UNIVERSITY OF MISKOLC

Faculty of Earth and Environmental Sciences and Engineering

Institute of Exploration Geosciences

Department of Geophysics

Geostatistics assisted well-logging inversion method developments

PhD THESIS

By

MOATAZ MOHAMED GOMAA ABDELRAHMAN

Scientific supervisor:
Prof. Dr. Norbert Péter Szabó

MIKOVINY SÁMUEL DOCTORAL SCHOOL OF EARTH SCIENCES
Head of the Doctoral School: Prof. Dr. Péter Szűcs

Miskolc, 2025
HUNGARY

The topic of the Candidate's thesis – robust inversion of wireline logging data – is in the focus of international research. The method developments introduced by the Candidate in the thesis belong to the range of modern data processing tools including machine learning and advanced inversion approaches applied to in situ geophysical data. The goal of the PhD research is to have a more efficient interpretation of well logging data, by creating a fully automated inversion workflow. In addition to presenting the theoretical background of research, the PhD thesis gives answers to several practical questions and challenges.

The first step of the implementation is the application of the Most Frequent Value (MFV) method for automation of zonation and layer boundary detection, which is an important pre-requisite for a reliable inversion of well logs. First, the Candidate develops an MFV-based robust clustering method and then other approaches for boundary detection as a novel application of Hurst exponent and factor analysis. Then, he performs the estimation of petrophysical quantities in reservoir rocks, where he applies a new discretization scheme, the so-called series expansion-based inversion approach. This improves both the forward and inverse problems. In addition to local (depth-by-depth) inversion, traditionally used in the oil and gas industry, the Candidate uses the interval inversion technique, too. The MFV method and the series expansion-based interval inversion procedure was originally developed at the Department of Geophysics of the University of Miskolc by Prof. Ferenc Steiner and Prof. Mihály Dobróka, respectively. The Candidate followed the research idea of his predecessors and supervisor and further developed them to give new scientific results in practical application.

In the framework of the proposed inversion methodology, the Candidate increases the stability of the often numerically poorly conditioned inverse problem. He introduces the golden section- and singular value decomposition-based inversion method for a more efficient and

faster processing of well logging data. He introduces new unknowns such as zone parameters, which are traditionally obtained from laboratory measurements or, in the absence thereof, arbitrarily selected. The conventional inversion approach treating the zone parameters as fixed values can highly increase the uncertainty of the estimation results. In the Candidate's PhD thesis, the harmful effect of modeling errors is reduced by setting the zone parameters as unknown, which improves the accuracy of the key parameters of the reserve calculation, namely the determination of the volume characteristics such as porosity, clay content, water saturation and matrix volumes, too. The new robust inversion technique presented in the PhD thesis serves for a more accurate and reliable reservoir characterization in hydrogeophysics, geothermal-, conventional and unconventional hydrocarbon exploration providing powerful tools of identification and evaluation of the reservoir rocks and a better reserve calculation.

The Candidate's continuous efforts towards scientific research, his creativity, and the results presented in this PhD thesis prove the scientific knowledge and the suitability of the Candidate for independent research. In my opinion, the Candidate's results, especially those related to the estimation of layer thicknesses, zone and volumetric parameters in a joint inversion procedure, are worth publishing in Scimago ranked international journals. The feasibility of the developed inversion method has been tested both on synthetic and real data, both in Hungary and Egypt, and validated also by core laboratory measurements. Herewith, I certify that the Candidate's PhD dissertation contains only valid data, and the presented results represent the Candidate's own work. In my opinion, the PhD thesis is fully adequate in scope and quality required by the Mikoviny Sámuel Doctoral School of Earth Sciences at the University of Miskolc. Based on all the above, I recommend the public defense to be carried out to complete a successful process of acquiring the PhD title.

11 May 2025, Miskolc

Prof. Dr. Norbert Péter Szabó

supervisor

CONTENTS

LIST OF FIGURES

List of tables

Chapter 1: Introduction

An accurate understanding of subsurface formations is a cornerstone of geophysical and hydrogeological research, with profound implications for resource exploration, environmental sustainability, and infrastructure planning. Borehole logging data, which offers critical insights into the composition, structure, and fluid content of the Earth's subsurface, serves as a primary tool for unraveling the complexities hidden beneath the surface. However, interpreting this data is no simple task (Zhang et al. 2024). Traditional methods, such as the widely used K-means clustering technique and conventional inversion approaches, often struggle to deliver reliable results when faced with noisy datasets, outliers, or heterogeneous geological formations (Jahantigh and Ramazi 2025). These limitations can lead to misinterpretations, obscuring important details about lithology, aquifer properties, or reservoir characteristics (Mohammed et al. 2025). This thesis seeks to address these challenges by introducing innovative methodologies that enhance the precision, robustness, and reliability of subsurface characterization, offering a fresh perspective on geophysical data interpretation.

At the heart of this research lies the recognition that traditional methods for interpreting borehole log data often fall short in handling real-world complexities. For instance, the K-means clustering technique, while widely used, is highly sensitive to outliers and noise, which can distort results and lead to misleading classifications (Aliyuda et al. 2022). Similarly, conventional inversion techniques, which are employed to estimate reservoir parameters, frequently encounter issues with convergence stability and computational efficiency, particularly when dealing with noisy or incomplete datasets (Ikhwan et al. 2024). These shortcomings highlight the need for more adaptive and resilient approaches that can navigate the inherent uncertainties of geophysical data (Okunola 2024). Recent advancements in statistical methods and computational techniques have opened new doors for addressing these challenges, providing a foundation for the development of more robust and accurate interpretation tools (Aversana 2023).

One such advancement is the development of the Most Frequent Value (MFV) algorithm, a statistical method that has been promising tool in mitigating the effects of outliers and stabilizing data analysis (Kemp and Steiner 1991). By incorporating the MFV algorithm into clustering techniques, it becomes possible to achieve more reliable and consistent results, even in the presence of noisy or erratic data. Additionally, interval inversion methods, originally developed for oil exploration, have demonstrated superior performance in parameter

estimation compared to traditional local inversion techniques. These methods offer greater convergence stability and accuracy, making them well-suited for handling the complexities of subsurface data. Building on these advancements, this research introduces a series of innovative methodologies that integrate robust statistical techniques with advanced inversion approaches, creating a cohesive framework for accurate and reliable subsurface characterization.

The central idea driving this work is that the integration of robust statistical methods, such as the MFV algorithm, with advanced inversion techniques, can significantly improve the interpretation of borehole log data. This approach is grounded in the assumption that the MFV algorithm's ability to suppress outliers and stabilize centroid estimation will enhance clustering results, while hybrid inversion methods optimize parameter estimation by combining the strengths of different algorithms. Furthermore, it is assumed that these methodologies can be applied across a wide range of geological settings, from groundwater formations to hydrocarbon reservoirs and geothermal systems, making them versatile tools for subsurface exploration.

The research strategy revolves around the development and validation of three key methodologies. The first is an MFV-based clustering algorithm that incorporates Steiner-Cauchy weights to suppress outliers and stabilize centroid estimation. This algorithm introduces a novel distance metric, known as the Steiner distance, which automatically identifies the most relevant range of the data while filtering out noise caused by outliers. By using the MFV as a centroid update mechanism, the algorithm ensures robust and stable clustering results, even when starting with randomly initialized centroids. The second methodology is a hybrid inversion approach that combines the Damped Least Squares (DLSQ) and Singular Value Decomposition (SVD) methods. This approach dynamically adjusts damping factors to optimize convergence, offering a more efficient and accurate alternative to traditional inversion techniques. The third methodology is an automated workflow that integrates Hurst analysis, interval inversion, and the Csókás method for comprehensive formation evaluation. This workflow uses Hurst analysis to estimate the number of layers and predict their depth locations, interval inversion to derive petrophysical parameters, and the Csókás method to improve hydraulic conductivity estimation. Together, these methodologies provide a cohesive framework for accurate and reliable subsurface characterization.

What sets this research apart is the seamless integration of these techniques into a unified framework that addresses the limitations of traditional methods. For example, the MFV-based clustering algorithm introduces a novel distance metric and centroid update mechanism, significantly improving the robustness and stability of clustering results. The hybrid inversion approach dynamically adjusts damping factors to optimize convergence, offering a more efficient and accurate alternative to traditional inversion techniques. The automated workflow leverages Hurst analysis for lithological boundary detection and interval inversion for high-resolution parameter estimation, providing a comprehensive solution for subsurface characterization.

The motivation behind this research stems from the growing need for accurate and reliable methods to interpret complex geophysical data, particularly in the face of increasing environmental and resource challenges. Groundwater depletion, hydrocarbon exploration in complex reservoirs, and geothermal energy development all require precise subsurface characterization to inform decision-making and resource management. By developing methodologies that enhance the accuracy and efficiency of data interpretation, this research aims to contribute to the advancement of geophysical and hydrogeological sciences. The potential impact of this work extends beyond academic research, offering practical tools for industry professionals and policymakers. The proposed methodologies can improve the assessment of aquifer properties, enhance hydrocarbon reservoir characterization, and optimize geothermal resource evaluation. Furthermore, the automated nature of these workflows reduces the reliance on manual interpretation, minimizing subjectivity and increasing reproducibility.

The PhD thesis presents a series of innovative methodologies that address the limitations of traditional borehole logging data interpretation techniques. By integrating robust statistical methods, advanced inversion algorithms, and automated workflows, this research provides a comprehensive framework for accurate and reliable subsurface characterization. The results demonstrate the potential of these methodologies to transform geophysical and hydrogeological research, offering new insights into complex geological systems and supporting sustainable resource management. The development of these techniques represents a significant step forward in the field, with broad applications in groundwater assessment, hydrocarbon exploration, and geothermal resource evaluation. Through this work, I aim to provide the scientific community and industry professionals with powerful tools for understanding and managing the Earth's subsurface resources.

Chapter 2: Advancing Lithological Classification with a Robust Clustering Technique

Geophysical data processing and interpretation involves studying and characterizing the Earth's subsurface using physical measurements and models. In recent years, there has been growing interest in using computational techniques like machine learning for more objective and consistent lithological classification (Tiab and Donaldson 2004). A common approach is to first numerically characterize rock samples based on features like mineral composition, texture, etc. extracted from imaging data or spectroscopy. Then, these multivariate feature vectors can be input into clustering algorithms to automatically group similar rock types. Recently, machine learning has been implemented as a vital and useful tool for many different geophysical applications, even in research and industry. The applications of machine learning in borehole geophysics include identifying patterns and correlations between various recorded well-log data. Therefore, the two main problems of machine learning can be concluded as classification and regression (Lima et al. 2020). Furthermore, the application of machine learning can be extended to build physics-based models that can simulate oil ang gas reservoirs and groundwater aquifers (Terry et al. 2019). This enables faster scenario testing and quantification of uncertainty (Ian et al. 2009). Supervised and unsupervised learning are the two major categories of machine learning algorithms. The main difference lies in whether the algorithm is trained on labeled or unlabeled data. In supervised learning, the training data contains example input-output pairs, and the algorithm learns to model the mapping from input to output. Common supervised tasks include classification, where the outputs are discrete categories and regression, where the outputs are continuous numerical values. Supervised methods include linear models such as logistic regression, and nonlinear models such as neural networks, and nonparametric models such as decision trees and random forests (Pandey et al. 2020). Specifically, robustness in this context pertains to the algorithm's insensitivity to statistical distribution variations and its capacity to manage outlier data without compromise. A robust K-means approach is characterized by its ability to maintain stable and accurate clustering outcomes even in the presence of atypical data points or deviations from assumed statistical models (Zhang et al. 2024). Furthermore, statistical efficiency is a pivotal component of robustness, reflecting the algorithm's proficiency in achieving precise clustering results while optimizing computational resources, thereby ensuring their scalability and efficiency. By embedding robustness as a foundational feature, the proposed K-means variant can offer

enhanced versatility and dependability across diverse datasets and practical applications, leading to more consistent and reliable analytical outcomes.

In this Chapter, I introduce a novel robust clustering technique called MFV clustering, which is tailored for lithological classification. MFV clustering incorporates innovative statistical metrics of feature variability to overcome the limitations of conventional clustering techniques. The common weaknesses of conventional clustering techniques include heavy computation, such as hierarchical clustering, or initial location dependency, such as in *K-mean* clustering.

2.1. K-means cluster analysis

In the preprocessing step of clustering a dataset with greater than one dimension, the data must be standardized. Scale the data to have a mean of zero and a standard deviation of one. The aim of this step is to minimize the effect of differences in scales of variables. Without normalization, variations in scales have the potential to introduce bias during the clustering process because features with greater numerical ranges can dominate the distance calculations and lead to biased results. Through normalization, all dimensions contribute equally to the outcome, allowing the clustering algorithm to perform optimally and produce more significant results. This process allows the underlying patterns in the data to be better detected and understood.

K-means clustering is an exploratory statistical method that is used to separate the multidimensional data into a few homogenous groups. It is a widely used technique for uncovering hidden patterns in a dataset. K-means clustering algorithms rely on calculating the distance between each data object and randomly generated centroids. However, a priori knowledge of the number of centroids (which is the number of clusters) is required for *the K-means* clustering algorithm (Khan and Ahmad 2004). For initializing the multidimensional statistical procedure for clustering well logging data, let $A = \{d_i | i = 1, ..., L\}$ be a well-log data set of L data types, and $c = \{c_i | i = 1, ..., K\}$ be a set of K centroids. Consider a dataset with N data points in an L dimensional space. As an example, two points in the L dimensional data space can be written as $d^{(i)} = [d_1^{(i)}, \cdots, d_L^{(i)}]^{\mathrm{T}}$ and $d^{(j)} = [d_1^{(j)}, \cdots, d_L^{(j)}]^{\mathrm{T}}$, where $1 \leq i, j \leq N$. K-means clustering aims to group the points into clusters by minimizing the following objective function

$$O = \sum_{k=1}^{K} \sum_{i=1}^{N} dis(d_i, c_k), \tag{1}$$

where the measure of similarity can be specified as the distance between the data points and the cluster centroids ($dis(d_i, c_k)$). During the clustering process, we minimize the distance given in equation (1). The Minkowski distance assumes that different number of elements belong to each cluster. The L$_p$ norm characterizes the distance between the data objects and the cluster centroids as follows

$$D_p = \left[\sum_{k=1}^{K} \sum_{i=1}^{N_k} |d_i - c_k|^p \right]^{\frac{1}{p}}. \tag{2}$$

In case of clustering borehole datasets, the most common distances to be used are the L$_1$ and L$_2$ norms belonging to p=1 and p=2, respectively (Szabó et al. 2021). Consider data objects $\mathbf{d}^{(i)}$ and $\mathbf{d}^{(j)}$. The L$_2$ norm is the so-called the Euclidean norm or the Euclidean distance, it can be represented as follows

$$D_{Euc} = \sqrt{(d_k^{(i)} - c_k)^{\mathrm{T}}(d_k^{(j)} - c_k)}. \tag{3}$$

The measured data are usually contaminated by noise and outliers, therefore, the weighted distance or Mahalanobis distance is preferably used in such cases

$$D_{Mah} = \sqrt{(d_k^{(i)} - c_k)^{\mathrm{T}}\mathbf{W}^{-1}(d_k^{(j)} - c_k)}, \tag{4}$$

where $\mathbf{W}^{-1}$ is the inverse of the covariance matrix of standardized observed variables, which contains the data variances in its main diagonal. If the variables exhibit interdependence, it is imperative to consider the measure of correlation, even in cases where the variables' magnitudes differ, making their distances incomparable. The Euclidean norm proves most effective when the data noises adhere to a Gaussian distribution. However, in cases of non-Gaussian distributed data, even when outliers are present in the dataset, the Manhattan distance provides a more robust estimation. The location of the centroids is iteratively updated according to the mean of the assigned data points as follow

$$c_k = \frac{1}{N_k} \sum_{i=1}^{N_k} d_i^{(k)}. \tag{5}$$

where N_k represents the total number of elements being averaged in cluster k. The K-means clustering algorithm begins by randomly initializing K centroids (defined prior to the

iterative process). Then, it proceeds to compute the distances between each data point and the centroids, followed by grouping the data points based on the minimum distance. Subsequently, the algorithm updates the centroid locations and repeats this process until no further changes occur in the centroids' locations. K-means clustering does not guarantee to give the same results with randomly chosen initial clusters (initial centroid selection dependency), it can give a reliable results when the initial centers is close to the final solution (Khan and Ahmad 2004). Aside from that, outliers can significantly affect the position of centroids, potentially leading to the formation of incorrect clusters. K-means cluster analysis is quite sensitive to outliers and may produce suboptimal results in the presence of significant outliers. To overcome such weaknesses, much previous research tried to adjust the initial centroid selection step. One of these approaches was to take the mean of the entire data and randomly perturbing it K times (Qu 2025). Bradley and Fayyad (1998) proposed a procedure that relate the initial location of centroids to the joint probability density function of the data (pdf), where the condition is to generate centroids near to the mode of the joint pdf. Furthermore, some statistical methods can be used to remove the outliers to avoid the second weakness point.

2.2. MFV-based cluster analysis

The goal of clustering is to minimize the distance within each cluster and to maximize the distance between different clusters. Therefore, the weighted mean of the data can be more accurate than the arithmetic mean, especially in case of outlier's presence. The most frequent value method, introduced by Steiner (1991), is an automated method that assigns higher weights to data points that are closer to each other and lower weights to those that are farther to reduce the influence of outliers in clustering algorithms (Tolner et al. 2023). By emphasizing the importance of proximity, the method gives priority to connections between data points that fall within ranges of interest, making it less sensitive to outliers that are further away from the central groups (Akbar, Szabó, and Dobróka 2021). This approach can contribute to consistent outcomes. Additionally, updating the centroid positions will ensure an accurate convergence pathway, irrespective of the initially random locations, as the location of the centroids in the data space will be determined by the most frequent value (MFV) in the data. The automated iterative process calculates the MFV (location parameter) and dihesion, ε (scale parameter) simultaneously by a recursion formula. Firstly, MFV is considered as the mean or median of the data points, while ε is estimated from the range of the data points. In another words, the

greater value of ε means that all the data points take the same (high-value) weights (including the outliers) in the beginning (Steiner 1990)

$$\varepsilon \leq \frac{\sqrt{3}}{2}(\max(x_i) - \min(x_i)), \tag{6}$$

and in the j-th iteration MFV and ε are calculated from each other based on equations (7) and (8) as follows

$$\varepsilon_{j+1}^2 = \frac{3\sum_{i=1}^{q}\frac{(x_i - MFV_{q,j})^2}{(\varepsilon_j^2 + (x_i - MFV_{q,j})^2)^2}}{\sum_{i=1}^{q}\frac{1}{(\varepsilon_j^2 + (x_i - MFV_{q,j})^2)^2}}, \tag{7}$$

$$MFV_{q,j+1} = \frac{\sum_{i=1}^{q}\frac{\varepsilon_{j+1}^2}{\varepsilon_{j+1}^2 + (x_i - MFV_{q,j})^2}x_i}{\sum_{i=1}^{q}\frac{\varepsilon_{j+1}^2}{\varepsilon_{j+1}^2 + (x_i - MFV_{kq,j})^2}}, \tag{8}$$

in the framework of the MFV algorithm. The weighting function can be used to cluster the noisy dataset using the following weight equation using w_k^{St} Steiner-weights for giving robust parameter estimation (Szabó and Dobróka 2017)

$$w_k^{(St)} = \frac{\varepsilon^2}{\varepsilon^2 + e_k^2}. \tag{9}$$

I modified the traditional K-means clustering method by the MFV algorithm to introduce an innovative clustering technique called MFV-based cluster analysis. Figure 1 shows the workflow of the proposed algorithm of the MFV clustering. The MFV-based clustering method can solve both prementioned weaknesses at the same time. By substituting equation (9) into equation (6), Steiner distance can be introduced as follows

$$D_{St} = \sqrt{\frac{1}{\sum_{q=1}^{N} w_q^{St}}\sum_{k=1}^{N} w_k^{St}(d_k^{(i)} - d_k^{(j)})^2}. \tag{10}$$

The MFV-based clustering updates the location of the centroids according to the weighted Steiner mean as follows

$$c_i^{St} = \frac{1}{\sum_{q=1}^{N_k} w_q^{St}}\sum_{k=1}^{N_k} w_k^{St}d_k^{(i)}. \tag{11}$$
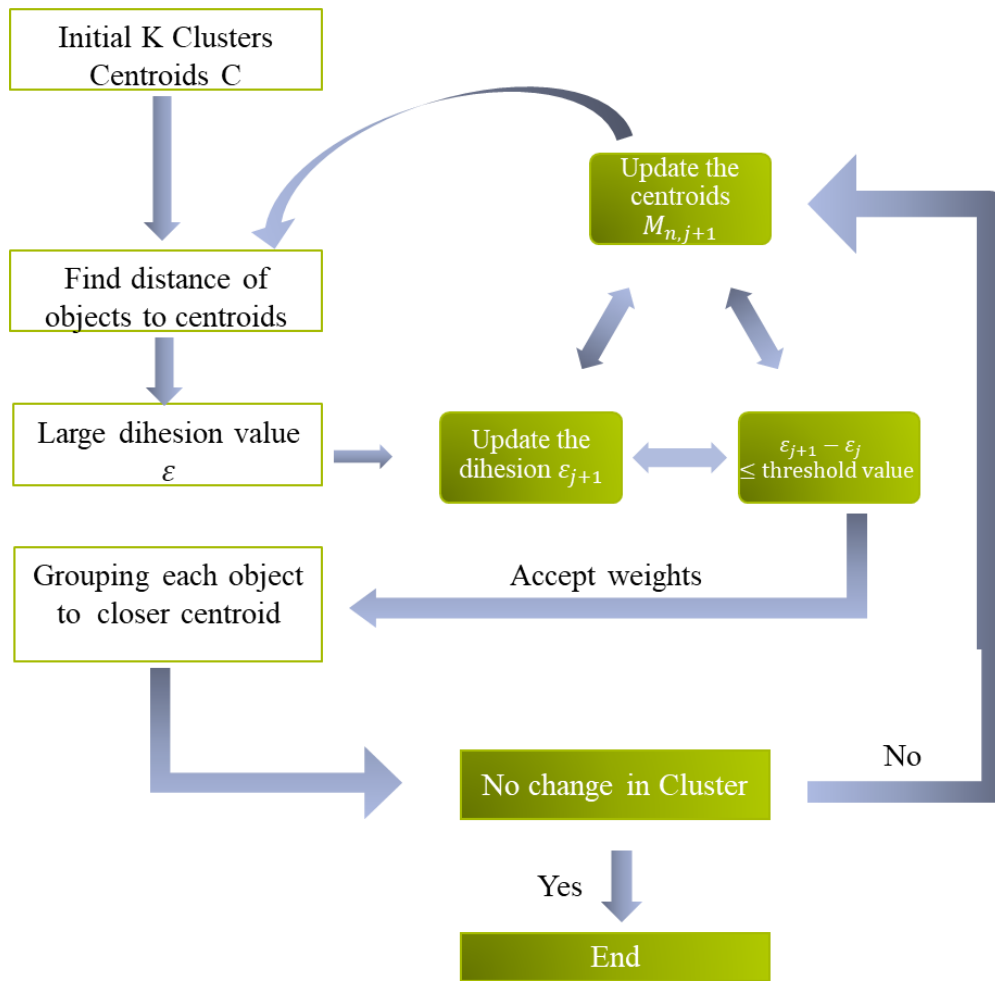
Figure 1. The flowchart of the MFV-based cluster method.

Comparative study between K-means and MFV-based clustering using synthetic data

To study the performance of the MFV-based clustering method, a statistical study was carried out using synthetic well-log data. The suggested model consists of four homogenous layers with different petrophysical parameters (Table 1). The calculated well-logging data reflects the presence of hydrocarbon and water bearing zones. Moreover, the volumetric lithological composition encompasses clay, carbonate, kerogen, and quartz constituents within the layers. To check the robustness of the MFV-based cluster method, the synthetic data were contaminated by 3% Gaussian distributed noise with artificial outliers' implementation. Regarding the geometrical information of the constructed model, the depth increment is 0.1 m, and the total depth is 45 m. Therefore, the calculated points are 450 points per well log. Figure 2 shows the synthetic well-logging data.

Table 1. The petrophysical model for synthetic data calculation. Denotations: water saturation (S_w), clay volume (V_{cl}), carbonate volume (V_c), and kerogen volume (V_k).

Layer	Thickness	Φ	S_w	V_q	V_{cl}	V_c	V_k

	(m)						
1	8	0.02	0.9	0.577	0.4	0.001	0.002
2	18	0.03	1.0	0.153	0.7	0.1	0.017
3	8	0.02	0.7	0.46	0.15	0.25	0.12
4	10	0.01	0.6	0.158	0.3	0.5	0.032

Denotations are porosity (Φ, [v/v]), water saturation (S_w, [v/v]), clay volume (V_{cl}, [v/v]), carbonate volume (V_c, [v/v]), and kerogen volume (V_k, [v/v]). Quartz volume is derived from the material balance equations $V_q = 1 - \Phi - V_{cl} - V_c - V_k$.
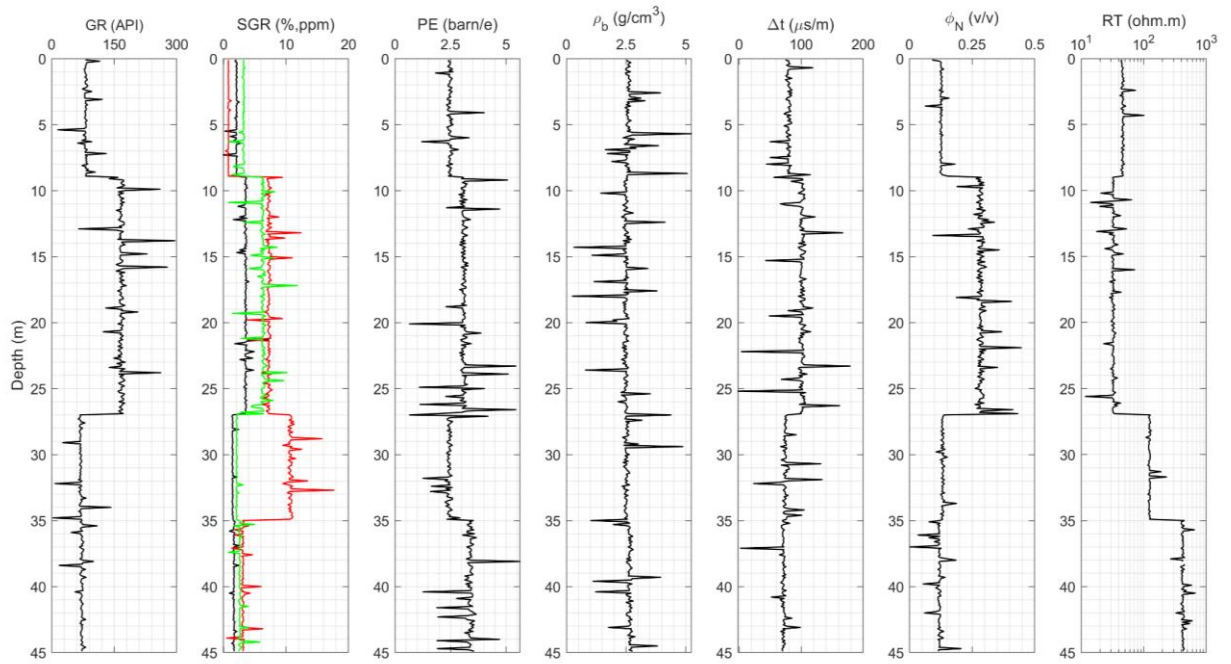
Figure 2. Synthetic well logging data contaminated with 3% Gaussian distributed noise. Denotations: GR [API] is natural gamma-ray log, SGR (%, ppm) is spectral gamma-ray log (black curve is potassium, the red curve is uranium, and green is thorium), Δt (μs/m) is the compressional acoustic slowness log, ϕ_N [v/v] is neutron log, ρ_b [g/cm³] is density log, PE [barn/e] is photoelectric absorption index, and RT [Ohm-m] is deep resistivity log.

The physical relationship between the calculated well-logging data and the predefined petrophysical parameters can be expressed through a set of equations known as response functions. Equations (12) to (21) represent the tool response functions used for calculating the synthetic well-log data (Alberty and Hashmy 1984)

$$GR = \rho_b^{-1}\left(V_k \, GR_k \rho_k + V_{sh} \, GR_{sh}\rho_{sh} + \sum_{i=1}^{n} V_{ma,i} GR_{ma,i}\, \rho_{ma,i} \right), \qquad (12)$$

$$K = \rho_b{}^{-1}\left(V_k K_k \rho_k + V_{sh} K_{sh}\rho_{sh} + \sum_{i=1}^{n} V_{ma,i}K_{ma,i}\, \rho_{ma,i}\right), \tag{13}$$

$$U = \rho_b{}^{-1}\left(V_k U_k \rho_k + V_{sh} U_{sh}\rho_{sh} + \sum_{i=1}^{n} V_{ma,i}U_{ma,i}\, \rho_{ma,i}\right), \tag{14}$$

$$Th = \rho_b{}^{-1}\left(V_k Th_k \rho_k + V_{sh} Th_{sh}\rho_{sh} + \sum_{i=1}^{n} V_{ma,i}Th_{ma,i}\, \rho_{ma,i}\right), \tag{15}$$

$$\rho_b = \emptyset\big[(\rho_w S_w) + \rho_g(1 - S_w)\big] + V_k\rho_k + V_{sh}\rho_{sh} + \sum_{i=1}^{n} V_{ma,i}\, \rho_{ma,i}\,, \tag{16}$$

$$\emptyset_N = \emptyset\big[(\emptyset_w S_w) + \emptyset_g(1 - S_w)\big] + V_k\rho_k + V_{sh}\rho_{sh} + \sum_{i=1}^{n} V_{ma,i}\, \emptyset_{ma,i}\,, \tag{17}$$

$$\Delta t = \emptyset[\Delta t_w S_w + (1 - S_w)\Delta t_g] + V_k\rho_k + V_{sh}\Delta t_{sh} + \sum_{i=1}^{n} V_{ma,i}\, \Delta t_{ma,i}\,, \tag{18}$$

$$P_e = \emptyset[S_w U_w + (1 - S_w)U_g] + V_k\rho_k + V_{sh}U_{sh} + \sum_{i=1}^{n} V_{ma,i}\, U_{ma,i}\,, \tag{19}$$

$$\frac{1}{\sqrt{R_d}} = \left[\frac{V_{sh}{}^{(1-0.5V_{sh})}}{\sqrt{R_{sh}}} + \frac{\sqrt{\emptyset}^m}{\sqrt{aR_w}}\right]\sqrt{S_w}^{\,n}, \tag{20}$$

$$\emptyset + V_k + V_{sh} + \sum_{i=1}^{n} V_{ma,i} = 1. \tag{21}$$

The exposition entails an elucidation of pertinent petrophysical properties and their relationship to logging data, characterized by a defined nomenclature. Specifically, $V_{ma,i}$ (v/v) signifies the fractional volume of the *i-th* matrix constituent, while V_k denotes the volume of kerogen. The count of mineral constituents, *n*, and the fractional volume of shale-exempt pore spaces ($\emptyset$, v/v) are established parameters. Likewise, the uninvaded zone features a discernible water saturation fraction denoted by S_w. The distinctive physicochemical characteristics of mud filtrate (*mf*), hydrocarbon (*h*), shale (*sh*), and the rock matrix (*ma*) are explicated by zone-specific parameters stipulated in equations (12) to (20). Notably, equation (21), denoted as the material balance equation, assumes the pivotal role of a constraint in the resolution of the inverse problem, particularly in the determination of quartz volume.

The comparison between the traditional K-means and MFV-based cluster includes testing the results stability and outliers suppress. Figure 3 shows the resulting clusters from

both techniques. The deduced clusters based on the Euclidean or Mahalanobis distances show a fair separation between different clusters. However, the amount of Gaussian noise and the high outliers' sensitivity causes the presence of false clusters labels. On the other hand, the weighted distance, Steiner distance, based clustering shows a high resistance to the outlier's presence and a highly smooth solution. As the purpose of clustering in borehole geophysics is to classify the dataset into different lithological rock types, the robustness of the cluster technique can be assessed by measuring the degree of perturbation of a data point that can suffer, while still being correctly assigned to its respective cluster.
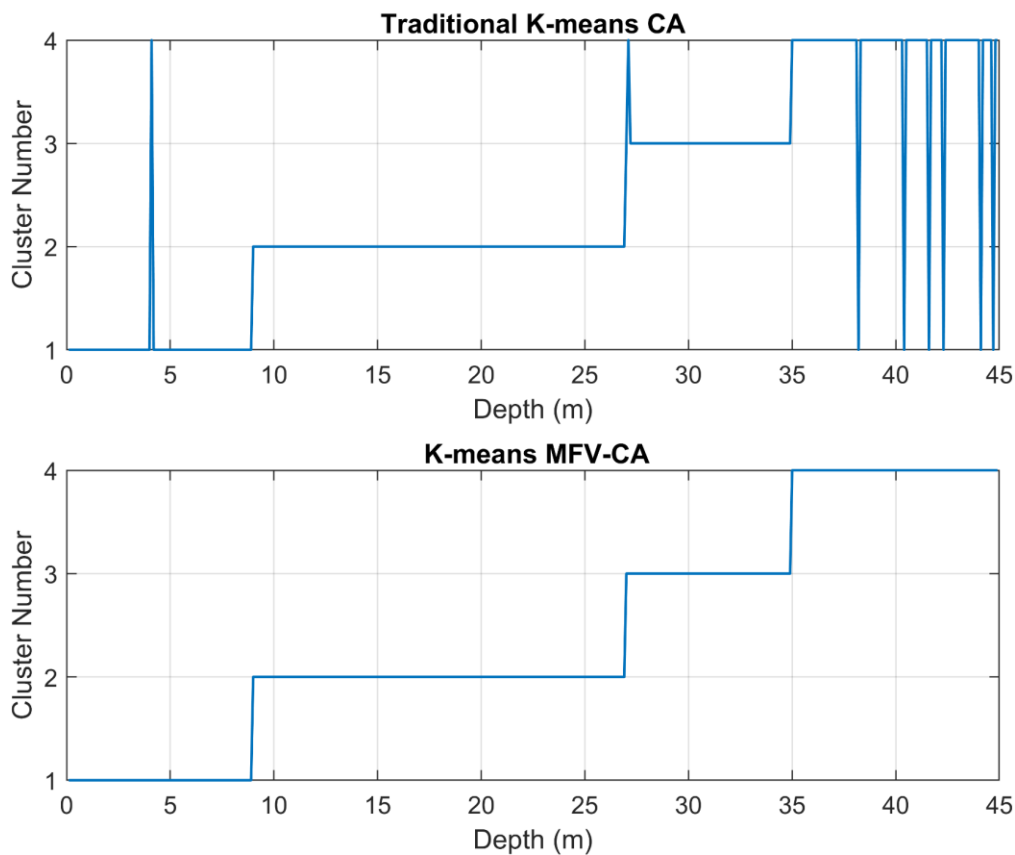
Figure 3. The resulting clusters from the traditional K-means clustering and the MFV-based clustering.

To assess the stability of the results of both clustering techniques and mitigate potential initial location dependency issues inherent to both methods, a comprehensive evaluation is carried out. Specifically, the test is repeated 100 times, with random selection of initial centroids. The arithmetic average is calculated to show that the convergence of the MFV-based cluster technique is more robust and not contingent on the initial location of the centroids. The MFV method exhibits a pronounced independence from the initial centroid location. Figure 4a vividly illustrates the remarkable stability of the results obtained through the MFV-based

approach. Conversely, in the case of the traditional clustering method, Figure 4b showcases a marked dependency on the initial centroid placement. Additionally, it underscores the considerable impact of outliers on the clustering outcome.

(a)   (b)
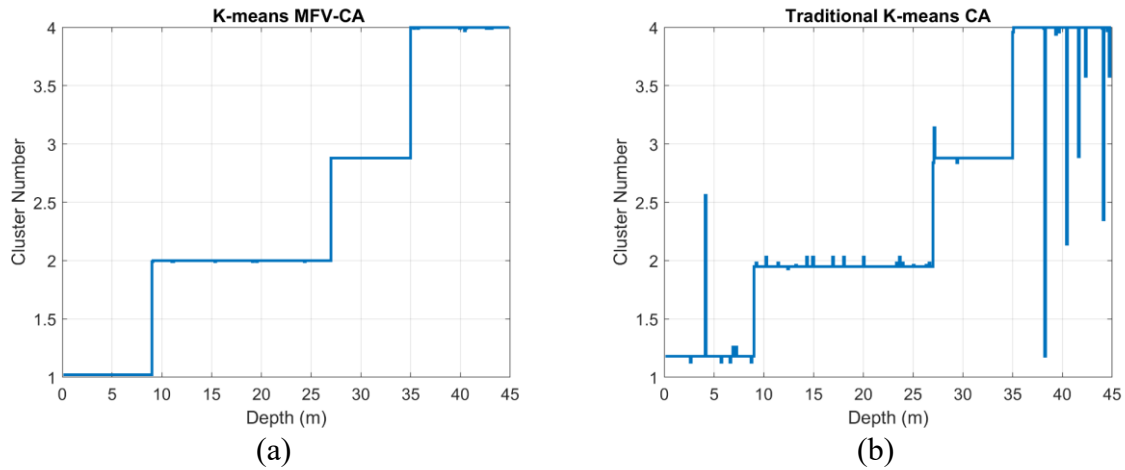
Figure 4. The arithmetic means of the results of 100 times-repeated tests (a) using the MFV-based clustering method, and (b) using the traditional K-means clustering method.

The descriptive statistics show that the range of Steiner distances is smaller than that of Euclidean or Mahalanobis distances. This implies that Steiner distances are less affected by outliers, making them more consistent when there are extreme values in the data. in other words, the sensitivity to the outliers is reduced in case of MFV-based clustering. Furthermore, the mean Steiner distance of 1.6 closely approximates the exact mode of the distance distribution function, whereas the Euclidean and Mahalanobis distances have a mean of 3.4 and 6.1, respectively. Additionally, the standard deviation of the Euclidean and Mahalanobis distances surpasses that of the Steiner distance. Finally, the sum of squared error favors the results obtained using Steiner distance. Table 2 provides a comprehensive summary of the descriptive statistics for both distances.

Table 2. The descriptive statistical parameters for Euclidean and Steiner distances.

	Mean	Standard deviation	Range	SSE
Euclidean distance	3.4	5.5	40.4	1522.3
Mahalanobis distance	6.1	11.9	92.9	2741
Steiner distance	1.6	1.5	9	547.7

Figure 5 shows the histogram for both distances, the histogram shows that the Euclidean and Mahalanobis distances have heavy tails compared to that of the Steiner distance. To sum up, the MFV-based clustering can solve the issue of initial location dependency and decrease the outliers' sensitivity. The findings of MFV-based clustering can delineate the lithological change in case of synthetic borehole geophysical data with high resolution and can be used for extracting the layer boundary locations.
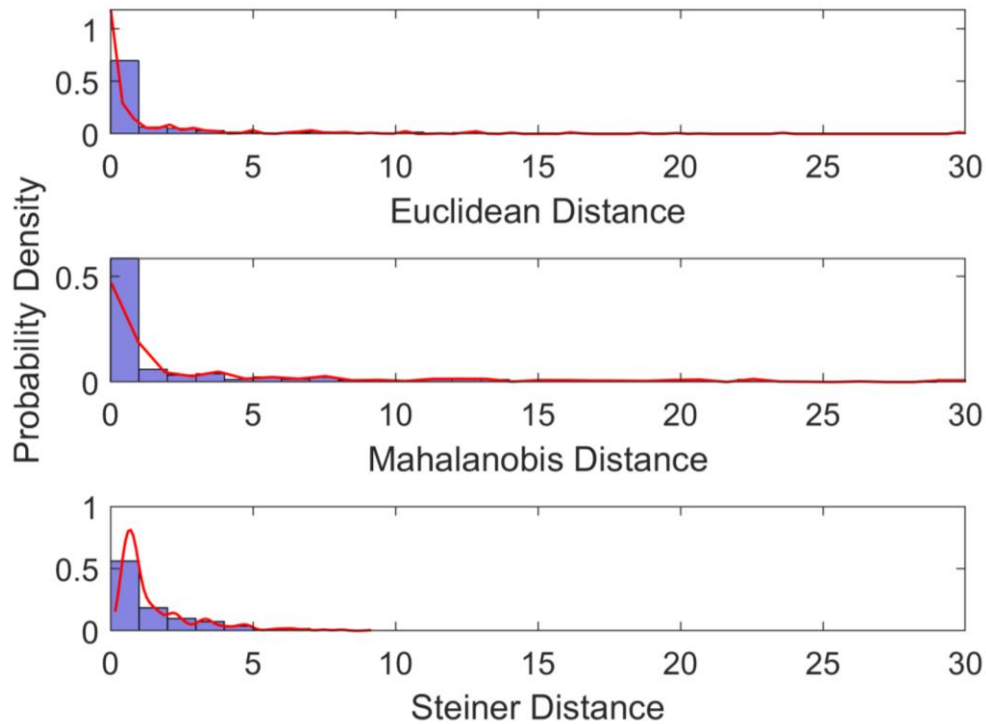
Figure 5. Frequency plot of the Euclidean and Steiner distances.

2.3. Field application of the MFV-based clustering technique

The MFV-based clustering technique has been used to classify the different lithological units of a field well-logging data set of hydrogeophysical exploration. The proposed clustering method has been applied on the Baktalórátháza-1 well drilled in Szabolcs-Szatmár-Bereg County, in northeast Hungary. The well logging dataset shows a response of shaly-sand sequence. The drilled well penetrated a sequence of layers fully saturated with water. The upper 80-100 m of the well was drilled into Pleistocene strata predominantly by sands. Sands of 100-160 m thickness had been deposited, followed by a shaly formation and 5-15 m thick coarse-grained beds. At a depth of 240 m, the border between the Pleistocene and Pannonian periods was detected (Szűcs et al. 2021). The cluster analysis was performed by three clusters, each

intended to delineate discrete categories representing varying levels of aquifer quality, specifically categorized as high, intermediate, and low.

The study identifies two clusters of aquifers: one with low natural gamma ray intensity, indicating high sandstone quality, and the other with high gamma ray intensity, indicating low shale content. Both clusters show a decrease in aquifer quality with depth. The MFV-based clustering method provides a higher-resolution depiction of layer continuity, particularly in deeper sections, whereas the traditional K-means approach erroneously identifies high-quality aquifers. The findings align with grain-size analysis of core samples, indicating that grain sizes remain relatively uniform within the deeper stratigraphic sequence. This confirms the robustness of the proposed clustering methodology. Figure 6 shows a 3D cross-plot representing the three clusters with respect to their GR, neutron-neutron, and gamma-gamma values. Figure 7 shows the clusters outcome for each depth point in vertical representation to show the continuity of the extracted lithological information. The last two tracks from the right side in Figure 7 shows the results from the traditional K-means and MFV-based cluster analysis, respectively.
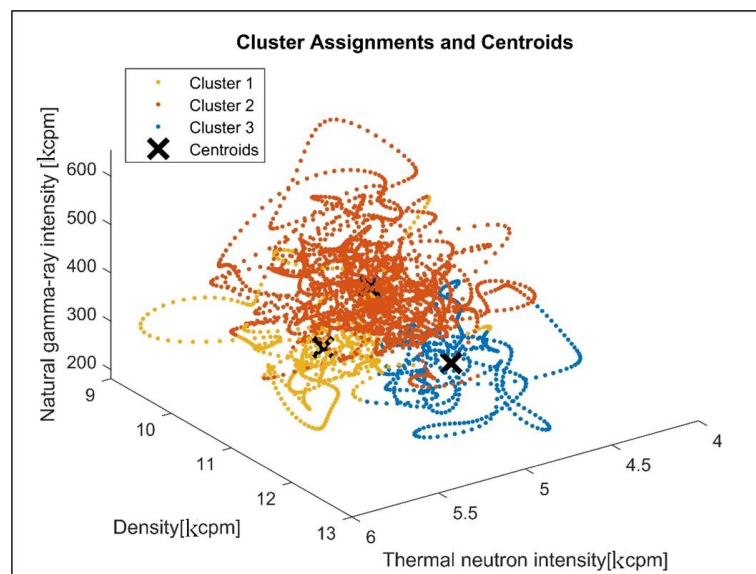
Figure 6. 3D cross-plot shows the MFV-based clustering results in a data space partition.
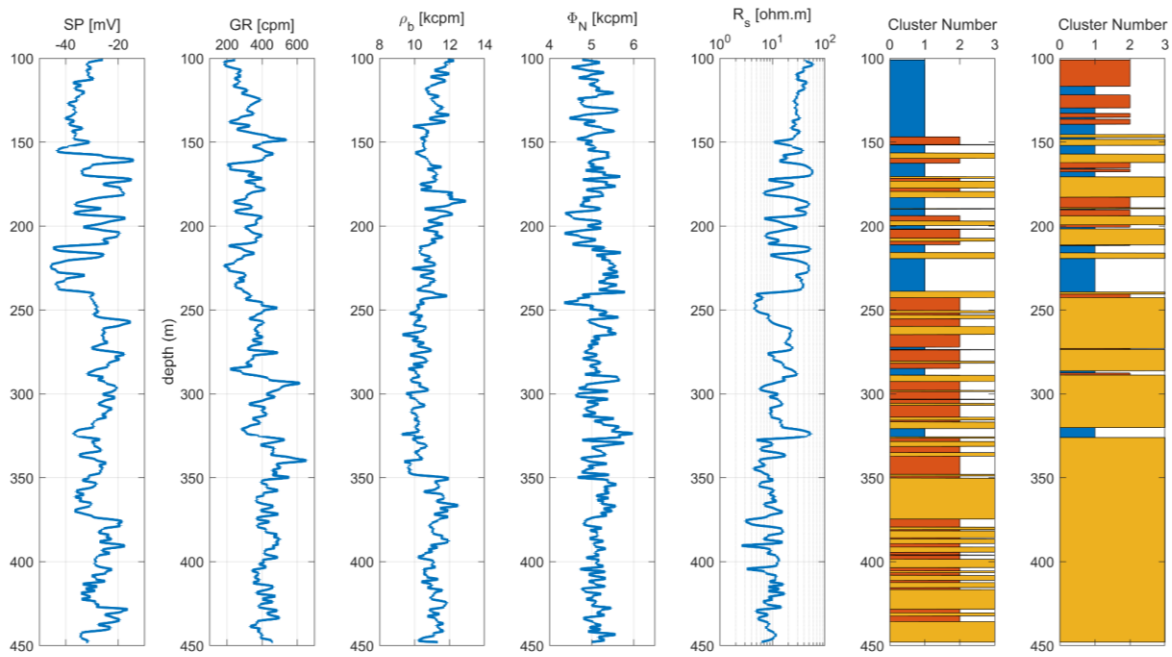
Figure 7. The depth distribution of clusters from both the K-means and MFV clustering methods.

2.4. Summary of results

The application of the K-means clustering technique on borehole log data resulted in spurious cluster labels and heightened susceptibility to outliers. To address this limitation, I introduced an innovative clustering method that integrates the iterative Most Frequent Value (MFV) algorithm with the traditional approach. This algorithm strategically incorporates the Steiner-Cauchy weights to assess the distances between data objects and their centroids and employs the MFV as a centroid update. The viability of this novel technique was assessed using synthetic data augmented with 3% Gaussian distributed noise and artificial outliers.

The statistical evaluation of the MFV-based clustering method demonstrated robustness against outlier sensitivity, as well as a remarkable stability of results even with random initialization of centroids. Moreover, the proposed clustering approach was applied to in situ borehole geophysical data. The MFV-based method emerged as a highly proficient tool for precise characterization of aquifer layers, particularly in deeper strata, surpassing the capabilities of the conventional K-means clustering approach. Its capacity to furnish a high-resolution depiction of layer continuity and to accurately discern regions of low-quality aquifers represents a substantial advancement in groundwater assessment. This proficiency was further affirmed by the congruence with grain size analysis findings and the consistent performance observed with synthetic data.

Thesis 1.

I have developed a robust cluster analysis method that can group well-logging data reliably by using the most frequent value (MFV) algorithm. The MFV-CA clustering algorithm can give the same cluster results with randomly chosen initial centroids. Besides that, the MFV-based clustering method effectively suppresses the outlying data during the classification of data objects. The optimal locations of the centroids are estimated by the MFV algorithm instead of the arithmetic mean of the cluster elements. I applied a weighting distance metric based on the Steiner weight function that so-called Steiner distance. I proved that Steiner distance can automatically identify the important range of the data, filtering out the noise caused by outliers. It retains the full scope of the dataset, ensuring all real data is preserved while minimizing distortion from extreme outlying values. In the latter case, I improved the results of rock typing in a Hungarian groundwater formation.

Chapter 3: A Meta-Algorithmic Framework for Regularization of Well Logging Inverse Problem

To improve the efficiency of the exploration techniques nowadays, advanced and robust processing methods should be introduced. The well-logging data interpretation is a vital source of information because it gives knowledge about rock characterization surrounding the location of drilling (Doveton 2001). The reservoir characterization procedure includes determined different types of parameters such as petrophysical parameters (i.e., porosity, rock matrix volumes, and fluid saturation), zone parameters (i.e., Archie's constants, shale parameters, matrix and pore-water parameters), and geometrical parameters (i.e., layer-boundary and thickness). The traditional methods of reservoir characterization suffer from the uncertainty in data interpretation, limited resolution, and difficulty in capturing complex geological features (Abordán and Szabó 2021). Besides that, the findings of the traditional methods suffer from cumulative errors and high correlation between the model parameters. Traditional well logging interpretation faces cumulative errors due to factors like data acquisition quality, calibration issues, relying on assumptions, and interpreter experience. These errors, beginning with tool malfunctions to variances in expert interpretation, will lead to significant inaccuracies in understanding subsurface conditions. This limitation can hinder the reliability of reservoir characterization and affect the decision-making procedure. Besides the deterministic (traditional) methods, geophysical inversion methods are widely used nowadays for accurate and reliable borehole logging interpretation (Dobróka et al. 2009).

Well-logging inverse problem is an iterative algorithm that is used to predict the reservoir parameters using the recorded logging data. Practically, the characterization of the reservoir parameters is carried out depth-by-depth point along the logged interval in the so-called local inversion. The local inversion used the linearized optimization algorithm to address the nonlinear relationship between the logging data and petrophysical parameters at certain depth point. Therefore, that concept introduces a poorly overdetermined inverse problem, that restricted the number of the unknowns to the number of recorded log data (Dobróka and Szabó 2015). In addition, the layer boundaries are unknown and cannot be estimated by local inversion. To overcome the limitations of the local inversion, a series expansion-based inversion can be used to estimate the reservoir characteristics within a predefined interval. This inversion procedure is called interval inversion (Dobróka et al. 2016). In contrast to the local inversion, the unknowns of the interval inversion are the series expansion of the basis functions

that represent the change in petrophysical parameters within a certain depth interval. The significant overdetermination ratio (as I can set significantly lower number of expansion coefficients than data) guarantees high accuracy and reliability in parameters estimation, moreover, it allows the implementation of further parameters in the inverse problem.

In both inversion techniques, the hyper-parameters that control the optimization problem are so important to guarantee the stability of the prediction procedure. The damping factor of the Damping Least Square (DLSQ) scheme is used to overcome the ill-posed problem (Nhu 2022; Levenberg 1944) . An ill-posed problem arises from the high condition number of the sensitivity matrix $\mathbf{G}$ (Jacobian matrix). In the context of inverse problem solution, a high condition number of the matrix $\mathbf{G}^T \text{Covd}^{-1} \mathbf{G}$ indicates that the problem is ill-conditioned, and the inverse of the previous matrix product cannot be calculated without regularization (Gavin 2019).

The sensitivity matrix's high sensitivity indicates that even minor changes in model parameters can cause significant fluctuations in predicted outcomes. This amplification makes it difficult to obtain a stable, accurate inverse solution without regularization, which introduces additional information or constraints. (Van Rijn and Hutter 2018). When an inversion problem is ill-posed due to a high condition number, it can lead to unstable solutions. This instability means that small errors or uncertainties in the observed data can result in large errors in the estimated model parameters. In practical terms, this can make it very difficult to obtain reliable and meaningful results from the inversion process. To address this issue, various techniques are employed, including regularization methods. These techniques help stabilize the inversion procedure by adding constraints or penalties to the optimization problem, which can mitigate the effects of an ill-posed problem (Prabhu 2018).

3.1. Theoretical overview of linearized inversion methods

The well-logging data and estimated parameters are connected by the so-called tool response functions. These response functions are characterized by their nonlinear nature. The linearized inversion procedure involves approximating the nonlinear relationship between the observed data and the estimated parameters using Taylor series truncated at the first order derivative (Menke 1984). By reducing the problem to sequences of linear problems, one can take advantage of well-established numerical methods for solving linear systems. This makes the overall process more computationally tractable and facilitates the convergence to a solution

(Menke and Eilon 2015). Let $\vec{m}_0$ be the starting model including the petrophysical parameters to be determined by inversion

$$\vec{m} = \vec{m}_0 + \delta\vec{m}, \tag{22}$$

where $\delta\vec{m}$ is the model correction vector. The chosen initial model parameters $\vec{m}_0$ will be refined iteratively until the stopping criterion is met. In other words, the $\vec{m}$ model vector will be calculated by modifying $\vec{m}_0$ with $\delta\vec{m}$ in the subsequent iteration step.

3.1.1. Linearized local (depth-by depth point) inversion

Depth-point-by depth point inversion in well logging involves analyzing and interpreting subsurface properties at each individual depth point, allowing for high-resolution insights into local geological variations and characteristics. To formulate the linearized local inversion, the vector of M number of model parameters and the N number of data vector can be given as

$$\vec{m} = \{m_1, m_2, \cdots, m_M\}^T, \tag{23}$$

$$\vec{d}_{obs} = \{d_1{}^{(obs)}, d_2{}^{(obs)}, \cdots, d_N{}^{(obs)}\}^T, \tag{24}$$

where T denotes the symbol of transpose. By using Taylor series truncated at first order derivative, the relationship between the data and model parameters can be approximated in the following form

$$d_{ek} = g_k(\vec{m}_0) + \sum_{j=1}^{M} \left(\frac{\partial g_k}{\partial m_j}\right) \delta m_j, \quad (k = 1, 2, \cdots, N) \tag{25}$$

By introducing the Jacobi matrix as

$$G_{kj} = \left(\frac{\partial g_k}{\partial m_j}\right)_{\vec{m0}}, \tag{26}$$

equation (26) can be written in vectorial form as

$$\vec{d}_e = \vec{d}^{(0)} + \mathbf{G}\delta\vec{m}, \tag{27}$$

This leads to a linear set of equations

$$\delta\vec{d}_e = \mathbf{G}\delta\vec{m}, \tag{28}$$

the objective function of the linearized inversion procedure is to reduce the error between the calculated and measured data

$$\vec{e} = \vec{d}^m - \vec{d}^0 - \mathbf{G}\delta\vec{m}, \tag{29}$$

when having more data than unknowns, in the case of DLSQ scheme, the model update equation can be introduced as follows

$$\delta\vec{m} = (\mathbf{G}^{\mathrm{T}}\mathbf{G} + \lambda\mathbf{I})^{-1}\mathbf{G}^{\mathrm{T}}\delta\vec{d}^{m}, \tag{30}$$

where λ is the damping factor (Levenberg 1944). Figure 8 shows the local inversion workflow.
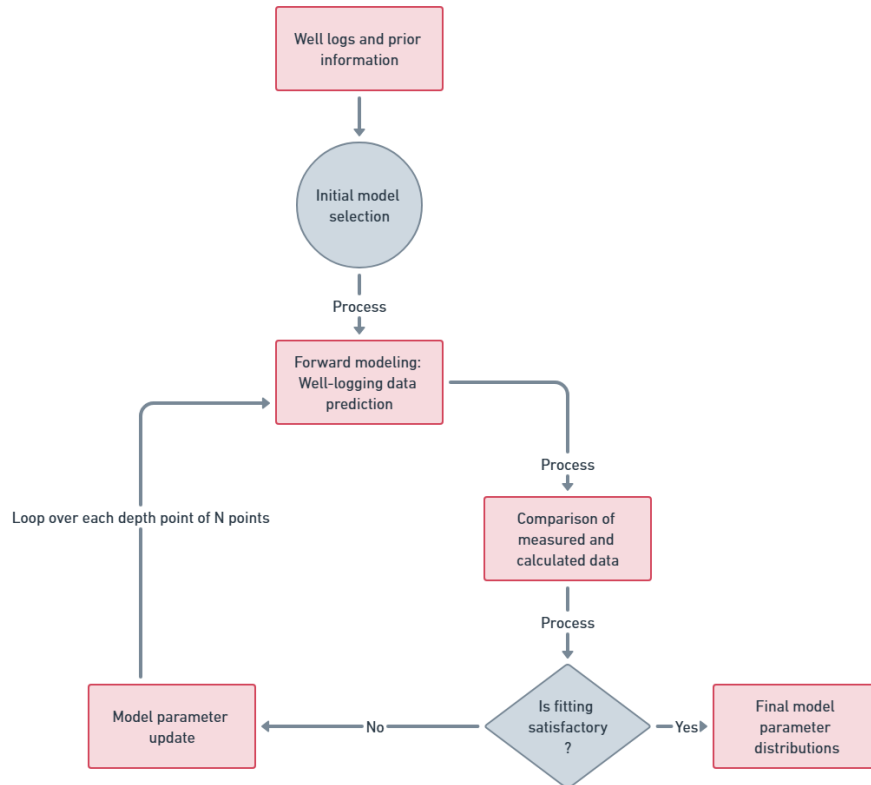
Figure 8. Local inversion workflow.

3.1.2. Linearized interval inversion

Since the number of the measured well-logs data is usually marginally more than the number of predicted model parameters at certain depth point, the local inversion has low overdetermination ratio that causes some limitation regarding to the accuracy and reliability of the estimated model parameters (Szabó and Dobróka 2020). An approach called interval inversion algorithm was introduced by the Geophysical Department of Miskolc University to invert data of greater interval simultaneously in one inversion procedure (Dobróka et al. 1991). The series expansion-based inversion algorithm interprets the response function as a depth dependent function. The relationship between the k-th observed data and model parameters at z depth can be written as follows

$$d^{(obs)}{}_k(z) = g_k(m_1(z), m_2(z), \dots, m_M(z)), \tag{31}$$

where petrophysical parameters in equation (31) are modeled using continuous functions that necessitate discretization for numerical computation. The discretization of the *i-th* spatial dependent model parameter can be written in the following form (Dobróka 1993)

$$m_i(z) = \sum_{q=1}^{Q} B_q^{(i)} \Psi_q(z), \tag{32}$$

where $B_q^{(i)}$ is the *q-th* discretization coefficient of series expansion polynomial function of Q degree of *i-th* model parameter, while Ψ_q is the *q-th* basis function that considered as known quantities. The basis function of the homogenous layer-wise model can be represented by the combination of two Heaviside functions that define a certain depth interval

$$\Psi_q(z) = u(z - Z_{q-1}) - u(z - Z_q), \tag{33}$$

where Z_{q-1} and Z_q are the upper and lower boundaries of the *q-th* interest interval, respectively. This combination of step functions introduces the basis function as zero in whole the depth points except the in the *q-th* layer where the basis function equal to 1. In the upcoming chapters I will introduce an innovative inversion algorithm that can predict the layer boundary coordinates and provide it as a priori information to the petrophysical parameters inversion phase. By substituting equation (32) into equation (31), the well-logging data can be calculated along the observed interval

$$d^{(obs)}{}_k(z) = g_k\left(B_1^{(1)}, \dots, B_Q^{(1)}, \dots \dots, B_1^{(M)}, \dots, B_Q^{(M)}, z\right). \tag{34}$$

On the contrary to the local inversion, the unknowns of the interval inversion are the series expansion coefficients $\vec{B}$. By using the DLSQ algorithm, the initial model can be iteratively refined, in case of interval inversion, using equation (35). The interval inversion workflow can be seen in Figure 9.

$$\delta B_q^{(i)} \Psi_q(z) = (\mathbf{G}^T\mathbf{G} + \lambda\mathbf{I})^{-1} \mathbf{G}^T \delta d^{B_q^{(i)}} \Psi_q(z), \tag{35}$$
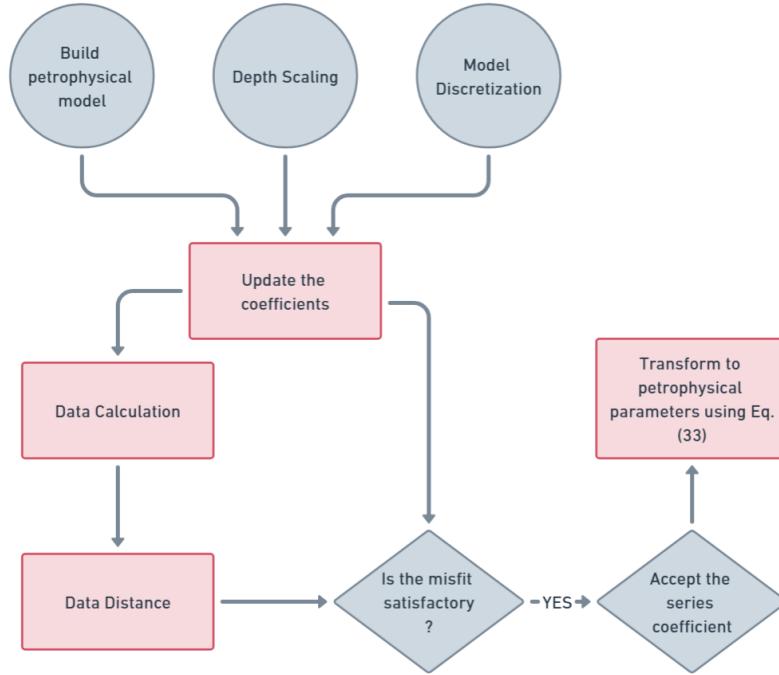
Figure 9. The workflow of the interval inversion.

3.2. Meta-algorithm assisted regularized interval inversion method

The inversion process uses hyperparameters like the damping factor to stabilize convergence to the optimal solution. The damping factor is a positive number decreased by iterations until near zero, as the initial model is affected by data noise. The ill-posed problem can be represented mathematically using the condition number

$$K = \frac{\lambda_{max}}{\lambda_{min}},$$

(36)

The terms λ_{max} and λ_{min} denote the $\mathbf{G}^{\mathrm{T}}\mathbf{G}$ matrix's maximum and minimum eigenvalues. An ill-posed problem is characterized by a high condition number, whereas a well-posed problem has a small condition number for the system of solving equations. Consequently, selecting an appropriate damping factor can mitigate instability and enhance the condition number. Abdelrahman et al. (2021) and Heriyanto and Srigutomo (2017) showed the feasibility of extracting the damping factor from the decomposition of the inverted matrix, in which the damping factor is a function of the eigenvalues of the decomposed sensitivity factor. The linearized interval inversion technique was combined with two search strategies—Golden Section Search (GSS) and Singular Value Decomposition (SVD)—to develop enhanced algorithms that aid in stabilizing the optimization process. This combination assists in finding the most suitable damping factor needed to address ill-posed problems, which are marked by a

34 | P a g e

high condition number, unlike well-posed problems that have a low condition number in their equations.

3.2.1. Golden Section Search (GSS) assisted interval inversion

The one of the unimodal optimization algorithms is the Golden Section Search (GSS) method. It can be called an interval reduction optimization process. Furthermore, the unimodal can be defined as a continuous function over an interval $[a,b]$, if there is one optimum point included in the same interval $[a,b]$ (Noroozi et al. 2022).

The GSS algorithm serves two primary objectives. The first aims to determine the optimal damping factor within a given interval, while the second seeks to achieve this with the fewest number of functions call possible. The theoretical foundation of interval reduction optimization methods, such as the bisection method or equal interval method, start by selecting a midpoint between a and b, denoted as $m = (a + b)/2$ for the bisection method, and introduce a small positive value of d, then, define two points, $X_1 = (m - d)/2$ and $X_2 = (m + d)/2$. If $f(X_1) < f(X_2)$, then the interval is reduced to $[a, X_1]$. Otherwise, the reduced interval becomes $[X_2, b]$. Figure 10 shows the intuition of the equal interval method (Davim 2008).
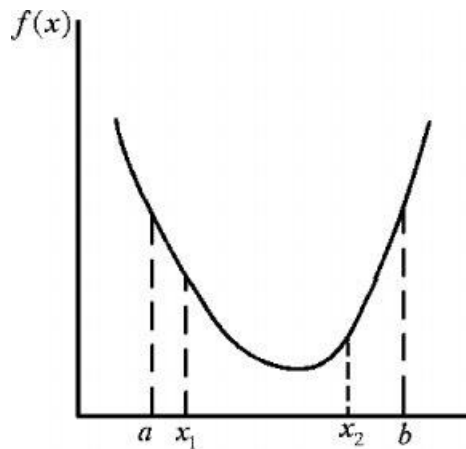
Figure 10. Equal interval method intuition.

The equal interval method experiencing prolonged computation times to find the optimum solution (Chang 2015). On the other hand, the GSS method can overcome this problem by choosing three points X_L, X_1 and X_u ($X_L < X_1 < X_u$) along the x-axis with corresponding values of the function $f(X_L)$, $f(X_1)$, and $f(X_u)$, respectively. Since $f(X_1) > f(X_L)$ and $f(X_1) > f(X_u)$ (Yazıcı et al. 2021). The maximum must lie between X_l and X_u. Now a fourth point denoted by X_2 is chosen to be between the larger of the two intervals of $[X_L, X_1]$ and $[X_1, X_u]$. The intermediate points X_1 and X_2 are chosen such that, the ratio of the distance from these

points to the boundaries of the search region is equal to the golden ratio which is computed as $(\sqrt{5}-1)/2$ times the difference between X_u and X_L (Pejic and Arsic 2019). Subsequently, the integration between the GSS and interval inversion algorithm proceeds to evaluate the inversion loop, extracting the data distances associated with the lower, upper, and intermediate points. Finally, a quality check is made to see if the difference between X_u and X_L falls below a specified threshold value. If this condition is met, the damping factor is determined as $(X_u + X_L)/2$. At this point, the outer loop is terminated, and the model with the optimal damping factor is accepted. This comprehensive algorithm guides the process of conducting the Golden Section Search, ultimately leading to the identification of the most suitable damping factor.

3.2.2. Singular Value decomposition (SVD) assisted interval inversion

As was mentioned in section 2.2, the ill-posed problem is based on the condition number which is function of the minimum and maximum eigenvalues of the $\mathbf{G}^T\mathbf{G}$ matrix. Meju (1992) introduced an iterative algorithm based on factorizing the sensitivity matrix (Jacobian matrix) into three other matrices as follows:

$$\mathbf{G}(B^i) = \mathbf{U}\mathbf{S}\mathbf{V}^T, \qquad (37)$$

where $\mathbf{U}$ is (NxQ_M) data eigenvector, $\mathbf{V}$ is (Q_MxQ_M) model parameters eigenvector, and $\mathbf{S}$ is (Q_MxQ_M) matrix with eigenvalues in its diagonal. The eigenvalues of $\mathbf{G}(B^i)$ are positive numbers with $i \leq M$ (number of parameters). The ill-posed problem will arise in case of small eigenvalues of the term $\mathbf{G}(B^i)^T\mathbf{G}(B^i)$ of equation (36) in section 2.1.1. Therefore, the SVD scheme recommends adding a positive bias to the eigenvalues of that term. By substituting equation (38) into Equation (36)

$$\delta B_q^{(i)} = (\mathbf{V}\mathbf{S}^2\mathbf{V}^T + \lambda\mathbf{I})(\mathbf{G}^T)\delta d^{B_q^{(i)}}, \qquad (38)$$

by modifying the diagonal non-zero values using the damped factor (Ekinci and Demirci 2008), the first term of equation (38) can be written as

$$(\mathbf{V}\mathbf{S}^2\mathbf{V}^T + \lambda\mathbf{I}) = (\mathbf{V}\,\mathrm{diag}(\eta_i{}^2)\mathbf{V}^T) + \lambda^2\mathbf{I} = \mathbf{V}\,\mathrm{diag}(\eta_i{}^2 + \lambda^2)\mathbf{V}^T, \qquad (39)$$

where ηi is the *i-th* eigenvalue, the following equation shows the inverse of equation (39)

$$(\mathbf{V}\,\mathrm{diag}(\eta_i{}^2 + \lambda^2)\mathbf{V}^T)^{-1} = \mathbf{V}\,\mathrm{diag}\left\{\frac{1}{\eta_i^2+\lambda_i^2}\right\}\mathbf{V}^T, \qquad (40)$$

Consequently, the final equation to update the model parameters provided by can be written as

$$B^{i+1} = B^i + \mathbf{V} \operatorname{diag}\left\{\frac{\eta_i}{\eta_i^2 + \lambda_i^2}\right\} \mathbf{U}^\mathrm{T} \delta d^{B^i}. \tag{41}$$

The damping factor is determined through a process of comparing the data misfit of the L-th test at any given iteration with that of the $(L+1)$-th test. If an improvement in misfit is observed, the new factor is accepted. This technique is known as the RIDGE method, which involves augmenting the eigenvalues of the diagonal matrix with a positively biased value (Arneson and Hersir 1988). This adjustment is calculated using the following equation

$$\lambda = \eta_L \, \Delta X^{\frac{1}{L}}, \tag{42}$$

where ΔX can be given by:

$$\Delta X = \frac{X_{R-1} - X_R}{X_{R-1}}, \tag{43}$$

where X_{R-1} and X_R are the previous and present iterations misfit.

3.3. Reliability and error estimation

The accuracy and reliability of inversion outcomes are established by examining how uncertainties in data that are observed propagate into the estimated model parameters. It is accomplished by covariance analysis, which measures variances and correlations among variables. Specifically, the model parameter covariance matrix is found from the data covariance matrix by using a generalized inverse matrix (Equation 44). This inverse matrix represents the structure of the forward model and contains any regularization or damping techniques applied during inversion to stabilize solutions and dampen noise.

$$\operatorname{cov}\mathbf{m} = \mathbf{G}^{-g} \cdot \operatorname{covd} \cdot \mathbf{G}^{-g\mathrm{T}}, \tag{44}$$

Uncertainties in data measurements are known, typically represented by the diagonal elements of the data covariance matrix and can be propagated through the inversion to provide estimates of model parameter uncertainties. These parameter uncertainties are the diagonal terms of the model covariance matrix, each of which is the variance of a particular parameter. The standard deviation of a parameter, which is the square root of its variance, gives the direct measure of confidence in its estimated value (Equation 45). These standard deviations give a measure of the degree to which each parameter is responsive to the quality and quantity of input data. In order to additionally assess the reliability of inversion results, one can calculate the correlation coefficient between pairs of model parameters (Equation 46). The coefficient measures the

extent of linear dependency between parameters. Values near +1 or -1 indicate a high interdependency, which is often a sign of instability or lower reliability of the solution.

$$\sigma_{mi} = \sqrt{cov\left(\mathbf{m}_{ii}\right)}, \qquad (45)$$

In Singular Value Decomposition (SVD)-based error estimation, the stability and reliability of inversion results are enhanced by decomposing the forward modeling operator into its principal components. This decomposition isolates the influence of data noise on model parameters by distinguishing between well-constrained and poorly resolved features in the solution. Small singular values, which amplify noise and destabilize results, are systematically damped through regularization, adjusting their contribution to the inverse operator based on their magnitude relative to a damping parameter. The SVD framework directly links the data uncertainties to model parameter uncertainties via the eigenvectors and damped singular values, allowing the calculation of the model covariance matrix. This matrix quantifies both the variance of individual parameters (indicating their sensitivity to data noise) and the correlations between parameters (highlighting interdependent uncertainties). By adaptively tuning the damping during inversion, SVD balances resolution and stability, ensuring robust uncertainty quantification while mitigating overfitting to noisy data.

$$\text{corr}(\mathbf{m})_{ij} = \frac{cov(\mathbf{m})_{ij'}}{\sqrt{cov(\mathbf{m})_{ij} cov(\mathbf{m})_{jj}}} \qquad (46)$$

In error estimation using Singular Value Decomposition (SVD), the stability and reliability of inversion solutions are enhanced by breaking down the forward modeling operator into its dominant components. The breakdown isolates the influence of data noise on model parameters by separating well-constrained and poorly resolved features within the solution. Small singular values, broadcasting noise and destabilizing the solutions, are consistently damped by regularization—scaling their effect on the inverse operator as a function of size compared to a damping parameter. The SVD formalism places the data errors in direct relation with the errors in the model parameters via the eigenvectors and damped singular values, and thus the model covariance matrix can be computed. This matrix quantifies both the variance of single parameters (their sensitivity to data noise) and parameter correlations (indicating interdependent uncertainties). By adaptive damping tuning during inversion, SVD breaks the resolution vs. stability trade-off to generate solid uncertainty quantification without overfitting noisy data.

$$\text{cov}\mathbf{m} = \mathbf{V} \, \text{diag}\left\{\frac{\eta_i}{\eta_i^2 + \lambda_i^2}\right\} \mathbf{U}^T \text{covd } \text{U diag}\left\{\frac{\eta_i}{\eta_i^2 + \lambda_i^2}\right\} \mathbf{V}^T \tag{47}$$

3.4. Synthetic modeling experiments

To test the feasibility of the inversion meta-algorithm, I generated noise free, and noise contaminated synthetic well-logging datasets as quasi measured inputs. The initial forward model can be constructed using the following equations (Alberty and Hashmy 1984):

$$\text{GR} = \rho_b{}^{-1}\left(V_{sh}\, GR_{sh}\rho_{sh} + \sum_{i=1}^{N} V_{ma,i} GR_{ma,i}\, \rho_{ma,i}\right), \tag{48}$$

$$\rho_b = \emptyset\left[\rho_{mf} - 1.07(1 - S_{xo})(\alpha_0\rho_{mf} - 1.24\rho_h)\right] + V_{sh}\rho_{sh}$$
$$+ \sum_{i=1}^{N} V_{ma,i}\, \rho_{ma,i}\,, \tag{49}$$

$$\emptyset_N = \emptyset\left\{\begin{array}{c}\emptyset_{N.mf} - (1 - S_{x0})C_{cor} \\ -2\emptyset(1 - S_{x0}) - S_{hf}(1 - 2.2\rho_h) \\ .\,[1 - (1 - S_{x0})(1 - 2.2\rho_h)]\end{array}\right\} + V_{sh}\rho_{sh} + \sum_{i=1}^{N} V_{ma,i}\, \emptyset_{ma,i}\,, \tag{50}$$

$$P_e = \frac{1.07}{\rho_b + 0.19}\left[\emptyset S_{x0}U_{mf} + \emptyset(1 - S_{x0})U_h + V_{sh}U_{sh} + \sum_{i=1}^{N} V_{ma,i}\, U_{ma,i}\right], \tag{51}$$

$$\frac{1}{\sqrt{R_d}} = \left[\frac{V_{sh}{}^{(1-0.5V_{sh})}}{\sqrt{R_{sh}}} + \frac{\sqrt{\emptyset}^m}{\sqrt{aR_w}}\right]\sqrt{S_w}^n, \tag{52}$$

$$\frac{1}{\sqrt{R_s}} = \left[\frac{V_{sh}{}^{(1-0.5V_{sh})}}{\sqrt{R_{sh}}} + \frac{\sqrt{\emptyset}^m}{\sqrt{aR_{mf}}}\right]\sqrt{S_{x0}}^n, \tag{53}$$

$$\emptyset + V_{sh} + \sum_{i=1}^{N} V_{ma,i} = 1. \tag{54}$$

The zone parameters that are used for creating synthetic well-logging datasets can be seen in Table 3.

Table 3. Zone parameters are used for the forward modeling of borehole geophysics (Szabó and Dobróka 2020).

Well-log	Zone Parameter	Symbol	Value	Dimensional unit
Natural gamma-ray intensity	Sand	GR_{sd}	10	API
	Shale	GR_{sh}	154	API
Gamma-gamma (bulk density)	Sand	ρ_{sd}	2.65	g/cm^3
	Shale	ρ_{sh}	2.54	g/cm^3
	Mud filtrate	ρ_{mf}	1.02	g/cm^3
	Hydrocarbon (gas)	ρ_h	0.2	g/cm^3

Neutron-porosity	Sand	$\Phi_{N,sd}$	-0.04	v/v
	Shale	$\Phi_{N,sh}$	0.31	v/v
	Mud filtrate	$\Phi_{N,mf}$	0.95	v/v
	mf correction coefficients	C_{cor}	0.69	-
	Residual hydrocarbon	S_{hrf}	1.2	-
Deep resistivity	Shale	R_{sh}	1	Ohm.m
	Pore water	R_w	0.5	Ohm.m
	Mud filtrate	R_{mf}	0.28	Ohm.m
	Cementation exponent	m	1.5	-
	Saturation exponent	n	1.8	-
	Tortuosity factor	a	1	-
Photo-electric absorption index	Sand	U_{sd}	4.8	Barn/cm^3
	Shale	U_{sh}	9	Barn/cm^3
	Mud filtrate	U_{mf}	0	Barn/cm^3
	Hydrocarbon(gas)	U_h	0	Barn/cm^3

The synthetic data is calculated based on the three layers suggested model. The synthetic data represents well-logs calculated of sandstone layers filled with hydrocarbon and water and separated by a shale layer. 20-degree Legendre polynomials were used to fit the synthetic model. The calculated well-log data vector consists of natural gamma-ray, bulk density, neutron-porosity, compressional sonic transit-time, shallow and deep resistivities, and photoelectric absorption logs. The Oil-Water Contact (OWC) suggested being located at 6 m.

3.4.1. Synthetic data-driven GSS-assisted interval inversion

The synthetic data was inverted using the combined DLSQ and GSS-based interval inversion algorithm. Figure 11 shows the initial model and the synthetic well-logging data based on the suggested model parameters. The calculated dataset is contaminated by 5% Gaussian distributed noise. Figure 12 shows petrophysical parameters used for calculating both the synthetic data (black solid lines) and the initial model (dashed red lines). (porosity, volume of sand, volume of shale, water saturation of uninvaded zone, and water saturation of the invaded zone). In Figure 13, the fitting between the synthetic data and the predicted data using the GSS-based interval inversion is illustrated. Figure 14 shows the fitting of the predicted model parameters to the actual model parameters of the synthetic model. The inversion procedure is stopped after 40 iterations to check the convergence trends for the conventional inversion and GSS-based inversion procedures.
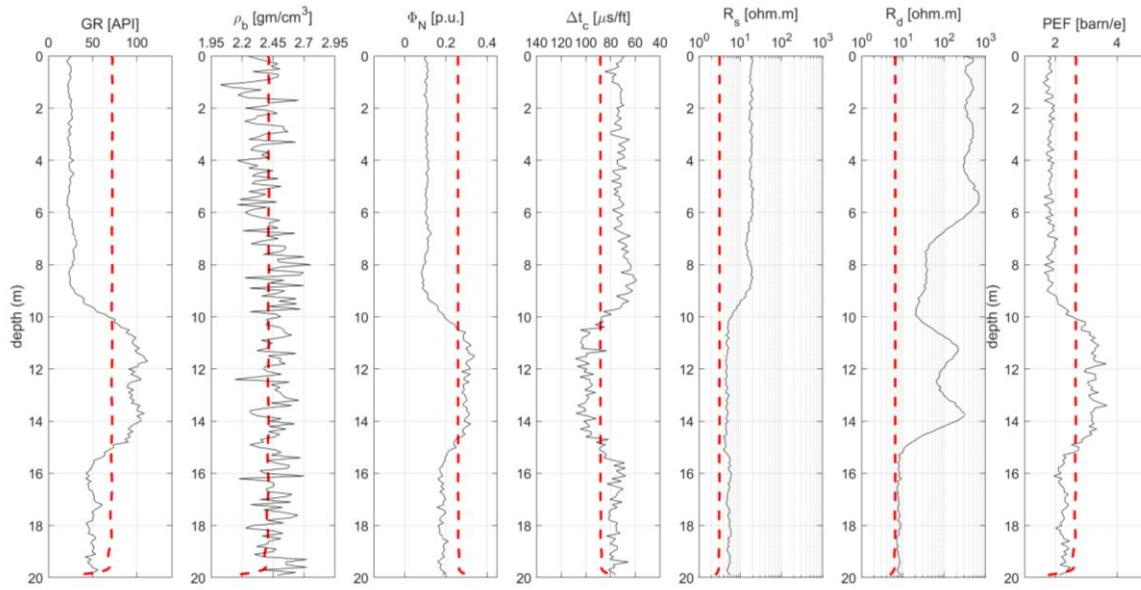
Figure 11. 5% Gaussian noise-contaminated synthetic data; the red dashed lines represent the data calculated using the initial model, while the solid black lines represent the synthetic data.
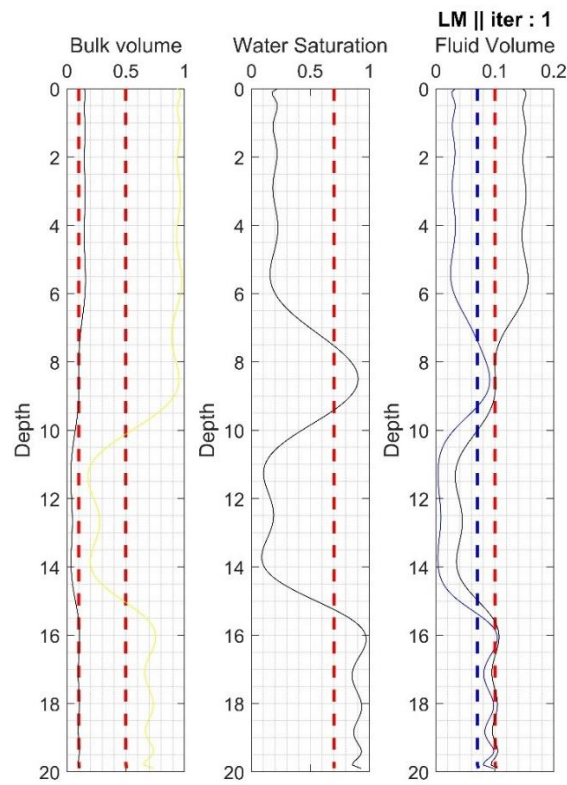
Figure 12. The model parameters used for calculating the synthetic data, and the initial model parameters; the red dashed lines represent the initial model, while the solid black lines represent the synthetic model approximated by Legendre polynomials.
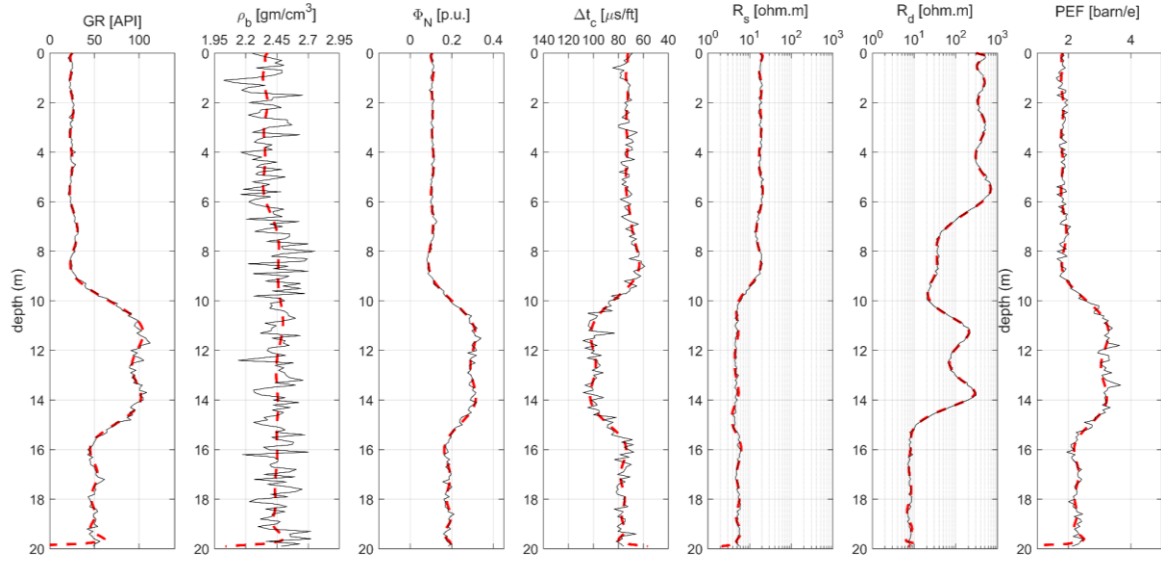
Figure 13. Fitting between the synthetic data and the calculated data (after 40 iterations); the red dashed lines represent the calculated data, while the solid black lines represent the synthetic data.
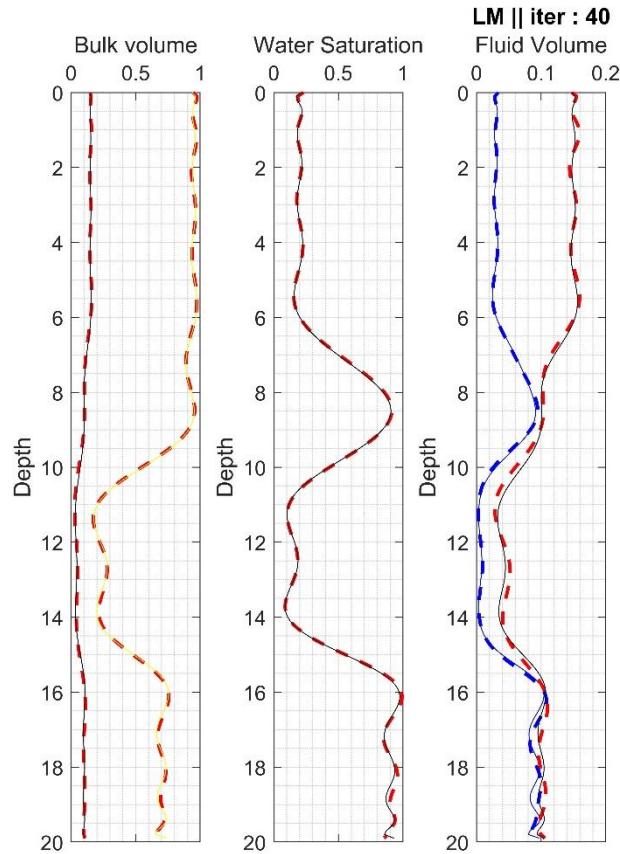
Figure 14. The synthetic and the predicted model parameters (after 40 iterations); the red dashed lines represent the predicted model, while the solid black lines represent the synthetic (target) model.

Considering the parameter count and polynomial degree, the discretized model parameters involve 105 unknowns, derived from 5 times (20 plus 1). Simultaneously, the

overdetermination ratio can be computed by multiplying the logging data per depth (*NL*), which is 7, by the total number of data points (depth points, *N*), which is 200. This yields 1400 data points. Consequently, the overdetermination ratio for the proposed synthetic model is approximately 13, calculated as 1400 divided by 105. To quantitively assess how accurate the model is, the misfit or discrepancy between measured and calculated data from a model is calculated and expressed as data distance. By calculating the data distance, one can assess the effectiveness of the inversion process. A smaller data distance indicates a better fit between the model and the observed data, suggesting that the inversion process is working well. Conversely, a larger data distance points to significant differences between the model predictions and the actual measurements, indicating the need for further refinement of the model or inversion parameters to achieve a more accurate representation of the underlying phenomenon. The choosing of a high damping factor can cause a delay in the convergence patterns of the data distance, which represents an increase in the number of iterations. Moreover, the convergence of the data distance shows what I can call a plateau effect. This plateau data distance causes no change in the optimization process and directly affects the condition number and increases the damping effect, therefore, increases the number of the iteration. Comparatively, the GSS-based interval inversion exhibits a swift and smooth convergence of data distance, whereas the conventional interval inversion demonstrates a more uneven convergence pattern. In Figure 15, the data distance for both the noise-free dataset and the dataset contaminated with 5% Gaussian distributed noise is presented. Notably, a plateau effect is clearly observed. The convergence stability iterations differ notably between the conventional interval inversion applied to the noise-free and contaminated datasets, whereas the GSS-based method displays consistent convergence stability across both scenarios. Specifically, data distance convergence for the conventional interval inversion began to stabilize around the 35th iteration for the 5% Gaussian noisy dataset and the 22nd iteration for the noise-free dataset. In contrast, for the GSS-based interval inversion, data distance convergence began to stabilize at approximately the 10th iteration in both cases.
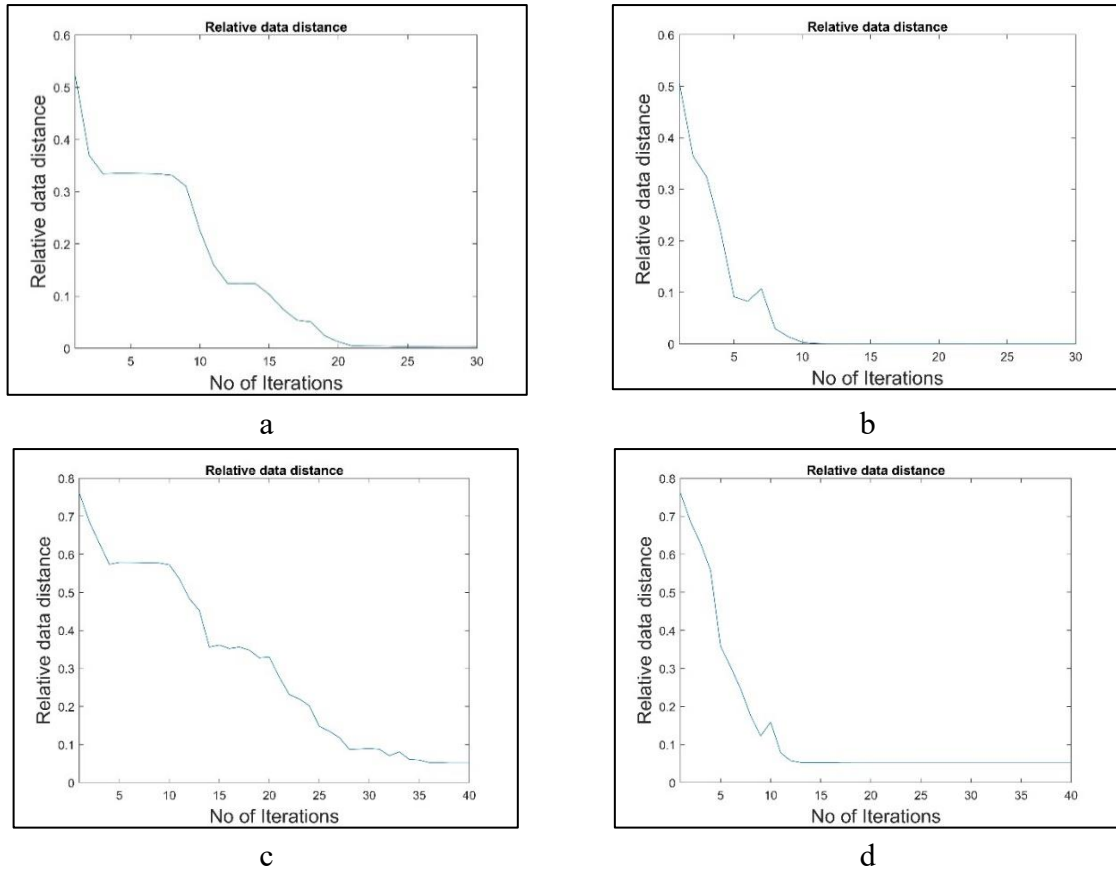
Figure 15. Data distance resulting from the conventional (a and c) and GSS-based interval inversion (b and d) procedures. Subfigures (a) and (b) pertain to the noise-free dataset, while (c) and (d) depict the contaminated dataset.

3.4.2. Synthetic data-driven SVD-assisted interval inversion

In the case of the SVD-assisted interval inversion, the synthetic data were calculated using the zone parameters of Table 3. The Heaviside function and Legendre polynomials were used to simulate petrophysical parameters changes within layers, resulting in a layer-wise homogeneous model consisting of four layers varying in lithological constituents, porosity, and saturation content. Figure 16 and Figure 18 show the fitting between the synthetic data (bold-black lines) and the calculated data (dashed-red lines) at iteration 1 and 60, respectively. Figure 17 and Figure 19 show the actual and the predicted petrophysical parameters at iteration 1 and 60, respectively. The SVD based inversion procedure shows a smooth convergent pattern, but it shows a rapid convergence until the data distance is lower to 5% and too slow convergent until reaching the zero data distance (Figure 20). The author suggests using the proposed algorithm for higher unknowns, integrating DLSQ and SVD schemes. The synthetic data from section 2.3.1 was used, but convergence development was time-consuming. The integration aims to automatically determine the optimal damping factor during inversion, using the last damping factor as prior information.
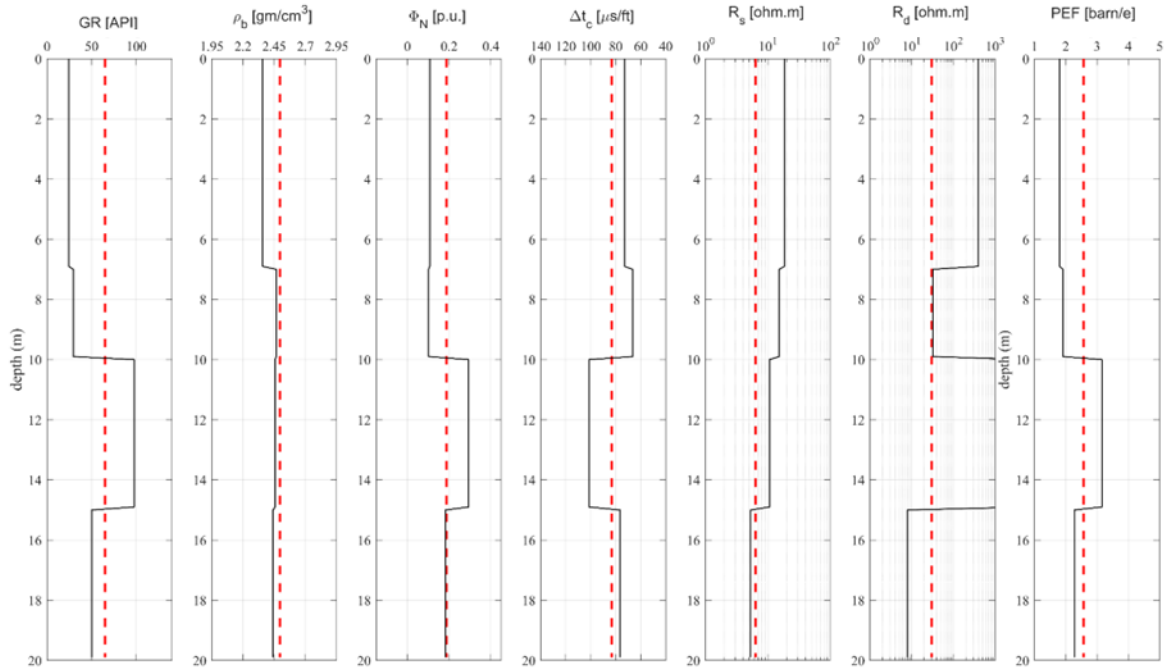
Figure 16. The synthetic (input) and calculated (output) data using the Heaviside basis functions; the red dashed lines represent the data computed using the initial model, while the solid black lines represent the synthetic data.
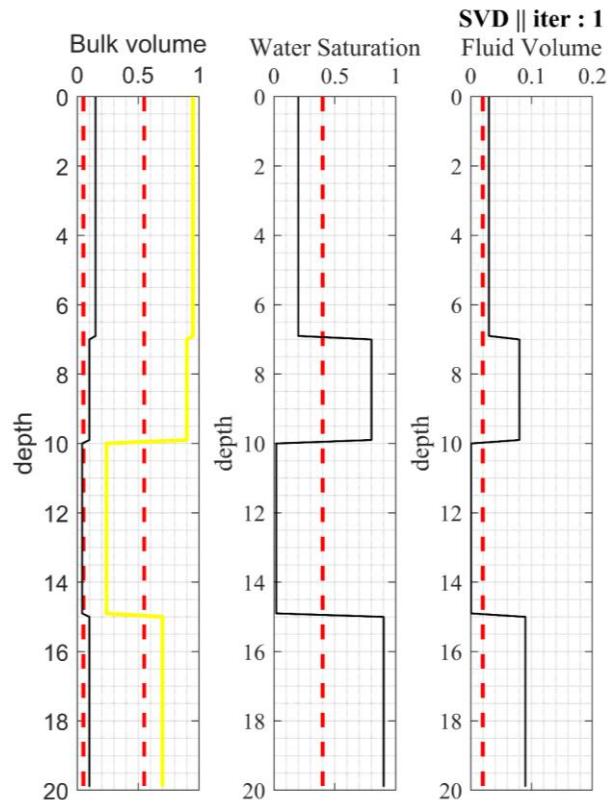
Figure 17. The synthetic and the predicted model parameters (iteration 1); the red dashed lines represent the initial model, while the solid black lines represent the synthetic model.
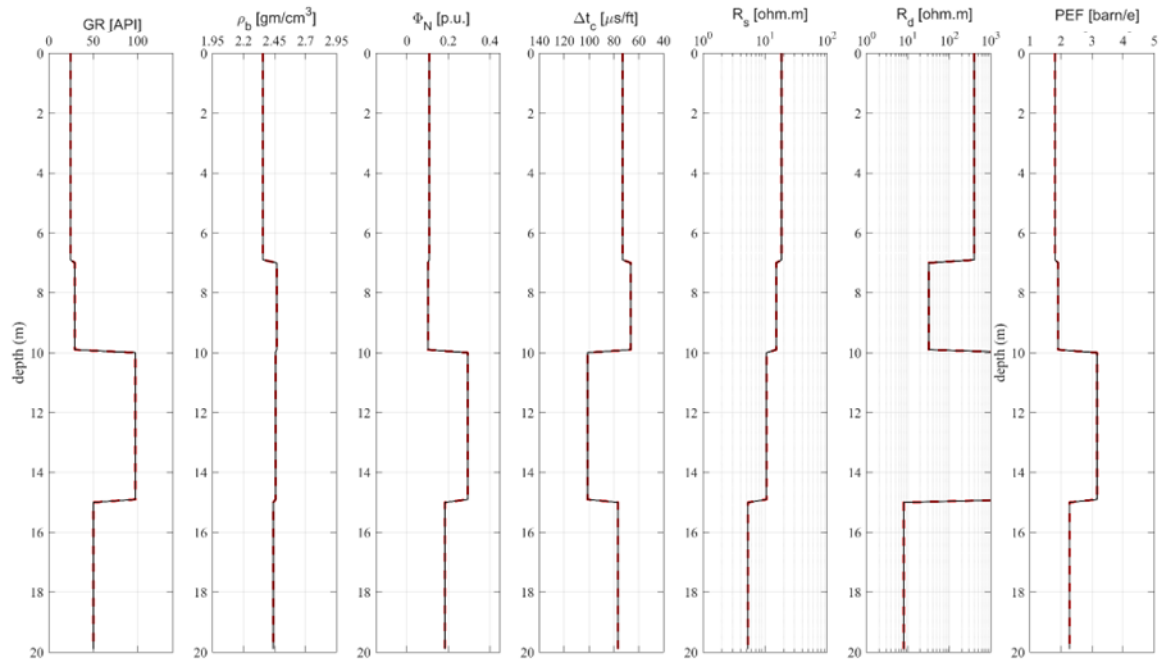
Figure 18. Fitting between the synthetic data and the calculated data (at iteration 60); the red dashed lines represent the calculated data over the estimated model, while the solid black lines represent the synthetic data.
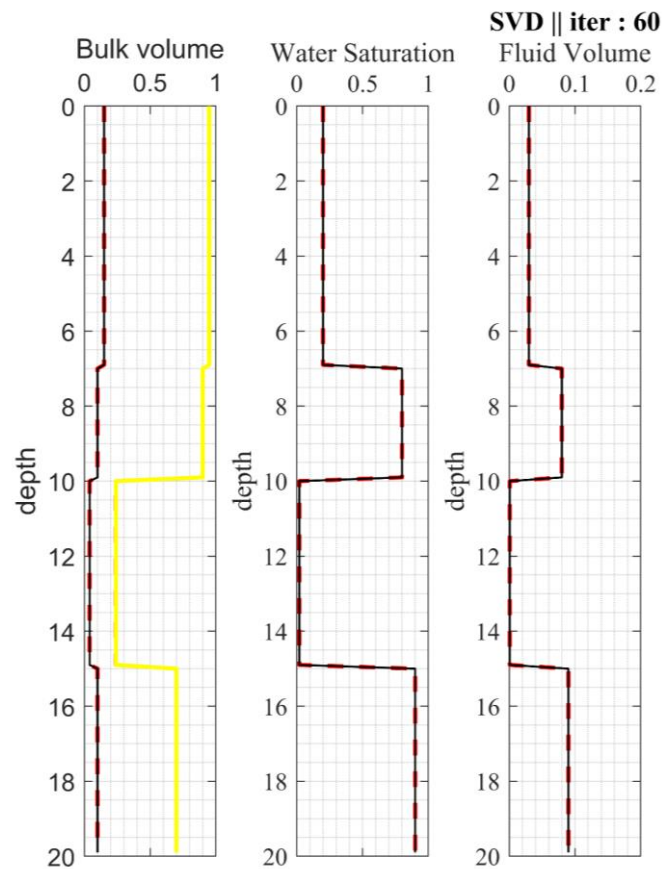
Figure 19. The synthetic (exactly known) and the predicted model parameters (at iteration 60); the red dashed lines represent the predicted model, while the solid black lines represent the synthetic model.
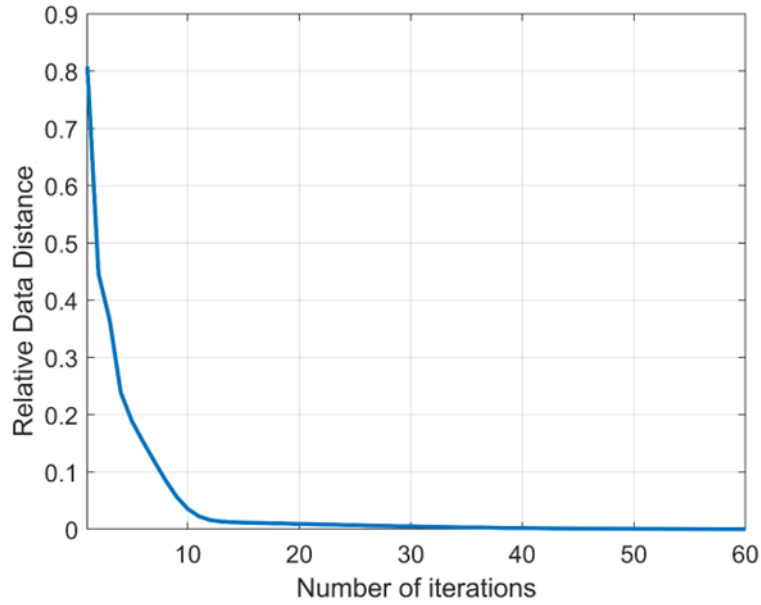
Figure 20. The convergence of data distance (the standalone SVD-based interval inversion).

As mentioned in section 2.3.1, the number of unknowns is 105, therefore, the product of the matrix $\mathbf{G^TG}$ will exhibit a size of 105-by-105, while in the previous case (Heaviside) the size of the same matrix is 3-by-3. It is obvious that the high degree polynomials will consume a high CPU time to extract the optimum damping factor by checking 105 eigenvalues. Therefore, I start the inversion procedure with SVD-based algorithm that automatically damps the inversion procedure by testing the eigenvalues of the inverted matrix. Then, after the convergent reaches the 25 % data distance, I change for the fast DLSQ algorithm (Figure 21).
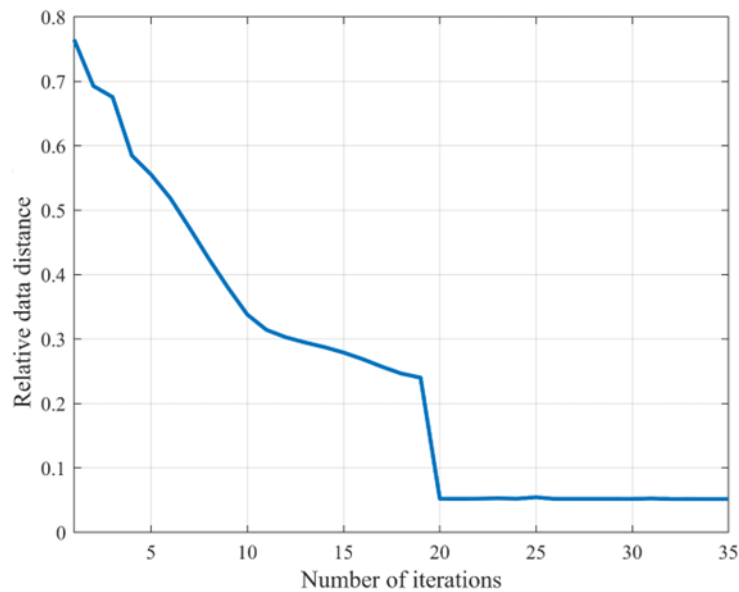
Figure 21. The convergence of data distance for the hybrid SVD-based interval inversion procedure.

3.5. Field application of the meta-algorithm-based interval inversion methods

The feasibility of both inversion meta-algorithms has been checked using a well-logging dataset measured from a gas bearing reservoir of an Egyptian field located at the northwestern part of the Western desert of Egypt. The reservoir belongs to the Jurassic sandstone sequence with high degree of heterogeneity (Abdelrahman 2021; Wahdan et al. 2013). Figure 22 and Figure 23 show the fitting between the field data and the calculated data at iteration 1 and 60, respectively. The thickness of the zone of interest is 37 m. Figure 24 shows that the reservoir consists mainly of sandstone layers with some shale laminations that affect both the storage and flow capacity.
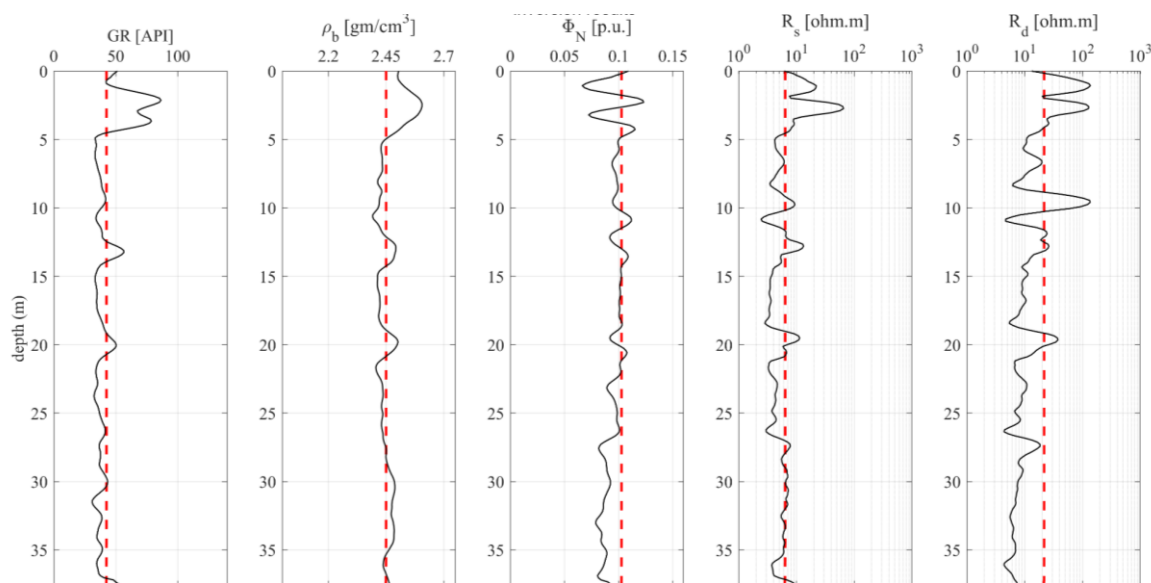
Figure 22. SVD-DLSQ inversion result. Fitting between the field data and the calculated data (iteration 1); the red dashed lines represent initial calculated data, while the solid black lines represent the field data.

The outcomes of the study reveal a noteworthy observation: reservoir parameters exhibit variations that are contingent upon the quality of the reservoir. This implies that these crucial parameters are not consistently uniform, but rather, they adapt and adjust in response to the varying degrees of reservoir quality. This insight underscores the dynamic nature of reservoir characteristics, emphasizing the need for tailored assessment strategies that account for these fluctuations.
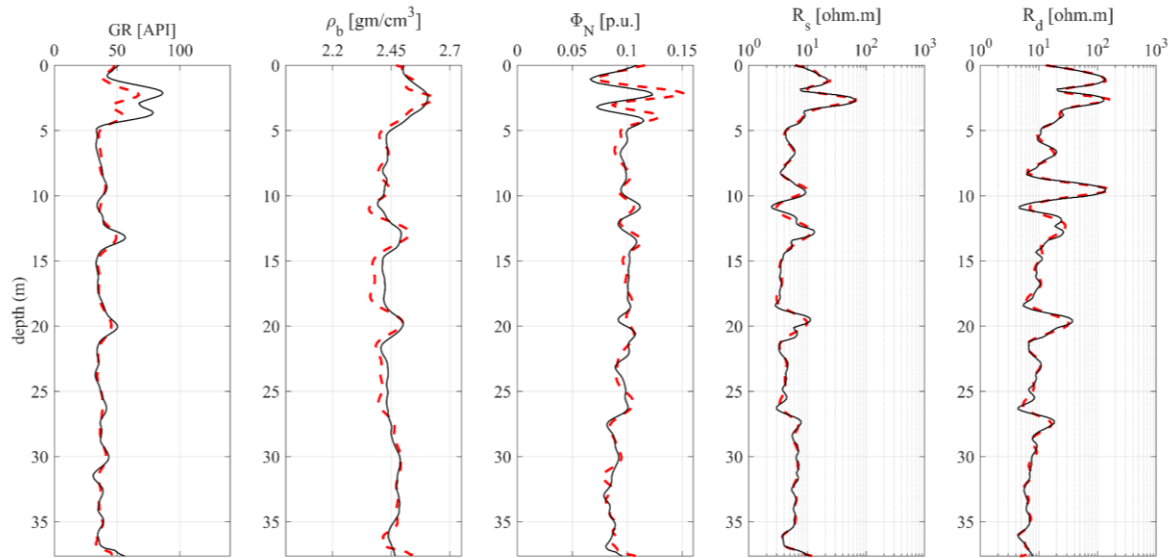
Figure 23. SVD-DLSQ inversion result. Fitting between the field data and the calculated data (iteration 60); the red dashed lines represent the initial calculated data, while the solid black lines represent the field data.
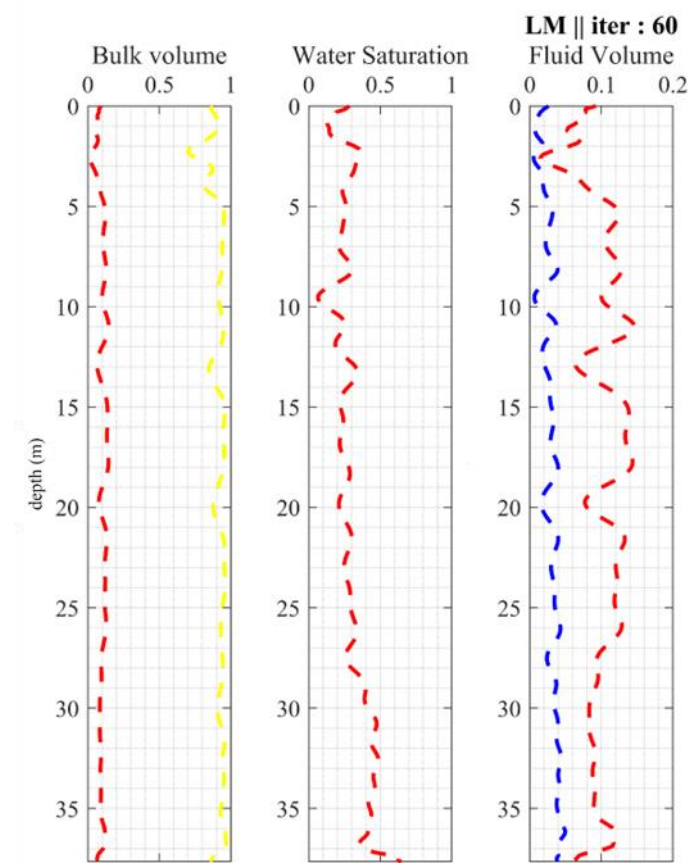
Figure 24. The predicted model parameters from SVD-DLSQ inversion method (iteration 60).

The final predicted parameters using the SVD- and GSS-based inversion methods are almost the same with a slightly smaller data distance in the case of the DLSQ-SVD-based

inversion. However, what stands out is the convergence trajectory of both the data distance and the number of iterations required to achieve stabilization. Figure 25a illustrates the data distance convergence for the DLSQ algorithm, while Figure 25b depicts the data distance convergence for the GSS-based algorithm. Additionally, Figure 25c demonstrates the convergence of the SVD-DLSQ-based algorithm. In the case of the standalone DLSQ algorithm, the convergence pathway exhibits not a steady convergence, with stability achieved at iteration 40. The use of the GSS based algorithm could reduce the number of iterations to lower than 25 iterations, moreover, the convergence trajectory shows a smooth convergence. Furthermore, employing the SVD-DLSQ method reveals a notably smooth and rapid convergence pattern, with stabilization occurring as early as iteration 9, demonstrating a consistently regular trajectory.
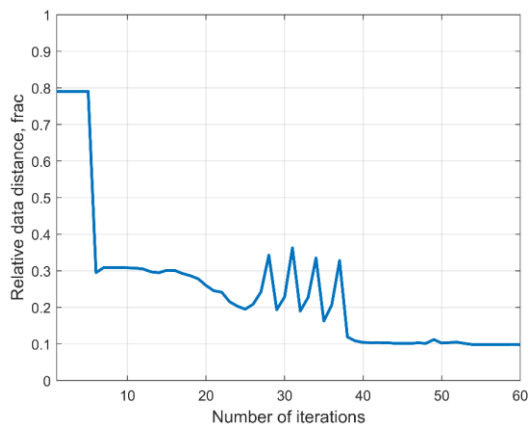
(a)
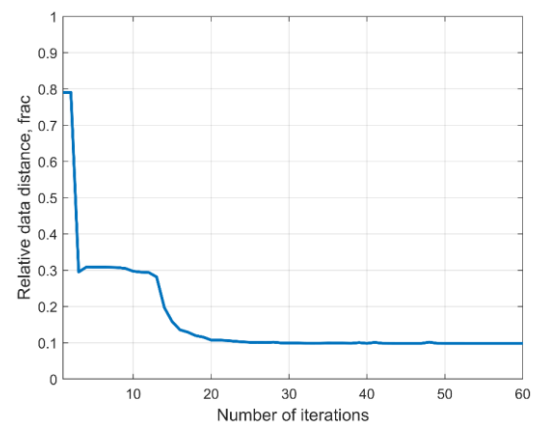
(b)
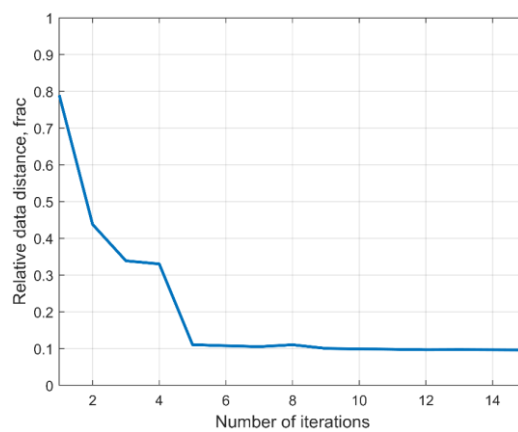
(c)

Figure 25. Data distance convergence curves for (a) DLSQ algorithm, (b) GSS-based algorithm, and (c) SVD-DLSQ algorithm.

The data distance of the initially calculated data is around 79% for all the inversion algorithms. The DLSQ algorithm shows the irregular data distance from 25 to 38 iterations. On the other hand, the GSS algorithm could avoid those irregularities, thus it could converge faster, but the integration between the SVD method and the DLSQ inversion algorithm could avoid the iteration consuming as well as the irregulars. For studying the effect of data variance on the solution of the inverse problem, I adopt the following values of standard deviation (data accuracy) for the observed well-logs: $\sigma_d(GR)=0.0484$, $\sigma_d(\rho_h)=0.0484$, $\sigma_d(\phi_N)=0.005$, $\sigma_d(R_s)=0.002$ and $\sigma_d(R_s)=0.05$. Figures (26 and 27) illustrate the results of both GSS-based DLSQ and SVD methods in petrophysical parameter estimation, with notable differences in reliability metrics. The SVD approach produced smoother error curves with narrow errors margin, compared to the Gss-based algorithm, which exhibits tighter but wider parameter error bounds. The correlation matrices reveal fundamental differences in parameter interdependence, with SVD showing generally lower correlation values compared to GSS-based algorithm, indicating that GSS-based algorithm may underestimate parameter relationships. This observation is quantitatively supported by the mean spread values—a scalar reliability measure ranging from 0 to 1, with GSS-based DLSQ yielding 0.16 versus SVD's 0.10. The higher mean spread in SVD suggests more comprehensive uncertainty quantification, making it potentially more suitable for robust reservoir characterization. These differences highlight how algorithm selection significantly impacts uncertainty estimation in well log analysis, with SVD based algorithm offering a more conservative but potentially more reliable assessment of parameters uncertainty.

3.6. Summary of results

This chapter is devoted to the investigation of inversion algorithms designed for estimating reservoir characteristics. It emphasizes the crucial importance of hyperparameters such as the damping factor in maintaining convergence stability during the optimization process. A comparison of the GSS-based and traditional interval inversion approaches demonstrates that the GSS approach has a noticeably smoother convergence pattern. This discrepancy is highlighted further by the studying of a plateau effect in the traditional technique, which affects convergence stability and necessitates a higher number of iterations, particularly in scenarios involving noisy datasets.

Furthermore, the research investigates a novel integration of the DLSQ and SVD schemes with the goal of improving computational efficiency. This integrated inversion

approach strategically leverages the strengths of both methods, using SVD to automatically determine an optimal damping factor in the early stages of inversion and DLSQ with the last known damping factor value for further refinement as the solution approaches its optimum state. These meta-algorithms' practical use is proven using well-logging data from a gas-bearing reservoir in an Egyptian field. This validation procedure confirms the dynamic character of reservoir data, arguing for customized evaluation methodologies to account for these intrinsic changes. Results demonstrated that SVD produced smoother error curves with narrower error margins compared to GSS-based methods, which exhibited wider parameter error bounds. Correlation matrices showed SVD generated lower correlation values than GSS-based algorithms, suggesting the latter may underestimate parameter relationships. This was further supported by mean spread values (a reliability measure from 0-1), with GSS-based DLSQ at 0.16 versus SVD's 0.10. The smaller value of mean spread in SVD indicates more focused uncertainty quantification, potentially making it more suitable for robust reservoir characterization.

Finally, this study gives essential insights into the complexities of reservoir parameter estimation and highlights the efficiency of various inversion approaches in generating consistent and accurate findings.
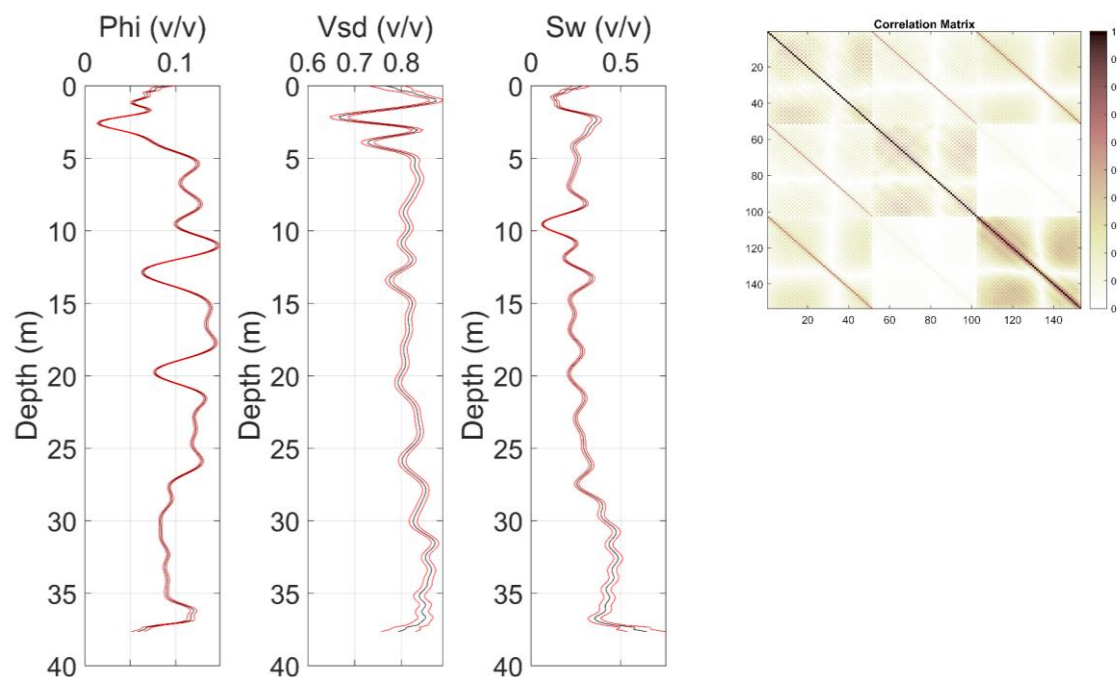
Figure 26. Quality check of petrophysical parameters estimated by GSS-based interval inversion procedure, separately. Solid Black lines represent the estimated values of porosity Φ, volume of sand V_{sd}, and water saturation S_w. Solid red lines show the error bounds of

petrophysical parameters calculated from the standard deviations σ of inversion estimates (left), the correlation matrix between the predicted parameters can be seen on the right side of the figure.
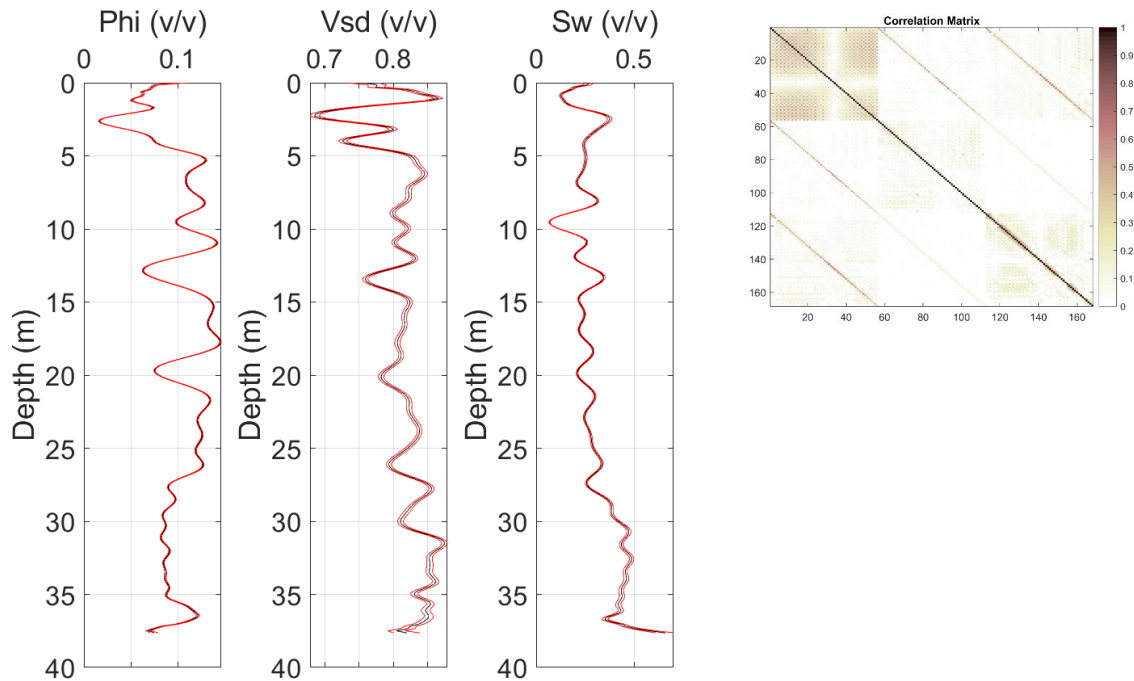
Figure 27. Quality check of petrophysical parameters is estimated by SVD-based interval inversion procedure, separately. Solid Black lines represent the estimated values of porosity Φ, volume of sand V_{sd}, and water saturation S_w. Solid red lines show the error bounds of petrophysical parameters calculated from the standard deviations σ of inversion estimates (left), the correlation matrix between the predicted parameters can be seen on the right side of the figure.

Thesis 2.

I have developed two combined well logging inversion algorithms to improve reservoir parameter estimation. For achieving good convergence, hyperparameters, particularly the damping factor, must be accurately estimated. A comparison between GSS-based and DLSQ interval inversion methods showed the former has superior convergence smoothness and noise resistance. I suggested the first hybrid of DLSQ and SVD methods to achieve optimal computational efficiency. This composite inversion method uses SVD for automatic selection of the initial damping factor and DLSQ for improving convergence. When applied to the well-log data of an Egyptian gas reservoir, the algorithms yielded clear information about the reservoir. The error estimation showed that SVD gave very good results with smaller

confidence intervals and lower values of parameter correlation. These results show the importance of having specific techniques for effective reservoir appraisal.

Chapter 4: Developments using the Hurst Exponent as Fractal Analysis for Unearthing Hidden Patterns

4.1. Introduction

Fractal analysis has been commonly utilized in geophysical exploration. For instance, it can be used in gravity and magnetic methods to characterize the causative sources. Furthermore, it can be used to characterize earthquakes in seismology (Ouadfeul et al. 2012). In petrophysics, the fractal analysis can be used to monitor the reservoir properties continuity and lithological classification (Lozada-Zumaeta et al. 2012). Nowadays, the use of fractal analysis has been included in seismic facies recognition. The wavelet based fractal analysis can be used to waveform classifier and facies recognition (López and Aldana 2007). In this chapter, from the side of method development, I show a new alternative for data processing by integrating the Hurst exponent as a fractal analysis method with interval inversion.

The proposed methodology uses multivariate statistical methods for dimensionality reduction, identifying influential data types and extracting unmeasurable information. Factor analysis reduces observed variables to factors reflecting linear combinations of input variables factor analysis, as demonstrated by Szabó and Dobróka (2013), can uncover interrelationships between statistical variables taken from observed data and some fundamental petrophysical features that are valid in diverse geological contexts. In their analysis, they discovered a strong association between the first statistical component describing a large portion of the variance of the input data and shale volume in Hungarian and North American wells. In groundwater investigations, the evaluation of shaly formations has various precedents (Szabó et al. 2014; Dennis and Lawrence 1984; Neasham 1977), including the calculation of hydraulic conductivity based on near-surface geophysical data (Niwas and Celik 2012). In this work, I analyze the fractal properties of the first factor log using Hurst exponent analysis and utilize it to determine the depth of layer boundaries. Several ways for recognizing faces using fractal analysis have already been described. López and Aldana (2007) enhanced lithofacies classification using Wavelet Based Fractal Analysis and Wavelet Classifier. To determine depositional patterns and trends, the natural gamma ray and porosity logs were studied. The high fractal dimensions were discovered to be related to shales, whereas the low ones were shown to relate to sands. Hernandez-Martinez et al. (2013) employed a multifractal Hurst analysis to identify electrofacies and highlight the intricacies of well-logging data. Hurst

exponent is connected to Hurst (1951) rescaled range analysis (R/S) approach, which was responsible for dam construction along the Nile River. Hurst was motivated by Einstein's random process description, which is referenced in Milton and Okamoto (2018).

Hurst examined the variation of water levels around the mean overtime and sought to determine if the volatility was random or had a concealed regular pattern. The suggested R/S approach can discriminate between entirely random and associated time series. T. Li et al. (2021) demonstrated that the analysis of fractals in well-logging data may be extended to fracture diagnosis, and he concludes that the degree of fracture development is linked to the increase in the acoustic log and dual lateral differential fractal dimensions. The high-frequency energy information, according to the proper scale, indicates notably high values in the fractured zone.

The proposed algorithm consists of a sequence of phases, starting with the Hurst exponent to identify lithology changes and layer boundaries for reliable petrophysical parameter estimation, followed by interval inversion, an indirect interpretation approach (Alberty and Hashmy 1984). I use a joint inversion procedure to estimate the petrophysical parameters of several layers. The calculation of the scale and depth dependent Hurst exponents using instantaneous slope rather than average slope allows for tracking the homogeneity of a subinterval and, as a result, recognizing the subinterval borders. This technique aims to develop an overarching framework to investigate the inhomogeneity inside each layer and detect an accurate location of the layer boundaries using the FA-Hurst exponent integration. An extra algorithm aids in the calculation of a hydraulic conductivity log generated from the inversion data. Csókás (1995) proposed an assessment approach based on the Kozeny-Carman equation that determines hydraulic conductivity solely from well logs. The workflow's practicality was validated using both synthetic well logs and field data gathered in East Hungary, and the estimation findings were confirmed using core data.

4.2. Methodology and Workflow Intuition

4.2.1. Factor analysis of well logs

As the aim of the Hurst exponent analysis is to extract the geometrical information of the layers, the first step aims to collect all the lithological information from the well-logging data set in one log to perform the fractal analysis. I used the fact that the first factor is highly correlated with the shale volume (Szabó and Dobróka 2018). The components in the input

dataset account for most variation, reducing the original issue's dimensionality. Well logs provide independent information, and partial correlations with measured logs can reveal input parameters significantly affecting generated components. The most significant advantage of this technique is that other petrophysical characteristics that are directly unmeasurable in the borehole may be connected to and deduced from them based on these linkages. I decompose the original data matrix $\mathbf{D}$ of size N-by-K into the sum of two matrices

$$\mathbf{D} = \mathbf{F}\mathbf{L}^{\mathrm{T}} + \mathbf{E}, \tag{55}$$

where $\mathbf{F}$ is the N-by-M matrix of factor scores, $\mathbf{L}$ is the K-by-M matrix of factor loadings, and $\mathbf{E}$ is the N-by-K matrix of residuals (N is the number of observed depth points along the processed interval, K is the number of measured well-log types, M is the number of extracted factors). The factor loadings (or weights) practically indicate the partial correlation coefficients between the measured logs and the factor scores; the latter generates a well log showing the depth variation of a statistical factor. Given that the factor score matrix is orthogonal ($\mathbf{F}\mathbf{F}^{\mathrm{T}}=\mathbf{I}$, where $\mathbf{I}$ is the identity matrix), the correlation matrix of the standardized input variables is as follows

$$\mathbf{R} = \mathrm{N}^{-1}\mathbf{D}\mathbf{D}^{\mathrm{T}} = \mathbf{L}\mathbf{L}^{\mathrm{T}} + \mathbf{\Psi}, \tag{56}$$

where $\mathbf{\Psi}$ denotes a diagonal matrix of error variances. The primary diagonal of the reduced correlation matrix $\mathbf{L}\mathbf{L}^{\mathrm{T}}$ contains the commonalities, which are the variance parts of common components. There are numerous techniques for determining factor loadings and scores, such as using the spectral decomposition of the reduced correlation matrix, Maximum Likelihood, soft computing approaches, etc. For example, the maximum likelihood technique derives both the factor scores and loadings in one statistical operation by using the following objective function, which expresses a weighted divergence between the measured data matrix and the matrix computed by the common factors. Bartlett (1937) developed a linear solution based on factor loadings that allows factor scores to be calculated as follows

$$\mathbf{F}^{\mathrm{T}} = (\mathbf{L}^{\mathrm{T}}\mathbf{\Psi}^{-1}\mathbf{L})^{-1}\,\mathbf{L}^{\mathrm{T}}\mathbf{\Psi}^{-1}\mathbf{D}^{\mathrm{T}}, \tag{57}$$

the first column of matrix $\mathbf{F}$ contains the scores of the first factor, which create a well log that has been proven to be largely sensitive to lithology features of hydrocarbon formations, i.e., shale content, in various oilfields (Szabó et al. 2014). In this work, I assume that the relationship between the first component and shale volume may be applied similarly in clastic groundwater

formations as well. Jöreskog (2007) proposed a rapid non-iterative approach for estimating factor loadings

$$\mathbf{L} = (\mathbf{diag}\Sigma^{-1})^{-\frac{1}{2}} \Omega(\Gamma - \Theta\mathbf{I})^{-\frac{1}{2}} \mathbf{U},$$ (58)

where Γ is the diagonal matrix of the sample covariance matrix's Σ first R eigenvalues. The eigenvalues of the matrix may be in its columns, and $\mathbf{U}$ indicates an arbitrarily selected M-by-M orthogonal matrix. Jöreskog's constant (Θ) is used to calculate the number of factors. The initial factor loadings are investigated to determine the degree of correlation between the relevant factor and all types of well-logs used in this study. In this method, the most significant wireline log types in calculating shale volume may be identified. I rotate the extracted factors (i.e., orthogonally change the factor loadings) to determine the best correlation between the factors and petrophysical attributes (i.e., shale volume) using Kaiser's (1958) varimax technique. If the rotated first component is a suitable shale indicator, I may use it to estimate the shale volume (for example, out of inversion) or as input for the R/S technique to estimate the Hurst exponent.

4.2.2. Hurst analysis of well logs

As a specialty, I use the first factor log to focus the lithological information derived jointly from all well logs, rather than directly on the measured logs, as indicated by Hernandez-Martinez et al. (2013). This study results in enhanced lithology identification, which will be useful in the subsequent automated inversion technique in our workflow to explain the Hurst exponent calculation process. Let us assume that vector $\mathbf{d}$ represents one column of data matrix $\mathbf{D}$. The data vector's i-th component is $d_i = \mathbf{d}(z_i)$, which represents the i-th datum of a particular well log. Index i traverses 1, 2, …, N, where N is the total number of measured data points down a borehole. Let us partition the original dataset into S equal length L subsets, where $L x Z | 0 x 1, | N*x$ is an integer. I compute partial summation datasets

$$x_j = \sum_{k=1}^{j}(x_k - \bar{x}_s),$$ (59)

where $\bar{x}_s$ is the mean of each L-length subgroup. The dataset's range relative to the mean within each subseries x_j is determined as $R_s = max(x_j) - min(x_j)$, which is then rescaled by the associated standard deviation σ_s, yielding $R_s/\sigma_s = (R/S)_s$. The rescaled range must be estimated over S randomly chosen subsets of length L with varying sizes, and the average of these ranges must be determined. If the stochastic process is associated with the observed

variable scales with a specific $I \in (I_{min}, I_{max})$ (some portion of N), the average of the above ratios can be related to I through a power law

$$\overline{(R/S)}_s \cong I^H, \tag{60}$$

where H is the Hurst exponent and is the fractal measure of the correlations in the sample. Mandelbrot and Wallis (1969) state that parameter H must always be bigger than or equal to zero and that a value of $H=0.5$ indicates a white-noise process. High H values suggest that the dataset has more autocorrelation. When $H>1.5$, the stochastic character of long-range autocorrelations is called into doubt, and a deterministic backdrop is assumed. The scale and depth-dependent Hurst exponent, derived using the instantaneous slope of the log-log relationship between the rescaled range $(R/S)_s$ vs. the scale of the subintervals, is used in this work to provide a more rigorous description of the scaling behavior. Furthermore, the trends of the local Hurst exponents reflect comparable behaviors in the innovative inquiry, which was done on the lithological representative factor analysis and statistically analyzed to address the distinctions between fluctuation-type and intensity. The Hurst exponent as a function of scale and depth $H(I,z)$ is examined to find sub-intervals in rock units. It allows you to investigate the fractal properties of the first-factor log. I picked the scale interval between $I_{min}=1$ and $I_{max}=100$ with a 5-increment based on test runs. The depth-dependent Hurst exponent highlights the various patterns of the first component. In addition to the Hurst exponent, I suggest additional statistical computations for more precise sub-zone boundary detection. During the analysis, the scale of distinct patterns might change, affecting the extraction of fractal properties. To account for it, I add a Z-scoring variable $Z(I,z)$ as a function of scale and depth, which assesses the departure of the Hurst exponent from its mean at a given depth. According to the principle of superposition, the values of $Z(I,z)$ are very near to each other at the same facies interval but rather far apart at separate facies intervals. To emphasize the border contrast between distinct lithological units, I propose calculating the cumulative total of the scores $Z(I,z)$ and the mean of the Hurst exponent $H(I,z)$, respectively. Finally, geometrical information will be introduced to the interval inversion as a priori information.

4.2.3. Theory of Csókás method

The predicted parameters from the interval inversion (as described in section 2.2.1) can be used to calculate the hydraulic conductivity (K_h) continuously using the Csókás method. The hydraulic conductivity is an important parameter to describe the quality of the aquifer, where the K_h represents how easily the water flows through the pore spaces of the formation. The

Kozeny-Carman model is the most extensively used approach for estimating hydraulic conductivity (Amaefule et al. 1993)

$$K_h = \frac{\rho_w g}{\mu} \frac{d^2}{180} \frac{\emptyset^3}{(1-\emptyset)^2} \, ,$$

(61)

where $\emptyset$ (v/v) is the formation porosity, ρ_w (g/cm^3) and μ (Ns/m^2) are the pore water density and viscosity, d (cm) is the dominant grain diameter, and g (cm/s^2) is the normal gravity acceleration. Sieve analysis may be used to assess the prevalent grain size

$$d = \frac{d_{10}+d_{60}}{2} \left(\frac{d_{10}}{d_{60}}\right)^{\frac{1}{2}} \, ,$$

(62)

where d_{10} (cm) and d_{60} (cm) are grain diameters measured at 10% and 60% cumulative frequencies of the grain-size distribution curve, respectively. Csókás (1995) assumed a relationship between the dominant grain size and the resistivity formation factor (F), as proposed by (Alger 1971). The formation factor is calculated as $F=R_0/R_w$, where R_0 is the resistivity of the fully saturated formation and R_w is the resistivity of the pore water in the same formation (Archie 1952). Csókás improved equation (61) by incorporating Alger's (Alger 1971)empirical equation

$$K_h = C \frac{\emptyset^3}{(1-\emptyset)^4} \frac{\left(\lg\frac{R_0}{R_w}\right)^2}{\left(\frac{R_0}{R_w}\emptyset\right)^{1.2}} \, ,$$

(63)

where C denotes a site-specific constant. In weakly sorted sediments with a formation factor of less than 10, equation (63) is valid. Using just well logs, the Csókás technique provides a continuous (in situ) assessment of hydraulic conductivity, transmissivity, critical flow velocity, specific surface of grains, and water production. It is limited to evaluating unconsolidated groundwater formations (with low formation factor values); however, it is based on relatively inexpensive well-logging observations to decrease the expenses of rock collection, pumping tests, and laboratory analyses. Figure 28 shows the workflow of the proposed algorithm.
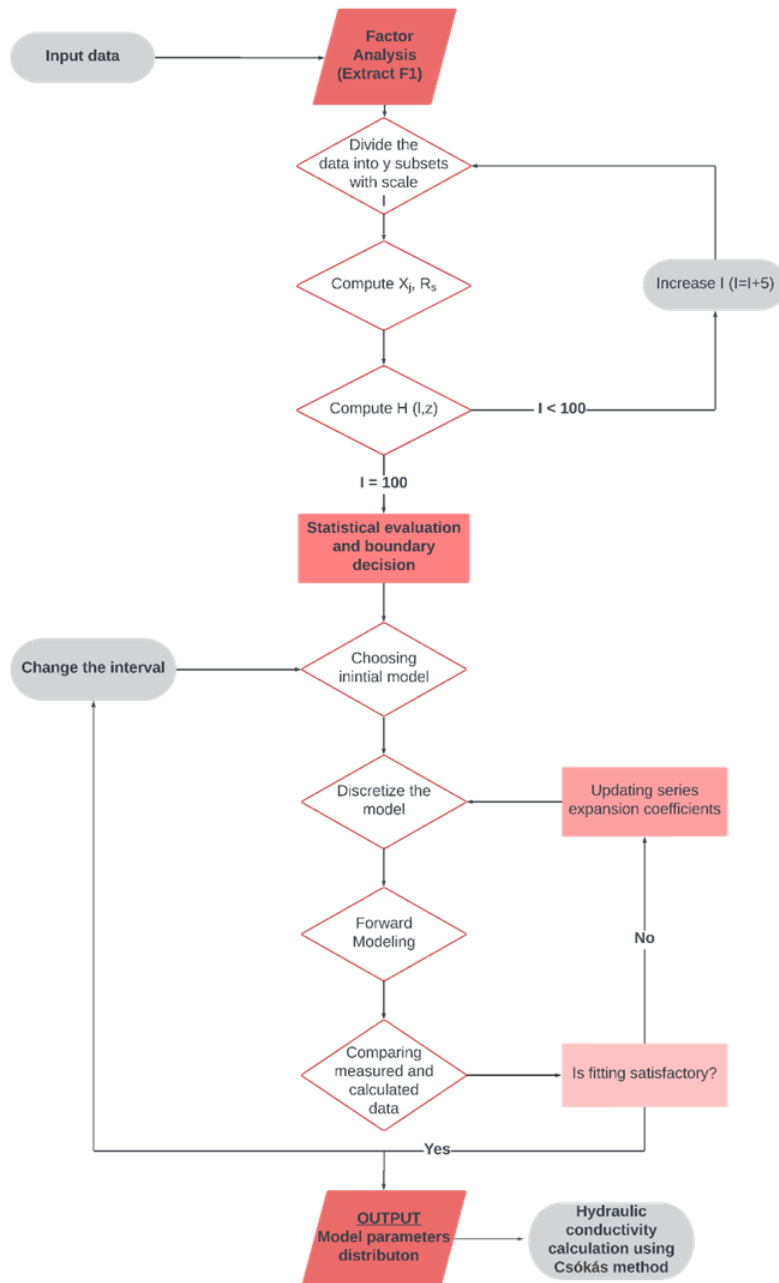
Figure 28. The statistical workflow for hydraulic conductivity estimation using the integration between fractal characteristics analysis and interval inversion.

4.3. Synthetic modeling test

The feasibility of the proposed statistical algorithm has been tested using synthetic data contaminated by 5 % Gaussian distributed noise. Equations from 45 to 50 were used to calculate the synthetic data according to suggested model parameters ($\emptyset$, V_{sh}, V_{sd}) that consist of four homogeneous layers fully saturated with water ($S_{x0}=S_w=1$). The water-bearing formations consist of alternating sequences of shale sand layers, exhibiting variations in their respective thicknesses. The synthetic well-logging data contain readings from GR, ρ_b, $\emptyset_N$, Δt,

R_s, R_d logs over a 20 m (relative) depth interval sampled at 0.1 m distance (Figure 29). Given that I have four layers with three petrophysical parameters per layer, in addition to that, the geometrical characteristics represent three boundaries separated between the layers. Therefore, the total number of unknowns, denoted as Q, is 15. Concurrently, the well logging dataset contains a total of N=1200 points. By utilizing the material balance equation, I can drive the sand volume, to reduce the number of unknowns to Q^*=11. Furthermore, the integration between the Hurst exponent method with the interval inversion method for identifying the layers' boundaries can reduce the number of unknowns to Q^{**}=8.
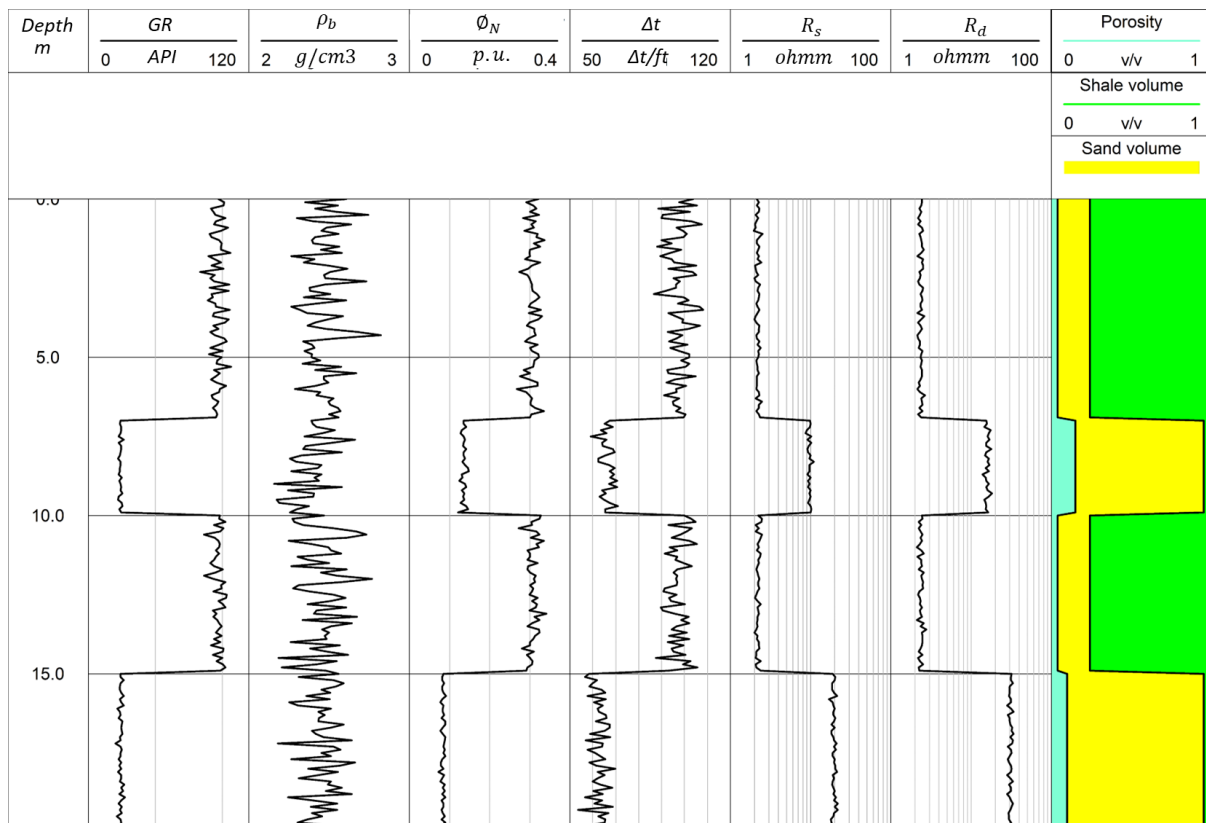
Figure 29. Synthetic wireline logs contaminated with 5 % Gaussian distributed noise (tracks 1-6) and the exactly known layer parameters (tracks 7-8) as input for testing the proposed well-log-analysis workflow.

The proposed workflow involves factor analysis for dimensionality reduction and identifying lithological characteristics. The first factor, which accounts for over 90% of data variance, has a high correlation with lithological logs, indicating lithological distinctions. The wireline logs are sensitive to shale volume variations, making it an effective indicator.

Table 4. Rotated factor loadings derived by factor analysis of synthetic wireline logs.

Well-log	First factor	Second factor	Third factor

Natural gamma-ray intensity	0.99	0.06	−0.06
Bulk density	0.04	0.99	6×10^{-3}
Neutron porosity	0.97	0.02	−0.23
Acoustic travel-time	0.96	0.05	−0.11
Shallow resistivity	−0.8	0.04	0.46
Deep resistivity	−0.9	0.03	0.43

In the second step of the workflow, the Hurst exponent analysis is applied to the first factor scores. As depicted in Figure 30, the scale-depth dependent Hurst exponent $H(I,z)$ accurately identifies three layers' boundaries at depths of 7, 10, and 15 m. Notably from left to right, the calculated mean of the Hurst exponent, along with the cumulative sum of mean and cumulative sum of scoring $Z(I,z)$ show a break pattern in front of the exact boundaries' location. The Hurst exponent's mean fluctuates around 0.5 at layer borders and lower than zero within each layer, indicating background noise. The scale range between 10 and 70 ensures a depth window of 0.74 m for the depth-dependent Hurst exponent. Circular anomalies have an amplitude value of 0.5, causing a distinctive line deflection at various boundaries' depths.
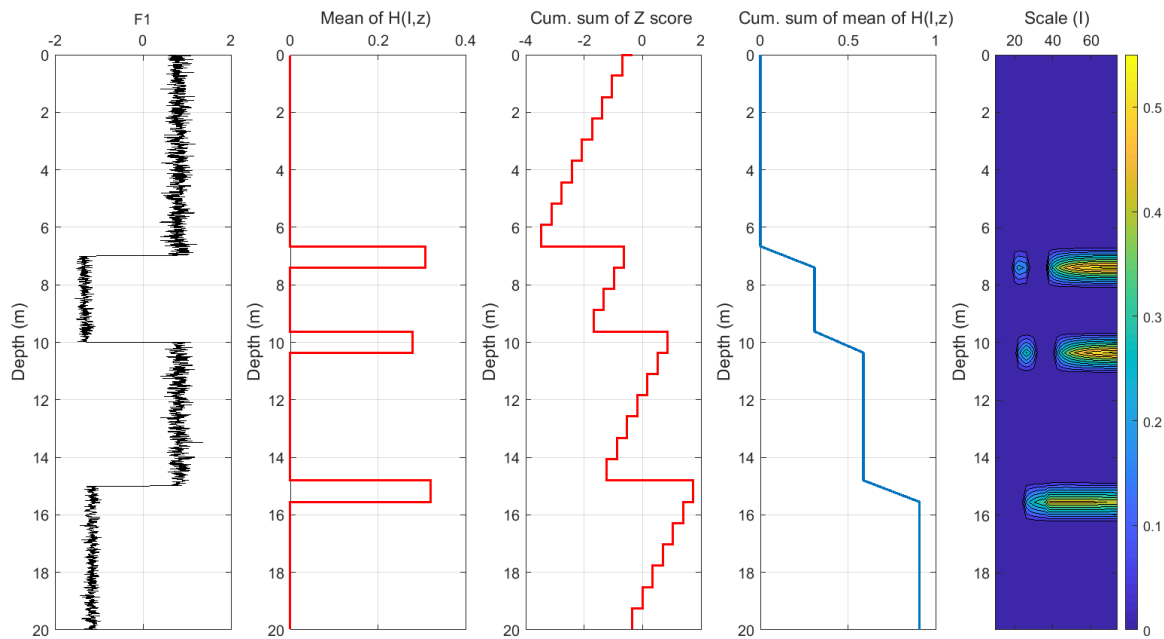
Figure 30. Result of Hurst analysis of the first factor log (F_1) extracted by factor analysis of synthetic well logs contaminated with 5 % Gaussian distributed noise. Tracks 2-5 show the changes of the Hurst exponent and its related quantities at each layer boundary.

In the third step of the workflow, the noisy synthetic well logs undergo inversion to estimate the porosity, shale volume, and sand content of the groundwater formations. To characterize the four-layered homogeneous model, I employ a combination of unit step (Heaviside) functions for series expansion (refer here to the equation). The basis function for

the q-th layer, denoted as $\Psi_q (z)$, is defined as $u\left(z - Z_{q-1}\right) - u\left(z - Z_q\right)$. The depth coordinates of the layer boundaries are automatically determined from the results of the Hurst analysis: Z_0=0 m (fixed), Z_1=7 m, Z_2=10 m, Z_3=15 m, Z_4=20 m (fixed). The convergence of the inversion procedure is stable and quick, as demonstrated by the consistent decrease in data misfit calculated by equation (60), which decreases from 76% to 5% (as shown in Figure 31)

$$\delta = \left[N^{-1} \sum_{k=1}^{N} \left(\frac{d_k^{(m)} - d_k^{(c)2}}{d_k^{(m)}} \right) \right]^{\frac{1}{2}}. \tag{64}$$

The inversion results are illustrated in Figure 32. A commendable alignment is observed between the quasi-measured well logs (illustrated by the black curves) and those computed using the estimated model parameters, which are constant within each layer (represented by the dashed red lines).
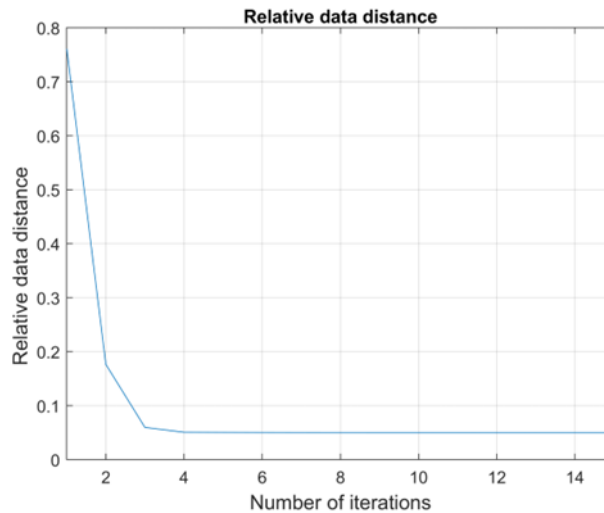
Figure 31. Development of convergence during the interval inversion of synthetic data contaminated by 5 % Gaussian distributed noise.

In the last track of figure 32, both the predicted (dashed red lines) and exactly known petrophysical parameters (depicted by the black lines) are displayed. The disparity between them is small and not discernible on the logs; on average, it amounts to less than 0.001 (v/v). This remarkable level of agreement between the target and estimated model is attributed to the employed discretization scheme and the large data-to-unknowns ratio. The demonstrated feasibility of the interval inversion method in groundwater formations stands as a testament to its potential for practical field applications.
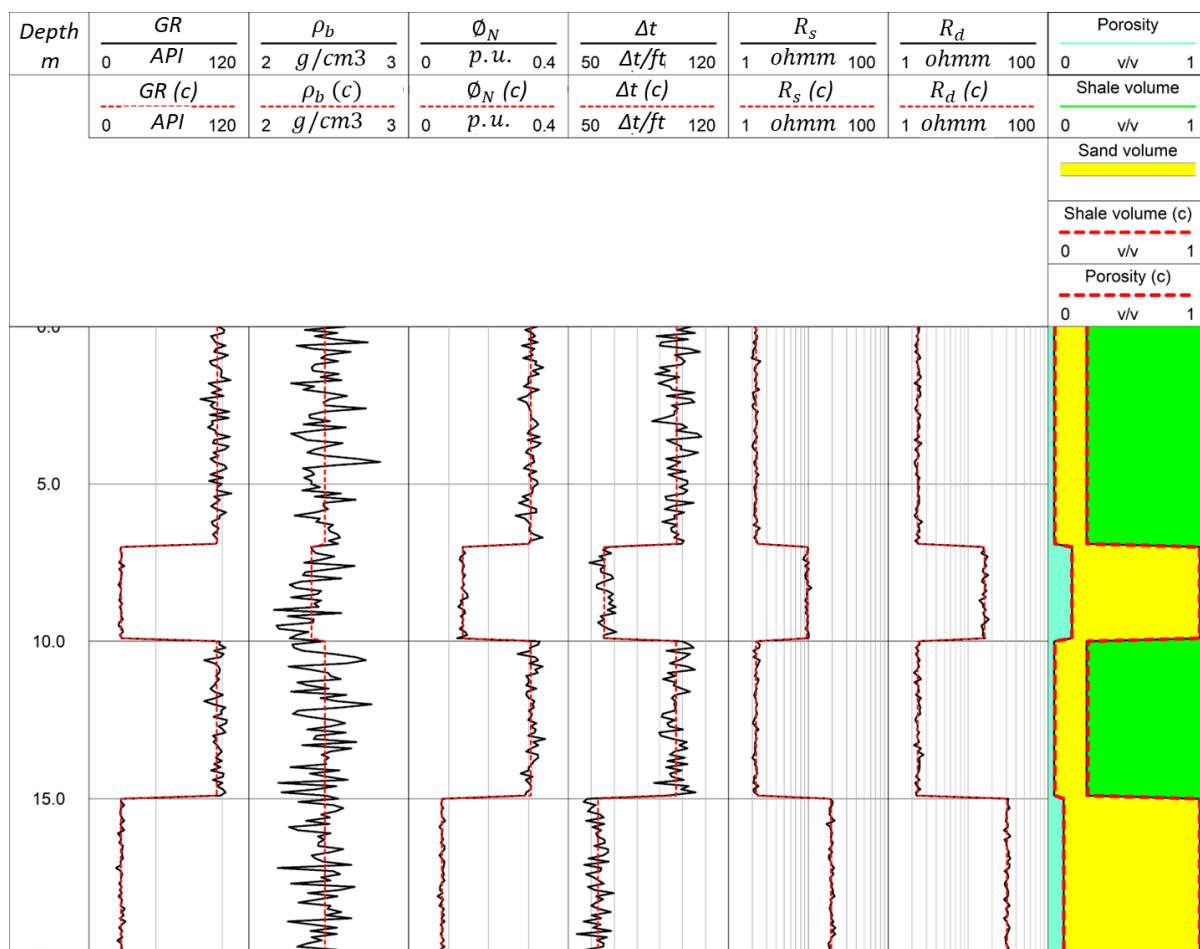
Figure 32. Result of interval inversion of synthetic well-logs contaminated with 5 % Gaussian distributed noise. Tracks 1-6 show the fit between the observed (black line) and calculated (red dashed line) data and track 7 includes the known and estimated model parameter distributions assuming a petrophysical model of homogeneous layers.

4.4. Hydrogeophysical Application to Field Data

The proposed method is applied to Cenozoic aquifers found in the Baktalórántháza-1 well in Szabolcs-Szatmár-Bereg County in North-East Hungary (Figure 33). The drill-hole penetrates fully saturated Pleistocene and Miocene age clastic formations. Surface geophysical measurements and well-logging operations did not reveal any indicators of hydrocarbons throughout the exploration phase. The borehole, on the other hand, was deemed suitable for creating thermal water. The upper 80-100 m of the shallow region are predominantly composed of Pleistocene sediments, mostly sands with varied grain sizes, according to vertical electrical sounding measurements. Borehole geophysics provided more detailed knowledge of the shaly sand sequence's underlying rocks. Sand predominates between depths of 100-160 m, according to wireline records, followed by a shaly sequence. A 5-15-meter-thick coarse-grained bed was also identified.

Figure 33. The map of Hungary including the location of Baktalórántháza-1 well indicated with the red circle.

The observed dataset encompasses typical well logs along with laboratory data obtained from 97 rock samples, which were taken at relatively frequent intervals along the investigated interval. The interval under consideration spans from 100 to 450 m. Within this range, measurements for SP (Spontaneous Potential), GR (Natural Gamma Ray), neutron-thermal neutron intensity (NN), gamma-gamma intensity (GG), and R_s (Shallow Resistivity) data were collected at 0.1 m intervals along the well. Core data is also available, specifically hydraulic conductivity and shale volume, which were derived from grain size analysis conducted on the rock specimens. In the initial step of the proposed workflow, I processed the aforementioned wireline logs using factor analysis. This analysis yielded two factors extracted from the borehole dataset and their respective factor loadings are detailed in Table 5. The factor loading matrix indicates that logs are primarily sensitive to lithology, such as SP, GR, and the saturation sensitive R_s log shows the strongest influence on the first factor. On the other hand, the second factor is correlated with the gamma-gamma tool, providing insights into porosity of groundwater formations.

Table 5. Rotated factor loadings derived by factor analysis of well logs recorded in Baktalórántháza 1 well.

Well-log	First factor	Second factor
Spontaneous potential	0.63	0.03
Natural gamma-ray intensity	0.67	−0.02
Gamma-gamma intensity	−0.2	0.98
Neutron-neutron intensity	−0.09	−0.14
Shallow resistivity	−0.99	−0.11

The Hurst exponent is used to quantify the degree of autocorrelation of the calculated first factor (F_1). The Hurst exponent ranges from 0 to 1.5, with changes in this value corresponding to variations in the F_1 log. A scale-depth dependent Hurst exponent varies between 0 and 0.6, indicating a random walk or Brownian series. Transition zones between two intervals in the data often display characteristics of Brownian motion, suggesting a mixed zone with properties gradually changing. Intervals with Hurst exponent values higher than 0.5 suggest persistence, with an increase in the well-log value more likely to be followed by another increase in the short term and a decrease followed by another decrease. This is crucial for detecting transgression and regression patterns within major sequences. The separation between different zones depends on the variation of the mean values of the function H(I,z). The results indicate that boundaries between different zones exhibit a consistent value of approximately H=0.5, while the pattern may vary within each layer. Five zones (Zone 1 to Zone 5) are identified, with boundaries situated at depths of 145, 210, 255, and 330 m, respectively. The mean of the Hurst exponent presents a mirrored pattern for the F_1, which could potentially be utilized as a facies identification technique in future analyses (Figure 34).
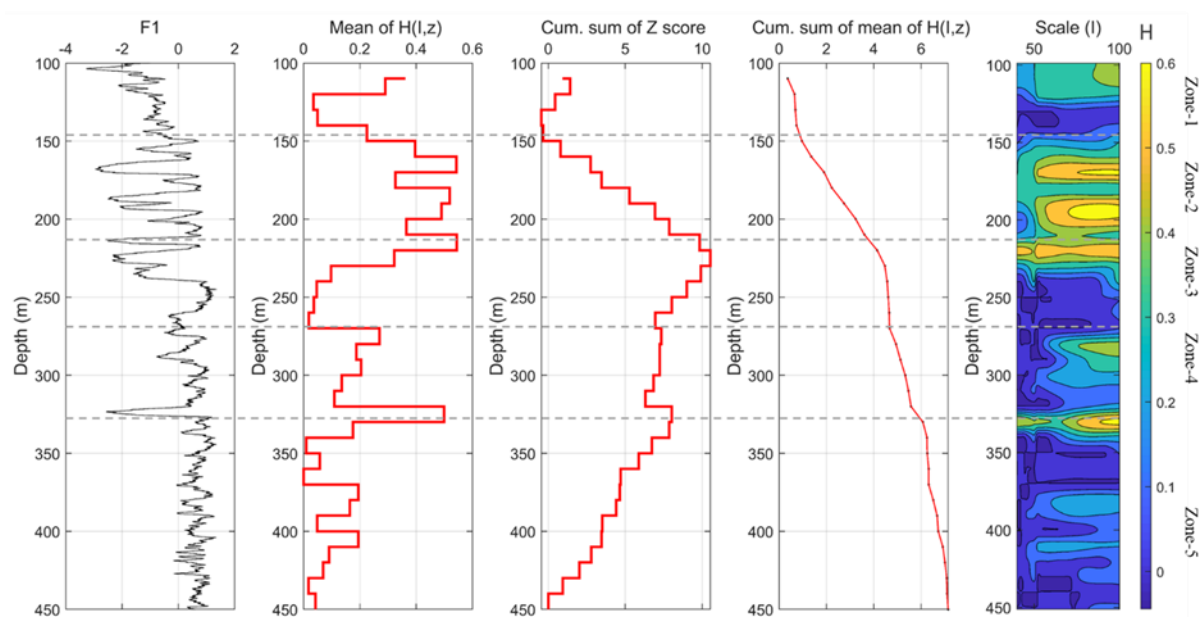
Figure 34. Hurst analysis of the first-factor log (F_1) extracted by factor analysis of well-logs measured in Baktalórántháza 1 well. Tracks 2-5 show the changes of the Hurst exponent and its related quantities at each identified lithological boundary.

In the third step of the statistical procedure, I invert the borehole logs using the zone boundaries indicated by the preceding Hurst analysis. I selected the model parameters for the five zones (as shown in Table 6) assuming a fully saturated formation. The first model is built

using the material balance equation, which calculates the sand volume depending on the previous two parameters (derived from inversion). In Figure 35, the calculated logs created using these model parameters are shown alongside the observed well logs. The calculated data for each pre-defined zone is shown by the red curve. I assume that petrophysical parameters stay constant inside each zone during the inversion process.

Table 6. Starting values of petrophysical parameters and those estimated by interval inversion assuming five homogeneous zones in Baktalórántháza 1 well.

Ordinal number of zones	Upper depth of zone	Initial value of porosity (v/v)	Initial value of shale volume (v/v)	Estimated value of porosity (v/v)	Estimated value of shale volume (v/v)
1	100	0.07	0.2	0.16	0.13
2	145	0.08	0.3	0.14	0.19
3	210	0.07	0.2	0.19	0.20
4	255	0.2	0.2	0.11	0.26
5	330	0.2	0.2	0.12	0.34
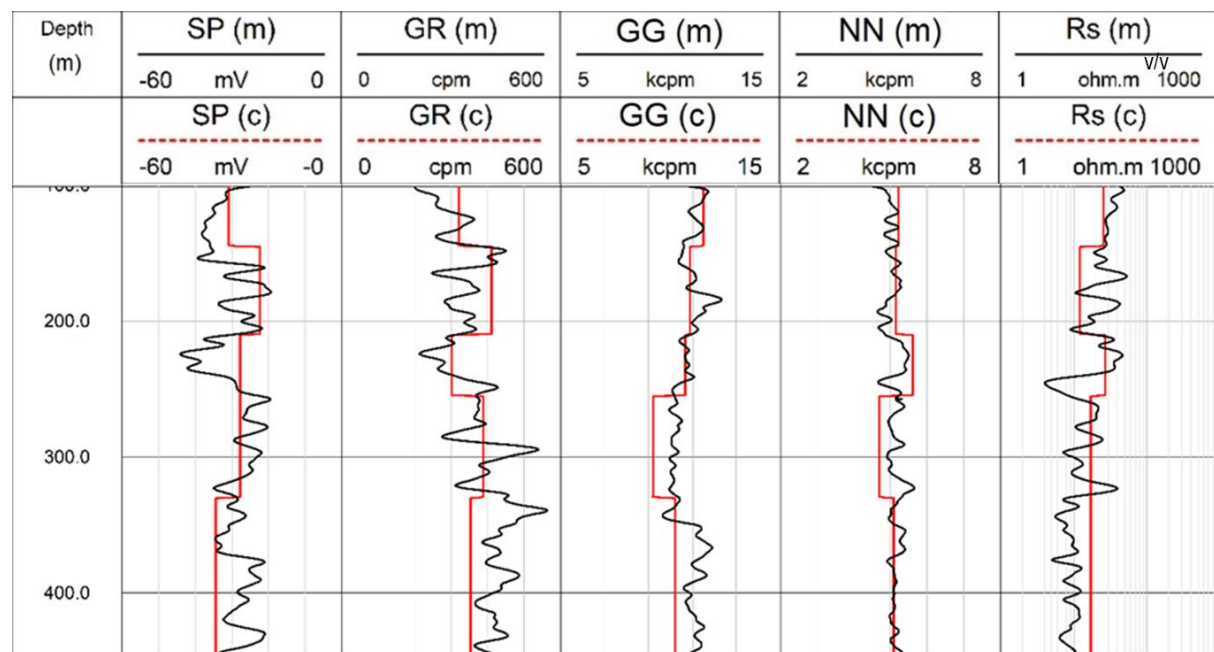
Figure 35. Observed wireline logs (black curve) and the calculated logs (red curve) using the initial model obtained from Hurst analysis assuming five homogeneous zones along the processing interval.

I use an orthogonal set of polynomials as basis functions to characterize the depth variation of petrophysical parameters within the five zones in greater detail. It is possible to estimate the petrophysical properties of inhomogeneous strata in a particular zone using this method. The basis function of the q-th zone is $\Psi_q(z) = P_{q-1}(z)$, where P_{q-1} is the q-th degree Legendre polynomial. I employ Legendre polynomials of 40-th degree for each model

parameter and have Q=402 unknowns versus N=3,470 data. The interval inversion has a data-to-unknowns ratio of 8.6, which is more than the local inversion, which has a ratio of just 5/2. Figure 36 verifies the inversion procedure's stability. The data misfit decreases from 24 % to 7 % in a convergent inversion procedure.
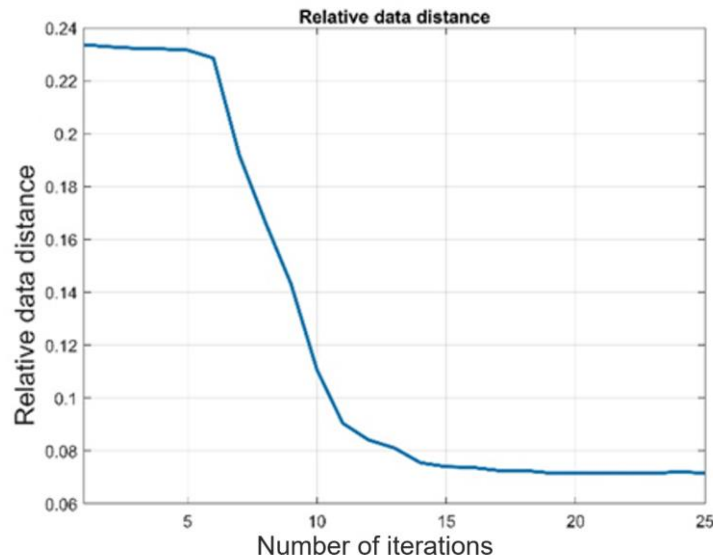
Figure 36. Development of convergence during the interval inversion of borehole logging data measured in Baktalórántháza-1 well.

The inversion results are shown in Figure 37. They consistently indicate a decrease in porosity percentage towards the sections of the logs. At the time there is an increase in shale content as we go deeper. The differences between the calculated and measured wireline logs can be attributed to factors. First, there's the influence of data noise which is inherent to these measurements. Additionally, when using series expansion-based interval inversion methods with approximations it may be challenging to detect layers due to resolution limitations. Since tool response functions approximate equations, there could be modeling errors that I cannot precisely determine. Furthermore, during the modeling step zone parameters are set arbitrarily which can introduce uncertainties in both data misfit and estimation errors for inverted petrophysical parameters. To evaluate how accurate our estimations of porosity and shale volume are, I use Equation (60). I adopt the following values of standard deviation (data accuracy) for the observed well-logs: σ_d(SP)=0.0484, σ_d(GR)=0.0484, σ_d(GG)=0.005, σ_d(NN)=0.002, and σ_d(R$_s$)=0.05. The confidence intervals for estimated model parameters are presented in Figure 38. It becomes clear that shale volume can be estimated with accuracy within 0.01 (v/v), while our predictions for porosity also have a range between 0.01 to 0.03 (v/v). Further evidence supporting the accuracy of estimating shale volume can be found

through laboratory measurements, Figure 38. In the track of the diagram, a remarkable correlation can be seen between the log indicating shale volume and the core data, for each zone.
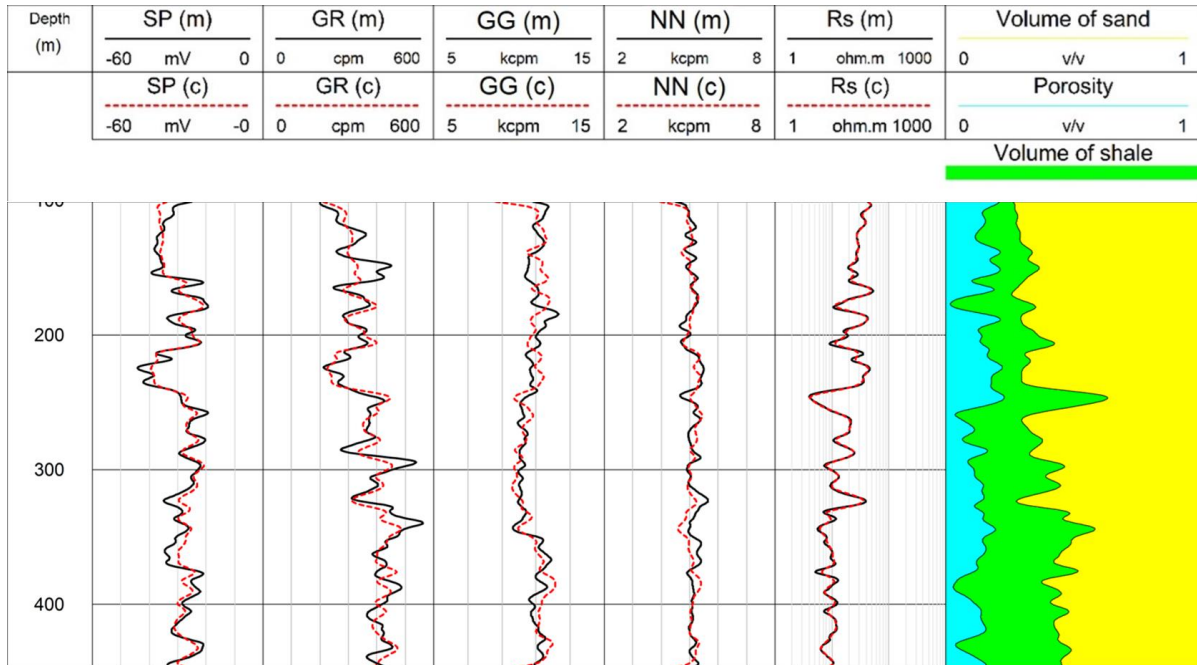
Figure 37. Result of interval inversion of borehole logging data measured in Baktalórántháza 1 well. Tracks 1-5 show the fit between the observed (black line) and calculated (red dashed line) data, track 6 includes the estimated model parameter distributions assuming a petrophysical model of 5 inhomogeneous zones.
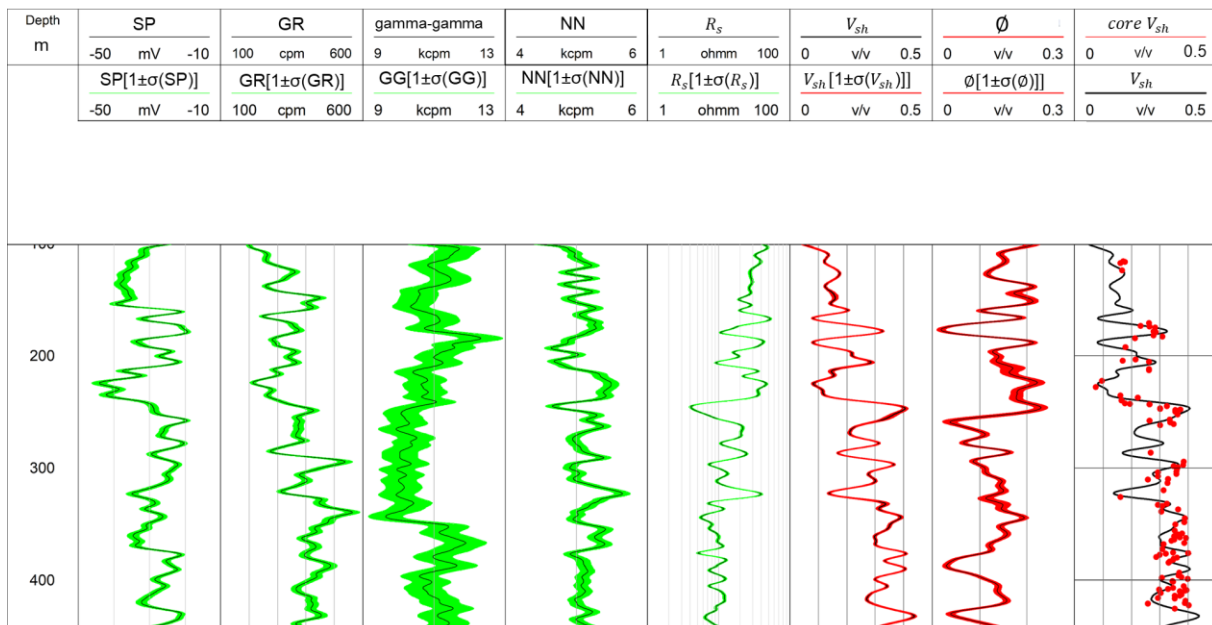
Figure 38. The porosity and shale content logs estimated by interval inversion of well-logs observed in Baktalórántháza 1 well. Black lines demonstrate the estimated petrophysical parameter, while the red lines show their error bounds derived from the model covariance matrix. Track 3 shows the validation of the predicted shale volume log with core-measured shale volume data.

The Csókás approach is used in the final phase of the process to estimate the vertical distribution of hydraulic conductivity along the well. As inputs to equation (63), I use the porosity log produced from inversion and the measured resistivity log. Furthermore, I derive the typical grain-size values from the grain-size distribution curves presented at the depth levels where rock samples were obtained. The shale volume and hydraulic conductivity logs coincide admirably, as shown in Figure 39, with their values fitting closely with those obtained from laboratory experiments.
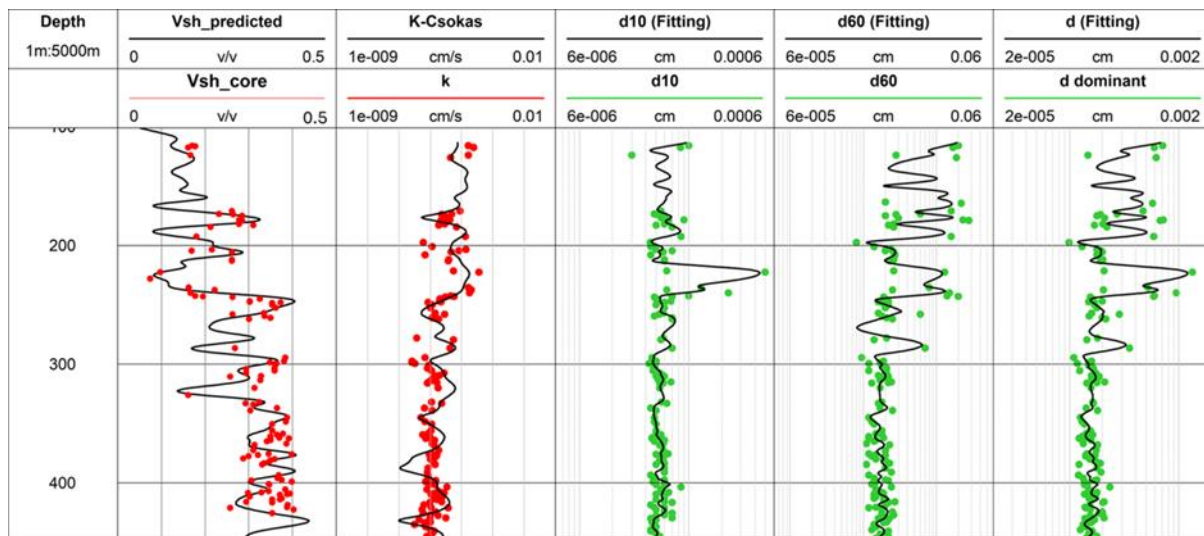
Figure 39. Shale volume (track 1) and hydraulic conductivity (track 2) logs estimated by interval inversion of well-logs (solid lines), characteristic grain sizes (tracks 3-5), and the core measured values of these parameters (red circles) in the Baktalórántháza 1 well.

The workflow's accuracy and reliability depend on input data uncertainty, which can be limited by instrumentation and ambient sounds. Inversion unknowns are often linked, increasing the risk of unstable processes and unclear results. Addressing this can be achieved by including credible priori knowledge on the groundwater formation. Determining polynomial degree can ensure correct vertical resolution but reduce overdetermination ratio and estimate accuracy. A trade-off between data-to-unknowns ratio and estimation accuracy is necessary for trustworthy information on petrophysical features of groundwater formations. Setting zone parameters is crucial for accurate match between observations and forecasts.

4.5. Summary of results

This chapter presents a fully automated formation evaluation procedure based on the joint application of all suitable borehole logs recorded in groundwater wells. The workflow includes innovative techniques, such as Hurst analysis, which provides an estimation of the number of layers and an approximate prediction of their depth location. The Hurst method is

based on statistical quantities extracted from raw data by factor analysis, revealing hidden information about lithology. The study emphasizes the analysis of the scaling interval, which is crucial for achieving better vertical resolution. Interval inversion, adapted from oil exploration, is applied for the first time in groundwater formations. Quality check is implemented through the application of the depth-dependent covariance matrix, calculating the uncertainty of inversion results. The proposed inversion method reliably derives petrophysical characteristics and allows fitting any basic function to model parameters. The inversion results agree with independent lab analysis. The porosity log as the output of inversion serves as input for the Csókás method, improving hydraulic conductivity estimation by determining the local relation between dominant grain size and formation factor. The site-specific constant in the Csókás formula should be estimated using regression relations. The proposed workflow may assist in improved assessment of other types of groundwater systems, providing reliable and quality-checked information for hydrogeophysicists and hydrogeologists.

Thesis 3.

I have developed a new integrated method based on the Hurst exponent and interval inversion for processing well logs. The conducted workflow consists of a series of innovative interpretation techniques that are used simultaneously to improve the evaluation of groundwater formations. The algorithm first extracts the first factor being a good lithology indicator. Then, Hurst analysis is applied to the first factor log to differentiate the lithology types. I proved that the scale and depth-dependent Hurst exponent exhibits a mean value of 0.5 at the exact location of the layers' boundaries with a decreasing pattern within the layers themselves. After the automated determination of the layer boundaries, interval inversion of the well logs provides the basic volumetric parameters such as porosity, shale volume and matrix volume. I proved the feasibility of the inversion algorithm using synthetic and field hydrogeophysical datasets. I extended the proposed algorithm to calculate the hydraulic conductivity using the Csókás method. I successfully validated the predicted petrophysical parameters using core laboratory measurements.

Chapter 5: Fully Automated Algorithm for Petrophysical and Zone Parameter, and Layer-Thickness Estimation with Robust Clustering and Local Inversion

5.1. Scientific background

Well log data offers a plentiful source of information beneficial for the interpretation of geometrical properties, such as layer thickness and dip, and petrophysical properties such as porosity, water saturation, and rock matrix composition. Traditionally, the reconstruction of layer thickness is troublesome, particularly when applying local inversion methods solely. Previous discussions had highlighted some of the techniques that were formulated to supplement conventional techniques of inversion, including novel algorithms that blend optimization with analysis techniques such as fractal analysis. While such techniques succeed in predicting necessary parameters, they do so in a step-by-step process and not concurrently.

This chapter introduces a novel algorithm for the simultaneous prediction of geometrical and petrophysical parameters through a combined optimization procedure. Through the incorporation of robust clustering approaches, in particular the most frequent value (MFV) clustering covered in Chapter 1, and the point-by-point and interval inversion techniques covered in Chapter 2, the method enhances the effectiveness of parameter estimation. This integration is consistent with the contemporary progress observed in the interpretation of geophysical data, where machine learning is applied (Ikhwan, Sastranegara, & Anggraini, 2024; Nunes et al., 2020; Terry et al., 2019; Shahriari et al., 2020). Research by Li and Sun (2016) and Spichak (2020) used clustering to enhance magnetic data inversion and particle swarm method optimization, respectively, illustrating the usefulness and strength of these intersections of technology.

Likewise, geophysical data, such as well logs, display fundamental parameters of zones like Archie's constants and physical properties of shale, fluid, and matrix. These constants, being constants within layers, but not necessarily between layers, underline the need for accurate, layer-specific estimates (Szabó, 2011). The understanding of these parameters relates directly to constructing high-resolution static reservoir models, key to successful dynamic modeling, reserves estimation, and optimization of well production. The extent to which poorly understood variables can be incorporated into the inversion process depends on

overdetermination ratios of well logs, clarified through theoretical models established by Dobróka et al. (2009b). In the local inversion case, this ratio has a maximum value of 1.5 (Abdelrahman & Hassan, 2022).

Following earlier chapters, the integration of MFV clustering with point-to-point inversion enables layer boundary definition—this advance now incorporates zone-specific parameter calculation, along with petrophysical measurements. Yet, there is an awareness of local inversion resolution constraints. Thus, the present algorithm reduces unknown parameters from potentially laborious datasets to three fundamental components, utilizing conservation principles to represent the storage, influx, and removal of fluids in the reservoir. This strategy not only simplifies the model but also increases the analytical accuracy, allowing for precise estimations based on both simulated datasets and actual field data. Finally, this advanced technique incorporates the vast amount of data held in well logs, using it to refine the inversion process and achieve greater precision in interpretation. The efficacy of this integrated method is tested by comparing results from synthetic case studies, comparing constant parameter layers with variable ones, to prove its suitability for real-world application.

5.2. Methodology and Inversion Workflow

The relationship between the data and the unknowns has been stated in chapter 2 as a nonlinear relationship between the well logging data and the model parameters such as petrophysical (m), zone (O), and layer-boundary coordinates (Z). Mathematically, the relationship can be expressed as follows

$$\boldsymbol{d} = \boldsymbol{g}(\boldsymbol{m}, \boldsymbol{O}) \,, \tag{65}$$

This nonlinear relationship can be approximated using Taylor series truncated at first order, where the general form of the Taylor approximation that includes all the unknowns can be written as follows:

$$\emptyset_k = g_k(m_0) + \sum_{i=1}^{p} \left(\frac{\partial g_k}{\partial m_i}\right) \Delta m_i + \sum_{j=1}^{o} \left(\frac{\partial g_k}{\partial O_j}\right) \Delta O_j. \tag{66}$$

where the first term on the right is the theoretical data calculated using the initial model parameters (m_0), followed by the derivatives of data with respect to the petrophysical, zone, and geometrical parameters, respectively.

In this chapter, I will introduce two automated algorithms. In the first algorithm, the integration between the MFV-clustering and the local (point-by point) inversion can automatically determine both the petrophysical and geometrical parameters, and the zone parameters will be considered as known constants. In the second algorithm, the previous algorithm is improved to calculate zone parameters in addition to petrophysical values. For the synthetic data set, I initially estimate seven well log data points, which might include parameters such as resistivity, porosity, density, gamma ray response, acoustic speed, neutron porosity, and water saturation. Similarly, for the field data, I base my estimates on five well log parameters, potentially a subset of those used in the synthetic data case.

This detailed approach provides leverage the critical information provided by well log measurements while simplifying the inversion process to optimize parameter estimation and enhance interpretational accuracy. To evaluate the performance of this enhanced approach, results are compared across two distinct synthetic cases: Inversion A, where the zone parameters remain constant across all layers, and Inversion B, where the parameters vary between layers. Besides that, I assumed a constant value of these zone parameters in the whole processing interval, while in the other scenario there are different zone parameters of each layer. Therefore, the proposed algorithm can be used for analyzing the outcomes of MFV-clustering each iteration to determine the continuity and, hence, the borders of the layers. The absolute difference between two clusters labeled L_i and L_j for two consecutive data points d_i and d_j is defined as Δ. I group labels with delta values that are less than or equal to zero based on a predetermined threshold value. This grouping results in the formation of a set S, which consists of labels L_i that meet the criteria (delta=threshold) for some other labels L_j in the vertical well-logging dataset. A layer is defined as a collection of datapoints d_i that matches L_i labels in S. In the inner loop of the meta-algorithm for the gathered data points in each layer to further predict layer parameters. Then, the preceding approach will be performed to confine the inversion findings to the real layers' coordinates derived from the MFV-clustering convergence. The objective function of the cluster analysis is expressed by the data distance between the calculated and measured data. Figure 40 shows the workflow of the proposed algorithm. It is obvious that the algorithm's convergence is subjected to the finding of the new petrophysical parameters and the zone parameters beside the defined layers boundaries.
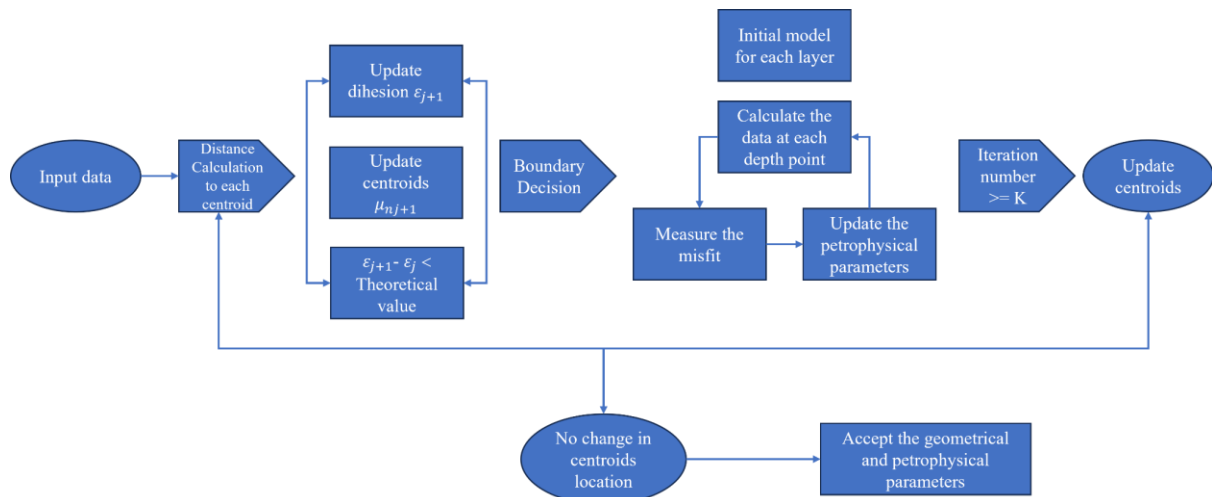
Figure 40. The schematic diagram shows the workflow of the cluster analysis assisted inversion algorithm.

The inversion meta-algorithm outlined here leverages a fusion of MFV-Clustering and Point-by-Point Inversion, aimed at automatically identifying key petrophysical and geometrical parameters. At the outset, it treats zone parameters as known, unchanging values, which streamlines the assessment process as the algorithm evaluates MFV-clustering results to pinpoint continuity and determine the edges of layers. The method involves computing the absolute difference, termed delta, between clusters (L_i and L_j) for each successive data point (d_i and d_j). Subsequently, the algorithm groups label by applying a delta rule, specifically joining those with deltas not exceeding a set threshold. This process culminates in the creation of set S, which comprises labels (L_i) that meet the threshold requirements in relation to other labels (L_j) in the vertical well-logging records. A layer is then characterized as a set of data points (d_i) matching L_i labels within set S.

5.3. Synthetic Modeling Test Using Local Inversion

5.3.1. Priori zone parameters

The feasibility of the proposed meta-algorithm has been proved using synthetic data contaminated with 1 % Gaussian distributed noise. The synthetic model consists of three inhomogeneous layers. The first and third layers are mainly composed of sandstone; both are partially saturated by water and hydrocarbon; the first layer is saturated by hydrocarbon more than the last one. The two sandstone layers are separated by a shaly sandstone layer. Figure 41 shows the fitting between the predicted and measured well logs. It is worth mentioning that the choosing for the initial model is constrained to an initial threshold data distance value to

guarantee that the initial model will start from anywhere in the model space with a misfit lower than 1. Figure 42 shows the initial values of petrophysical parameters.
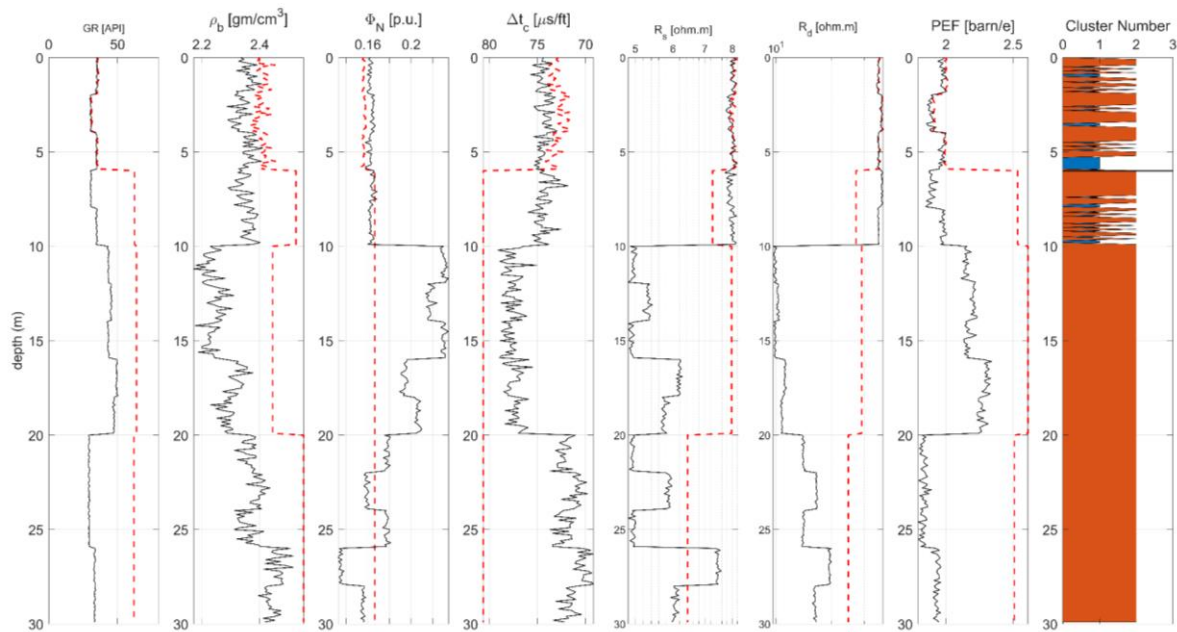
Figure 41. Fitting between the synthetic data and the calculated data (iteration 1); the red dashed lines represent the calculated data, while the solid black lines represent the synthetic data.
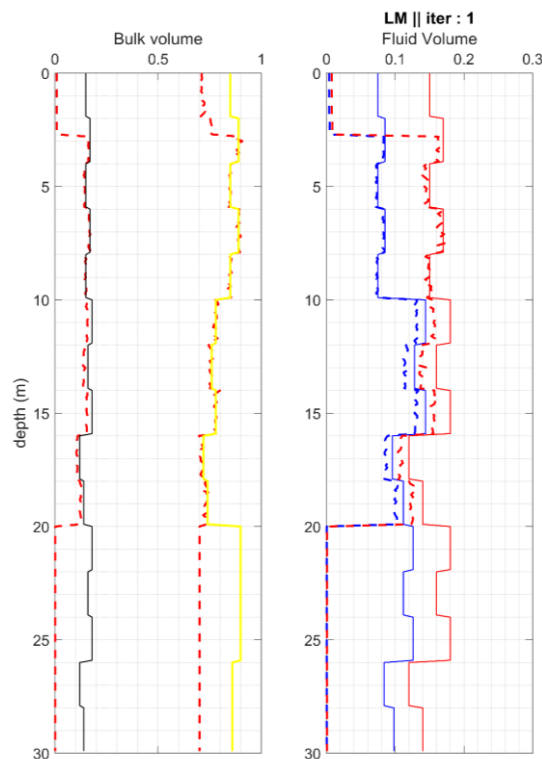
Figure 42. The predicted (dashed lines) and exactly known (solid lines) petrophysical parameters of the gas-bearing reservoir at iteration 1.

By splitting each layer into smaller, discrete subzones, the technique is supposed to account for variation within each layer. This method enables a more detailed examination of the petrophysical and geometrical features inside each layer. The use of subzones shows that the algorithm can capture fine-scale fluctuations and heterogeneity in well-logging data, which can be critical for correct geological formation characterization. The threshold is set at 10 points or 1 meter in this context, suggesting that the algorithm considers changes in data values less than this threshold to be within the same layer. If the absolute difference (delta) between clusters L_i and L_j for successive data points d_i and d_j is less than or equal to 10 points or 1 meter, they are clustered together and considered part of the same layer. This threshold aids in recognizing continuous segments in well-logging data, allowing for more precise layer boundary identification. This threshold was most likely chosen based on domain expertise and the unique characteristics of the well-logged data being evaluated. It is a significant factor that determines the performance and outcomes of the algorithm. Figure 43 shows the fitting between the calculated and actual synthetic data, while Figure 44 shows the final prediction results. The algorithm converges at 1 % data distance which is proportional to the percentage of the data noise (Figure 45). Figure 46 shows the clustering results at iteration 1 and 10.
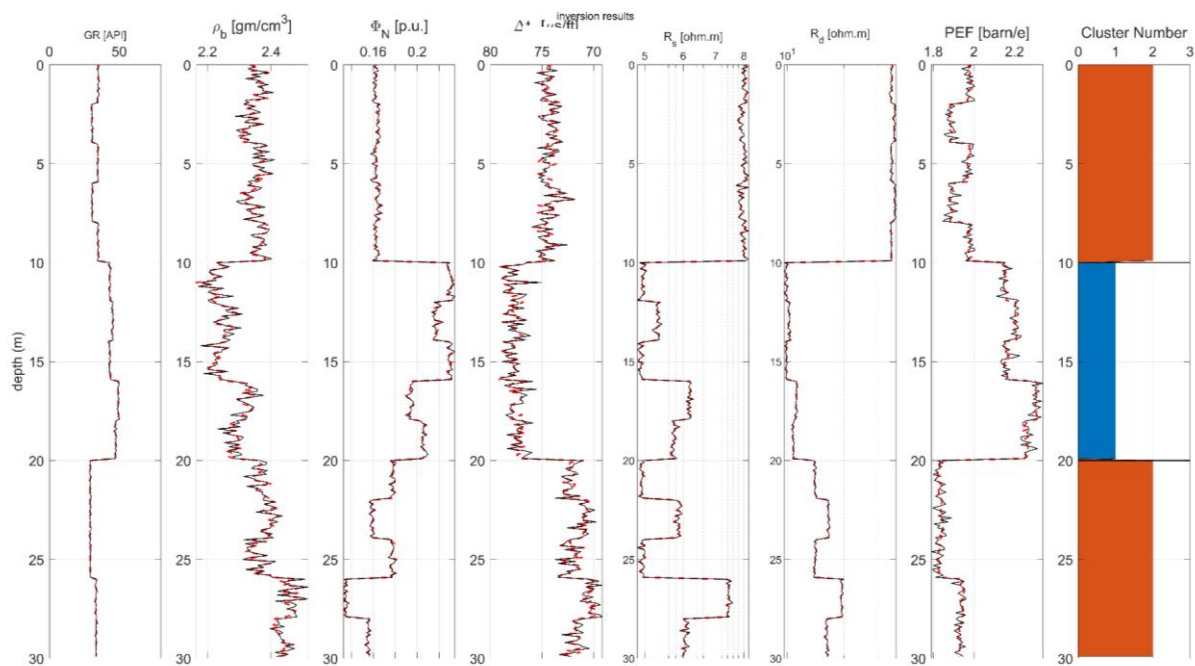
Figure 43. Fitting between the synthetic data and the calculated data (after 10 iterations); the red dashed lines represent the calculated data, while the solid black lines represent the synthetic data.
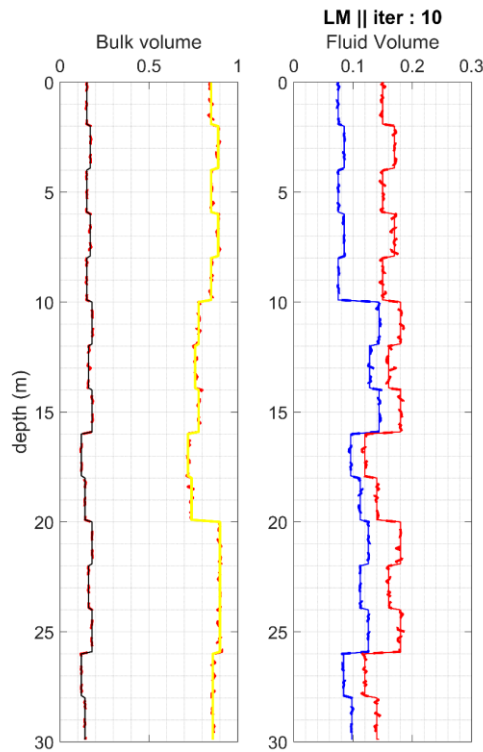
Figure 44. The predicted (dashed lines) and actual (solid lines) petrophysical parameters of the gas-bearing reservoir at iteration 10.

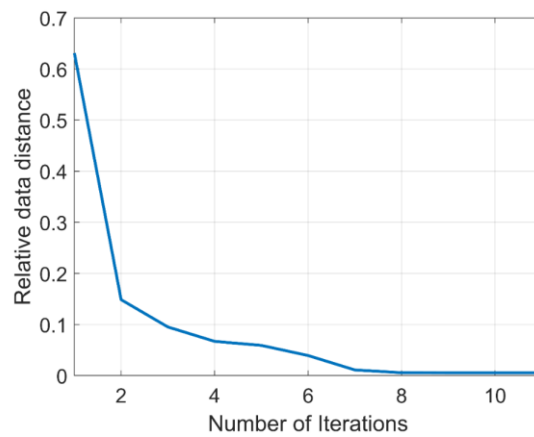**Figure 45.** Development of convergence during the local inversion of synthetic data contaminated by 1 % Gaussian distributed noise.
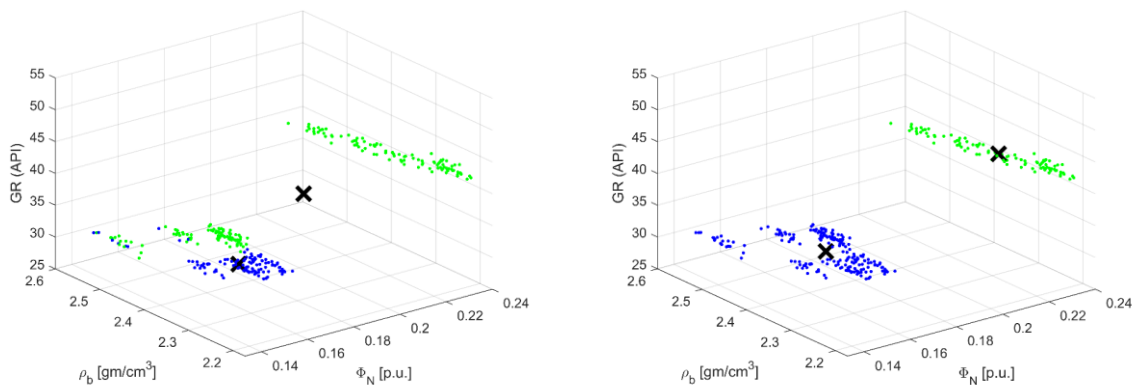
Figure 46. MFV-clustering results at iteration 1 (a) and iteration 10 (b), black crosses indicate the defined cluster centers.

5.3.2. Estimation of zone parameters

I used the same synthetic model that was used in chapter 4 for calculating the synthetic data. However, I increased the number of unknowns to 7 such as the density of shale and cementation exponent. In that case the density of shale and the cementation exponent were the unknown as zone parameters. Besides that, I assumed a constant value of these zone parameters in the whole processing interval, while in the other scenario there are different zone parameters of each layer. Within the context of zone parameters' evaluation alongside petrophysical parameters, the organization of the Jacobian matrix in figure (47) outlines the differences between the varying parts of the model.
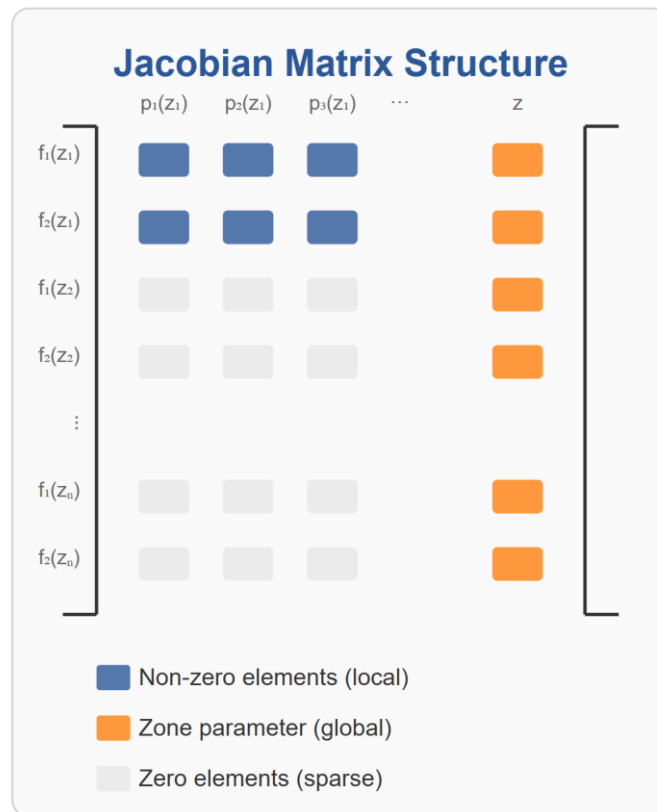
Figure 47. Jacobian matrix structure in case of zone parameters implementation.

This matrix contains a choice of non-zero elements indicated in blue color, which represent local petrophysical parameters such as ($p_1(z_1)$), ($p_2(z_1)$), and ($p_3(z_1)$). These are the specific values that display variation within the respective specified zone. Further, the orange sections

represent zone parameters that have a global attribute and therefore influence all layers and are symbolized by the parameter (z). The lighter zones of the matrix, represented in gray, represent zero elements, revealing areas where the influence of some parameters is minimal or zero. This organized method of the Jacobian matrix enables effective computation through the concentration of fundamental parameters impacting the well logging data and differentiating between locally changing and globally invariable factors. The synthesis of these specialized and generalized parameters into the Jacobian matrix provides an overall understanding of subsurface properties, elevating the accuracy and reliability of petrophysical and zone parameter estimates.

Figure 48 shows the initial location of the cluster centroids (left), and the final location of the centroids (right). Figures 49 and 50 the fitting between the calculated and measured data at iteration 1 and initial guess of the petrophysical and zone parameters, respectively.
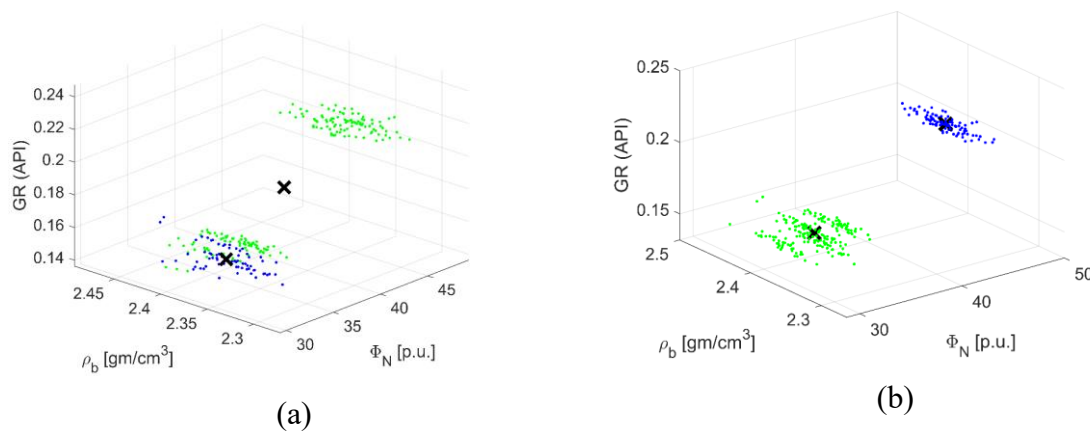
Figure 48. MFV-clustering results ("x" represents the cluster centroids) (a) initial guess (b)final location.
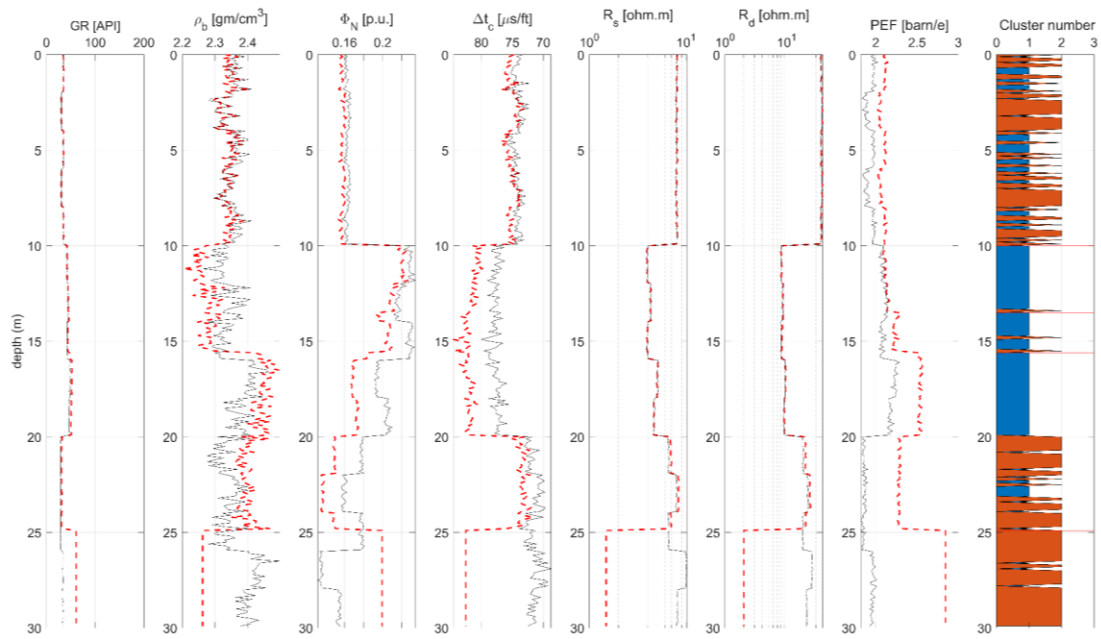
Figure 49. The fitting between the calculated and measured well log data sets at iteration 1 (1st scenario) for inversion A.
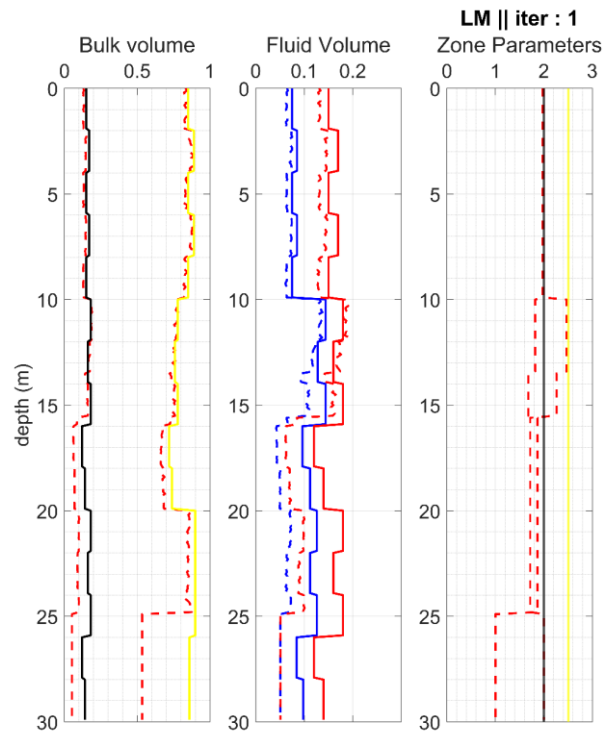
Figure 50. Initial guess of the petrophysical and zone parameters (1st scenario) for inversion A.

In the light of geological modeling and petrophysical analysis, the threshold of layer thickness refers to the minimum thickness a geological layer must have to be considered distinct and significant in the model. This threshold is crucial because it directly influences the

placement of layer boundaries and consequently, the total number of layers identified in the subsurface profile. Determining an appropriate threshold for layer thickness is a nuanced process that involves both geological understanding and practical considerations. It is often guided by the resolution of the available data and the smallest unit of interest that can be confidently distinguished given the measurement accuracy. For instance, a well log tool has a specific vertical resolution, typically in the range of tens of centimeters to a few meters, which sets a natural limit on the thinnest layers it can resolve. Layers thinner than this resolution may appear blended with adjacent layers, potentially obscuring their boundaries. The choice of threshold can significantly impact geological interpretation. Selecting a too-low threshold might lead to the identification of many thin layers, some of which could reflect noise rather than meaningful geological features. Conversely, a too-high threshold might oversimplify the model, missing critical thin beds such as those that might act as barriers or conduits to fluid flow. Further, geological factors like depositional environment, tectonic activity, and diagenetic processes naturally influence layer continuity and variability, which must be considered when setting a threshold. Ultimately, the threshold should be chosen to balance detail with clarity, ensuring that the model remains both geologically realistic and computationally feasible. This balance is essential for constructing models that are not only accurate but also useful for predicting subsurface behavior and guiding exploration or production strategies. In this study the suitable threshold number is 0.1 meters to be aligned with vertical resolution of logging tools. Figures 51 and 52 show the fitting between the calculated and measured data and the prediction of petrophysical and zone parameters at iteration 15, respectively. Figure 53 shows the convergence of the relative data distance.
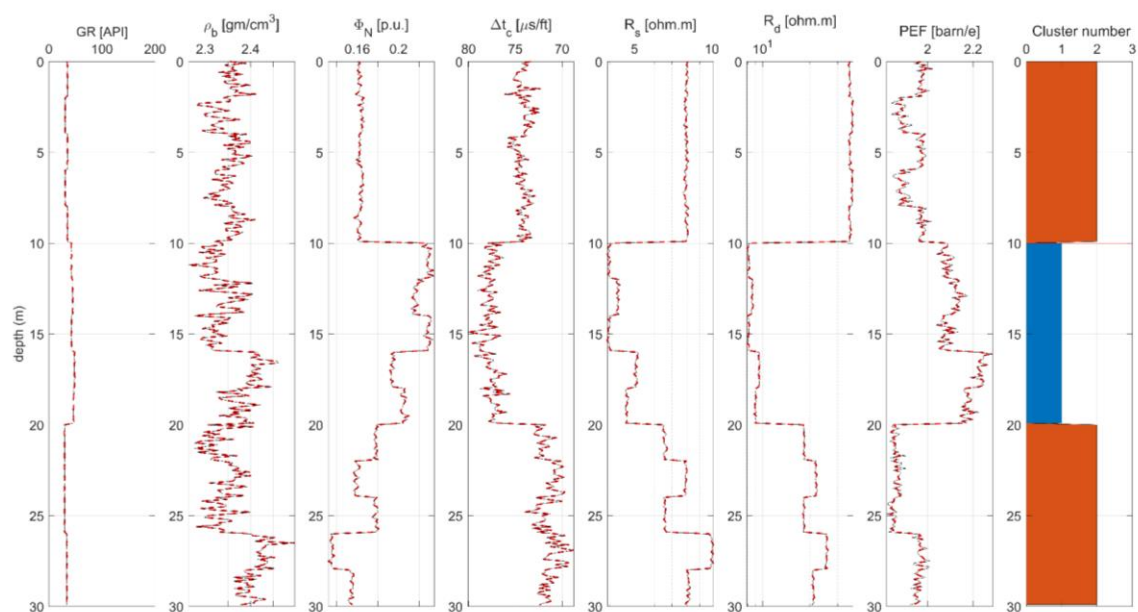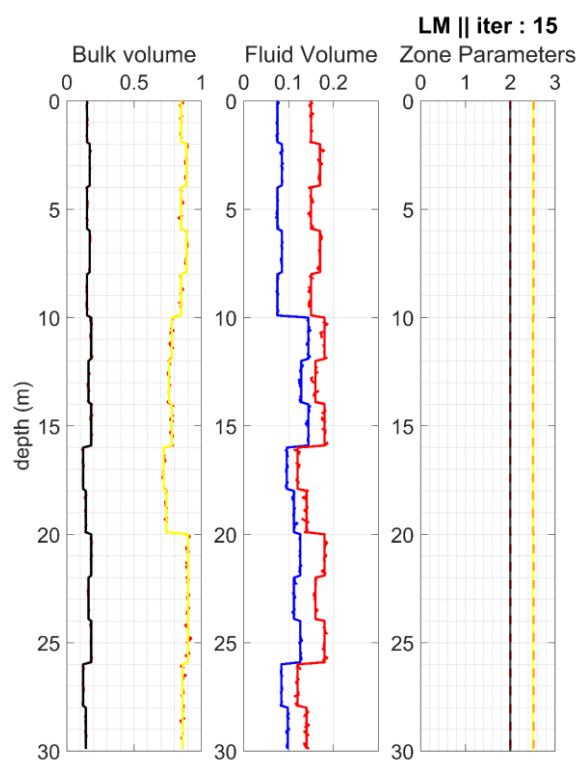
Figure 51. The fitting between the calculated and measured well log datasets at iteration 15 (1st scenario) for inversion A.

Figure 52. The predicted petrophysical and zone parameters at iteration 15 (1st scenario) for inversion A.

Figure 53. Development of convergence during the local inversion of synthetic well logging data using inversion A algorithm.

The following part shows the feasibility of applying inversion B algorithm in the second scenario, where the zone parameters are different from layer to layer. Figures 54 and 55 show the fitting between the calculated and measured data sets at iteration 1, and the initial guess of the petrophysical and zone parameters of each layer.

Figure 54. The fitting between the calculated and measured well log data sets at iteration 1 (2nd scenario) for inversion B.
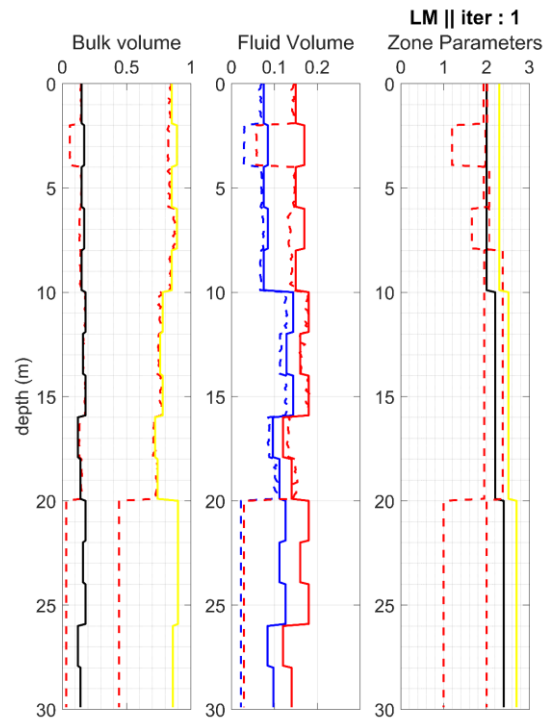
Figure 55. Initial guess of the petrophysical and zone parameters (2nd scenario) for inversion B.

Figures 56 and 57 show the fitting between the calculated and measured datasets and the predicted petrophysical and zone parameters of the (2nd scenario) inversion B algorithm at iteration 15.
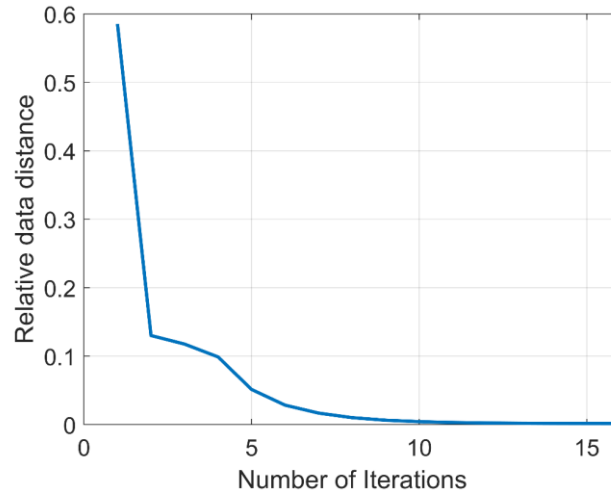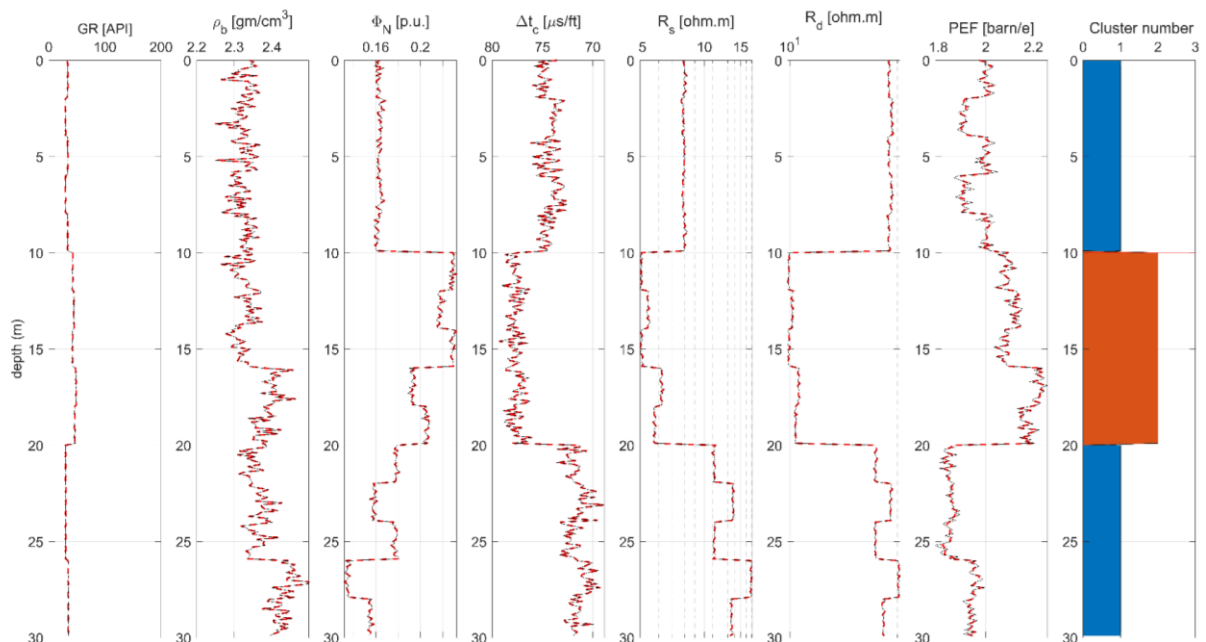

Figure 56. The fitting between the calculated and measured well log data sets at iteration 15 (2nd scenario) for inversion B.
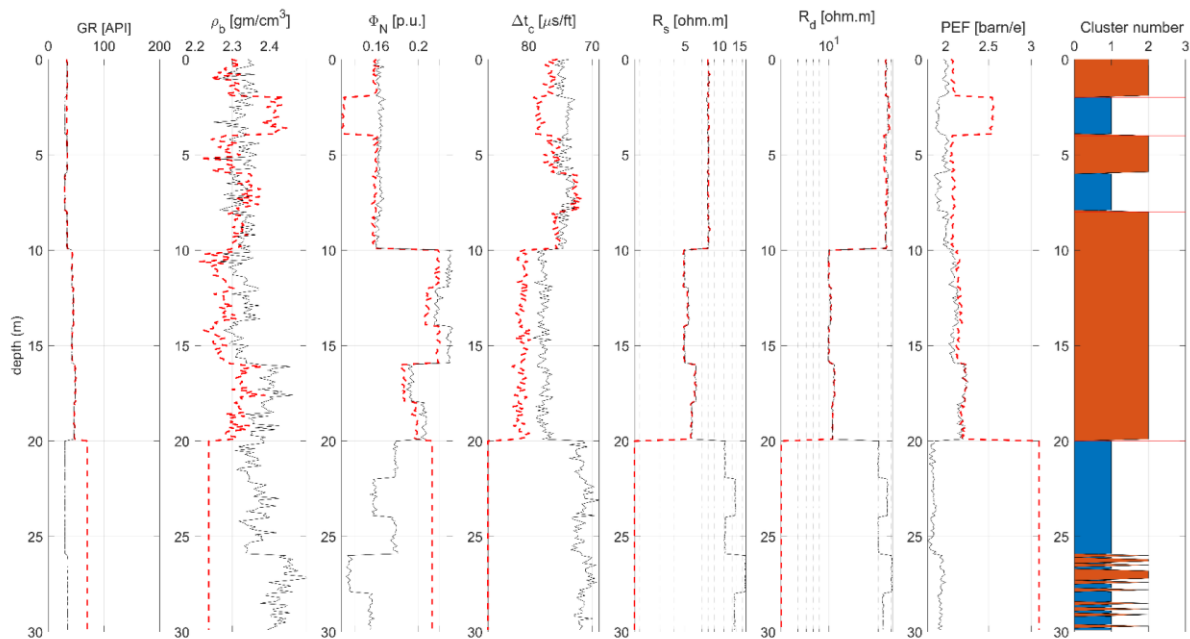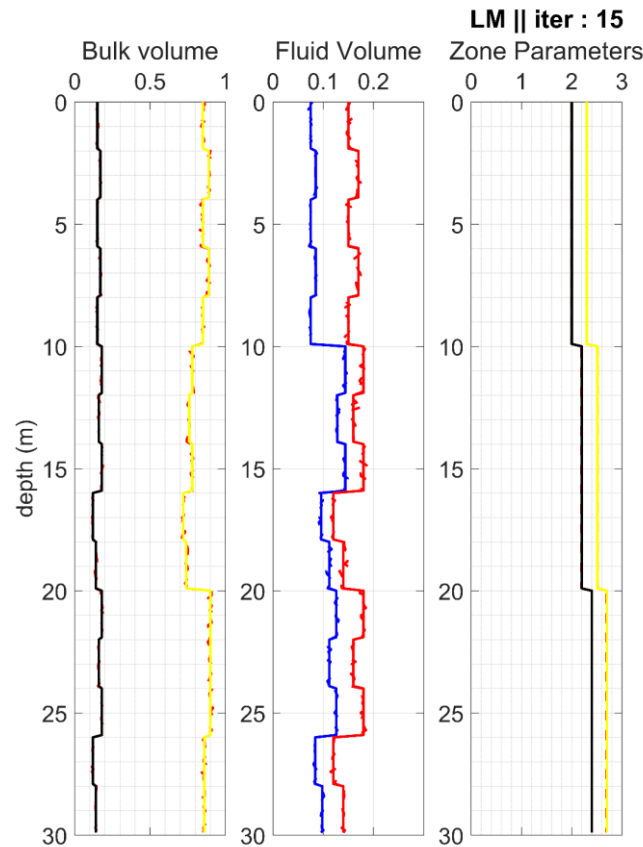
Figure 57. The predicted petrophysical and zone parameters at iteration 15 (2nd scenario) inversion B.

5.4. Hydrocarbon Field Case Study

5.4.1. Priori zone parameters

The proposed algorithm was tested using a hydrocarbon field data from Hungary. The recorded well-logging dataset depicts a sand-shale sequence response of thickness 25 m. To identify lithological units, a lithological analysis was carried out utilizing cross plots between well log data. The elbow method revealed that three lithological clusters were the ideal number (Figure 58). Within the geological formation under research, these three clusters indicate sandstone, shaly-sandstone, and shale strata. The cross plots allowed the well log data to be classified into three separate lithological groups based on the graphical correlations between different recorded parameters. This lithological categorization can help to understand the stratigraphic composition and attributes of the geological strata penetrated by the well, where the algorithm divides the sequence into 5 shale-sandstones intervals. Figure 59 illustrates the results of MFV-clustering at iteration 1. in a 3D cross-plot (x-axis: GR (API), y-axis: Bulk

density (g/cm^3), z-axis: Neutron-porosity (v/v)). Figure 60 shows the fitting between the calculated (red dashed) and measured (solid lines) data set at iteration 10. Figure 61 shows the predicted model parameters. Figure 62 shows the results of MFV-clustering.
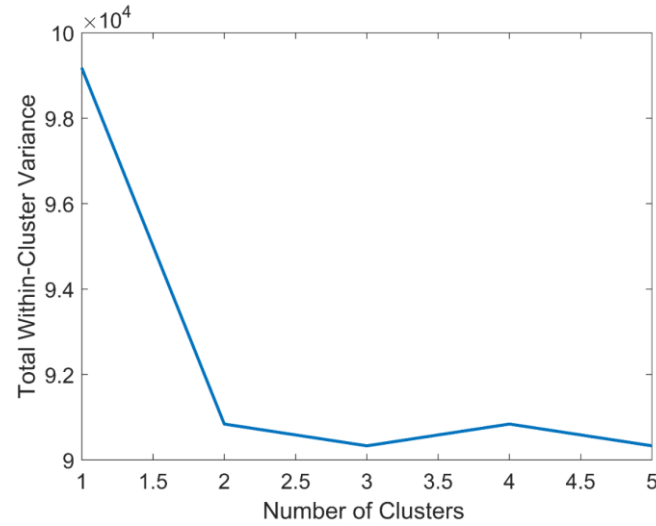
Figure 58. Elbow method for identifying the optimal number of clusters (the optimal number is three).
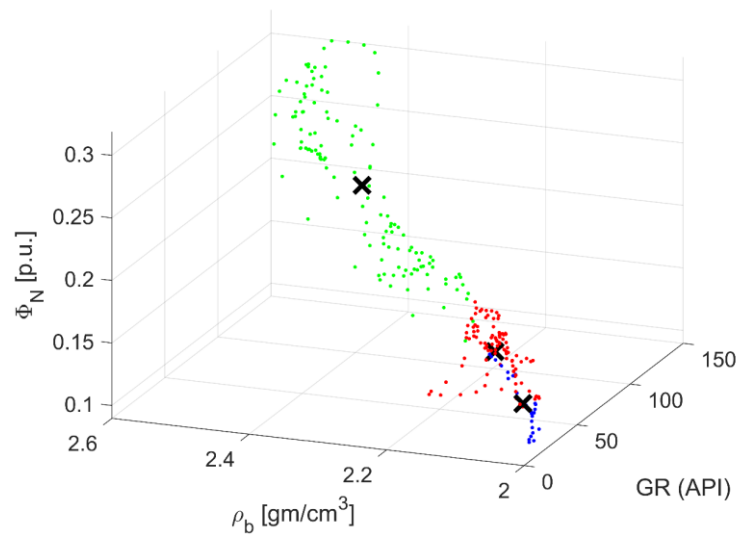
Figure 59. MFV-clustering results at iteration 1 ("x" represents the cluster centroids).
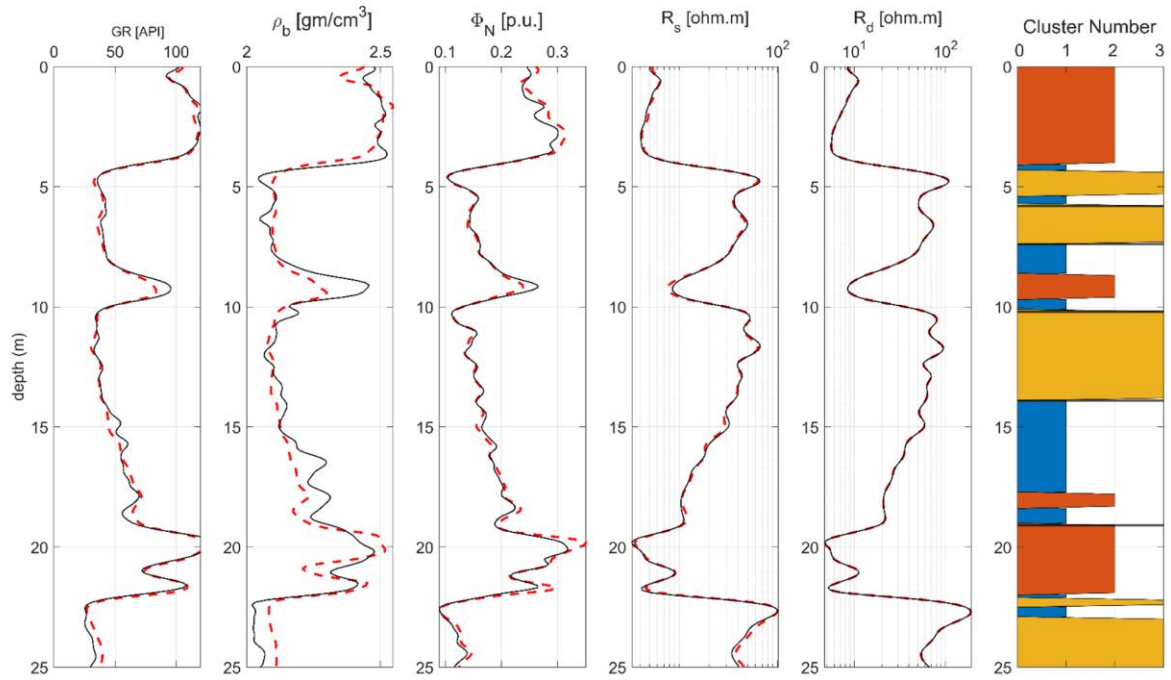
Figure 60. The fitting between actual (solid black line) and predicted (red dashed line) wireline logging data in the synthetic case (iteration 10).
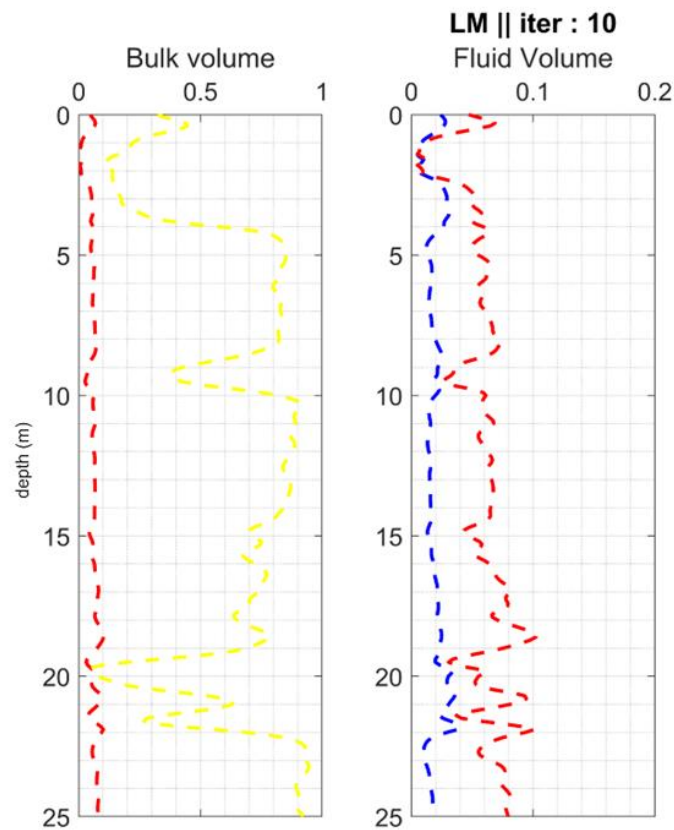
Figure 61. The predicted (dashed lines) petrophysical parameters at iteration 10.
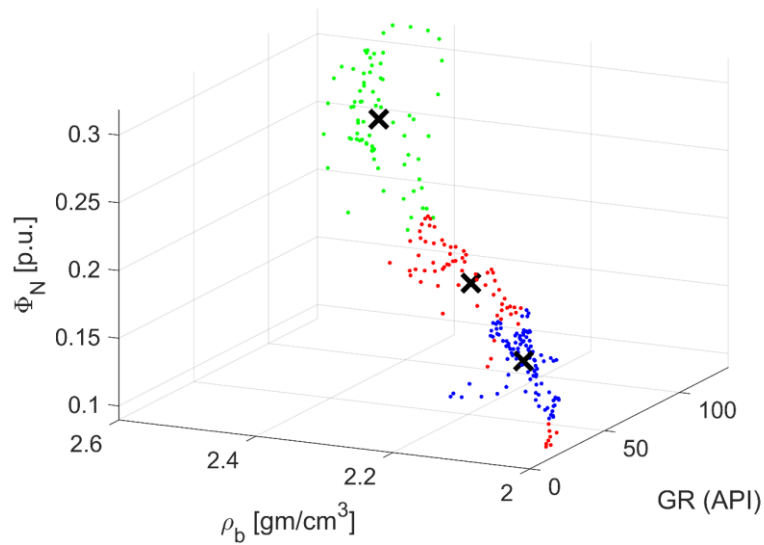
Figure 62. MFV-clustering results of iteration 1 ("x" represents the cluster centroids).

The misfit between the measured and calculated data converges from data distance of 0.6 at iteration 1 to 0.75 at iteration 10. Figure 63 shows the convergence of inversion algorithm.
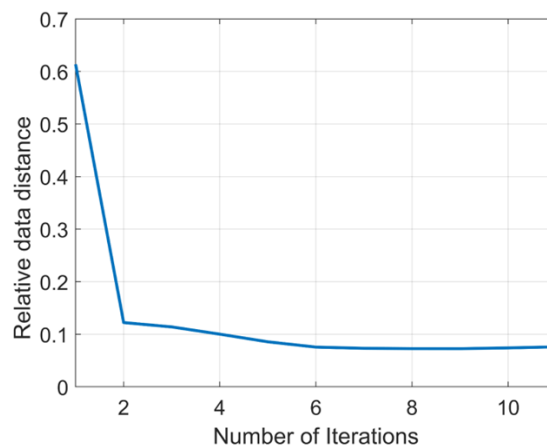
Figure 63. Development of convergence during the local inversion of real well logging data.

To check the stability of the inversion algorithm, the workflow was repeated 8 times, and the data distance convergence was recorded and plotted (Figure 64). The inversion algorithm chooses a different initial model each time, by turn the data distance starts from different values, and converges to the same end point.
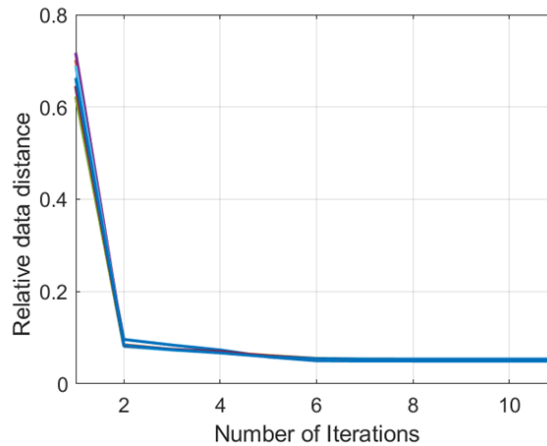
Figure 64. Relative data distance convergence repeated 8 times to check the algorithm stability.

5.4.2. Estimation of zone parameters

The estimation of zone parameters within each interval reveals an inverse relationship between shale resistivity and the cementation exponent. As the resistivity of shale increases, the cementation exponent tends to decrease, suggesting these parameters are indirectly linked in influencing petrophysical properties. The predicted reservoir characteristics and fluid distribution align with the expected rock types, indicating that hydrocarbon levels tend to diminish in regions with higher shale content, and increase in areas with lower shale content. Figure 65 shows the predicted petrophysical and zone parameters after 20 iterations. Figure 66 shows the convergence of the relative data distance at iteration 11 (the algorithm stopped at iteration 11 where no additional change happened). The predicted zone parameters show a range of cementation exponents between 1.8 to 2.7, while the shale resistivity shows a range between 3.8 to 5.9 ohmm. The resulting petrophysical and zone parameters show a proper fitting with actual well logging data. The zone parameters show an increase from shallow to deep. Specifically, the bulk density log, which shows a better fitting between the calculated and the actual data. Figure 67 shows the data distance convergence.
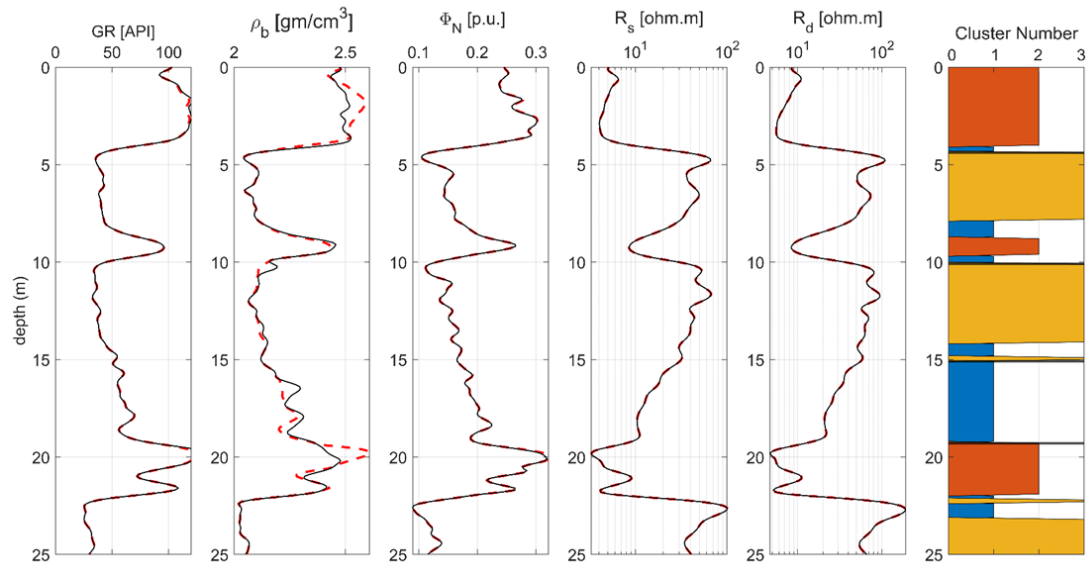
Figure 65. Shows the fitting between the calculated and measured data and the prediction of petrophysical and zone parameters at iteration 11.
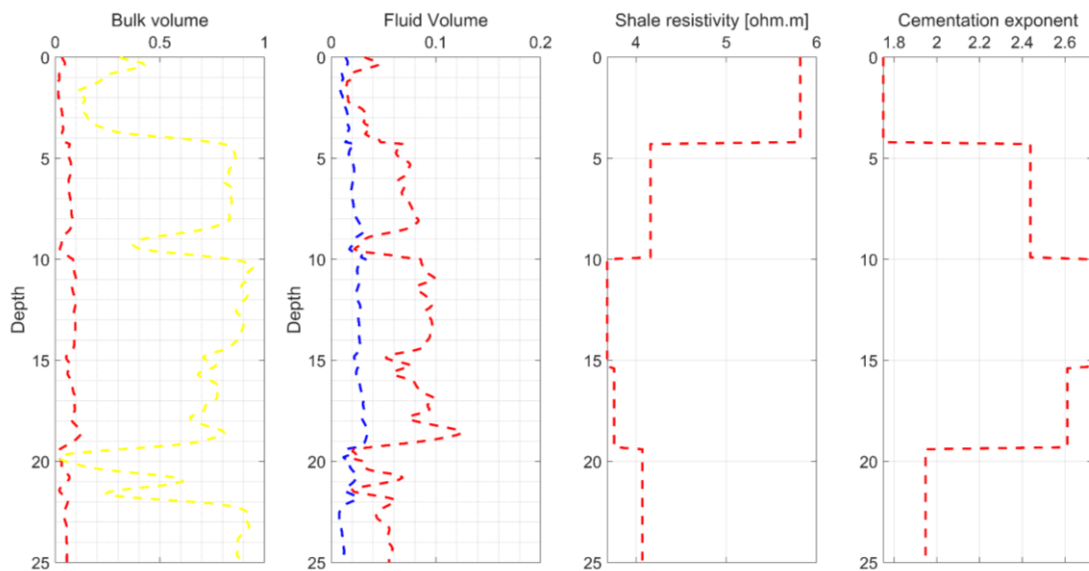
Figure 66. The predicted (dashed lines) petrophysical parameters at iteration 11 using integrated different zone parameters inversion scenario (inversion B).
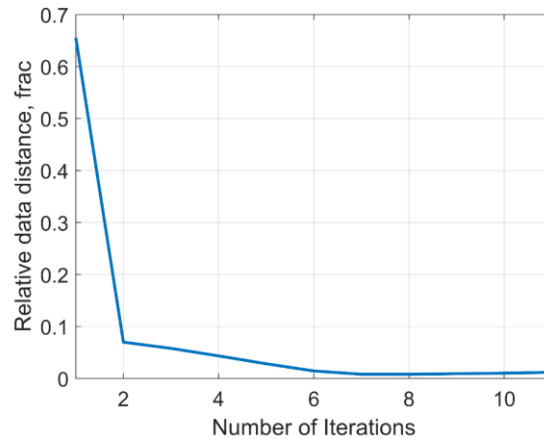
Figure 67. Development of convergence during the local inversion of real well logging data

5.5. Summary of results

In this study, an integration of the MFV clustering method and point-by-point inversion was developed and tested using both synthetic and field well logging datasets to evaluate petrophysical, zone, and geometrical attributes. The synthetic dataset was generated based on a 30 m sequence comprising two hydrocarbon and water-saturated sandstone layers flanking a shale layer, with 1% Gaussian noise added for complexity. The inversion algorithm demonstrated its efficacy by reducing the relative data distance curve from 52% to 1%, highlighting its accuracy in fitting measured and calculated data. For field data application, the algorithm analyzed a dataset containing gamma ray, density, neutron, and resistivity logs, clustering the data into three groups: sandstone, shale sandstone, and sandy shale layers based on shale volume. Petrophysical parameters within each layer revealed different sand layers partially saturated with hydrocarbons, which were interspersed with thin shale laminas influencing reservoir connectivity.

The automated method facilitated the estimation of volumetric parameters one layer at a time, contributing to the construction of a realistic static reservoir model. Validated by synthetic cases with constant and varying zones, the method demonstrated high accuracy under both uniform and heterogeneous subsurface conditions. Allowing zone parameters to vary improved predictions of shale volume and porosity, emphasizing the importance of localized tuning in the inversion process. Application of this method to a dataset from a Hungarian hydrocarbon

field confirmed its predictive capability regarding petrophysical properties and variations in zone parameters. Shale-rich zones exhibited low hydrocarbon content, and predictions for lithological patterns were consistent with actual formations. This approach not only enhances the estimation of petrophysical properties but also effectively quantifies zone parameters, underscoring its utility in reservoir characterization.

Thesis 4.

I present an automated inversion scheme utilizing MFV clustering and point-by-point inversion to predict geometrical and petrophysical parameters, improving log data interpretation. The initial algorithm assumes known zone parameters, while the advanced version handles unknown parameters, both constant and varying across layers. Validation with synthetic data sets showed robust clustering into lithological groups. Application to a Hungarian dataset, as depicted in the figure, demonstrates classification into rock types and parameter estimation, with bulk volume around 0.5, fluid volume 0.1, shale resistivity up to 6 ohmm, and cementation exponent ranging from 1.8 to 2.6. This approach enhances reservoir characterization by integrating layer-specific volumetric and zone parameters, aligning with calculated saturation profiles and advancing static reservoir modeling.

Chapter 6: Fully Automated Algorithm for Petrophysical, Zone Parameters, and Layer-Thickness Estimation with Robust Clustering Assisted Interval Inversion

In this chapter, both clustering and interval inversion play critical roles in interpreting well logging data. Clustering helps in discovering meaningful patterns from the complex, multidimensional datasets that characterize well logs, allowing for the identification of distinct zones and formations based on their unique characteristics. Once these zones are defined, I will introduce two distinct methodologies, taking advantage of the interval inversion techniques, and anticipate variations in petrophysical parameters, such as porosity and permeability, within them in case of known zone parameters and unknown zone parameters. This dual approach not only enhances the precision of subsurface characterization but also offers a cohesive understanding of how different geological layers interact, ultimately leading to more informed decision-making in resource exploration and extraction. The overall convergence of the relative data distance is a function of the misfit of the cluster as well as the interval inversion results.

The newly introduced algorithm addresses the complex challenge of polynomial degree and thickness dependency, where variations in polynomial degree correspond to changes in petrophysical parameters across defined thicknesses. The algorithm strategically determines the appropriate polynomial degree based on interval thickness or the coefficient of variation within each zone, enhancing its adaptability and precision. This novel approach integrates unsupervised pattern recognition with interval inversion, significantly transforming how subsurface data is interpreted. The algorithm can automatically predict the number of sub-intervals and, furthermore, it can adjust the degree of the polynomial in the homogeneous layer-wise modeling to ensure the best fit between the calculated and actual well-logging data. Thanks to the high data-to-unknown parameter ratio, I could implement all the unknown parameters in one optimization phase. The zone parameters include the textural properties of rocks, i.e., cementation exponent (m), and tortuosity factor (a), and the shale parameters, i.e., resistivity of shale, and natural gamma-ray intensity of shale. These parameters were dealt as constant within each subinterval and differ from one another; therefore, the interval inversion uses the Heaviside as a basis function to represent the continuity of the zone parameters. Finally, the change of the overdetermination ration is a function of the thickness of the defined interval, so a coefficient of variation has been calculated each iteration to explain the degree of the polynomial. Figure 68 shows the schematic diagram of the workflow of the proposed

inversion algorithm. The optimal cluster number is pre-defined from the elbow method in case of the real data.
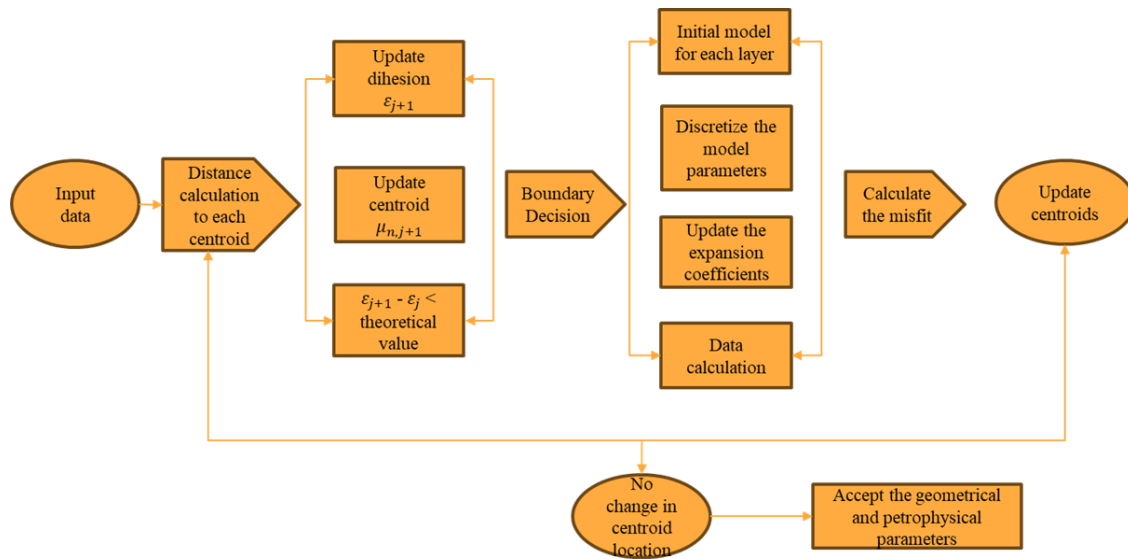
Figure 68. The schematic diagram shows the workflow of the proposed inversion algorithm.

6.1. Synthetic modeling test

6.1.1. Prior zone parameters

In this section, I test the change of the lithology as well as the amount of noise added to the inverted data. Therefore, I used the same set of synthetic well logging data but with increasing the amount of Gaussian distributed noise to 2, 5 and 7 %. Besides that, I increased the amount of shale in case of the data contaminated with 5 and 7 % noise. Figure 69, and Figure 70 show the fitting and the initial and estimated model parameters at iteration 1.
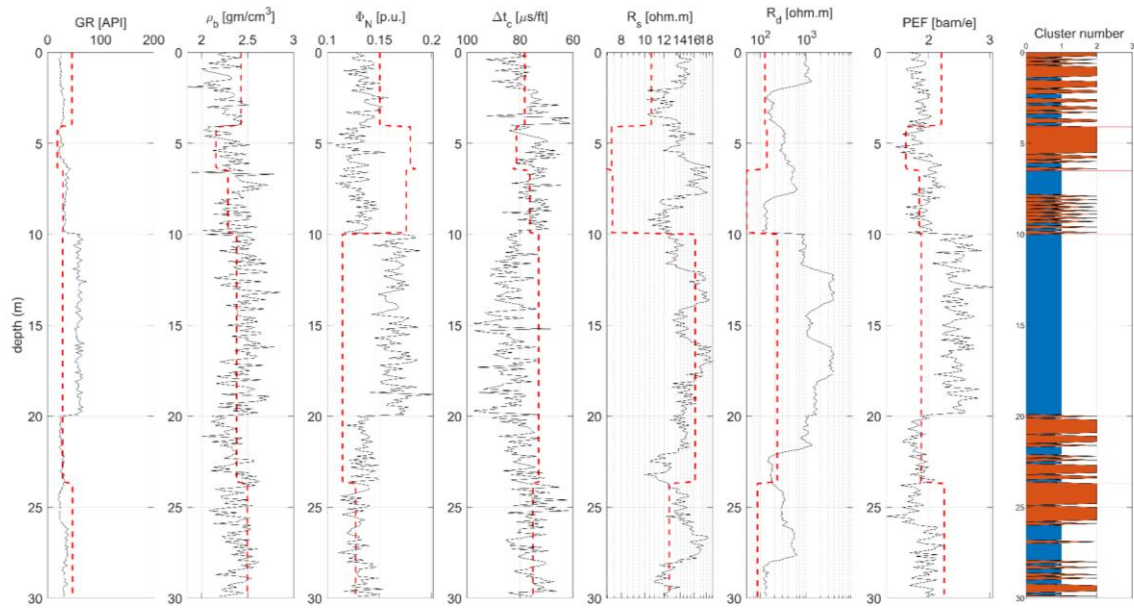
Figure 69. Synthetic wireline logs (black curve) and the calculated logs (red curve) using the initial model obtained from cluster analysis at iteration 1 (7% Gaussian noise).
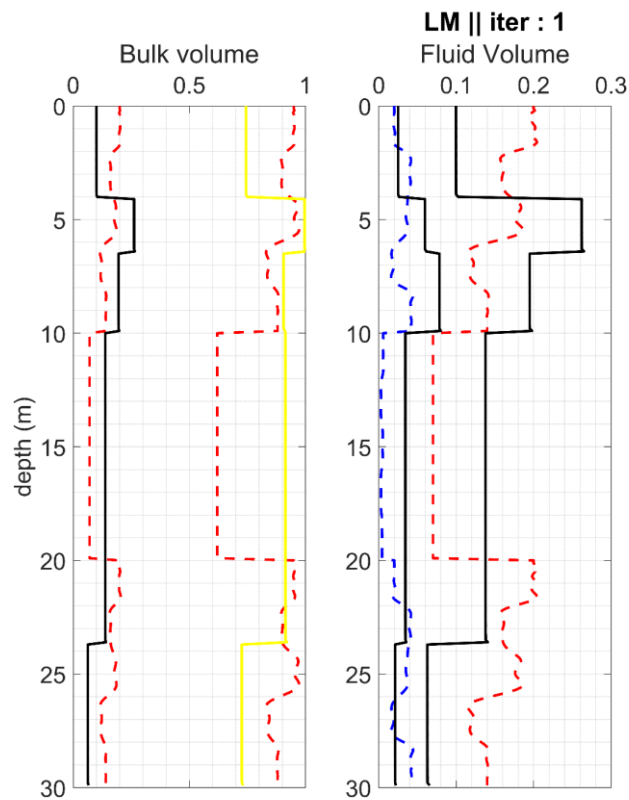
Figure 70. The predicted (dashed lines) and exactly known (solid lines) petrophysical parameters of the synthetic data contaminated by 7% Gaussian distributed noise at iteration 1.

Figure 71 (a), (b), and (c) show the fitting results in case of the 2, 5, and 7 % Gaussian noise contamination, respectively. It's clear from the Figures the effect of noise specifically on

density log, where the boundary between different layers disappears in case of a higher noise dataset.
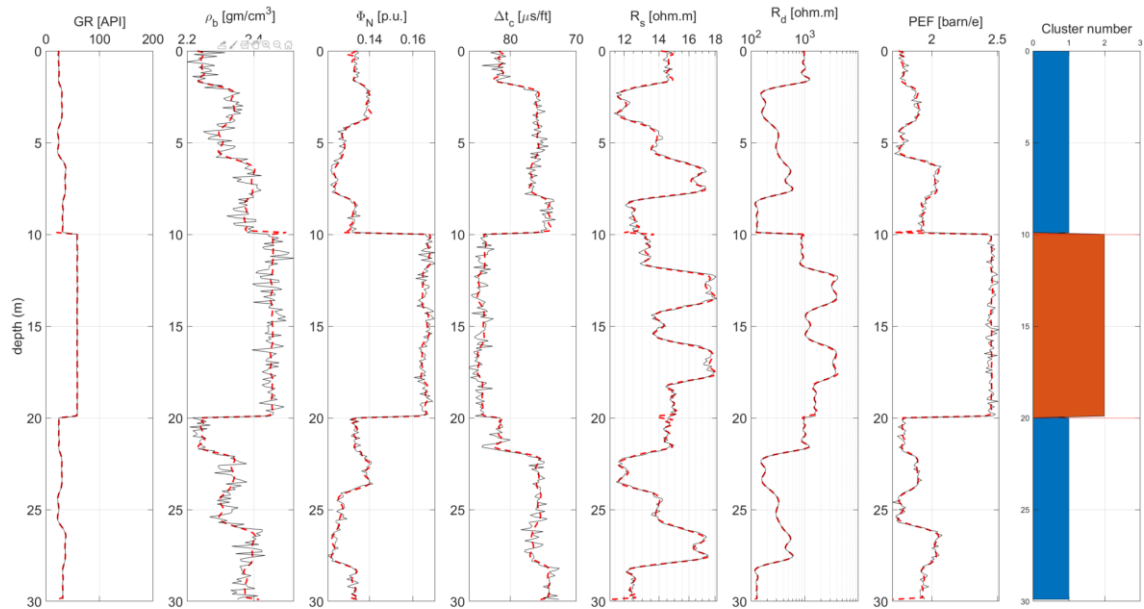
(a)
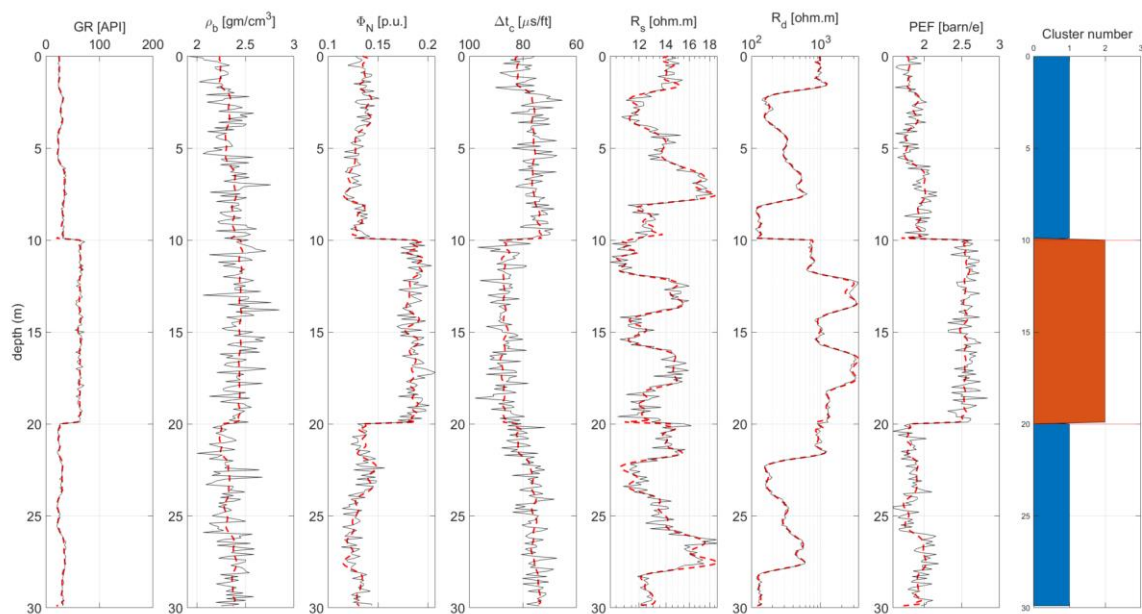
(b)
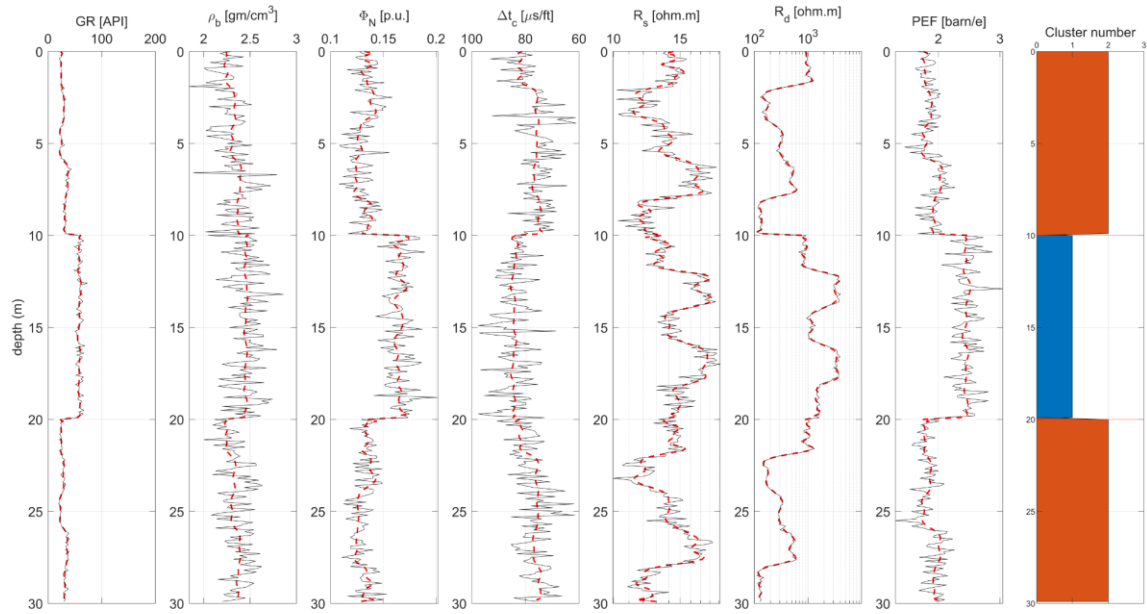
(c)

Figure 71. Fitting between the calculated and synthetic data. (a) 2% noise, (b) 5% noise, and (c) 7% Gaussian distributed noise.

Figure 72 (a) and (b) show the initial location of the centroids of MFV-clustering loop. The model under consideration prominently displays a discernible partition into two distinct lithological clusters. In scenarios characterized by an increased level of noise contamination, there is a noteworthy surge in shale content within the shale layer, reaching a substantial 50%. This augmentation in shale content not only exerts an influence on the inversion procedure but also accentuates the differentiation between the clusters. Consequently, this amplification contributes to a heightened clarity in the demarcation of layer boundaries. The intricate interplay between shale content variations and their impact on the clustering dynamics underscores the nuanced complexity inherent in the model's lithological composition.
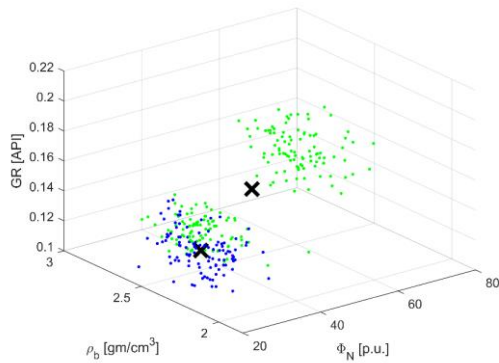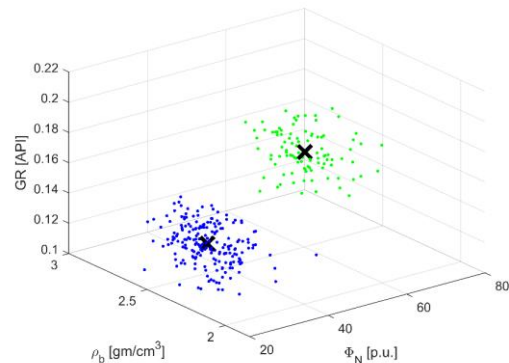
(a)   (b)

Figure 72. MFV-clustering results ("x" represents the cluster centroids) (a) initial guess (b)final location.

Figure 73 shows the convergence of data distance. The data distance stabilizes the amount of contamination noise. The analysis of convergence patterns reveals an important relationship between the iteration counts of the main loop, which manages clustering, and the inner loop, which performs inversion, particularly in response to varying levels of noise in the data. Convergence patterns indicate how swiftly and efficiently an algorithm reaches stability, and these patterns are influenced by the iterations in both loops. The main loop works on organizing data into clusters by continually refining centroids and updating cluster assignments, a process that can be significantly affected by noise as it tends to obscure data patterns. In contrast, the inner loop focuses on adjusting model parameters to align the predictions with observed data, requiring precision that is often sensitive to noise levels. Noise acts as random variability within the data, potentially complicating both the clustering and inversion processes. When noise levels rise, the clusters interference increases. Besides that, iterations might be necessary to achieve stable convergence as the algorithm works to distinguish meaningful signals from background noise. The observed relationship suggests that for inversion stability, the main loop should iterate approximately half as frequently as the inner loop. This balance ensures that while the inversion process effectively refines parameters to match expected patterns, the clustering process remains efficient in adapting to these adjustments. By understanding and applying this interrelationship, the iterative algorithm can be optimized to converge effectively, even in the presence of noise, enhancing both the reliability and accuracy of the results generated from noisy datasets.
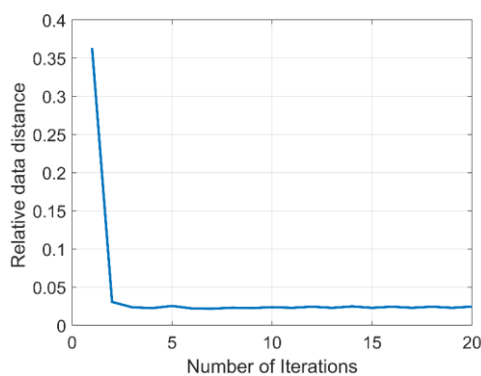
(a)
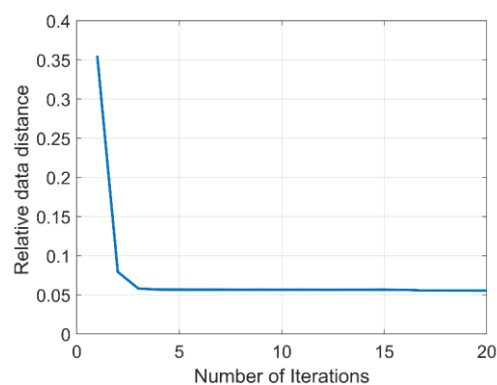
(b)
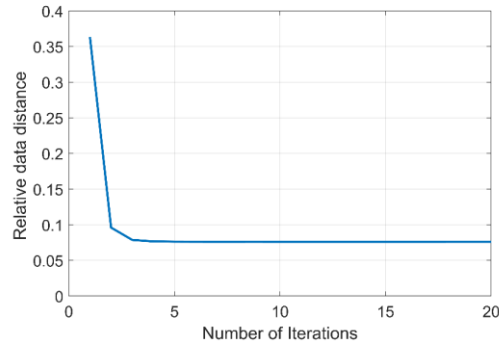
(c)

Figure 73. The convergence of the relative data distance. (a) 2% noise, (b) 5% noise, and (c) 7% Gaussian distributed noise.

The inversion algorithm could detect the coordinate of the boundary quite accurately until 7 % noise but failed for the higher noise percentage. Figure 74a-c shows the predicted parameters of each dataset.
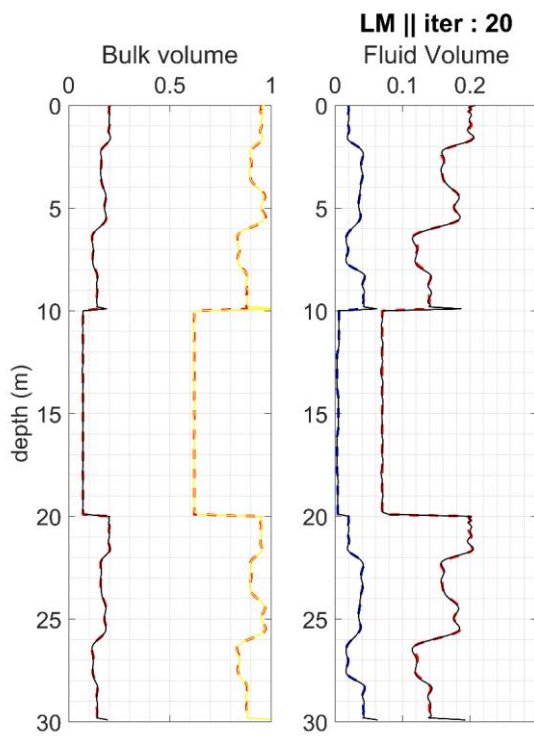
(a)
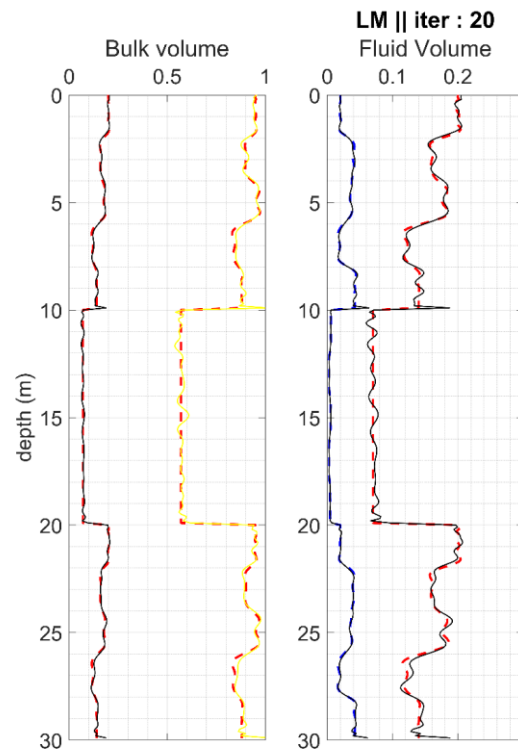
(b)
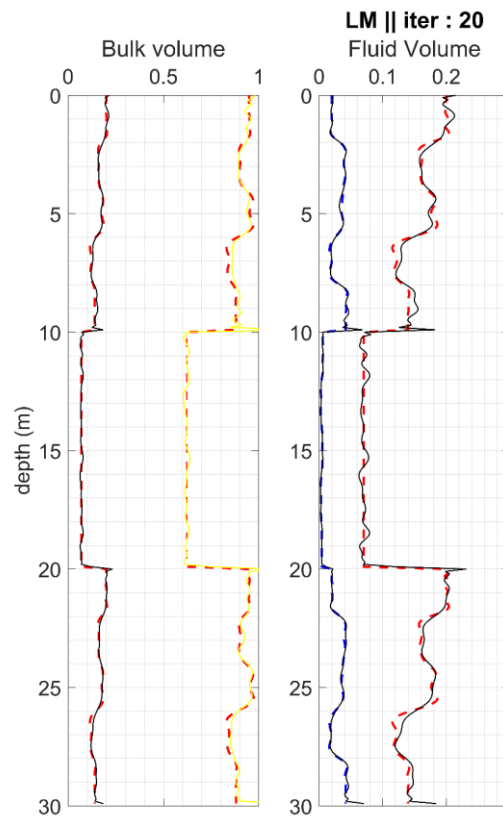
(c)

Figure 74. The exact known (dashed lines) and predicted (solid lines) petrophysical parameters of the gas-bearing reservoir at Iteration 20. (a) 2% noise, (b) 5% noise, and (c) 7% Gaussian distributed noise.

6.1.2. Estimation of zone parameters

In case of optimize zone parameters, sensitivity study is an important tool to show how the change of model parameters can influence the change of the calculated data. Four representative layers were modeled: clean hydrocarbon-filled sand (layer 1), tight sand (layer 2), water-saturated sand (layer 3), and sandy shale (layer 4). This allowed testing of sensitivity of both textural and shale parameters across a range of conditions.

Figure 75 summarizes how the petrophysical properties of each layer were systematically varied. Examining the parameter sensitivities for this diverse set of scenarios informs the selection of significant terms to include during subsequent optimization workflows by revealing those with the most pronounced influence on outcomes given underlying geology. The sensitivity metrics quantify the relative response arising from adjustments to textural attributes versus shale measures under analogous perturbations. By understanding sensitivities for models approximating potential reservoir architectures, parameters most essential for

constraining estimates can be identified before calibrating interpretations through history matching based on actual well data. The parameter sensitivity function is given as follows:

$$S(m_i) = \frac{\partial d_l}{\partial m_i} * \frac{m_i}{d_l(m_i)} \tag{62}$$

Table 7. Petrophysical model for computing parameter sensitivity functions.

	Porosity (v/v)	Volume of sand (v/v)	Water saturation (v/v)
Layer 1	0.15	0.80	0.20
Layer 2	0.05	0.90	0.05
Layer 3	0.10	0.90	0.90
Layer 4	0.04	0.60	0.10

The sensitivity plot shows how changes in the various field parameters affect the results for the four modeled layers. Several important observations can be made. First, the cementation exponent and acoustic damping appear to be significantly affected in all cases. Meanwhile, the dense water shows little sensitivity except in the dense sand layer. Moreover, layer 2 exhibits unusual behavior compared to the other cases - parameters such as porosity and grain density for unconventional tight sands may exhibit different relationships than conventional reservoirs. The sensitivity to these heterogeneous constructions may need new investigation. In summary, cementation exponent and turbidity are always the most sensitive terms, water hardness always has a negligible effect, while unconventional fields reveal their underlying complexities by sensitivity discontinuities. The diagnostic power of sensitivity analysis is evident in how it reveals contrasts in fundamental parameter dependencies between standard and tight framework. A sensitivity study aims to identify parameters that significantly impact geophysical model outcomes. Key parameters like cementation exponent, shale density, and neutron density are chosen based on their sensitivity profiles. These parameters accurately represent subsurface conditions and make the model more robust. The study also evaluates the feasibility of the interval inversion method for multi-parameter estimation using synthesized well-logging data. Sensitivity analysis and localized inversion allow for the extension of the definition to obtain three additional zone parameters simultaneously.
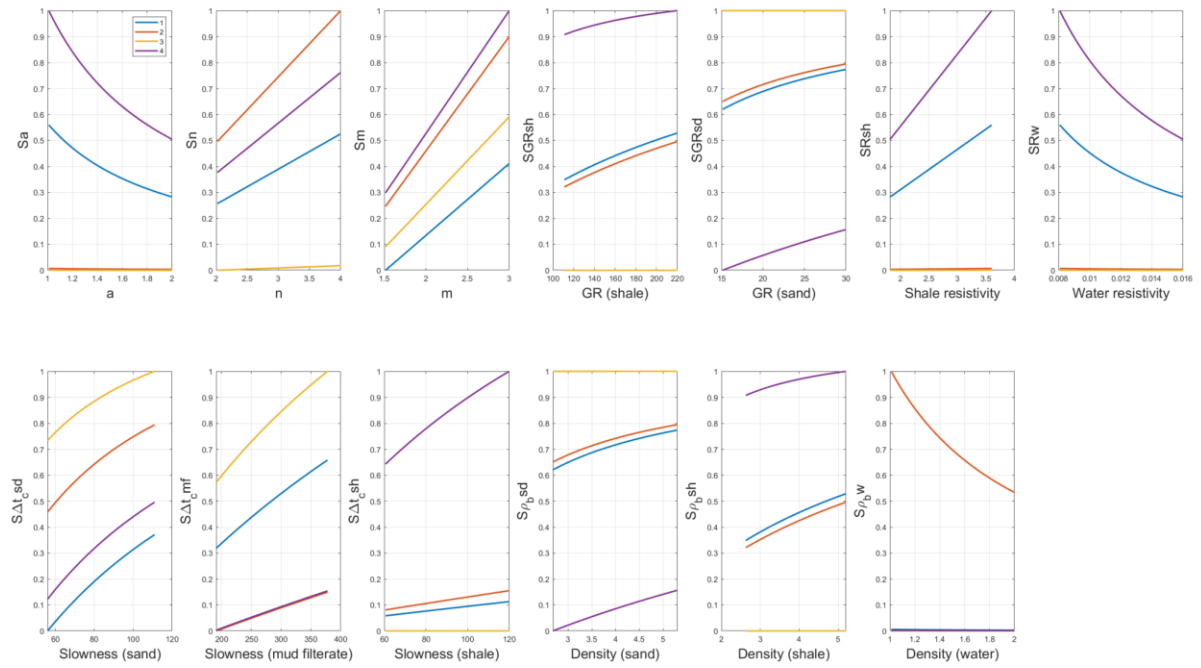
Figure 75. Sensitivity study of zone parameters within different four lithological layers, layer 1 (blue), layer 2 (orange), layer3 (yellow), and layer 4 (purple).
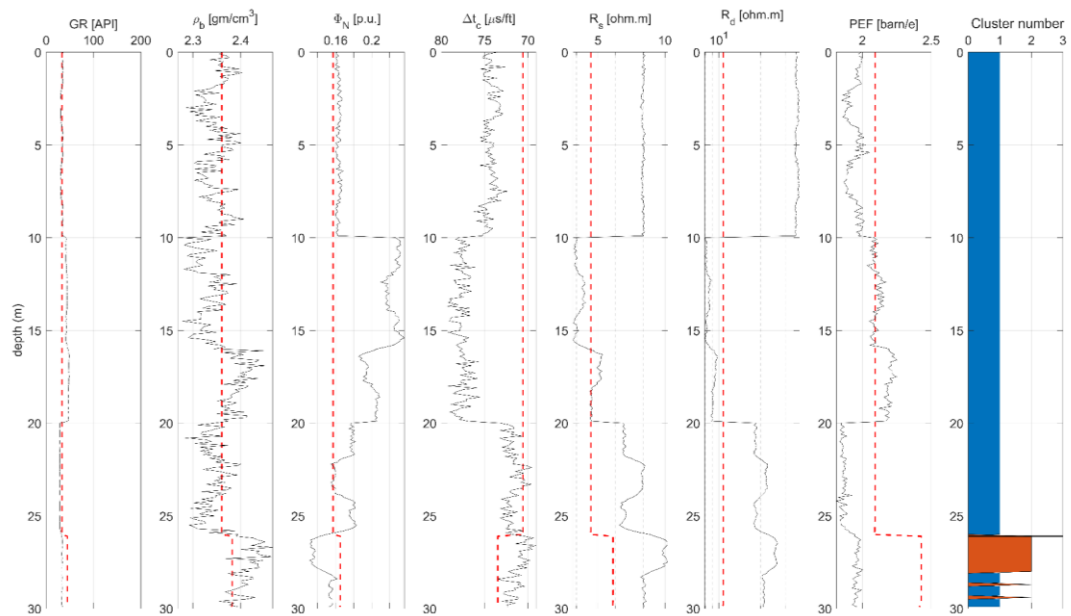
(a)
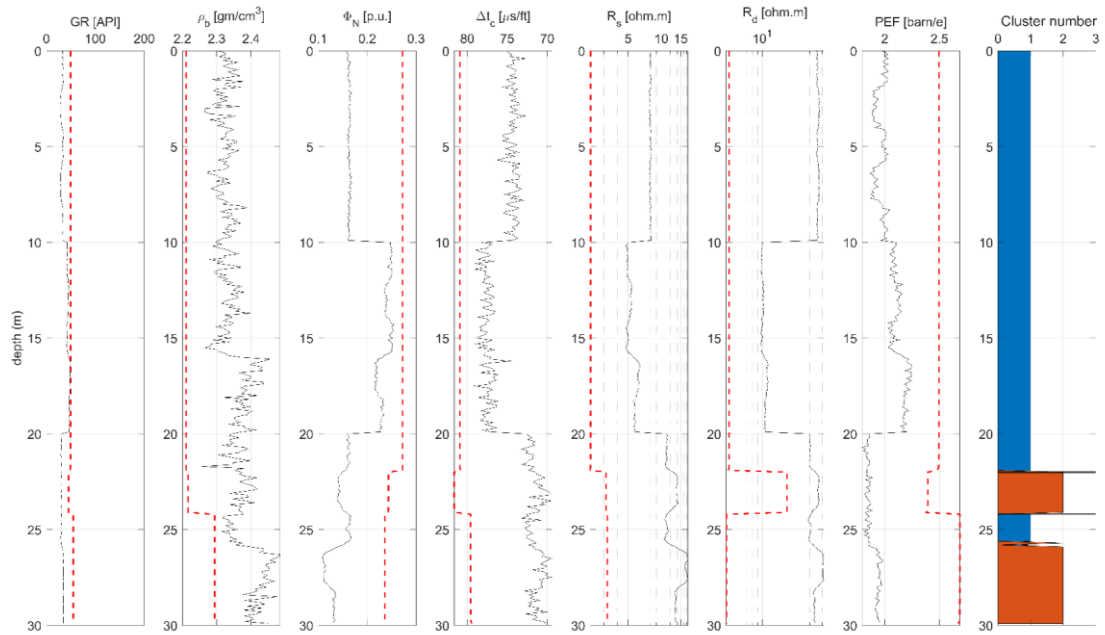
(b)

Figure 76. The fitting between synthetic (solid black line) and predicted (red dashed line) well logging data in the synthetic case of iteration 1, (a) inversion A (scenario 1), and (b) inversion B (scenario 2).

(a)
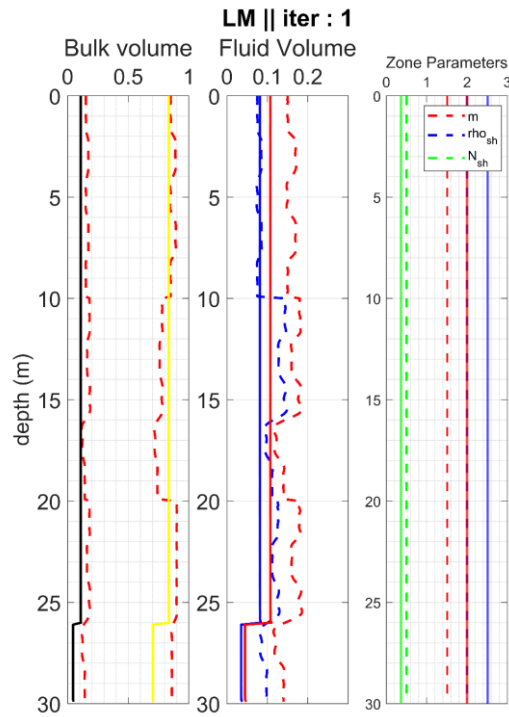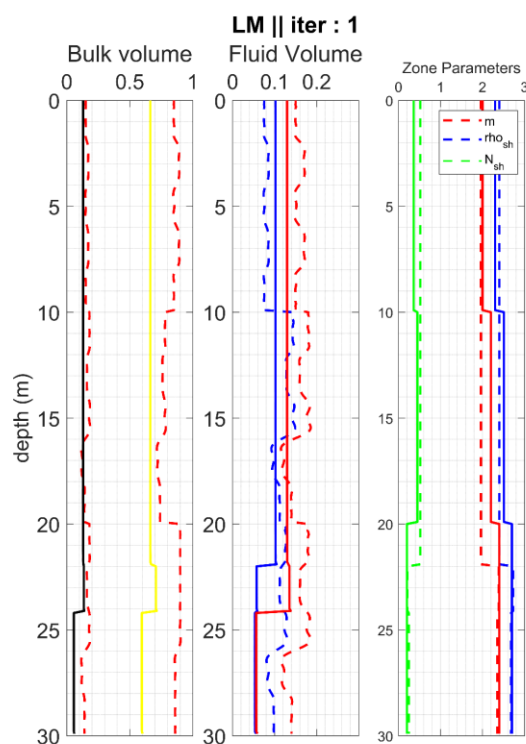
Figure 77. The exactly known (dashed lines) and predicted (solid lines) petrophysical parameters of the synthetic data at iteration 1. (a) inversion A (scenario 1), and (b) inversion B (scenario 2).
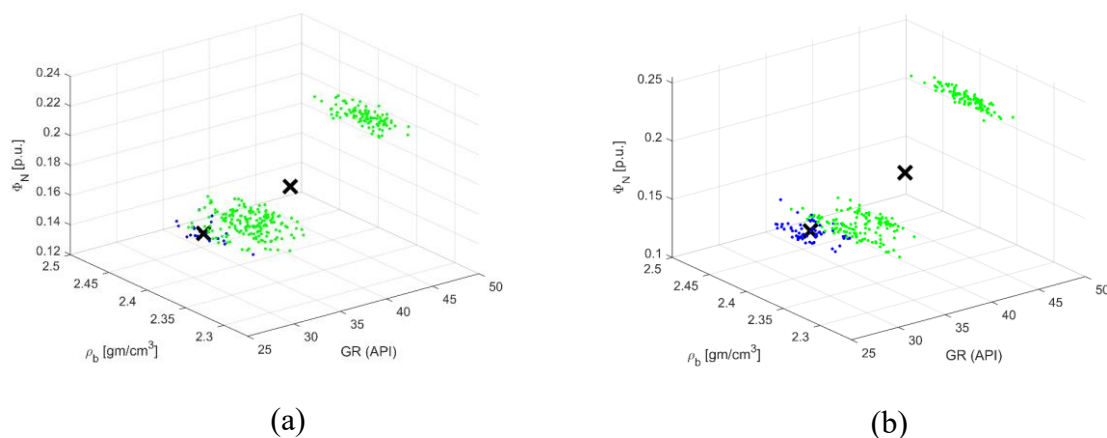
Figure 78. Location of cluster centroids at iteration 1, (a) scenario 1 (inversion A) and (b) scenario 2 (inversion B).

The inversion algorithm was used to invert the petrophysical parameters (porosity, rock constituents, and fluid saturation) and the three zone parameters (cementation exponent, density of shale, and neutron of shale) (figure 77) .With two layers interpreted, the total unknowns comprise: the 2 interface depths (layer boundary coordinates), plus 5 petrophysical properties interpreted per layer (e.g. porosity, saturation), plus 3 additional zone parameters

estimated per layer. This totals 10 unknowns. Figure 78 shows the cluster results at iteration 1 for both scenarios. Both parameter scenarios were run for 10 iterations, respectively. The results show that actual logging measurements are consistent with predictions. In addition, interpreted formation characteristics and development of the stratigraphic grouping are presented in Figure 79. The main observation is that the unsupervised clusters start from initial configure rations between the two cases. This means that the algorithm automatically determines zone partitions without set level boundary restrictions-satisfying one of the basic assumptions of unsupervised learning. The constant parameter conditions rapidly converge to the cluster positions in 2-3 iterations. Meanwhile, in the case of variable parameters, there is a gradual refinement of interfaces spanning roughly 5 iterations as specific properties are calibrated per layer (Figure 80). Finally, allowing different properties for history matching enhances the alignment between mathematical models and synthetic true data, improving accuracy. Figure 81 shows the convergence of the cluster phase. In summary, the iterative development verifies the continuity of the unsupervised multi-parameter inversion. The difference in performance between the constant and extreme parameters indicates that the variable formulation effectively implements the method for generating geologically reasonable solutions consistent with the synthetic model, as demonstrated by the final location of the cluster centroids.

(a)

(b)
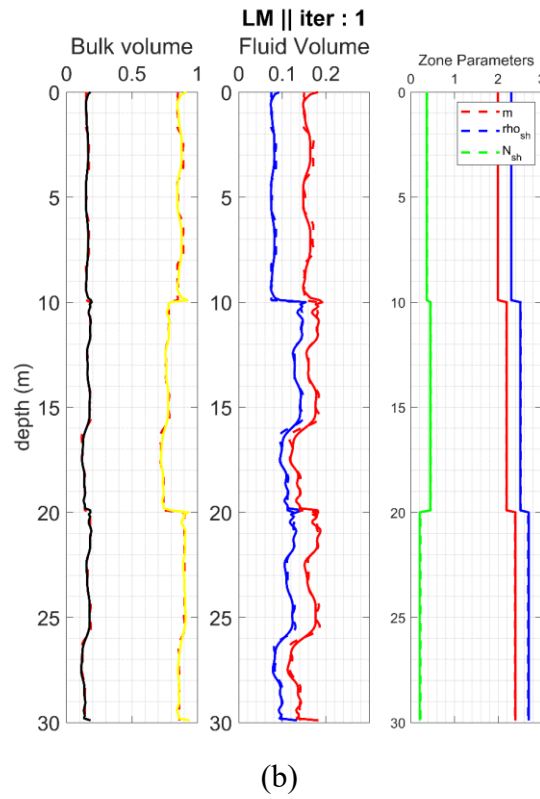
Figure 79. The exactly known (dashed lines) and predicted (solid lines) petrophysical parameters of the synthetic data at iteration 10. (a) inversion A (scenario 1), and (b) inversion B (scenario 2) .

(a)   (b)
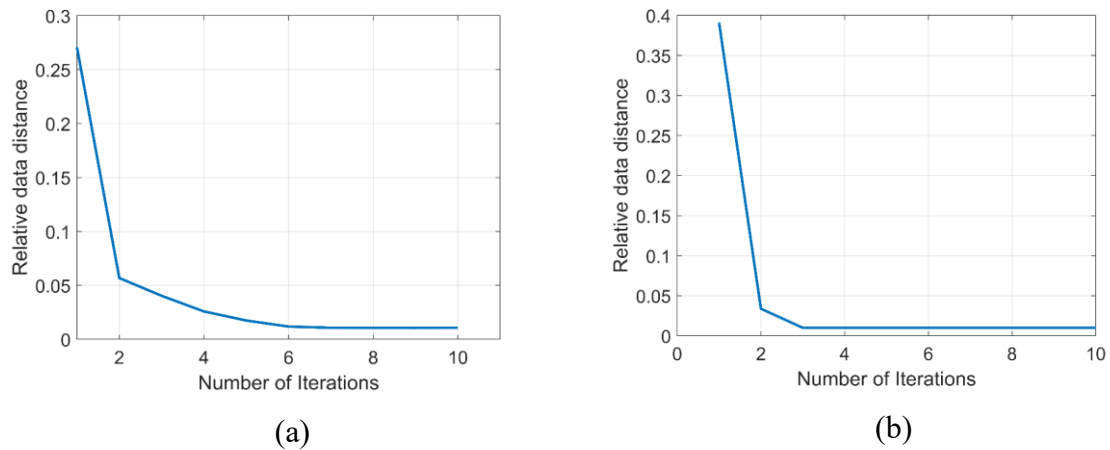
Figure 80. Data distance convergence.(a) inversion A (scenario 1), and (b) inversion B (scenario 2.
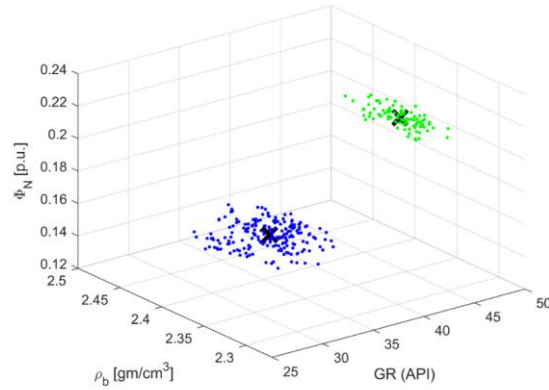
Figure 81. MFV-clustering results ("x" represents the cluster centroids).

6.2. Field Data Test

6.2.1. Hydrogeophysical field case study

For demonstrating the feasibility of the inversion method, I applied the algorithm on the hydrogeophysical data in chapter 3. In the case of field data, the cluster numbers were identified by using the elbow method. Figure (82) shows that the optimal cluster number is 3. Figure 83) shows the fitting between the calculated and actual well-logging data of the Baktalórántháza-1 well at iteration 1 at each defined zone according to the cluster results. The last track on the right side represents the initial results of the MFV-clustering phase. Figure 84 shows the initial guess of the petrophysical parameters within each zone. The algorithm successfully identified subzones in well-logging data, identifying the boundary between Pleistocene and Miocene sequences, and predicting petrophysical parameters, with sand dominating at 100-160 m depths and increasing shale content with depth figures (85 and 86).
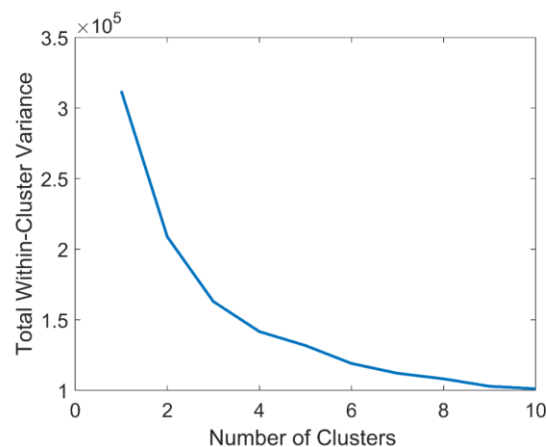
Figure 82. Elbow method for optimal number of clusters identification, the optimal number is 3.
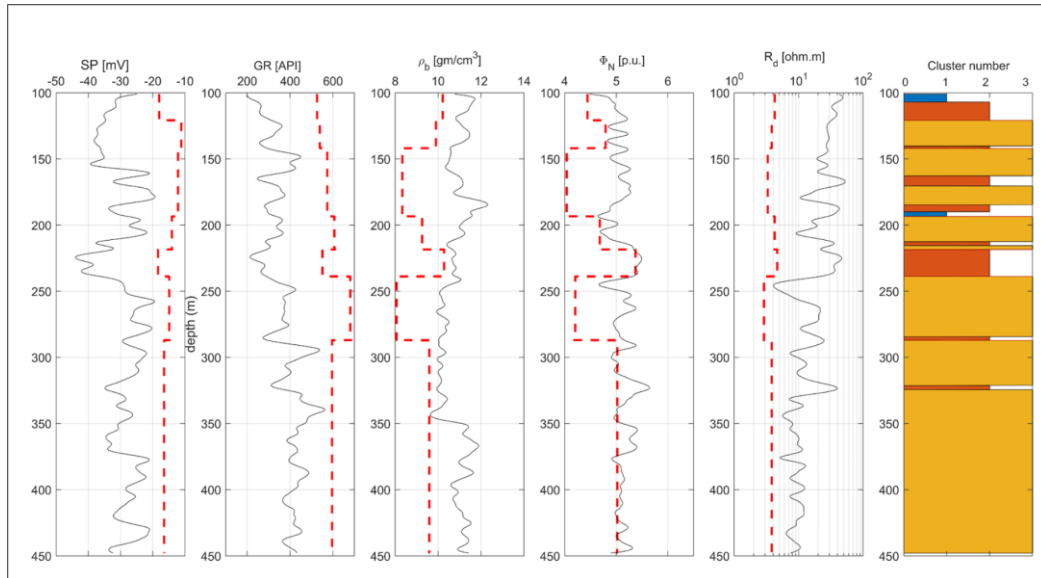
Figure 83. The fitting between the measured (solid black line) and predicted (red dashed line) wireline logging data (iteration 1).

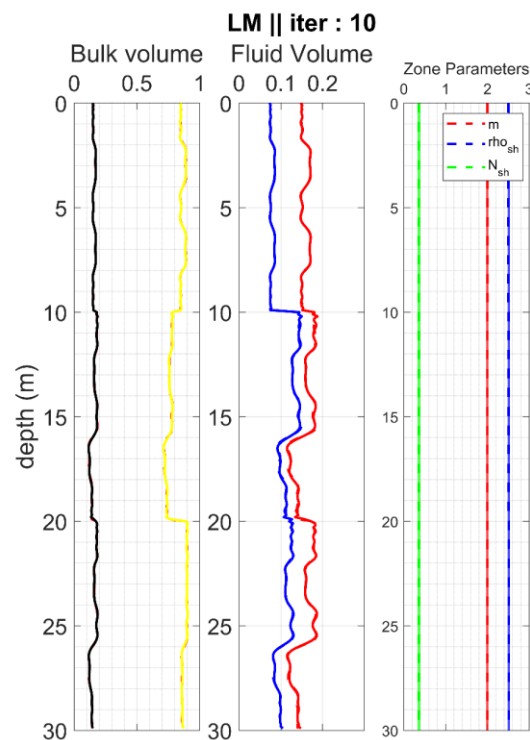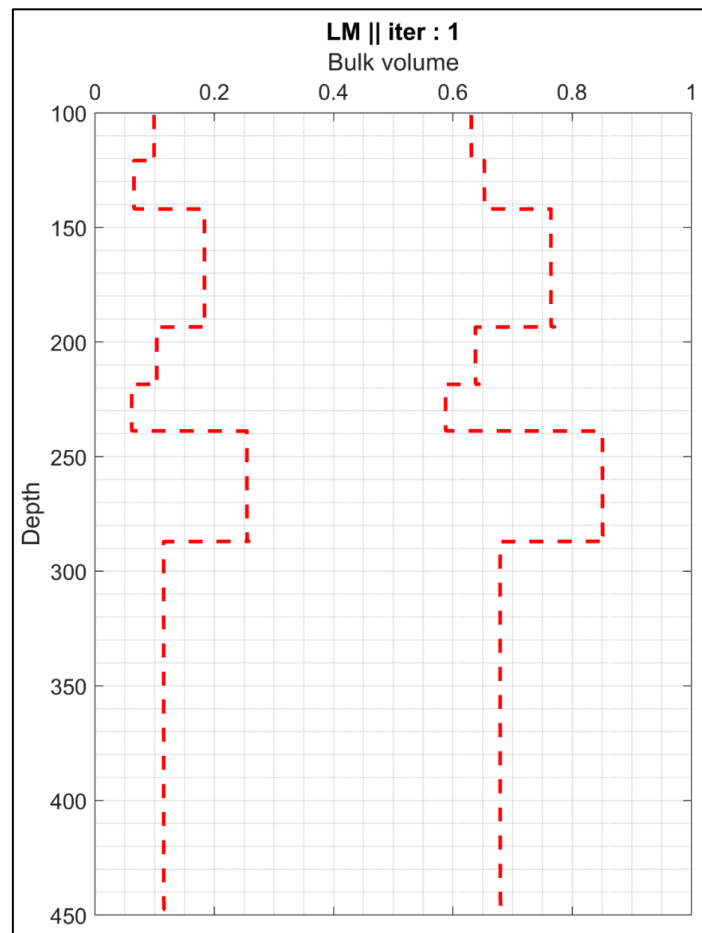**Figure 84.** The predicted (dashed lines) petrophysical parameters at iteration 1.
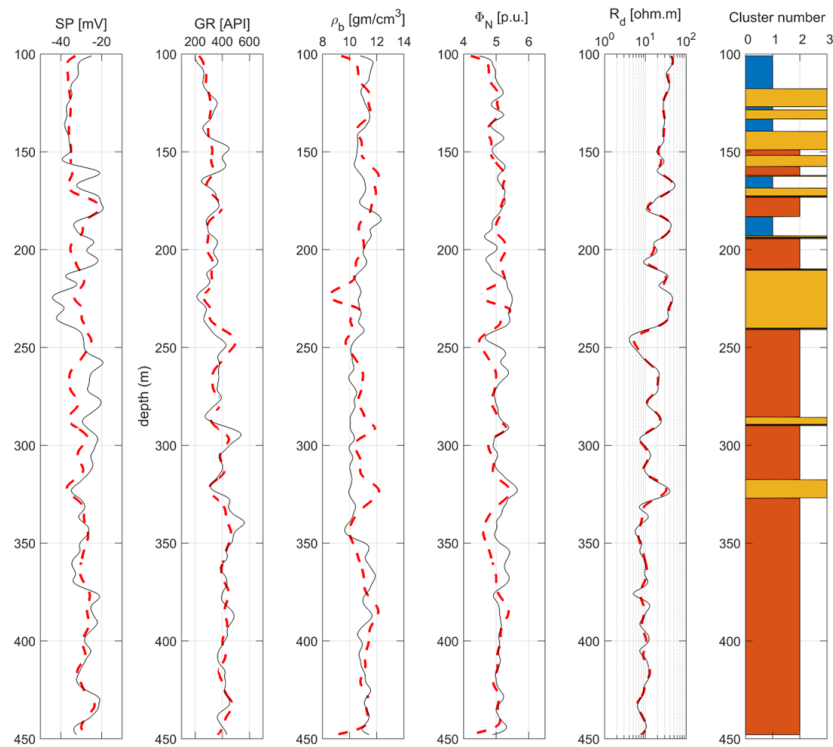
Figure 85. The fitting between measured (solid black line) and predicted (red dashed line) wireline logging data at the end of the inversion procedure.
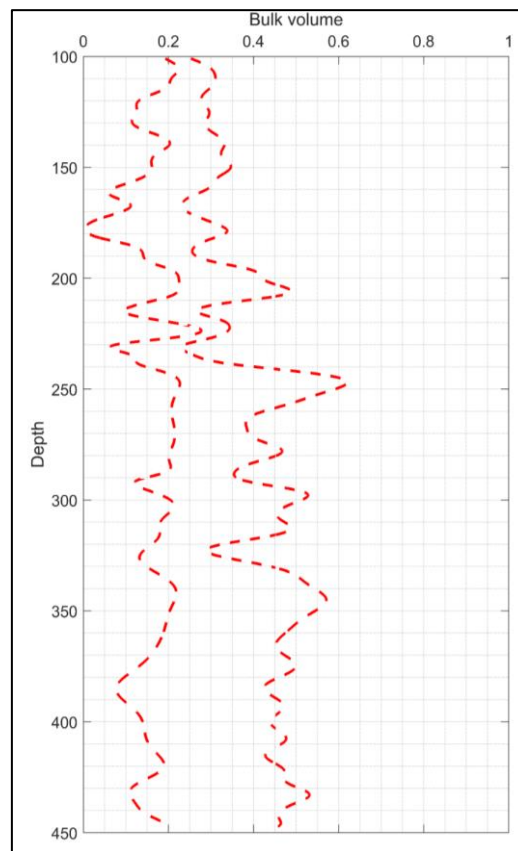
Figure 86. The estimated model parameter distributions at iteration 20.

6.2.2. Hydrocarbon field case study

A real well logging dataset from a tight sand gas deposit in northwest Egypt was used to evaluate the suggested procedure (the same well-logging dataset used in chapter 3). The wireline logging dataset documented a Jurassic sequence's parameter changes. To achieve the robust prediction of the predicted parameters, I studied the sedimentological features of the core samples that were taken from the Jurassic reservoir.

6.2.3. Sedimentological investigation

Core samples from sandstone layers reveal diverse sedimentary structures, reflecting the reservoir's heterogeneity (Figure 87). Sedimentological analysis reveals mudstone laminated, cross-bedded with drapes, and wavy-bedded patterns. SEM and petrography study reveal pore space cementation, with quartz overgrowth filling pore spaces, authigenic kaolinite filling inter-granular spaces, and fibrous illite covering quartz overgrowth. Illite suggests reduced reservoir permeability due to diagenesis (figure 88).

(a)

(b)

(c)

(d)
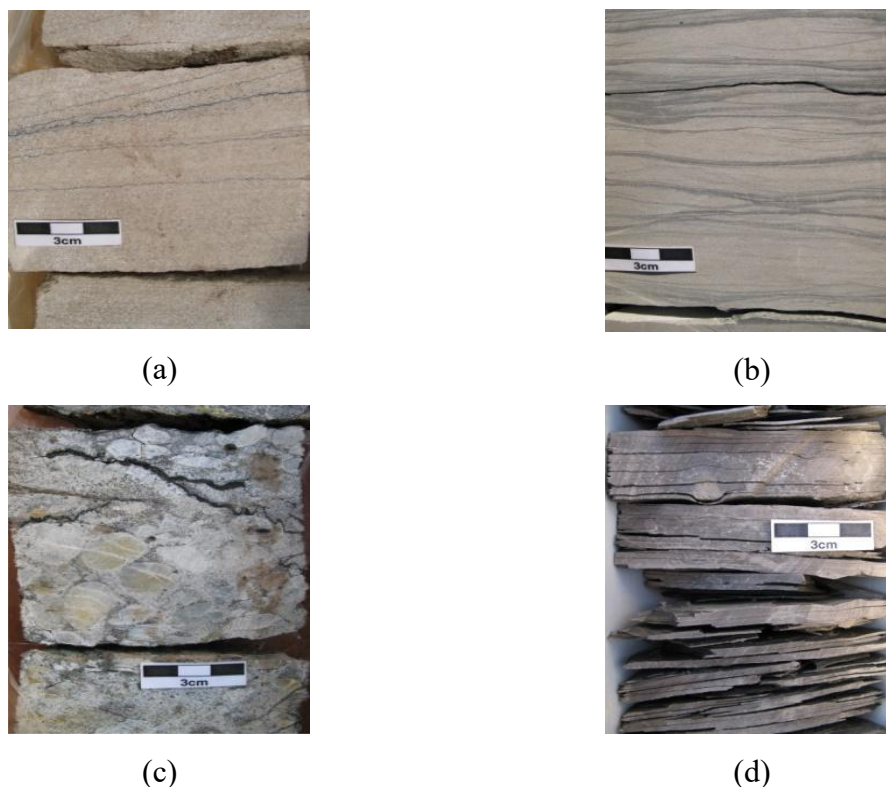
Figure 87. Core description of the collected samples. (a) S.S cross-bedded with drapes, (b) S.S wavy-bedded, (c) Clast supported conglomerate, and (d) Mudstone laminated.
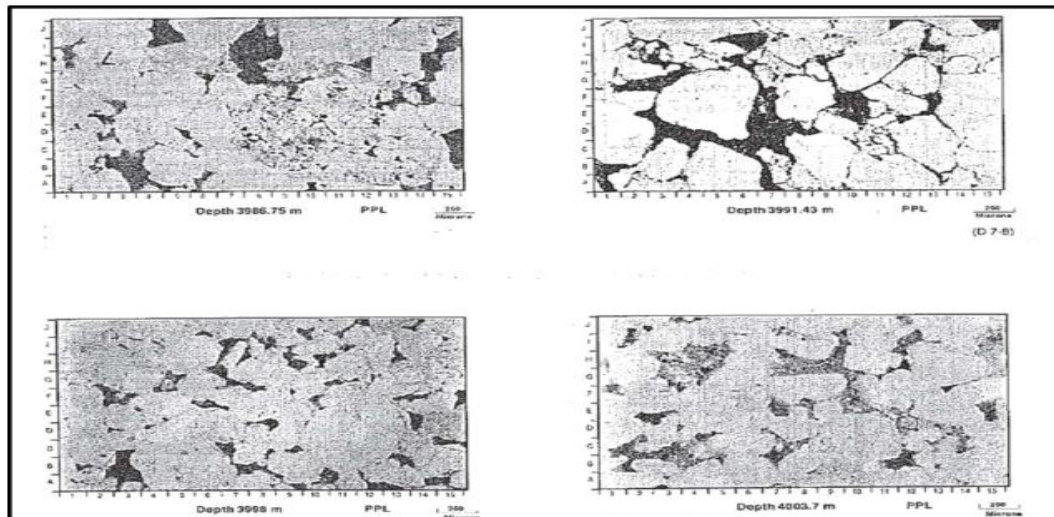
Figure 88. Authigenic kaolinite filling inter-granular pore spaces and QZ overgrowth with illite fibrous coating.

6.2.4. Field application of the automated algorithm

It was possible to estimate the differences between the data points using MFV-clustering. In addition, a boundary choice of two meters was made for converting the labels of the cluster findings into the bounds of the layers. In other words, count it at the same layer if the label change fell within a 2-meter range, and lay the boundary if the label change exceeded that range. The fitting between the calculated and measured data is displayed in Figure 89, where the predicted petrophysical parameters are displayed in figure 90. The first track displays the rescaled depth to represent the reservoir thickness from 0 to 37 m. The last track displays the labels that resulted from the cluster loop with the boundaries between different intervals. According to the suggested methodology, the reservoir is mostly made up of sandstone intervals with a little amount of shale sticks, which may have an impact on the reservoir's vertical homogeneity. The reservoir's initial five meters are a shale interval that decreased the fluid saturation and porosity. To enhance the fitting between the computed and measured datasets, the integration of MFV-clustering and interval inversion might isolate the shale layer and carry out a separate interval inversion about this shale layer.
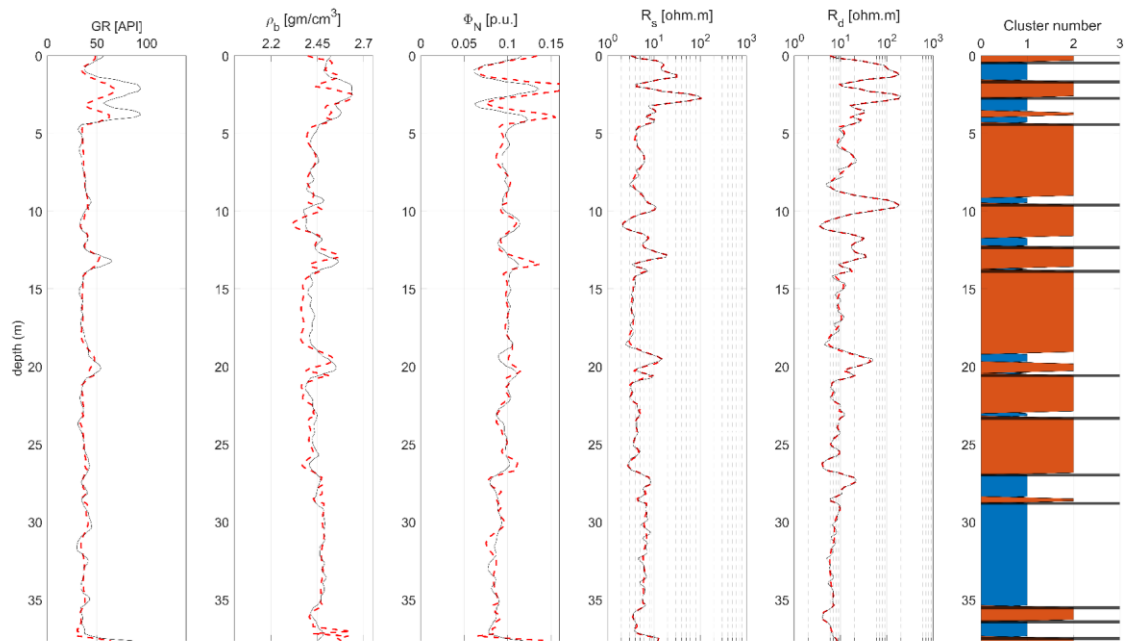
Figure 89. The fitting between field (solid black line) and predicted (red dashed line) well logging data.
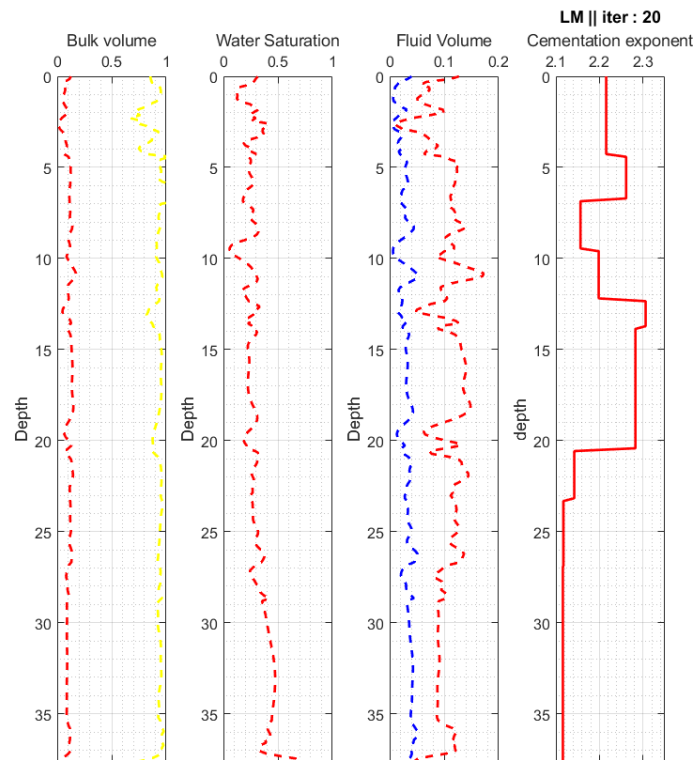
Figure 90. The estimated model parameter distributions using inversion A at iteration 20.

The data distance vs. iteration steps cross-plot was used to assess the stability of the fully automated inversion issue. The suggested interval is more stable in terms of the convergence process when compared to the cluster-defined based inversion, which used the whole recorded interval as a single layer. In addition, the data distance's stability of convergence is demonstrated by the MFV-cluster-assisted interval inversion at 7.5%, which is

less than the standard inversion's 10% convergence. The data distance convergence for both the traditional interval inversion and the MFV-cluster-assisted interval inversion is displayed in figure (91).
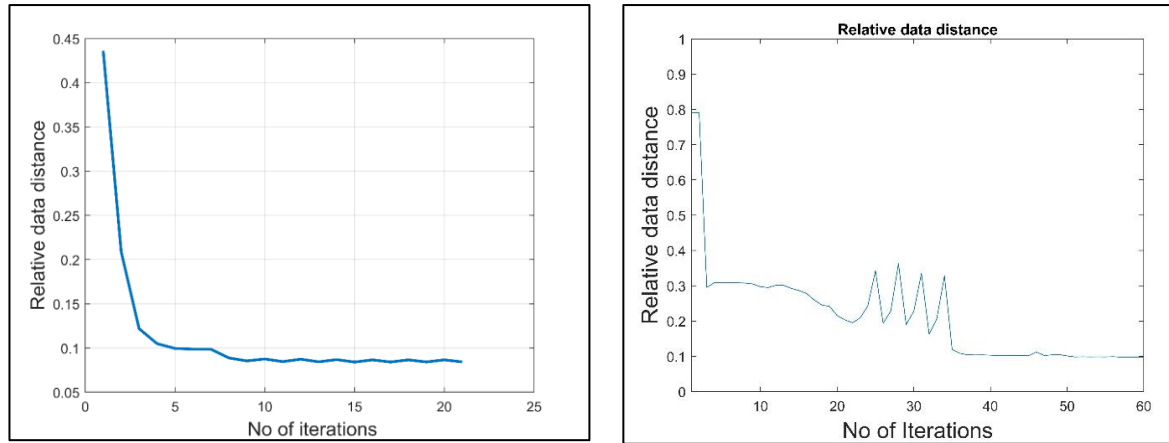
Figure 91. Data distance convergence plot. Cluster analysis-assisted interval inversion (left), and conventional interval inversion (right).

Figure (92) shows the error estimation for petrophysical parameters: ϕ (porosity), V_{sd} (shale volume), and S_w (water saturation) against depth. The same data accuracy values of chapter 3 were used for Quality check calculations of petrophysical. The plots show the prediction confidence deviation, as indicated by the red error lines around the estimated value of each parameter. Closely grouped error lines at lower depths for ϕ represent high confidence, and large deviations around the 20-meter depth are zones of greater uncertainty. Similar patterns for V_{sd} and S_w are seen with different confidence levels at different depths. The correlation matrix captures the interrelations among the parameters, where the diagonal entry is a perfect self-correlation. Lighter off-diagonal entries correspond to higher correlations or anti-correlations between different parameters, and these indicate large interdependencies that must be considered in modeling exercises. In figure 93, the introduction of the cementation exponent as an additional parameter alters the error limits and correlation behavior. Overall, the error limits appear tighter in some sections, indicating improved prediction quality with the new parameter. This added complexity, however, also introduces uncertainty into some depth ranges. The revised correlation matrix indicates variations in parameter interactions, reflecting the impact of the cementation exponent on such relationships. This adjustment highlights variation in model sensitivities with an emphasis on the intricate relationship between logging measurements and geological features. Generally, including this zone parameter enhances knowledge and introduces a comprehensive understanding of parameter interactions and uncertainties.
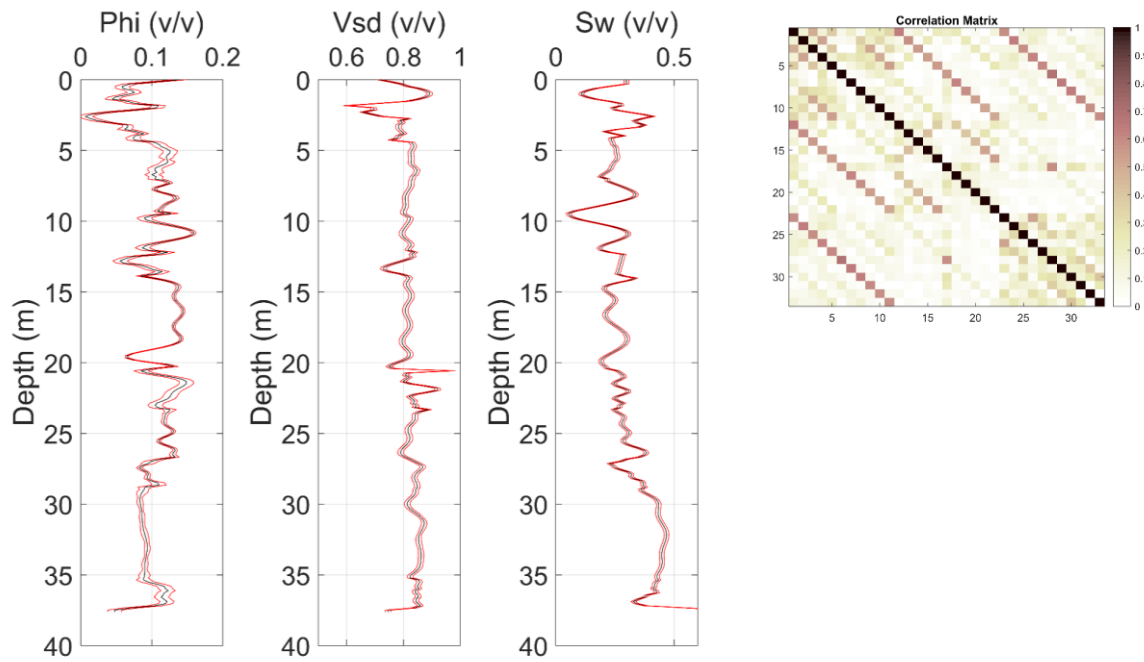
Figure 92. Quality check of petrophysical parameters is estimated by automatic interval inversion procedure, separately. Solid Black lines represent the estimated values of porosity Φ, volume of sand *Vsd*, and water saturation *Sw*. Solid red lines show the error bounds of petrophysical parameters calculated from the standard deviations σ of inversion estimates (left), the correlation matrix between the predicted parameters can be seen on the right side of the figure.
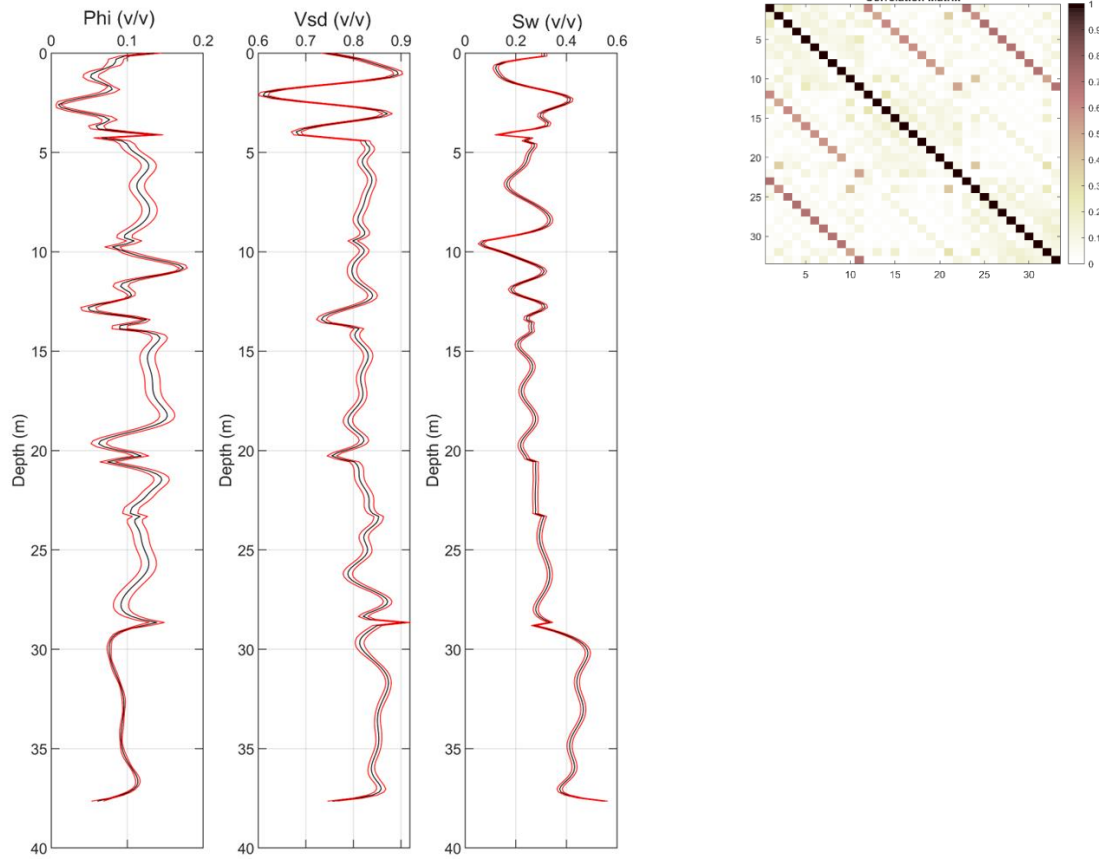
Figure 93. Quality check of petrophysical parameters is estimated by fully automatic interval inversion procedure, separately. Solid Black lines represent the estimated values of porosity Φ, volume of sand *Vsd*, and water saturation *Sw*. Solid red lines show the error bounds of petrophysical parameters calculated from the standard deviations σ of inversion estimates (left), the correlation matrix between the predicted parameters can be seen on the right side of the figure.

6.3. Summary of results

This chapter investigates the impact of data noise on lithology determination using synthetic borehole logging data from Egypt and Hungary. The study found that noise levels increased with increasing noise contamination, affecting the convergence of data distance. The algorithm performed accurately in identifying boundary coordinates up to 7% noise but struggled at greater noise percentages. A hydrogeophysical dataset from northeast Hungary was used for geothermal applications, successfully detecting subzones in well-logging data and identifying the boundary between Pleistocene and Miocene sequences. Surface geophysical measurements and well-logging operations did not reveal hydrocarbon indicators.

The chapter describes an automated machine learning-enabled interval inversion system for analyzing well logs data to improve petrophysical and zone parameter estimation. Parameter

sensitivity studies show that characteristics such as cementation exponent and tortuosity factor significantly impact outcomes across lithologies, including clean sands, tight sands, water sands, and shaly sands. The integrated machine learning clustering and interval inversion approach is illustrated using both simulated data sets and genuine well logs from an Egyptian tight sand gas reserve. The approach uses synthetic data to automatically define zone divisions across rock units and refine interfaces over numerous optimization rounds. Core sample analysis helps understand the intricacies of sedimentary structures and diagenetic changes that affect reservoir quality before inversion. The automated procedure separates a shale interval and several sandstone intervals, improving convergence stability compared to typical inversion procedures without machine learning. The proposed machine learning-powered interval inversion strategy enables efficient and consistent characterization of subsurface geological formations from conventional well logging data with minimal user intervention.

Thesis 6.

The dissertation introduces a novel method in petrophysical parameter prediction using a strongly overdetermined interval inversion process where zone parameters such as cementation exponent, tortuosity, and shale properties are optimized. It automated layer boundary identification and was field tested with synthetic data at noise levels of 2%, 5%, and 7%. Applied to northeast Hungarian hydrogeophysics data, it was found to be able to identify six distinct subzones and to rightly establish geological boundaries. Validation on an Egyptian hydrocarbon dataset showed successful classification of rock interfaces and estimation of parameters, with bulk volumes around 0.5, water saturation around 0.5, and fluid volume with a peak of 0.2. The machine-learning-based inversion gave a reduced data misfit of 7.5%, with improved convergence stability significantly. Zone parameters, including varying cementation exponent values, were estimated correctly, providing robust reservoir characterization.

Conclusions

In the first chapter, I outlined the motivation underlying this study, emphasizing the transformative potential of integrating inversion techniques and machine learning applications in geophysical interpretation. It examines their strengths in pattern recognition and parameter estimation, highlighting their potential in enhancing the precision and efficiency of interpreting complex well-logging data.

In the second chapter, I introduce an innovative clustering technique aimed at improving lithological classification in geophysical data analysis. This method employs the Most Frequent Value (MFV) algorithm, which addresses the shortcomings of traditional clustering like susceptibility to noise and initial centroid dependency. The MFV approach uniquely incorporates Steiner distances for weighted centroid adjustments, leading to enhanced accuracy and consistency. Testing this technique on synthetic well-log data, even with added noise and outliers, demonstrated its ability to maintain cluster integrity and offer reliable results. When applied to actual field data, this method accurately classified aquifer quality, providing a high-resolution understanding of subterranean layers. The outcomes were corroborated by core sample analyses, signaling a notable advancement in geological data interpretation methodologies.

In the third chapter, I focus on developing advanced techniques for improving the estimation of reservoir properties using well-logging data. This involves overcoming limitations in traditional methods, such as uncertainty and resolution issues, by introducing sophisticated algorithms. Central to these techniques is the management of key parameters, particularly the damping factor, which plays a crucial role in ensuring the stability of the estimation process. The integration of Golden Section Search (GSS) with interval estimation methods demonstrates a more consistent and efficient approach, especially in handling noisy datasets. Additionally, the novel combination of Damped Least Squares (DLSQ) and Singular Value Decomposition (SVD) methods is presented, aimed at enhancing computational efficiency. This hybrid approach initially leverages SVD for optimal parameter setting before transitioning to DLSQ for fine-tuning as the solution nears completion. The efficacy of these methods is validated through application to data from a gas-bearing reservoir in Egypt, revealing insights into the variable quality of reservoir characteristics and supporting the need for tailored assessment techniques.

In chapter fourth, I introduced a novel methodology for extracting hidden patterns within geophysical data by integrating the Hurst exponent, a measure of fractal analysis, with interval inversion. This approach begins with factor analysis to reduce dimensionality and highlight influential lithology-sensitive factors from well-log data. The Hurst exponent analysis is then applied to identify layer boundaries based on a scale and depth-dependent measure, effectively capturing lithological changes. The developed algorithm seamlessly integrates these findings into the interval inversion process, enhancing the reliability of petrophysical parameter estimates. I demonstrated the algorithm's efficacy using synthetic data and real-world hydrogeophysical datasets from Hungary. The workflow also extends to calculating hydraulic conductivity through the Csókás method, providing high-resolution insights into aquifer quality.

Chapter five introduces an enhanced algorithm for estimating petrophysical, zone, and geometrical parameters from well log data. The method integrates the Most Frequent Value (MFV) clustering with local inversion, providing detailed insight into reservoirs' static and dynamic characteristics. The algorithm's adaptability is demonstrated by increasing the number of unknowns to include zone parameters like shale density and cementation exponent. Synthetic modeling demonstrates its ability to fit both petrophysical and zone parameters simultaneously, proving its effectiveness across diverse geological settings. Field data application from a Hungarian petroleum site confirms the algorithm's ability to predict petrophysical parameters under varying zone conditions. An innovative approach combines unsupervised data pattern recognition with interval analysis for precise boundary detection in subsurface geological data. This method addresses polynomial degree and thickness dependency, ensuring accurate representation across varying lithological changes. Synthetic tests show the algorithm's proficiency in delineating boundaries under diverse noise conditions up to 7% contamination. Application to real-world data from Hungarian geothermal explorations demonstrates the approach's practical utility, enhancing the reliability of subsurface analysis and resource evaluation.

Chapter six presents an innovative approach that combines unsupervised data pattern recognition with interval analysis to achieve precise boundary detection in subsurface geological data. This method addresses the challenge of polynomial degree and thickness dependency, providing a dual-layered understanding of subsurface characteristics. Synthetic tests show the algorithm's proficiency in delineating boundaries under diverse noise conditions up to 7% contamination. Application to real-world data from Hungarian geothermal

explorations demonstrates the approach's practical utility, enhancing the reliability of subsurface analysis and resource evaluation. A novel approach is introduced for interpreting well-logging data through the integration of robust clustering and interval inversion techniques, leveraging machine learning to improve the estimation of subsurface properties. This automated method identifies the number of sub-intervals and dynamically adjusts polynomial degrees to enhance the fit between calculated and observed data. A sensitivity study highlights the importance of parameters like cementation exponent and tortuosity, influencing the algorithm's application in synthetic models and field datasets from an Egyptian gas reservoir. The approach successfully delineates distinct geological intervals, enhancing the accuracy of petrophysical predictions and demonstrating improved convergence stability.

Acknowledgments

I am deeply grateful to my supervisor, **Prof. Szabó Norbert Péter**, for his exceptional guidance, insightful suggestions, and unwavering support during my research. His expertise and dedication have been invaluable, and I feel honored to have worked under his mentorship.

I also wish to extend my sincere appreciation to the Department of Geophysics at Miskolc University. The department has provided a nurturing and intellectually stimulating environment that has greatly enriched my academic experience.

A special note of thanks goes to **Prof. Mihály Dobróka** for his steadfast support and encouragement, both in scientific endeavors and beyond. His profound insights have been instrumental in advancing my research, and his kindness has been a source of personal motivation.

I am also grateful to my professors from Egypt, particularly **Prof. Abdel Moktader Elsayed, Prof. Attia Mahmoud Attia, and Prof. Abdel Khaleq El Werr** for their support throughout my academic journey. Special thanks go to **Prof. Wael Zakarya** for his guidance and expertise in teaching MATLAB, which has significantly contributed to the development of my analytical and computational skills.

On a personal level, I want to express my gratitude to my family, **my father, mother, and sisters** for their unwavering love and support throughout this journey. You have been my anchor and constant source of encouragement.

To **my wife**, you hold a special place in my heart. Your love, patience, and understanding have been the foundation that allowed me to pursue and achieve my goals. To **my children**, you are the light of my life and my daily inspiration. I hope this achievement serves as a reminder that dedication and effort lead to great rewards.

Thank you all for being my strength and support.

References

Abdelrahman, M. M.G. 2021. "Lithofacies Classification Using Bayes Theorem Method : Case Study Western Desert, Egypt." *Multidiszciplináris Tudományok* 11 (1): 76–89. https://doi.org/10.35925/j.multi.2021.1.8.

Abdelrahman, M. M.G., N. P. Szabó, and M. Dobróka. 2021. "Petrophysical Parameters Estimation Using Levenberg-Marquardt and Singular Value Decomposition Inversion Schemes." *Multidiszciplináris Tudományok* 11 (5): 24–38. https://doi.org/10.35925/j.multi.2021.5.3.

Abordán, Armand, and Norbert Péter Szabó. 2021. "Machine Learning Based Approach for the Interpretation of Engineering Geophysical Sounding Logs." *Acta Geodaetica et Geophysica* 56 (4): 681–96. https://doi.org/10.1007/s40328-021-00354-4.

Akbar, M. N.A., N. P. Szabó, and M. Dobróka. 2021. "An Automated and Robust Solution of K-Means Cluster Analysis Based on Most Frequent Value Approach." *82nd EAGE Conference and Exhibition 2021* 3 (January): 1897–1901. https://doi.org/10.3997/2214-4609.202011546.

Alberty, M., and K. Hashmy. 1984. "Application of ULTRA to Log Analysis." In *SPWLA 25th Annual Logging Symposium 1984*, 1–17.

Alger, R. P. 1971. *Interpretation of Electric Logs in Fresh Water Wells in Unconsolidated Formation*. SPE Reprint Series 1.

Aliyuda, Kachalla, Aliyuda Ali, Bello Abdulwahab, Muhammed, and Jerry Raymond. 2022. "Machine Learning Algorithms for Predicting Reservoir Porosity Using Stratigraphic-Dependent Parameters," no. October. https://www.researchgate.net/publication/374505791.

Amaefule, J. O., M. Altunbay, D. Tiab, D. G. Kersey, and D. K. Keelan. 1993. "Enhanced Reservoir Description: Using Core and Log Data to Identify Hydraulic (Flow) Units and Predict Permeability in Uncored Intervals/ Wells." *Proceedings - SPE Annual Technical Conference and Exhibition* Omega (c): 205–20. https://doi.org/10.2523/26436-ms.

Archie, G. E. 1952. "Classification of Carbonate Reservoir Rocks and Petrophysical Considerations." *AAPG Bulletin* 36 (2): 278–98. https://doi.org/10.1306/3d9343f7-16b1-11d7-8645000102c1865d.

Arneson, K., and G. P. Hersir. 1988. "One Dimensional Inversion of Schlumberger Resistivity Soundings: Computer Program, Description and User's Guide:"

Aversana, Paolo Dell. 2023. "Multi-Physics Inversion and Machine Learning . Theory and Applications in Geophysics ML-Int RL-Pre RL-Inv," no. February.

Bartlett, M. S. 1937. "The Statistical Conception of Mental Factors." *British Journal of Psychology. General Section* 28 (1): 97–104. https://doi.org/10.1111/j.2044-8295.1937.tb00863.x.

Bradley, P.S., and U.M. Fayyad. 1998. "Refining Initial Points for K-Means Clustering." *Proceedings of the Fifteenth International Conference on Machine Learning,Madison,WI,* no. 24–27: 91–99.

Chang, Kuang-Hua. 2015. "Chapter 3 - Design Optimization." In *Design Theory and Methods Using CAD/CAE*, edited by Kuang-Hua Chang, 103–210. Boston: Academic Press. https://doi.org/https://doi.org/10.1016/B978-0-12-398512-5.00003-7.

Csókás, J. 1995. "Determination of Yield and Water Quality of Aquifers Based on Geophysical Well Logs (in Hungarian)." *Magyar Geofizika* 35 (4): 176–203.

Davim, J. Paulo. 2008. *Machining: Fundamentals and Recent Advances*. *Machining: Fundamentals and Recent Advances*. https://doi.org/10.1007/978-1-84800-213-5.

Dennis, C B, and Tony D Lawrence. 1984. "Log Evaluation Of Clastic Shaly Formations Using Corrected Rwa-Ratio Techniques ."

Dobróka, M., A Gyulai, T. Ormos, J. Csókas, and L. Dresen. 1991. "Joint Inversion of Seismic and Geoelectric Data Recorded in an Underground Coal Mine." *Geophysical Prospecting* 39 (5): 643–65. https://doi.org/10.1111/j.1365-2478.1991.tb00334.x.

Dobróka, M., and N. P. Szabó. 2015. "Well Log Analysis by Global Optimization-Based Interval Inversion Method." In *Artificial Intelligent Approaches in Petroleum Geosciences*, edited by Constantin Cranganu, Henri Luchian, and Mihaela Elena Breaban, 245–68. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-16531-8_9.

Dobróka, M., N. P. Szabó, E. Cardarelli, and P. Vass. 2009. "2D Inversion of Borehole Logging Data for Simultaneous Determination of Rock Interfaces and Petrophysical Parameters." *Acta Geodaetica et Geophysica Hungarica* 44 (4): 459–79. https://doi.org/10.1556/AGeod.44.2009.4.7.

Dobróka, M., N. P. Szabó, J. Tóth, and P. Vass. 2016. "Interval Inversion Approach for an Improved Interpretation of Well Logs." *Geophysics* 81 (2): D155–67. https://doi.org/10.1190/GEO2015-0422.1.

Dobróka, M. 1993. "Scientific Report for the Hungarian Oil and Gas Company."

Doveton, J. 2001. "All Models Are Wrong, but Some Models Are Useful: 'Solving' the Simandoux Equation." In .

Ekinci, Yunus Levent, and Alper Demirci. 2008. "A Damped Least-Squares Inversion Program for the Interpretation of Schlumberger Sounding Curves." *Journal of Applied Sciences* 8 (22): 4070–78. https://doi.org/10.3923/jas.2008.4070.4078.

Gavin, Henri P. 2019. "The Levenberg-Marquardt Algorithm For Nonlinear Least Squares Curve-Fitting Problems." *Duke University*, 1–19. http://people.duke.edu/~hpgavin/ce281/lm.pdf.

Heriyanto, M., and W. Srigutomo. 2017. "1-D DC Resistivity Inversion Using Singular Value

Decomposition and Levenberg-Marquardt's Inversion Schemes." In *Journal of Physics: Conference Series*. Vol. 877. Institute of Physics Publishing. https://doi.org/10.1088/1742-6596/877/1/012066.

Hernandez-Martinez, E., T. Perez-Muñoz, J. X. Velasco-Hernandez, A. Altamira-Areyan, and L. Velasquillo-Martinez. 2013. "Facies Recognition Using Multifractal Hurst Analysis: Applications to Well-Log Data." *Mathematical Geosciences* 45 (4): 471–86. https://doi.org/10.1007/s11004-013-9445-6.

Hurst, H. 1951. "Long-Term Storage Capacity of Reservoirs." *Transactions of the American Society of Civil Engineers* 116: 770–99.

Ian, H. Witten, Eibe Frank, and Mark A. Hall. 2009. *Data Mining Practical Machine Learning Tools and Techniques. International Encyclopedia of Education, Third Edition*. https://doi.org/10.1016/B978-0-08-044894-7.01318-X.

Ikhwan, Muhammad, Tofan R M Sastranegara, and Jayanti Anggraini. 2024. "The Use Of Machine Learning in Geothermal Reservoir Rock Typing : Indonesia Case Study," no. July: 16–18.

Jahantigh, Moslem, and Hamidreza Ramazi. 2025. "Integration of Airborne Geophysics Data with Fuzzy C-Means Unsupervised Machine Learning Method to Predict Geological Map, Shahr-e-Babak Study Area, Southern Iran." *Journal of Mining and Environment* 16 (1): 273–89. https://doi.org/10.22044/jme.2024.14173.2636.

Jöreskog, K. G. 2007. "Simultaneous Factor Analysis in Several Populations." In .

Kaiser, Henry F. 1958. "The Varimax Criterion for Analytic Rotation in Factor Analysis." *Psychometrika* 23 (3): 187–200. https://doi.org/10.1007/BF02289233.

Kemp, A., and F. Steiner. 1991. *The Most Frequent Value. Intro- Duction to a Modern Conception Statistics*. Budapest: Akademia Kiado.

Khan, Shehroz S., and Amir Ahmad. 2004. "Cluster Center Initialization Algorithm for K-Means Clustering." *Pattern Recognition Letters* 25 (11): 1293–1302. https://doi.org/10.1016/j.patrec.2004.04.007.

Levenberg, K. 1944. "A Method for the Solution of Certain Non-Linear Problems in Least Squares." *Quarterly of Applied Mathematics* 1 (278): 536–38.

Li, Tianyang, Ruiheng Li, Nian Yu, Zizhen Wang, and Ruihe Wang. 2021. "A Novel Approach Based On Feature Fusion For Fracture Identification Using Well Log Data." *Fractals* 29 (8). https://doi.org/10.1142/S0218348X2150256X.

Lima, L., N. Bize-Forest, A. Evsukoff, and R. Leonhardt. 2020. "Unsupervised Deep Learning for Facies Pattern Recognition on Borehole Images." *Offshore Technology Conference Brasil 2019, OTCB 2019*. https://doi.org/10.4043/29726-ms.

López, M., and M. Aldana. 2007. "Facies Recognition Using Wavelet Based Fractal Analysis and Waveform Classifier at the Oritupano-A Field, Venezuela." *Nonlinear Processes in*

Geophysics 14 (4): 325–35. https://doi.org/10.5194/npg-14-325-2007.

Lozada-Zumaeta, M., R. D. Arizabalo, G. Ronquillo-Jarillo, E. Coconi-Morales, D. Rivera-Recillas, and F. Castrejón-Vácio. 2012. "Distribution of Petrophysical Properties for Sandy-Clayey Reservoirs by Fractal Interpolation." *Nonlinear Processes in Geophysics* 19 (2): 239–50. https://doi.org/10.5194/npg-19-239-2012.

Mandelbr, B. b, and J R Wallis. 1969. "Computer Experiments With Fractional Gaussian Noises .1. Averages and Variances." *Water Resources Research* 5 (1): 228-.

Meju, M. A. 1992. "An Effective Ridge Regression Procedure for Resistivity Data Inversion." *Computers and Geosciences* 18 (2–3): 99–118. https://doi.org/10.1016/0098-3004(92)90079-7.

Menke, W. 1984. *Geophysical Data Analysis: Discrete Inverse Theory. Geophysical Data Analysis: Discrete Inverse Theory*. Academic Press Inc. https://doi.org/10.1016/0040-1951(86)90212-x.

Menke, W., and Z. Eilon. 2015. "Relationship Between Data Smoothing and the Regularization of Inverse Problems." *Pure and Applied Geophysics* 172 (10): 2711–26. https://doi.org/10.1007/s00024-015-1059-0.

Milton, S. R., and J. Okamoto Jr. 2018. "Application of Hurst Exponent (H) and the R/S Analysis in the Classification of FOREX Securities." *International Journal of Modeling and Optimization* 8 (2): 116–24. https://doi.org/10.7763/ijmo.2018.v8.635.

Mohammed, Musaab A A, Norbert P Szabó, and Péter Szűcs. 2025. "High - Resolution Characterization of Complex Groundwater Systems Using Wireline Logs Analyzed with Machine Learning Classifiers and Isometric Mapping Techniques." *Modeling Earth Systems and Environment* 1: 1–18. https://doi.org/10.1007/s40808-024-02263-1.

Neasham, John W. 1977. "The Morphology of Dispersed Clay in Sandstone Reservoirs and Its Effect on Sandstone Shaliness, Pore Space and Fluid Flow Properties." *Proceedings - SPE Annual Technical Conference and Exhibition* 1977-Octob. https://doi.org/10.2118/6858-ms.

Nhu, Vu Huu. 2022. "Levenberg–Marquardt Method for Ill-Posed Inverse Problems with Possibly Non-Smooth Forward Mappings between Banach Spaces." *Inverse Problems* 38 (1): 15007. https://doi.org/10.1088/1361-6420/ac38b7.

Niwas, Sri, and Muhammed Celik. 2012. "Equation Estimation of Porosity and Hydraulic Conductivity of Ruhrtal Aquifer in Germany Using near Surface Geophysics." *Journal of Applied Geophysics* 84: 77–85. https://doi.org/10.1016/j.jappgeo.2012.06.001.

Noroozi, Mohammad, Hamed Mohammadi, Emad Efatinasab, Ali Lashgari, Mahdiyeh Eslami, and Baseem Khan. 2022. "Golden Search Optimization Algorithm." *IEEE Access* 10 (April): 37515–32. https://doi.org/10.1109/ACCESS.2022.3162853.

Okunola, Abiodun. 2024. "The Role of Ensemble Learning in Improving Efficiency of Machine Learning Algorithms," no. December.

Ouadfeul, Sid-Ali, Leila Aliouane, and Amar Boudella. 2012. "Fractal and Chaos in Exploration Geophysics." In , edited by Sid-Ali Ouadfeul, Ch. 8. Rijeka: IntechOpen. https://doi.org/10.5772/53560.

Pandey, Yogendra Narayan, Ayush Rastogi, Sribharath Kainkaryam, Srimoyee Bhattacharya, and Luigi Saputelli. 2020. *Machine Learning in the Oil and Gas Industry. Machine Learning in the Oil and Gas Industry*. https://doi.org/10.1007/978-1-4842-6094-4.

Pejic, Dragana, and Milos Arsic. 2019. *Minimization and Maximization of Functions: Golden-Section Search in One Dimension*. Springer International Publishing. https://doi.org/10.1007/978-3-030-13803-5_3.

Prabhu. 2018. "Understanding Hyperparameters and Its Optimisation Techniques." 2018.

Qu, Bowen. 2025. "Bayes ' Theorem in Machine Learning : A Literature Review" 0: 26–31. https://doi.org/10.54254/2753-8818/86/2025.20329.

Rijn, Jan N. Van, and Frank Hutter. 2018. "Hyperparameter Importance across Datasets." In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2367–76. Association for Computing Machinery. https://doi.org/10.1145/3219819.3220058.

Steiner, F. 1990. "The Fundamentals of Geostatistics (in Hungarian)." *Tankönyvkiadó,Budapest, Hungary*.

Szabó, N. P., B. A. Braun, M. M.G. Abdelrahman, and M Dobróka. 2021. "Improved Well Logs Clustering Algorithm for Shale Gas Identification and Formation Evaluation." *Acta Geodaetica et Geophysica* 56 (4): 711–29. https://doi.org/10.1007/s40328-021-00358-0.

Szabó, N. P., and M. Dobróka. 2013. "Extending the Application of a Shale Volume Estimation Formula Derived from Factor Analysis of Wireline Logging Data." *Mathematical Geosciences* 45 (7): 837–50. https://doi.org/10.1007/s11004-013-9449-2.

Szabó, N. P., and M. Dobróka. 2017. "Robust Estimation of Reservoir Shaliness by Iteratively Reweighted Factor Analysis." *Geophysics* 82 (2): D69–83. https://doi.org/10.1190/GEO2016-0393.1.

Szabó, N. P., and M. Dobróka. 2018. "Exploratory Factor Analysis of Wireline Logs Using a Float-Encoded Genetic Algorithm." *Mathematical Geosciences* 50 (3): 317–35. https://doi.org/10.1007/s11004-017-9714-x.

Szabó, N. P., and M. Dobróka. 2020. "Interval Inversion as Innovative Well Log Interpretation Tool for Evaluating Organic-Rich Shale Formations." *Journal of Petroleum Science and Engineering* 186 (March). https://doi.org/10.1016/j.petrol.2019.106696.

Szabó, N. P., M. Dobróka, E. Turai, and P. Szűcs. 2014. "Factor Analysis of Borehole Logs for Evaluating Formation Shaliness: A Hydrogeophysical Application for Groundwater Studies." *Hydrogeology Journal* 22 (3): 511–26. https://doi.org/10.1007/s10040-013-1067-z.

Szűcs, P., N. P. Szabó, M. Zubair, and S. Szalai. 2021. "Innovative Hydrogeophysical Approaches as Aids to Assess Hungarian Groundwater Bodies." *Applied Sciences* 11 (5): 2099. https://doi.org/10.3390/app11052099.

Terry, Neil, Frederick D. Day-Lewis, John W. Lane, Jared J. Trost, and Barbara A. Bekins. 2019. "Geophysical Mapping of Plume Discharge to Surface Water at a Crude Oil Spill Site: Inversion versus Machine Learning." *Geophysics* 84 (5): EN67–80. https://doi.org/10.1190/geo2018-0690.1.

Tiab, D., and E.C. Donaldson. 2004. *Theory and Practice of Measuring Reservoir Rock and Fluid Transport Properties*. Edited by Gulf Professional Publishing. 2nd ed. Oxford.

Tolner, Ferenc, Sándor Fegyverneki, Balázs Barta, and György Eigner. 2023. "Robust Clustering Based on the Most Frequent Value Method." *Multidiszciplináris Tudományok* 13 (1): 141–53. https://doi.org/10.35925/j.multi.2023.1.11.

Wahdan, Maamoun, Hassan Abdel Rahman, and Taher El Zefzaf. 2013. "Petrographic and Diagenetic History of Creataceous/Paleozoic Reservoirs and Their Impact on The Permeability Distribution Obaiyed Field, Western Desert, Egypt." *North Africa Technical Conference and Exhibition*. https://doi.org/10.2118/164735-MS.

Yazıcı, İrfan, Ersagun Kürşat Yaylacı, and Faruk Yalçın. 2021. "Modified Golden Section Search Based MPPT Algorithm for the WECS." *Engineering Science and Technology, an International Journal* 24 (5): 1123–33. https://doi.org/10.1016/j.jestch.2021.02.006.

Zhang, Yan, Li Zhang, Zhenyu Lei, Fan Xiao, Yongzhang Zhou, Jing Zhao, and Xing Qian. 2024. "Unsupervised Machine Learning-Based Singularity Models : A Case Study of the Taiwan Strait Basin."