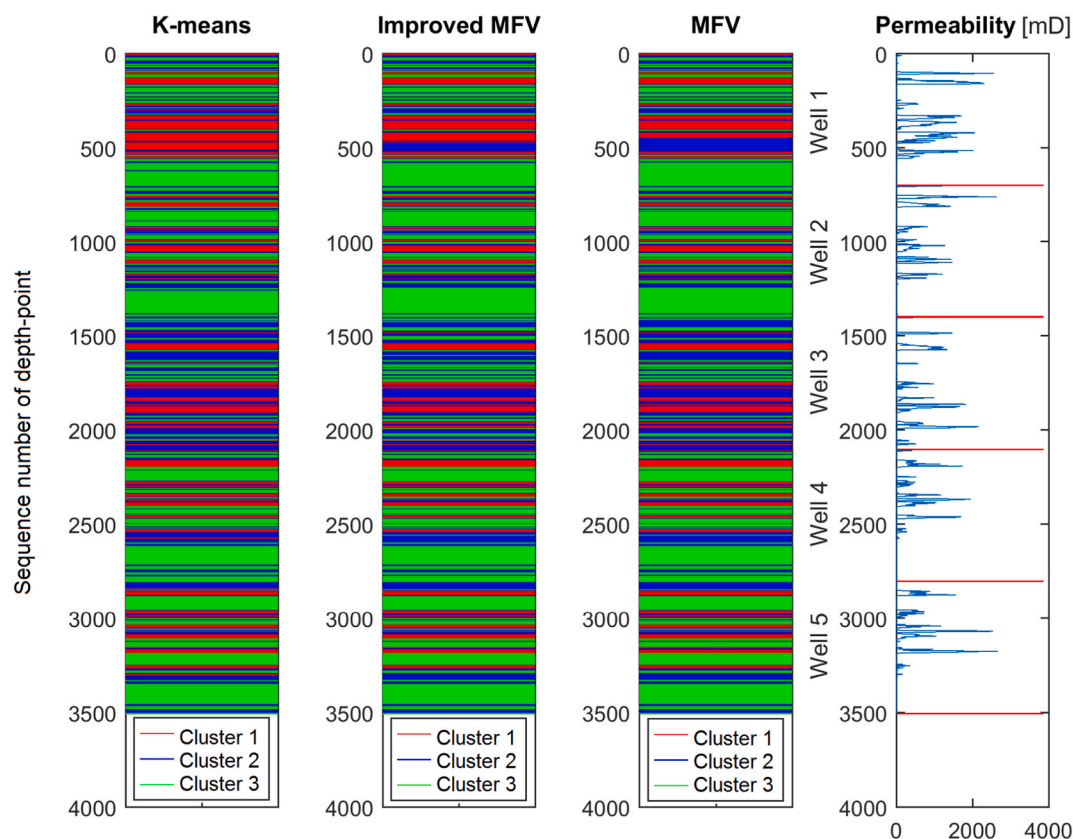
**Fig. 7.** Results of two-dimensional cluster analysis, (a) K-means CA procedure, (b) most frequent value-based CA process, cluster numbers are: 1 (gas-bearing layers), 2 (water saturated permeable formations), 3 (clayey and silty impermeable formations).

of absolute permeability (Fig. 8). The K-means, pure MFV-based and improved MFVs cluster analysis give almost the same results. They identify the gas zones at high permeability values (Cluster 1). Blue intervals (Cluster 2) can be associated with brine saturated formations with moderate/high permeability, while green ones (Cluster 3) show impermeable shaly formations. Minor differences in the vertical distribution of clusters can be observed in depth intervals of thin layers. The distance correlation between the identified clusters and permeability is  $-0.62$  (K-means CA),  $-0.63$  (MFV-CA),  $-0.64$  (improved K-MFVs CA).

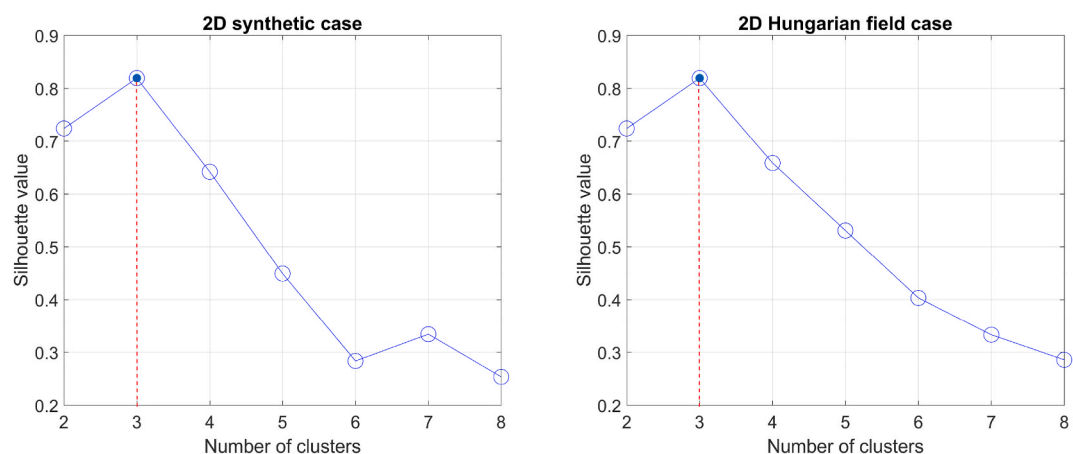
## 5. Discussion

The selection of the optimal number of clusters is of high importance, to which mathematical and geological assumptions should be considered simultaneously. When assuming too small number of clusters in non-hierarchical clustering, it may give results with poor vertical resolution. In the reverse case, specifying more groups than needed clustering may generate non-existing rock types. In the synthetic examples, we designated the number of clusters to be determined (sections 3.1-3.2). The well-known elbow technique using the SSE vs. cluster number plot gives a good approximation, which can be improved using Eq. (10) in calculating the values of SSE. We confirm our selections by applying the silhouettes suggested in Ref. [49]. The silhouette value shows the strength of cohesion of an object to its own cluster in the space of observed data, which ranges between  $-1$  and  $1$ . The resultant cluster configuration is appropriate for high values, in the case when the objects are like the others in the cluster and have poor connection to other clusters. By calculating the silhouette values using the city block (Manhattan) distance metric at different cluster numbers, we obtain 3 optimal clusters for both the 2D synthetic example (section 3.2) and the field case (section 4). The silhouette plots for these problems can be found in Fig. 9.

The quality of 2D clustering is checked by performing synthetic modeling experiments (section 3.2). The accuracy of measured data

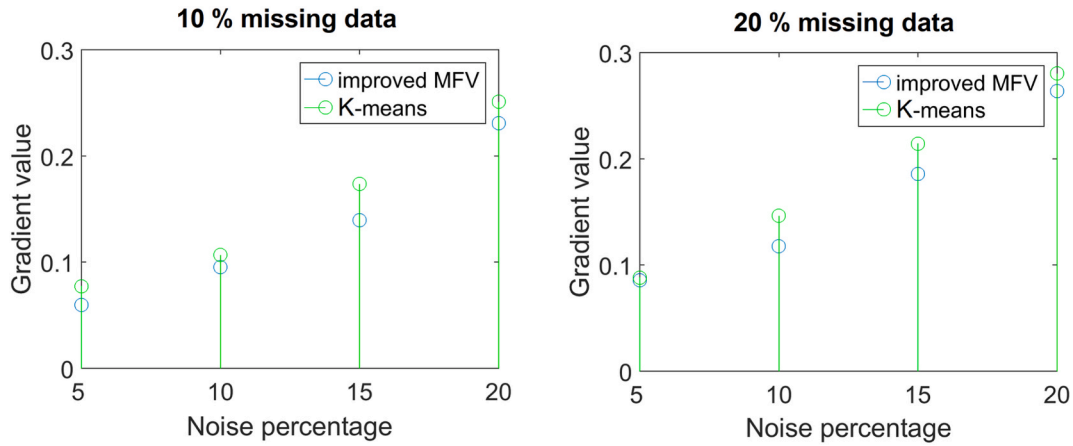


**Fig. 8.** One dimensional representation of clusters estimated by K-means CA (first track), K-MFVs CA (third track) and improved K-MFVs CA methods (second track), formation permeability derived from independent formation evaluation (last track) showing good correlation with the cluster analysis results.

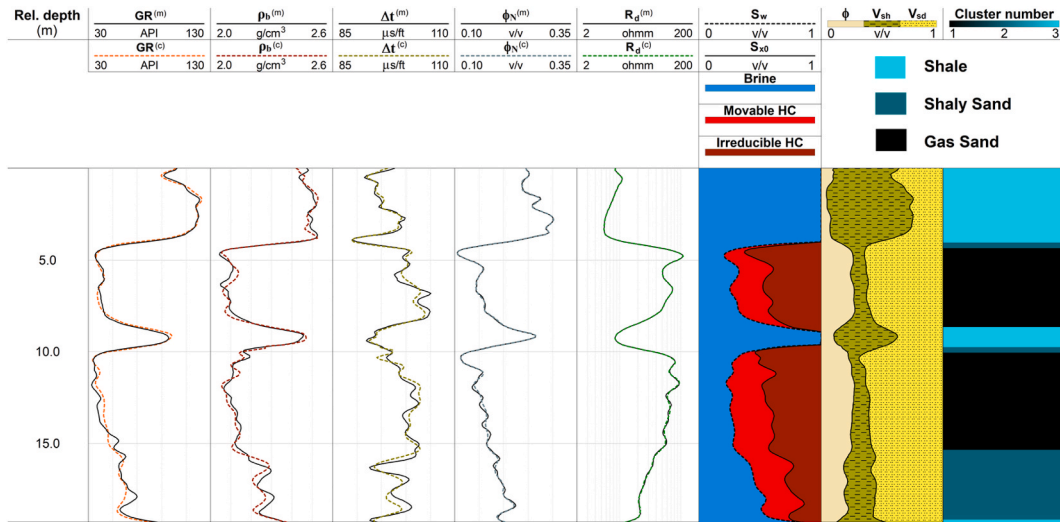


**Fig. 9.** Selection of optimal number of clusters in 2D clustering problems in this study using Rousseeuw's silhouettes method.

has a strong influence of the reliability of cluster definition. It is also shown that the resultant clusters can be compared to lithological characteristics or other important reservoir parameters (section 4). The correlation coefficient expresses the strength of these relations; however, it does not characterize the spatial resolution achieved by the given clustering process. This achievement strongly depends on the level (and statistical distribution) of the noise of input data. As another alternative to test the noise rejection capability of the 2D clustering methods, a simple metric for the characterization of the vertical variability of clusters is suggested. We define the following gradient value



**Fig. 10.** Quality check of cluster numbers by the gradient value derived from the K-means and K-MFVs CA results (ordinate axis) at different Gaussian distributed noise levels (axis of abscissae) and incompleteness of the input data matrix.

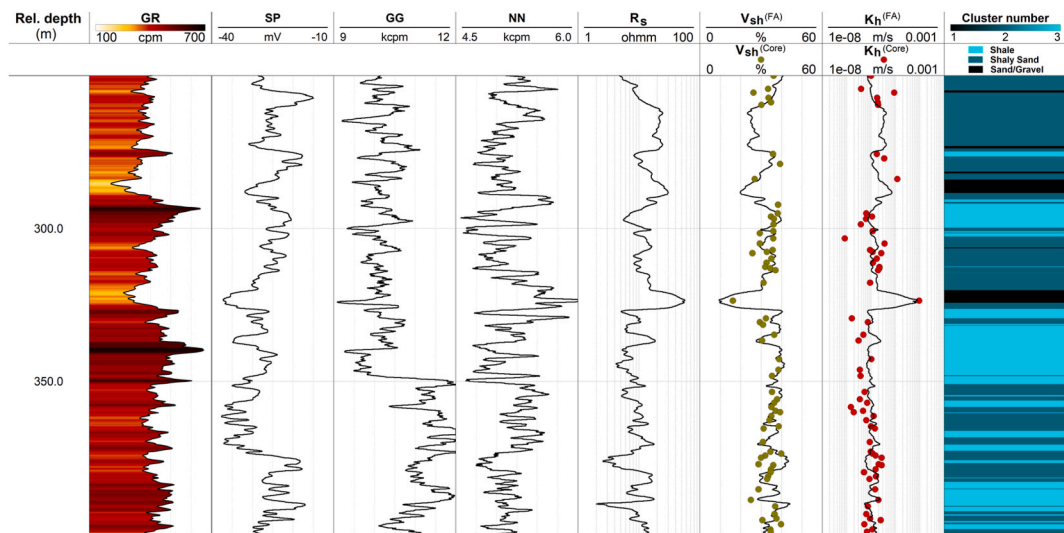


**Fig. 11.** Result of cluster analysis in Well-I, tracks 1–5 show the measured (m) and inversion derived calculated wireline logs (c), track 6 contains the saturations of water and hydrocarbon, track 9 shows the fractional volume of pore space, shale and sand, track 10 includes the cluster number log estimated by the K-MFVs CA procedure.

$$grad(\mathbf{v}) = \frac{\sum_{i=1}^N \sqrt{\left(\frac{dv}{dz}\right)_i^2}}{N}, \quad (16)$$

where  $v_i$  is the  $i$ -th element of the cluster number vector  $\mathbf{v}$  ( $i = 1, 2, \dots, N$ ). By assuming the same 1D model introduced in section 3.1 (track 8 contains the exact distribution of clusters in Fig. 2), we calculate synthetic wireline logs using Eqs. 11–15. We create different input datasets by adding 5%, 10%, 15% and 20% Gaussian distributed noise to the noiseless data, respectively. At the same noise level, we consider 1/20 and 1/10 data losses, too. After data imputation using the algorithm [17], the K-means CA, K-MFVs CA and the improved K-MFVs CA methods were run, separately. The gradient measures for the vectors of cluster numbers are plotted in Fig. 10 to display the extent of their variability. Each value represents the average of 10 computer run results in the graphs. The variability of the improved K-MFVs CA is shown in blue and it is green for that of the K-means CA procedure. The experiment shows that the K-MFVs CA method is less affected by the data noises, which confirms the results shown in Fig. 2. The result is assumed to be independent on the cluster number distribution of the exact model, the gradient value expresses the uncertainty of the result of cluster analysis (density of incorrect thin stripes on the well log of cluster numbers) obtained at different accuracy of input data.

The clustering results can be confirmed by those of independent formation evaluation methods. In the investigated Hungarian hydrocarbon field, the cluster number logs given by the MFVs CA approach can be directly compared to core data or petrophysical parameters obtained from quantitative interpretation of well logs. As earlier, we determined the spatial distribution of three clusters



**Fig. 12.** Result of cluster analysis in a Hungarian water well, tracks 1–5 show the observed wireline logs, track 6 contains the shale volume estimated by factor analysis and core measurements, track 7 shows the hydraulic conductivity predicted by factor analysis and core measurements, track 8 illustrates the cluster number log estimated by the K-MFVs CA procedure.

along the borehole. We validate the results of cluster analysis with inversion estimates given in a short (18 m thick) interval in Well-1. In addition to the measured logs studied in section 4, the neutron porosity ( $\varphi_N$ ) and the shallow resistivity log ( $R_s$ ) was incorporated into the inversion process as input data (Fig. 11). The aim of inversion was to optimize the model parameters of a shaly sandy structure while minimizing the misfit between the measured and predicted data. Porosity ( $\varphi$ ), shale volume ( $V_{sh}$ ) and water saturation in the invaded ( $S_{xo}$ ) and virgin zone ( $S_w$ ), respectively, was estimated by depth-by-depth inversion, while the movable and irreducible hydrocarbon saturation and sand volume ( $V_{sd}$ ) were derived using the material balance equation [46]. By comparing tracks 6–7 to track 8 in Fig. 11, one can see that the clusters correlate well to the lithology and fluid types. Cluster 1 represents gas-bearing sand beds, Cluster 2 shows shaly sandy intervals and Cluster 3 gives the shaly (non-reservoir) zones. The reliability of cluster analysis is confirmed by a good agreement between the observed data and those calculated using the model parameters estimated by inversion (tracks 1–5).

Core laboratory measurements can be used as independent source of information to validate the clustering results. As an example, we apply the K-MFVs CA method to a single borehole dataset collected from a thermal water well in East Hungary, approx. 100 km northeast from the investigated oilfield. A shallow Miocene formation is given composed of dominantly of shale with some porous and permeable sandy interbeds. The processed well logs are gamma-ray image (GR), spontaneous potential (SP), gamma-gamma intensity (GG), neutron-thermal neutron intensity (NN) and shallow resistivity ( $R_s$ ). Shale volume [50] and hydraulic conductivity ( $K_h$ ) [51] is estimated by factor analysis of all well logs (Fig. 12). These parameters were available also from the laboratory involving grain-size distribution measurements of rock specimens. The resultant clusters hold information on both the lithology and hydraulic characteristics. By evaluating the connection between the core data and the cluster number log, one can identify the good conductivity zones (Cluster 1) and separate them from low permeability (Cluster 2) and impermeable layers (Cluster 3).

## 6. Conclusions

We propose a robust clustering method, which performs automated well-to-well correlation using all available wireline logs. For assuring robustness, we define the weighted distance between the data objects and the centroid, and the weights are optimally chosen in an automated iterative procedure. The initial centroids are calculated as the most frequent value of the cluster elements and are chosen by a histogram-based selection technique. We apply the new clustering method for the identification of Hungarian hydrocarbon reservoirs using all available well logs in the exploration area. The clustering results can reveal lithological and petrophysical characteristics, which are essential for reliable reservoir modeling.

The MFV procedure has already been shown to be a more powerful tool than some classic statistical algorithms originating from the least-squares or maximum likelihood principle, which can be applied well for non-Gaussian error distributions of measurements. This paper shows a perspective to process big data with a quick tool in case of incomplete datasets. The CPU time requirement of the clustering algorithm is relatively high; however, it can be reduced by optimizing the control parameters of the weighting procedure and dividing the depth interval to be processed into parts. The performance of the K-MFVs cluster analysis is tested by using synthetic well logs showing good noise rejection capability. Its feasibility is also confirmed using field data measured in a Hungarian hydrocarbon field. The optimal number of clusters can be found by mathematical assumptions and involving site specific a priori geological/geophysical information. The well log analysis method is recommended as new alternative tool for a more robust and reliable rock typing in different kind of reservoir rocks. In the future, the clustering technique can be fruitfully applied for the identification of unconventional hydrocarbon resources and geothermal reservoirs to assist 2D/3D lithological/petrophysical modeling for these

complex lithologies. The clustering method as a preliminary data processing tool may also support well logging inversion procedures with reliable initial models.

### Credit author statement

Norbert Péter Szabó: Conceptualization, Methodology, Investigation, Visualization, Resources, Writing - Original Draft. Roland Kilik: Software, Methodology, Data curation, Validation. Mihály Dobróka: Methodology, Supervision, Writing - Review & Editing.

### Data availability statement

The data that has been used is confidential.

### Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

### Acknowledgments

The research was carried out in the Project No. K-135323 supported by the National Research, Development and Innovation Office (NKFIH), Hungary (first and third author). The research was partly funded by the Sustainable Development and Technologies National Programme of the Hungarian Academy of Sciences (FFT NP FTA) (second author). The authors thank for the permission of using the data to Anita É. Csoma, E&P Innovation Lead & Head of Subsurface, Hungarian Oil and Gas Company (MOL Group).

### Nomenclature

a	tortuosity factor (Archie's constant)
c	centroid of a cluster
cpm	counts per minute as measure of intensity ( $\text{kcpm} = 10^3 \text{ cpm}$ )
c	column vector of cluster centroids
d	scalar value of one measurement data
d	column vector of observed data
D	matrix of observed data as input of cluster analysis
f(d)	probability distribution function of data
GG	gamma-gamma intensity log (kcpm)
GR	natural gamma-ray intensity log (API)
GR <sub>sh</sub>	natural gamma-ray intensity of shale (API)
GR <sub>q</sub>	natural gamma-ray intensity of quartz (API)
GR <sub>c</sub>	natural gamma-ray intensity of calcite (API)
K	pre-defined number of clusters in non-hierarchical cluster analysis
K <sub>h</sub>	hydraulic conductivity log ( $\text{ms}^{-1}$ )
L	total number of logging instruments
m	cementation exponent (Archie's constant)
MFV, M	most frequent value of a statistical sample
n	saturation exponent (Archie's constant)
N	total number of depth points
NN	neutron-thermal neutron intensity log (kcpm)
P <sub>e</sub>	photoelectric absorption index log (barn/e)
P <sub>e,c</sub>	photoelectric absorption index of calcite (barn/e)
P <sub>e,g</sub>	photoelectric absorption index of hydrocarbon (gas) (barn/e)
P <sub>e,q</sub>	photoelectric absorption index of quartz (barn/e)
P <sub>e,sh</sub>	photoelectric absorption index of shale (barn/e)
P <sub>e,w</sub>	photoelectric absorption index of pore-water (barn/e)
R <sub>s</sub>	shallow resistivity log (ohmm)
R <sub>d</sub>	deep resistivity log (ohmm)
R <sub>sh</sub>	resistivity of shale (ohmm)
R <sub>w</sub>	resistivity of pore-water (ohmm)
SSE	Sum of Squared Error as measure of variation within the clusters
S <sub>g</sub>	hydrocarbon (gas) saturation of rock formation (v/v)
S <sub>w</sub>	water saturation in the uninvaded (virgin) zone (v/v)
S <sub>x0</sub>	water saturation in the invaded zone (v/v)

$\mathbf{v}$	one element of cluster number vector $\mathbf{v}$
$\mathbf{v}$	column vector of estimated cluster numbers
$V_c$	fractional volume of calcite (v/v)
$V_q$	fractional volume of quartz (v/v)
$V_{sd}$	fractional volume of sand (v/v)
$V_{sh}$	fractional volume of shale (v/v)
$z$	depth coordinate (m)
$\delta^{(M)}, \delta^{(St)}$	distance metrics used in cluster analysis (M = Manhattan, St = Steiner)
$\Delta t$	acoustic (P-wave) travel-time log ( $\mu\text{sft}^{-1}$ )
$\Delta t_c$	acoustic (P-wave) travel-time of calcite ( $\mu\text{sft}^{-1}$ )
$\Delta t_g$	acoustic (P-wave) travel-time of hydrocarbon (gas) ( $\mu\text{sft}^{-1}$ )
$\Delta t_q$	acoustic (P-wave) travel-time of quartz ( $\mu\text{sft}^{-1}$ )
$\Delta t_{sh}$	acoustic (P-wave) travel-time of shale ( $\mu\text{sft}^{-1}$ )
$\Delta t_w$	acoustic (P-wave) travel-time of pore-water ( $\mu\text{sft}^{-1}$ )
$\varepsilon$	dihesion as scale parameter of weight function $\varphi(d)$
$\varphi$	porosity of rock formation (v/v)
$\varphi_N$	neutron porosity log (%)
$\varphi(d)$	weight function of data
$\rho_b$	bulk density (gamma-gamma log) ( $\text{gcm}^{-3}$ )
$\rho_c$	density of calcite ( $\text{gcm}^{-3}$ )
$\rho_g$	density of hydrocarbon (gas) ( $\text{gcm}^{-3}$ )
$\rho_q$	density of quartz ( $\text{gcm}^{-3}$ )
$\rho_{sh}$	density of shale ( $\text{gcm}^{-3}$ )
$\rho_w$	density of pore-water ( $\text{gcm}^{-3}$ )
$\xi_{\text{eff}}(e)$	effective number of data playing significant role in computing MFV
$\Psi$	objective function used in estimating $\varepsilon$

## References

- [1] C. Mayer, A.M. Sibbit, GLOBAL, a new approach to computer-processed log interpretation, Proceedings of the 55th SPE Annual Fall Technical Conference and Exhibition 9341 (1980) 1–14, <https://doi.org/10.2118/9341-MS>.
- [2] M. Alberty, K. Hashmy, Application of ULTRA to Log Analysis. SPWLA 25th Annual Logging Symposium, SPWLA-1984-Z, 1984, pp. 1–17.
- [3] S.M. Ball, D.M. Chace, W.H. Fertl, The Well Data System (WDS): an advanced formation evaluation concept in a microcomputer environment, Proceedings of the SPE Eastern Regional Meeting 17034 (1987) 61–85, <https://doi.org/10.2118/17034-MS>.
- [4] W.B. Hemphkins, Multivariate Statistical Analysis in Formation Evaluation, SPE California Regional Meeting, San Francisco, USA, 1978, <https://doi.org/10.2118/7144-MS>, 7144-MS.
- [5] O. Nasraoui, C.-E.B. N'Cir, Clustering Methods for Big Data Analytics. Techniques, Toolboxes and Applications, Springer Cham, 2019, <https://doi.org/10.1007/978-3-319-97864-2>.
- [6] F.F. Chen, Y.S. Yang, M. Pervukhina, B.M. Clennell, J.A. Taylor, Clustering analysis for porous media: an application to a dolomitic limestone, J. Pet. Sci. Eng. 146 (2016) 770–776, <https://doi.org/10.1016/j.petrol.2016.07.031>.
- [7] V. Tavakoli, Geological Core Analysis. Application to Reservoir Characterization, Springer Cham, 2018, <https://doi.org/10.1007/978-3-319-78027-6>.
- [8] W. Wang, Z. Wang, J.Y. Leung, C. Kong, Q. Jiang, Petrophysical rock typing based on deep learning network and hierarchical clustering for volcanic reservoirs, J. Pet. Sci. Eng. 210 (2022), 110017, <https://doi.org/10.1016/j.petrol.2021.110017>.
- [9] K. Zhang, N. Lin, C. Fu, D. Zhang, X. Jin, C. Zhang, Reservoir characterisation method with multi-component seismic data by unsupervised learning and colour feature blending, Explor. Geophys. 50 (2019) 269–280, <https://doi.org/10.1080/08123985.2019.1603078>.
- [10] W.J. Teh, G.P. Willhite, J.H. Doveton, Improved Reservoir Characterization Using Petrophysical Classifiers within Electrofacies, SPE Improved Oil Recovery Symposium, Tulsa, USA, 2012, <https://doi.org/10.2118/154341-MS>.
- [11] W.J.M. Al-Mudhafar, M.A. Bondarenko, Integrating K-means clustering analysis and Generalized Additive Model for efficient reservoir characterization, 77th EAGE Conference and Exhibition (2015), <https://doi.org/10.3997/2214-4609.201413024>, 1–6.
- [12] W.J. Al-Mudhafar, E.M. Al Lawe, C.I. Noshi, Clustering Analysis for Improved Characterization of Carbonate Reservoirs in a Southern Iraqi Oil Field, Offshore Technology Conference, Houston, USA, 2019, <https://doi.org/10.4043/29269-MS>.
- [13] E. Sfidari, A. Kadkhodaie-Ilkhchi, S. Najjari, Comparison of intelligent and statistical clustering approaches to predicting total organic carbon using intelligent systems, J. Pet. Sci. Eng. 86–87 (2012) 190–205, <https://doi.org/10.1016/j.petrol.2012.03.024>.
- [14] H. Yang, H. Pan, H. Ma, A.A. Konaté, J. Yao, B. Guo, Performance of the synergetic wavelet transform and modified K-means clustering in lithology classification using nuclear log, J. Pet. Sci. Eng. 144 (2016) 1–9, <https://doi.org/10.1016/j.petrol.2016.02.031>.
- [15] A. Yadav, S.R. Nayak, S. Mondal, Agglomerative clustering to improve the resolution of pseudo well stochastic seismic inversion: a case study, J. Pet. Sci. Eng. 208C (2022), 109566, <https://doi.org/10.1016/j.petrol.2021.109566>.
- [16] N.P. Szabó, M. Dobróka, R. Kavanda, Cluster analysis assisted float-encoded genetic algorithm for a more automated characterization of hydrocarbon reservoirs, Intell. Control Autom. 4 (2013) 362–370, <https://doi.org/10.4236/ica.2013.44043>.
- [17] N.P. Szabó, K. Nehéz, O. Hornyák, I. Piller, Cs Deák, P.P. Hanzelík, Cs Kutasi, K. Ott, Cluster analysis of core measurements using heterogeneous data sources: an application to complex Miocene reservoirs, J. Pet. Sci. Eng. 178 (2019) 575–585, <https://doi.org/10.1016/j.petrol.2019.03.067>.
- [18] J. Jarzyna, et al., Shale gas in Poland, in: H. Al-Megren, R. Altamimi (Eds.), Advances in Natural Gas Emerging Technologies, IntechOpen, London, 2017, <https://doi.org/10.5772/67301>.
- [19] N.P. Szabó, B.A. Braun, M.M.G. Abdelrahman, M. Dobróka, Improved well logs clustering algorithm for shale gas identification and formation evaluation, Acta Geod Geophys 56 (2021) 711–729, <https://doi.org/10.1007/s40328-021-00358-0>.
- [20] N.P. Szabó, R. Valadez-Vergara, S. Tapdigli, A. Ugochukwu, I. Szabó, M. Dobróka, Factor analysis of well logs for total organic carbon estimation in unconventional reservoirs, Energies 14 (2021) 5978, <https://doi.org/10.3390/en14185978>.



- [21] N.P. Szabó, A genetic meta-algorithm-assisted inversion approach: hydrogeological study for the determination of volumetric rock properties and matrix and fluid parameters in unsaturated formations, *Hydrogeol. J.* 26 (2018) 1935–1946, <https://doi.org/10.1007/s10040-018-1749-7>.
- [22] J.A. Hartigan, M.A. Wong, Algorithm AS 136: a k-means clustering algorithm, *J. R. Stat. Soc. Ser. C Appl. Stat.* 28 (1979) 100–108, <https://doi.org/10.2307/2346830>.
- [23] M.C. Cowgill, R.J. Harvey, L.T. Watson, A genetic algorithm approach to cluster analysis, *Comput. Math. Appl.* 37 (1999) 99–108, [https://doi.org/10.1016/S0898-1221\(99\)00090-5](https://doi.org/10.1016/S0898-1221(99)00090-5).
- [24] M.T. Gallegos, G. Ritter, A robust method for cluster analysis, *Ann. Stat.* 33 (2005) 347–380, <https://doi.org/10.1214/009053604000000940>.
- [25] P. Filzmoser, R.G. Garrett, C. Reimann, Multivariate outlier detection in exploration geochemistry, *Comput. Geosci.* 31 (2005) 579–587, <https://doi.org/10.1016/j.cageo.2004.11.013>.
- [26] A.C. Atkinson, M. Riani, Exploratory tools for clustering multivariate data, *Comput. Stat. Data Anal.* 52 (2007) 272–285, <https://doi.org/10.1016/j.csda.2006.12.034>.
- [27] H.R. Samadi, R. Kimiaefar, A. Hajian, Robust earthquake cluster analysis based on k-nearest neighbor search, *Pure Appl. Geophys.* 177 (2020) 5661–5671, <https://doi.org/10.1007/s00024-020-02618-6>.
- [28] A. Ali, C. Sheng-Chang, Characterization of well logs using K-mean cluster analysis, *J. Pet. Explor. Prod. Technol.* 10 (2020) 2245–2256, <https://doi.org/10.1007/s13202-020-00895-4>.
- [29] A. Ali, C. Sheng-Chang, M. Shah, Integration of cluster analysis and rock physics for the identification of potential hydrocarbon reservoir, *Nat. Resour. Res.* 30 (2021) 1395–1409, <https://doi.org/10.1007/s11053-020-09800-6>.
- [30] F. Steiner, Most frequent value procedures (a shortmonograph), *Geophys. Trans.* 34 (1988) 139–260.
- [31] F. Steiner, The Most Frequent Value: Introduction to a Modern Conception of Statistics, *Akadémiai Kiadó*, 1991.
- [32] F. Steiner, Optimum Methods in Statistics, *Akadémiai Kiadó*, 1997.
- [33] J. Zhang, Most frequent value statistics and distribution of 7Li abundance observations, *Mon. Not. Roy. Astron. Soc.* 468 (2017) 5014–5019, <https://doi.org/10.1093/mnras/stx627>.
- [34] J. Zhang, Most frequent value statistics and the hubble constant, *Publ. Astron. Soc. Pac.* 130 (2018) 1538–3873, <https://doi.org/10.1088/1538-3873/aac767>.
- [35] P. Szűcs, F. Civan, M. Virág, Applicability of the most frequent value method in groundwater modeling, *Hydrogeol. J.* 14 (2006) 31–43, <https://doi.org/10.1007/s10040-004-0426-1>.
- [36] L. Völgyesi, G. Tóth, Improvement of QDaedalus measurements with continuous detection of environmental parameters, *Acta Geod Geophys* 56 (2021) 607–622, <https://doi.org/10.1007/s40328-021-00359-z>.
- [37] M. Dobróka, Á. Gyulai, T. Ormos, J. Csókás, L. Dresen, Joint inversion of seismic and geoelectric data in an underground coal mine, *Geophys. Prospect.* 39 (1991) 643–665, <https://doi.org/10.1111/j.1365-2478.1991.tb00334.x>.
- [38] Á. Gyulai, P. Szűcs, E. Turai, et al., Geoelectric characterization of thermal water aquifers using 2.5D inversion of VES measurements, *Surv. Geophys.* 38 (2017) 503–526, <https://doi.org/10.1007/s10712-016-9393-z>.
- [39] N.P. Szabó, G.P. Balogh, J. Stickel, Most frequent value-based factor analysis of direct-push logging data, *Geophys. Prospect.* 66 (2018) 530–548, <https://doi.org/10.1111/1365-2478.12573>.
- [40] M. Dobróka, H. Szegedi, On the generalization of seismic tomography algorithms, *Am. J. Comput. Math.* 4 (2014) 37–46, <https://doi.org/10.4236/ajcm.2014.41004>.
- [41] T.E. Dobróka, An MFV-based image processing filter and its application to seismic tomographic images, *Acta Geod Geophys* 56 (2021) 731–742, <https://doi.org/10.1007/s40328-021-00351-7>.
- [42] D.O.B. Nuamah, M. Dobróka, P. Vass, et al., Legendre polynomial-based robust Fourier transformation and its use in reduction to the pole of magnetic data, *Acta Geod Geophys* 56 (2021) 645–666, <https://doi.org/10.1007/s40328-021-00357-1>.
- [43] B.Á. Braun, A. Abordán, N.P. Szabó, Lithology determination in a coal exploration drillhole using Steiner weighted cluster analysis, *Geosci. Eng.* 5 (2016) 51–64.
- [44] N.P. Szabó, B.A. Braun, M.M.G. Abdelrahman, et al., Improved well logs clustering algorithm for shale gas identification and formation evaluation, *Acta Geod Geophys* 56 (2021) 711–729, <https://doi.org/10.1007/s40328-021-00358-0>.
- [45] R. Kilik, Histogram-based weighted median filtering used for noise reduction of digital elevation model data, *Acta Geod Geophys* 56 (2021) 743–764, <https://doi.org/10.1007/s40328-021-00356-2>.
- [46] M. Dobróka, N.P. Szabó, J. Tóth, P. Vass, Interval inversion approach for an improved interpretation of well logs, *Geophysics* 81 (2016) D155–D167, <https://doi.org/10.1190/geo2015-0422.1>.
- [47] N.P. Szabó, A. Abordán, M. Dobróka, Permeability extraction from multiple well logs using particle swarm optimization based factor analysis, *Int. J. Geom.* 13 (2022) 10, <https://doi.org/10.1007/s13137-022-00200-x>.
- [48] A. Timur, An investigation of permeability, porosity and residual water saturation relationships for sandstone reservoirs, *Log. Anal.* 9 (1968) 3–5.
- [49] P.J. Rousseeuw, Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, *J. Comput. Appl. Math.* 20 (1987) 53–65, [https://doi.org/10.1016/0377-0427\(87\)90125-7](https://doi.org/10.1016/0377-0427(87)90125-7).
- [50] N.P. Szabó, M. Dobróka, E. Turai, et al., Factor analysis of borehole logs for evaluating formation shaliness: a hydrogeophysical application for groundwater studies, *Hydrogeol. J.* 22 (2014) 511–526, <https://doi.org/10.1007/s10040-013-1067-z>.
- [51] N.P. Szabó, Hydraulic conductivity explored by factor analysis of borehole geophysical data, *Hydrogeol. J.* 23 (2015) 869–882, <https://doi.org/10.1007/s10040-015-1235-4>.