

Unsupervised machine learning assisted hydrogeophysical borehole logging inversion for robust aquifer characterization

M. Abdelrahman¹, N. Szabo¹

¹ University of Miskolc; ² Ain Shams University

Summary

This research proposes an integrated algorithm that uses an unsupervised machine learning technique, specifically the new K-mean clustering, for automatic aquifer characterization using hydrogeophysical borehole logging data. The MFV-cluster algorithm was employed to determine layer boundaries and petrophysical parameters automatically. The viability of the suggested process was evaluated using synthetic and field data, and it was found to be effective in distinguishing between various forms and providing a preliminary estimate for layer thicknesses. The integration between the new cluster technique and interval inversion can help with the automatic detection of both the geometrical and petrophysical parameters. The field data used in the study showed a shaly sand pattern response. The MFV clustering technique was applied to this field data and was able to distinguish between various forms and provide a preliminary estimate for layer thicknesses. The results of the statistical evaluation of synthetic data contaminated by 30 percent of outliers prove a high dependency on the initial location of the centroid. The interval inversion approach enhances the number of inverted data points by representing the petrophysical parameters as a continuous function.

Unsupervised machine learning assisted hydrogeophysical borehole logging inversion for robust aquifer characterization

Introduction

Machine learning and artificial intelligence have become increasingly popular in various industries due to their ability to analyze vast amounts of data and make accurate predictions. With the ever-growing amount of data available, machine learning and artificial intelligence are expected to continue to play a crucial role in shaping our world. The underground characteristics of an aquifer or groundwater system are monitored and characterized using hydrogeophysical wells. These wells usually combine hydrological and geophysical methods to gather information on the flow rate and conductivity of groundwater. This research proposed a new integrated algorithm for automatic aquifer characterization using hydrogeophysical borehole logging data. An unsupervised machine learning technique using new K-mean clustering was employed to automatically determine layers' boundaries and petrophysical parameters. The MFV-cluster algorithm demonstrated better grouping performance than the Euclidean norm. The viability of the suggested process was evaluated using synthetic and field data, and it was found to be effective in distinguishing between various forms and providing a preliminary estimate for layer thicknesses. The MFV clustering technique can be integrated with traditional or interval inversion techniques to reduce overdetermination ratios and directly use suitable zone parameters for each layer. The process of borehole inversion to estimate petrophysical parameters is time-intensive and requires significant manual input and prior information. The interval inversion was introduced to enhance the overdetermination ratio, but the choice of basis functions depends on geological complexity. The integration between the new cluster technique and the interval inversion can help with the automatic detection of both the geometrical and petrophysical parameters. In addition, this research study throws light on the availability of extending this concept to the full prediction of the zone parameters of the aquifer. The field data used in the study showed a shaly sand pattern response. The well's top 80–100 m was drilled into Pleistocene layers primarily composed of sandstones. Shale was found to be the underlying rock of the sandy sequence, according to borehole records. After the 100–160 m thick sands, a shaly formation and 5–15 m thick coarse-grained layers had been formed. The boundary between the Pleistocene and Pannonian eras was found at a depth of 240 meters. Most of the Pannonian shaly complex was composed of bituminous, clayey sand, gravel, clayey sediment, clayey marl, and clay. The MFV clustering technique was applied to this field data and was able to distinguish between various forms and provide a preliminary estimate for layer thicknesses.

Method and/or Theory

1. Non-hierarchical cluster analysis of well logs

Cluster analysis is a statistical technique that helps to group objects in a dataset based on how similar they are to each other using a chosen distance metric. The primary objective is to create groups that are distinct from each other and contain similar objects. This process aims to achieve a high level of homogeneity within a group and a significant difference between groups. To state the statistical problem, the two multivariate objects from a population of N random variables $X^1, \dots, X^N$ are represented by vectors $x^{(i)}$ and $x^{(j)}$. The Mahalanobis distance can imply the use of the weighted distance

$$D_w = \sqrt{\frac{1}{\sum_{q=1}^N w_q} \sum_{k=1}^N w_k (x_k^{(i)} - x_k^{(j)})^2} \quad (1)$$

with the weight ordered to the k -th datum. Steiner (1991) assigned equal weight to the values within the framework of the Most Frequent Values method (MFV)

$$w_k^{St} = \frac{\varepsilon^2}{\varepsilon^2 + e_k^2} \quad (2)$$

where the ε^2 dihesion is derived automatically by an inner iterative procedure from the dataset. If the noise in the data follows a Gaussian distribution, using the arithmetic mean of the coordinates of the members of the cluster to define the centroid can yield satisfactory results. However, in cases where the

data noise deviates from the Gaussian distribution, a more robust centroid definition can be achieved by implementing Steiner weights

$$c_i^{St} = \frac{1}{\sum_{q=1}^{M_i} w_q^{St}} \sum_{p=1}^{M_i} w_p^{St} x_p^{(i)} \quad (3)$$

The results of the statistical evaluation of synthetic data contaminated by 30 percent of outliers prove a high dependency on the initial location of the centroid. Fig. 1 shows the comparison between both clustering methods in stability.

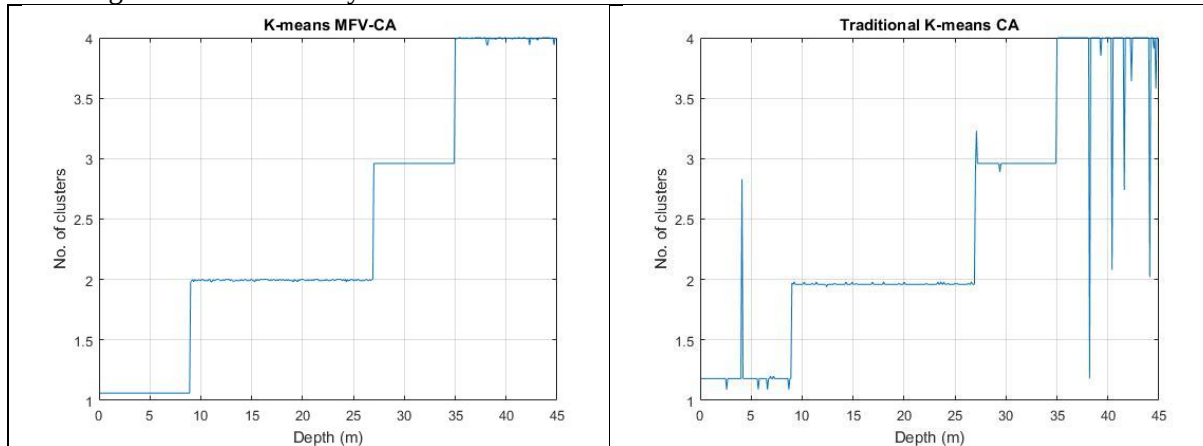


Figure 1 The average value obtained from performing both clustering methods 100 times.

2. Petrophysical parameter prediction using interval inversion

The inversion procedure is based on the existence of a relationship, known as the response function, between the estimated parameters and the data. This relationship is fundamental to the concept of inversion and can be utilized to compute the well-logging data from the petrophysical model. Szabó and Dobróka (2020) invented a so-called series expansion-based inversion or interval inversion. The goal of the interval inversion approach is to enhance the number of inverted data by representing the petrophysical parameters in a specific depth range as a continuous function. To achieve this, the interval inversion problem utilizes a discretized series expansion of petrophysical parameters. The data column vector is then represented as depth-dependent calculated data.

$$d_j^{(c)}(z) = g_j(m_1(z), m_2(z), \dots, m_M(z)) \quad (4)$$

The discretization of the petrophysical parameters can be represented by the formula where M refers to the number of predicted parameters at a certain depth z . The discretized model parameters can be represented as follows:

$$m_i(z) = \sum_{q=1}^{Q_i} B_q^{(i)} \Psi_q(z) \quad (5)$$

where m_i represents the i -th model parameter, $B_q^{(i)}$ is the q -th expansion coefficient, and $\Psi_q(z)$ is the basis function. The choice of type and degree of the basis function depends on the degree of heterogeneity and the change of the parameters within the interval that is extracted automatically from the MFV clustering. In the frame of the series expansion concept, the model vector can be represented as follows (Dobróka et al. 2016):

$$\mathbf{m} = [B_1^{(1)}, \dots, B_Q^{(1)}, \dots, B_1^{(M)}, \dots, B_Q^{(M)}]^T \quad (6)$$

The integration between the robust clustering technique with the interval inversion algorithm was tested on both synthetic and hydrogeophysical field data. Fig. 2 shows the predicted layers' boundaries and fitting between calculated data and synthetic data, and the convergence of the MFV-based clustering. The constructed synthetic model consists of three layers. The model comprised two layers of sandstone, with a layer of clay sandwiched in between them. The following figures show the results of the fully automatic interpretation. Fig. 3 shows the initial model parameters (left) regarding the first guess of the layers' boundaries and the final model parameters (right) regarding the exact layers' boundaries identification.

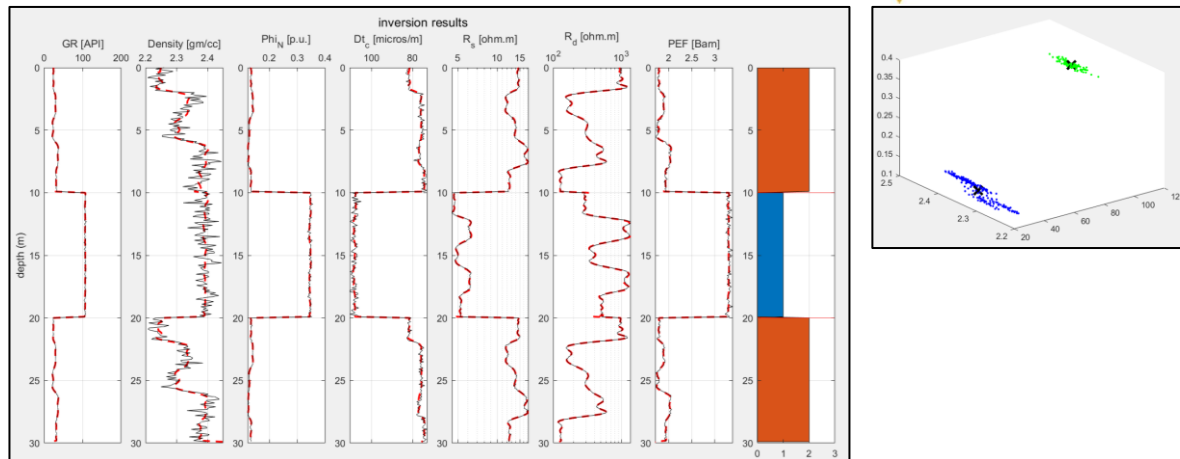


Figure 2 Final centroid location and boundaries (red lines) detection and fitting between calculated and measured.

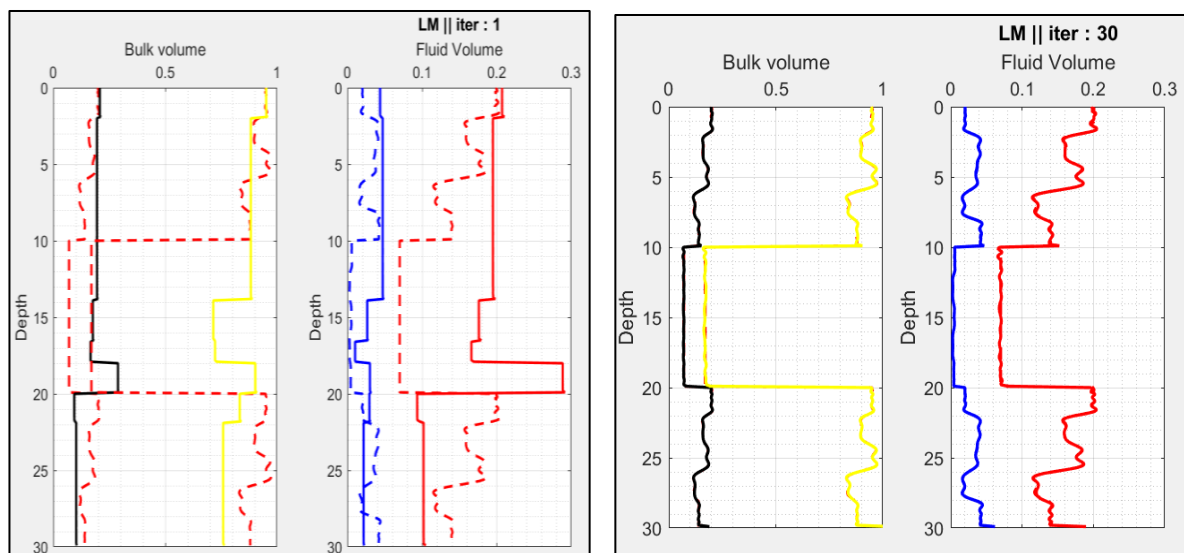


Figure 3 Initial (Left) and final (Right) model prediction regarding the extracted layers number.

The proposed algorithm was implemented on the Baktalórántháza-1 well, which is located in Szabolcs-Szatmár-Bereg County in northeast Hungary. The drilling site is situated approximately 2.1 km away from Baktalórántháza village, and the maximum depth reached by the measurements was 1197.70 m. The bottom hole temperature was recorded as 53 degrees Celsius. The borehole shows promising potential for hydrological exploration and is considered a productive thermal water well. Analysis of the well-logging data indicates a shaly sand sequence. According to the borehole logs, the underlying rock of the sandy succession is shale. Sands ranging from 100 to 160 m in thickness were deposited, followed by a shaly formation and 5-15 m thick coarse-grained beds. The Pannonian shaly complex is mainly composed of clayey sand, gravel, clayey silt, clayey marl, and bituminous. Fig. 4 shows the initial guess of porosity and shale volume at iteration 1 (left) and the final predicted porosity and shale volume at iteration 15 (right). The results show an increase in the shale volume with depth, while it shows a decrease in porosity with depth.

Conclusions

In conclusion, cluster analysis is a useful statistical technique for grouping similar objects in a dataset, with the primary objective of creating distinct groups that contain similar objects. The interval inversion

approach is a useful tool for petrophysical parameter prediction and involves representing the parameters in a specific depth range as a continuous function. The integration of the robust clustering technique with the interval inversion algorithm has been tested on both synthetic and hydrogeophysical field data, with promising results. The proposed algorithm was implemented on the Baktalórátó-1 well in Hungary, and the analysis of the well-logging data indicated a shaly sand sequence. The predicted parameters show an increase in the shale volume toward the Pannonian complex. Overall, the integration of robust clustering with interval inversion provides a powerful tool for analysing complex datasets in geosciences.

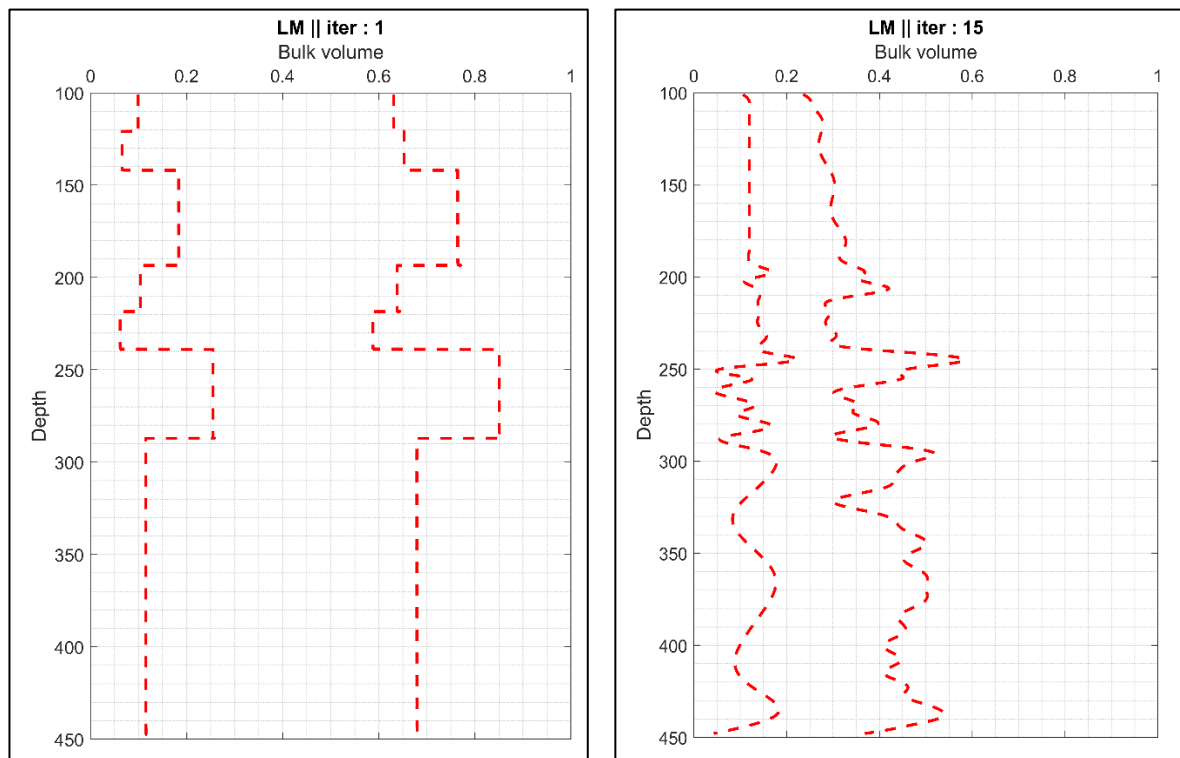


Figure 4 Initial (Left) and final (Right) model prediction regarding the extracted layers number.

Acknowledgments

The research was carried out in project No. K-135323 is supported by the National Research, Development, and Innovation Office (NKFIH).

References

- 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.
- Steiner, F. 1991. *The Most Frequent Value. Intro- Duction to a Modern Conception Statistics*. Budapest: *Akademia Kiado*.
- 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(November): 106696.