Machine learning based approach for the interpretation of engineering geophysical sounding logs

Armand Abordán^{1,2} · Norbert Péter Szabó^{1,2}

Received: 18 May 2021 / Accepted: 29 July 2021 / Published online: 12 August 2021
© The Author(s) 2021

Abstract

In this paper, a set of machine learning (ML) tools is applied to estimate the water saturation of shallow unconsolidated sediments at the Bátaapáti site in Hungary. Water saturation is directly calculated from the first factor extracted from a set of direct push logs by factor analysis. The dataset observed by engineering geophysical sounding tools as special variants of direct-push probes contains data from a total of 12 shallow penetration holes. Both one- and two-dimensional applications of the suggested method are presented. To improve the performance of factor analysis, particle swarm optimization (PSO) is applied to give a globally optimized estimate for the factor scores. Furthermore, by a hyperparameter estimation approach, some control parameters of the utilized PSO algorithm are automatically estimated by simulated annealing (SA) to ensure the convergence of the procedure. The result of the suggested ML-based log analysis method is compared and verified by an independent inversion estimate. The study shows that the PSO-based factor analysis aided by hyperparameter estimation provides reliable in situ estimates of water saturation, which may improve the solution of environmental and engineering problems in shallow unconsolidated heterogeneous formations.

Keywords Factor analysis · Particle swarm optimization · Simulated annealing · Direct push logging · Hyperparameter estimation

1 Introduction

Borehole geophysical data contain valuable information about the physical characteristics of the investigated subsurface (Everett 2013). The measured raw data can be processed and turned into petrophysical parameters by several different approaches. The simplest is the deterministic modeling, which is often based on a single empirical equation that transforms an observed variable into a petrophysical parameter (e.g., shale volume estimation along boreholes based only on the natural gamma-ray intensity log). A

✉ Armand Abordán
gfaa@uni-miskolc.hu

¹ Department of Geophysics, University of Miskolc, 3515 Miskolc-Egyetemváros, Hungary

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

more advanced way to process geophysical data is inverse modeling (Zhdanov 2015). This approach combines numerical mathematics and optimization theory to derive the physical parameters of the investigated geological structures from the measured data. Geophysical inverse problems can be solved by linearized or global optimization methods (e.g., simulated annealing (SA), genetic algorithm (GA), particle swarm optimization (PSO)). However, in all cases, geophysical inversion is based on a scientific understanding, a set of equations (i.e., response functions) that describe the relationship between the observed data and the petrophysical parameters of the subsurface.

Inverse problems are solved by updating a starting model iteratively until the synthetic data calculated on the model fits the measured data (Menke 2012). This optimization task is usually solved by linearized methods (e.g., the least squares method). These provide a fast and satisfactory solution given that there is a good initial (starting) model. However, the application of these methods to large scale (multivariate) problems is often problematic due to the complexity of the objective function. During the minimization of the objective function (that measures the misfit between the measured and calculated data) these linearized methods use a gradient-based search, which means that they stabilize in the nearest local minimum of the objective function.

The above-mentioned global optimization methods use random search instead, therefore are capable to get out of local minima of the objective function. Thus, they can provide a reliable and convergent solution independently of the chosen starting model. For this reason, global optimization techniques are widely used in geophysics (Sen and Stoffa 2013), e.g., full-waveform Rayleigh-wave inversion (Xing and Mazzotti 2019), inversion of magnetotelluric data (Wang et al. 2012), two-dimensional inversion of magnetic data (Liu et al. 2018).

One of their disadvantages is that their computational requirements far exceed those of linearized methods (Kaikkonen and Sharma 2001), but due to the rapid development of computing, this is no longer a problem nowadays. Furthermore, it is worth mentioning that based on a single program run, global optimization methods cannot provide information on the error of parameter estimation like linearized methods. However, with different hybrid solutions, one can combine linearized methods with global methods to create efficient, two-phase algorithms. In general, complex inverse problems can be initialized by some global method to avoid the algorithm from stabilizing in a local optimum of the objective function. Then, when the procedure is close enough to the optimal solution, it can be switched to a linearized method so that the computation of estimation error of derived petrophysical parameters becomes possible and the runtime of the algorithm is reduced as well. Such hybrid solutions were suggested by Soupios et al. (2011) for seismic inversion, Chunduru et al. (1997) for 2D resistivity profiling data and by Szabó and Dobróka (2020) for the non-linear well logging inverse problem.

Because these global methods do not use linearization, they do not require supplementary information (e.g., derivatives or a good starting model), but their convergence is greatly influenced by their control (hyper) parameters (e.g., combination and parameters of genetic operators in GA, initial temperature, cooling schedule and parameter perturbation in SA or inertia weight and learning factors in PSO). However, through hyperparameter estimation, it is possible to automatically select the optimal values for these parameters with a secondary optimization algorithm, thus guaranteeing the convergence of the procedure. Based on a similar principle, the efficiency of geophysical inversion methods can also be increased by hyperparameter estimation. For example, the value of parameters otherwise treated as constants (e.g., zone parameters in case of

well logging inversion) in the response functions can be optimized by incorporating an additional program loop (Dobróka et al. 2016).

2 Machine learning tools used in geophysics

In addition to inversion-based data processing methods, another approach is to use machine learning (ML) from the toolbox of artificial intelligence (Dramsch 2020). The aim of ML is to create such algorithms that can improve their own effectiveness by utilizing the experience gained during their operation. Such as artificial neural network, deep learning, cluster analysis or fuzzy logic. The frequently used regression analysis (linear, non-linear or logistic) is also considered an ML tool. The use of these tools is gaining ground in the processing and interpretation of geological and geophysical data in recent years (Caté et al. 2017). Some examples from the literature are seismic interpretation (Wang et al. 2018), seismology (Kong et al. 2019), fault detection (Araya-Polo et al. 2017), electrical resistivity tomography (Vu and Jardani 2021) and well logging inversion (Szabó 2018). The use of these methods is advantageous in the absence of the previously mentioned scientific background (mathematical relationship does not necessarily exist between the petrophysical parameters and observed data) or the theoretical description is so complex (e.g., extremely long runtime) that we must disregard the exact description.

In recent decades a wide variety of machine learning tools have been developed. To select the one that is most suitable for solving a given problem, it is crucial to understand how they work. Therefore, it is best to categorize the different ML methods based on their learning style. This way we can sort them into three main categories. The most commonly applied tools are based on supervised learning. In this case there is exact information (inputs and outputs) for training the algorithm. In addition to the input data, we also know what kind of results we can expect. Thus, supervised learning-based methods provide exact parameters (i.e., numerical labels) such as porosity or categorical labels (e.g., rock type). Some of the most often used ML approaches based on supervised learning are regression analysis and classifiers.

In case of unsupervised learning, there is no exact information in the dataset regarding the output. Meaning that it uses data without labels and therefore the procedure has to group the observed data by finding structures within the dataset. A frequently applied tool that uses unsupervised learning is cluster analysis, which classifies the input data based on some distance metric. Clustering is often applied on geophysical datasets, e.g., for rock typing based on wireline logging data (Ali and Sheng-Chang 2020). Dimension reduction methods also utilize the unsupervised learning approach (e.g., factor analysis or principal component analysis). Big datasets are often difficult to handle therefore it can be advantageous to reduce the size (dimensionality) of the problem. By keeping most of the information contained in the original dataset (i.e., statistical sample), the same phenomenon can be described with fewer variables. Thus, the new variables contain the essential features of the investigated object as well as possible new properties that cannot be measured directly. Furthermore, by removing the error factors, it can even be used to improve the signal-to-noise ratio. In factor analysis, a large number of inter-related or independent variables are replaced by a smaller number of uncorrelated variables, where the resulting new variables cannot be measured directly. The applicability of the dimensionality reduction methods comes from the fact that the newly extracted variables often show a strong correlation with different petrophysical parameters, thus

they can be used e.g., for lithological classification (Puskarczyk et al. 2019) or for quantitative estimation of petrophysical parameters (Szabó 2011).

The last category of ML methods based on learning style is the semi-supervised approach. A combination of the previous two cases, it can be used when only part of the database contains information about the output (i.e., labeled data) while some of the data does not have all the necessary information available (i.e., unlabeled data) for supervised training of the system. Semi-supervised learning ensures that the latter data is not wasted and contributes in some way to the design of the system. An example of a semi-supervised based system can be found in Li et al. (2019) for lithology recognition using a generative adversarial network.

In the following sections, an ML-based system is suggested for direct-push log analysis, which enables the estimation of water saturation in shallow unconsolidated heterogeneous formations.

3 Water saturation estimation in the shallow subsurface by factor analysis

A direct push (DP) logging technology named as engineering geophysical sounding was developed in Hungary (Fejes and Jósa 1990) based on cone penetration testing (CPT). Besides cone resistance and sleeve friction, DP logging can also measure the same parameters routinely recorded by wireline logs including gamma-ray intensity, bulk density, neutron porosity and resistivity. It enables the characterization of the shallow unconsolidated subsurface with high vertical resolution down to approximately 50 m with a highly mobile equipment. These measurements are done by advancing steel rods into the ground without drilling, therefore there is less disturbance of the subsurface, which is also advantageous considering geophysical measurements.

The different DP technologies can be used to solve problems of contamination mapping, environmental risks assessment and ground-water investigations (Dietrich and Leven 2006). By processing direct push logs, one can derive quantitative information about the composition of shallow unconsolidated sediments, such as clay content, porosity or water saturation (Balogh 2016). In this paper, an ML-based statistical approach is developed for the quantitative analysis of direct-push logs.

Factor analysis, as mentioned before is an unsupervised ML tool for describing several measured quantities with fewer uncorrelated variables. In case of DP logging data, the measured logs are the input variables, which are processed jointly to extract new factors. Here, the derived factors can be looked at as factor logs, which can be related to petrophysical parameters by regression analysis (Szabó et al. 2018). In this study, water saturation (S_w) of shallow unconsolidated formations is estimated based on the first factor log extracted from a direct push logging dataset.

As a preliminary step, DP logs need to be standardized to serve as input for factor analysis. Then they are collected into a matrix **D**, where individual columns contain the recordings of different logging tools. In this paper, the processed logging tools are the natural gamma-ray intensity, GR (cpm), cone resistance, $RCPT$ (MPa), bulk density, DEN (g/cm³), neutron porosity, $NPHI$ (v/v) and resistivity, RES (ohmm).

$$\mathbf{D} = \begin{pmatrix} D_{11} & D_{12} & \cdots & D_{1K} \\ D_{21} & D_{22} & \cdots & D_{2K} \\ \vdots & \vdots & \vdots & \vdots \\ D_{i1} & D_{i2} & \cdots & D_{iK} \\ \vdots & \vdots & \vdots & \vdots \\ D_{N1} & D_{N2} & \cdots & D_{NK} \end{pmatrix}, \quad (1)$$

where K is the number of applied DP logging tools and N shows the number of measured depth points in the given sounding hole. The solution of factor analysis is based upon the decomposition of data matrix $\mathbf{D}$

$$\mathbf{D} = \mathbf{F}\mathbf{L}^T + \mathbf{E}, \quad (2)$$

where $\mathbf{F}$ is an N -by- M matrix of factor scores (i.e., factor logs) where M denotes the number of computed factors ($M < K$), $\mathbf{L}$ is a K -by- M matrix of factor loadings, which shows the correlation relationship between the measured DP logs and the newly extracted factors and $\mathbf{E}$ is an N -by- K error matrix. Based on the model of factor analysis in Eq. (2), the derived factors are essentially the weighted sums of the measured direct push logs. The first column of matrix $\mathbf{F}$ (i.e., the first factor) describes most of the data variance of the measured dataset and therefore generally bears the most significance for data interpretation. By assuming that the factors are linearly independent, the correlation matrix is given by

$$\mathbf{R} = N^{-1}\mathbf{D}^T\mathbf{D} = \mathbf{L}\mathbf{L}^T + \mathbf{\Psi}, \quad (3)$$

where $\mathbf{\Psi}$ is a diagonal matrix of specific variances that is independent of the common factors. For the estimation of the factor loadings in matrix $\mathbf{L}$, Jöreskog (2007) suggested the following non-iterative formula

$$\mathbf{L} = (\text{diag}\mathbf{S}^{-1})^{-1/2}\mathbf{\Omega}(\mathbf{\Gamma} - \theta\mathbf{I})^{1/2}\mathbf{U}, \quad (4)$$

where $\mathbf{\Gamma}$ is the diagonal matrix of the first M number of sorted eigenvalues of the sample covariance matrix $\mathbf{S}$, $\mathbf{\Omega}$ denotes the matrix of the first M number of eigenvectors, $\mathbf{I}$ is the identity matrix, $\mathbf{U}$ is an arbitrary M -by- M orthogonal matrix and θ is an adequately chosen constant that is to be slightly smaller than 1.

Once factor loadings are available, the factor scores can be estimated by a maximum likelihood method (Bartlett 1937)

$$\mathbf{F}^T = (\mathbf{L}^T\mathbf{\Psi}^{-1}\mathbf{L})^{-1}\mathbf{L}^T\mathbf{\Psi}^{-1}\mathbf{D}^T. \quad (5)$$

As an advanced approach, in this study, factor scores are estimated by means of global optimization. We currently utilize the metaheuristic particle swarm optimization (PSO) for giving an estimate to the factors scores. This PSO-based solution of factor analysis is referred to as FA-PSO (Abordán and Szabó 2018). In this approach, factor analysis is treated as an inverse problem, thus Eq. (2) must be reformulated

$$\mathbf{d} = \tilde{\mathbf{L}}\mathbf{f} + \mathbf{e}, \quad (6)$$

where the standardized measured DP logs are represented as a KN length column vector $\mathbf{d}$, factor loadings are given in a NK -by- NM matrix $\tilde{\mathbf{L}}$, factor scores are also gathered in a column vector $\mathbf{f}$ of MN length and $\mathbf{e}$ is the KN length residual vector (Szabó and Dobróka 2018). To initialize this metaheuristic solution, measured DP logs are first collected into

the column vector $\mathbf{d}$, then the matrix of factor loadings $\tilde{\mathbf{L}}$ can be estimated by Eq. (4). For getting more meaningful factors, the varimax rotation is applied on the loadings (Kaiser 1958). Then these factor loadings are kept constant, while the optimal values of factor scores $\mathbf{f}$ is approximated by PSO. For measuring discrepancy, the following L_2 norm based objective function is used, which is minimized to estimate the optimal values of factors scores

$$E = \sqrt{\frac{1}{NK} \sum_{i=1}^{NK} (d_i^{(m)} - d_i^{(c)})^2}, \quad (7)$$

where $d_i^{(m)}$ and $d_i^{(c)}$ are the i th standardized measured and calculated direct push logging data, respectively. The calculated data comes from the multiplication $\tilde{\mathbf{L}}\mathbf{f}$ from Eq. (6), while $\mathbf{d}$ stores the measured DP logs in the same equation. The before mentioned multiplication of factor loadings and factor scores permits the calculation of synthetic DP logs, which can be looked at as the solution of the forward problem.

For solving this optimization problem PSO is applied which is a global optimization method inspired by the social behavior of animals. The original method was developed by Kennedy and Eberhart (1995). PSO is often applied for its relatively low computational requirements and easy implementation compared to other optimization methods such as the genetic algorithm (Holland 1975). It is a population based technique where each particle (i.e., possible solution) searches for the optimal solution by adjusting its own position in the search space by taking into account its own best position and the whole swarm's best position in every iteration step. This searching mechanism of PSO is governed by Eqs. (8–9). Having an n -dimensional optimization problem, the i th particle's position in the search space can be represented by the vector $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{in})^T$ and its velocity by vector $\mathbf{v}_i = (v_{i1}, v_{i2}, \dots, v_{in})^T$. Every particle of the swarm has a memory of its best position, which is continuously updated during the iterations and is stored in vector $\mathbf{p}_i = (p_{i1}, p_{i2}, \dots, p_{in})^T$ for the i th particle. The position and velocity update equations are

$$\mathbf{x}_i(t+1) = \mathbf{x}_i(t) + \mathbf{v}_i(t+1), \quad (8)$$

$$\mathbf{v}_i(t+1) = w\mathbf{v}_i(t) + r_1c_1(\mathbf{p}_i(t) - \mathbf{x}_i(t)) + r_2c_2(\mathbf{g}(t) - \mathbf{x}_i(t)), \quad (9)$$

where $t=1, \dots, T$ is the current iteration step, T is the last iteration and $i=1, 2, \dots, S$ shows the particle index and S is the population size. In Eq. (9) r_1 and r_2 denote random variables uniformly distributed in 0 to 1 and w is an inertia weight (Shi and Eberhart 1998) that was introduced to balance between global and local search. Vector $\mathbf{g}$ stores the very best position found by the swarm until the current iteration step. It is continuously updated in each iteration and helps the swarm to find the global optimum of the objective function. Acceleration factors c_1 and c_2 are positive constants, where c_1 is the cognitive scaling parameter and c_2 is the social scaling parameter, both generally set as 2 (Kennedy and Eberhart 1995).

Since PSO is a metaheuristic method, the chosen control (hyper-) parameters have a great effect on its performance (Zhang et al. 2005). To increase the reliability of the searching mechanism in finding the global optimum, the automatic selection of acceleration factors c_1 and c_2 is developed. It is carried out in an additional program loop with the help of simulated annealing (Metropolis et al. 1953) as depicted in Fig. 1.

For the outer program loop, we initialize the values of both c_1 and c_2 hyperparameters as 1 and then let SA select the optimal values automatically for the current optimization

Fig. 1 Flowchart of the FA-PSO method aided by hyperparameter estimation

problem. In every SA iteration step their value is adjusted by adding a small b parameter to both values. Parameter b is randomly generated in each iteration from $-b_{\max}$ to $b_{\max}$, where $b_{\max}$ is also reduced iteratively as $b_{\max} = b_{\max} \times \varepsilon$, where ε is a constant smaller than 1.

Once the new c_1 and c_2 parameters are selected, PSO is run to find the optimal values of factor scores $\mathbf{f}$. If the difference in energy (ΔE) in two successive iteration steps of the SA program loop according to the objective function in Eq. (7) is negative (i.e., PSO was more effective with the new hyperparameters), then the current values of parameters c_1 and c_2 are accepted and the iterations are continued. If the difference in energy is positive (i.e., new control parameters decreased the efficiency of PSO), then the accepting probability of the new c_1 and c_2 is given by $P_a = \exp(-\Delta E/T^*)$, where T^* is the current temperature of the system. Temperature is reduced logarithmically according to $T^{(\text{new})} = T_0 / \lg(1 + q)$ as suggested by Geman and Geman (1984), where q is the number of iterations computed so far and T_0 is the starting temperature of the system. The new hyperparameters are accepted only if a random number generated from 0 to 1 is smaller than P_a . This mechanism of accepting worse solutions prevents SA from being stuck in a local optimum near the starting model and thus allows for the optimal selection of parameters c_1 and c_2 for PSO.

4 Field tests

4.1 One-dimensional application

To test the suggested hyperparameter estimation assisted factor analysis, a direct push logging dataset is used that was measured in Báticaapáti, Southwest Hungary. The dataset contains a total of 12 sounding holes with the natural gamma-ray intensity, GR (cpm), cone resistance, $RCPT$ (MPa), bulk density, DEN (g/cm³), neutron porosity, $NPHI$ (v/v) and resistivity, RES (ohmm) logs measured in the upper 20–28 m of unconsolidated loessy-sandy layers. The sounding holes are located along a 550 m long profile approximately 50 m from each other.

First, a one-dimensional application is shown for sounding hole 7 (SH7) where logging data is available for the interval of 0.5–27.7 m. To start off the procedure, DP logging data is standardized and then factor loadings are estimated by Eq. (4). The rotated factor loadings by the varimax algorithm are shown in Table 1. The first factor explains 71% of the total data variance, while the other 29% is explained by the second factor.

According to the computed factor loadings, the first factor correlates most with the bulk density, neutron porosity and resistivity logs, while the second factor is most influenced by the cone resistance and natural gamma-ray intensity logs. Then these factor loadings remain unchanged for the remainder of the procedure. To give an estimate to the factor scores $\mathbf{f}$ by PSO, first, a random population of 60 particles (each representing

Table 1 Rotated factor loadings estimated by the FA-PSO method in SH-7

Direct-push logs	Symbol	Factor 1	Factor 2
Natural gamma-ray intensity	GR	0.03	−0.61
Cone resistance	RCPT	−0.09	0.74
Bulk density	DEN	0.89	0.15
Neutron-porosity	NPHI	0.82	−0.18
Resistivity	RES	−0.89	0.18

a solution candidate for vector $\mathbf{f}$ is generated with uniform distribution within the search space previously defined by solving Eq. (5). In this instance, the limits of factor scores are set as -5 to 5 . The inertia weight w for Eq. (9) is initialized as 1 and then adjusted in each iteration based on $w^{(\text{new})} = w^{(\text{old})} \times \alpha$, where α is a damping factor set to 0.99 , while hyperparameters c_1 and c_2 are automatically selected by SA in the outer program loop. SA is run for 50 iteration steps to find the optimal values of c_1 and c_2 (Fig. 2).

The generated hyperparameters are plugged into PSO in each SA iteration step to test their effectiveness. With the currently set c_1 and c_2 parameters, PSO is run for 2000 iterations to given an estimate to the optimal values of factor scores in vector $\mathbf{f}$. During the PSO runs, all 60 particles are adjusted according to Eqs. (8)–(9), and the objective function defined in Eq. (7) is recalculated with the new values of factor scores $\mathbf{f}$ in every iteration step. Figure 2 on the right depicts the final data distance of each PSO run with the corresponding c_1 and c_2 control parameters (Fig. 2 on the left). It can be seen that after approximately 15 iterations, SA finds the optimal c_1 and c_2 parameters and thereafter their value somewhat stabilizes and the data distance reached by PSO does not decrease any further. The optimal values of c_1 and c_2 in this case are found to be 1.54 and 2.29 , respectively, which somewhat differ from the default values of 2 and 2 , respectively (Kennedy and Eberhart 1995).

Fig. 2 Convergence of the FA-PSO method (on the right) by altering control parameters c_1 and c_2 by SA (on the left) in SH-7

Once the pre-defined number of iteration steps is reached, the final values of the factor scores (represented by the particle with the lowest data misfit) are accepted as the final solution. In this instance, the final data distance reached by PSO with the automatically selected hyperparameters at the end of the procedure was 0.45. Then the first factor (F_1) can be used to estimate the water saturation (S_w) of the penetrated formations by regression analysis (Szabó et al. 2012). In this paper, an exponential relationship is assumed between the first factor log and water saturation (Szabó et al. 2018)

$$S_w = ae^{(bF_1)} + c, \quad (10)$$

where a , b and c are area specific regression coefficients. As a reference for regression analysis the S_w result log taken from the complete quality controlled inversion (Drahos 2005) was applied.

Figure 3 on the left depicts the exponential relationship between the first factor log and water saturation along sounding hole 7 based on Eq. (10). For this DP logging dataset, the regression coefficients with 95% confidence bounds are found to be $a=0.8084$ [$a_{min}=0.6445$, $a_{max}=0.9723$] and $b=0.1996$ [$b_{min}=0.1606$, $b_{max}=0.2387$] and $c=-0.2046$ [$c_{min}=-0.3652$, $c_{max}=-0.0440$]. On the right of Fig. 3, water saturation estimated by local inversion and factor analysis is plotted, where the high Pearson's correlation coefficient ($r=0.98$) indicates that the two variables are nearly linearly proportional and thus shows the reliability of the FA-PSO based water saturation estimation.

The results of the FA-PSO method in SH-7 are depicted in Fig. 4 along with the input direct push logs. The first five tracks contain the measured logs, track 6 contains the first (purple) and second factor (red) logs and the last track shows the water saturation estimates by inverse modeling (green) and by the FA-PSO method (blue) utilizing the first factor log.

It can be seen that the two water saturation estimates are almost identical, which validates the applicability of factor analysis for the processing of direct push logs and can serve as an independent tool for estimating water saturation in the shallow unconsolidated subsurface.

Fig. 3 The exponential relationship between the first factor and water saturation in SH-7 (on the left), water saturation derived by local inversion and factor analysis (on the right)

Fig. 4 The results of the FA-PSO method in SH-7

4.2 Two-dimensional application

The processing of direct push logs by factor analysis can also be carried out in two-dimensions, which permits the simultaneous processing of DP logs recorded in neighboring sounding holes. Thus, a 2D section of water saturation can be estimated in one interpretation phase from the DP logs recorded in multiple sounding holes (Szabó et al. 2018). First, gather the measured DP logs in vector $\mathbf{d}^{(h)}$ defined in Eq. (6) from the h th hole ($h=1,2,\dots,H$). Then the model of factor analysis can be extended for multiple sounding holes as

$$\begin{pmatrix} \mathbf{d}^{(1)} \\ \vdots \\ \mathbf{d}^{(h)} \\ \vdots \\ \mathbf{d}^{(H)} \end{pmatrix} = \begin{pmatrix} \tilde{\mathbf{L}}^{(1)} & 0 & 0 & 0 & 0 \\ 0 & \ddots & 0 & 0 & 0 \\ 0 & 0 & \tilde{\mathbf{L}}^{(h)} & 0 & 0 \\ 0 & 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & 0 & \tilde{\mathbf{L}}^{(H)} \end{pmatrix} \times \begin{pmatrix} \mathbf{f}^{(1)} \\ \vdots \\ \mathbf{f}^{(h)} \\ \vdots \\ \mathbf{f}^{(H)} \end{pmatrix} + \begin{pmatrix} \mathbf{e}^{(1)} \\ \vdots \\ \mathbf{e}^{(h)} \\ \vdots \\ \mathbf{e}^{(H)} \end{pmatrix}, \quad (11)$$

where $\tilde{\mathbf{L}}^{(h)}$ denotes the matrix of factor loadings and $\mathbf{f}^{(h)}$ denotes the vector of factor scores computed for the h th sounding hole. Since there are N_h number of logged depth points in the h th sounding hole, the total number of depth points is $N^* = N_1 + N_2 + \dots + N_H$. The matrix of factor loadings in Eq. (11) is calculated similarly to the one-dimensional case by Eq. (4), and the optimal factor scores are estimated by PSO. Here it should be noted that the two dimensional case differs from the one dimensional case, because here measured DP logs are processed jointly from several sounding holes together while assuming that the same factor loadings are applicable for the whole exploration area. Once the factor logs are derived and are related to water saturation by regression analysis, they can be interpolated

(e.g., by kriging) between the sounding holes to derive the map of water saturation for the investigated area.

To test this two-dimensional approach for water saturation estimation by factor analysis, data is collected from 12 sounding holes (SH-1 to SH-12), which are located along a 550 m long profile, approximately 50 m apart from each other. The natural gamma-ray intensity, GR (cpm), cone resistance, $RCPT$ (MPa), bulk density, DEN (g/cm³), neutron porosity, $NPHI$ (v/v), resistivity, RES (ohmm) logs are all available along the penetrated sounding holes to serve as the input of the two-dimensional factor analysis. The total number of DP logging data from the 12 sounding holes combined is 15,500. By extracting 2 factors, the rotated factor loadings are given in Table 2. The first factor describes 72% of the total data variance, while the remaining 28% is described by the second factor.

The calculated factor loadings for the full dataset of all 12 sounding holes are essentially the same as the factor loadings estimated only in SH-7 (Table 1). The first factor correlates best with the bulk density, neutron porosity and resistivity logs, while the second factor is most influenced by the cone resistance and natural gamma-ray intensity logs. The remainder of the two-dimensional procedure is identical to the one-dimensional case. Factor loadings in Table 2 are fixed, and then by solving Eq. (5) for the factor scores, the boundaries of the search space for PSO is defined as -5 to 30 . Due to the larger dataset and thus increased number of unknowns, PSO here requires 7500 iterations to find the optimal values of the factor scores by utilizing 60 particles. Hyperparameters c_1 and c_2 are again automatically selected by SA in 50 iteration steps. The minimal data distance reached by PSO in finding the optimal values of factor scores by utilizing the SA derived hyperparameters is depicted in Fig. 5 on the right for all 50 iteration steps with the corresponding c_1 and c_2 parameters which are shown in Fig. 5 on the left.

It can be seen that after approximately 12 iterations, SA finds the optimal values of c_1 and c_2 , thereafter their value somewhat stabilizes and the data distance reached by PSO does not decrease any further. The optimal values of c_1 and c_2 are found to be 1.33 and 2.13, respectively. The final data distance reached by PSO with these parameters is 0.50. Then the resultant first factor log can be related to the water saturation of the investigated area by regression analysis based on Eq. (10) as seen in Fig. 6 on the left. As a reference for regression analysis, the quality checked local inversion derived water saturation values are used. In this instance, the regression coefficients with 95% confidence bounds are found to be $a=0.8544$ [$a_{min}=0.8086$, $a_{max}=0.9001$] and $b=0.1982$ [$b_{min}=0.1880$, $b_{max}=0.2085$] and $c=-0.2258$ [$c_{min}=-0.2706$, $c_{max}=-0.1810$]. On the right of Fig. 6 water saturation estimated by local inversion and factor analysis is plotted, where the high correlation coefficient ($r=0.97$) indicates that the two estimates are nearly linearly proportional and thus shows the reliability of the two dimensional factor analysis for water saturation estimation.

By interpolating the estimated water saturation logs between the sounding holes, one can derive the map of the water saturation of unsaturated formations along the processed profile

Table 2 Rotated factor loadings estimated by the FA-PSO method in sounding holes 1–12

Direct-push logs	Symbol	Factor 1	Factor 2
Natural gamma-ray intensity	GR	0.03	− 0.50
Cone resistance	RCPT	− 0.08	0.70
Bulk density	DEN	0.85	0.20
Neutron porosity	NPHI	0.77	− 0.24
Resistivity	RES	− 0.88	0.16

Fig. 5 Convergence of the FA-PSO method (on the right) by altering control parameters c_1 and c_2 by SA (on the left) for multiple sounding holes

Fig. 6 The exponential relationship between the first factor and water saturation in sounding holes 1–12 (on the left), water saturation derived by local inversion and the two dimensional factor analysis (on the right)

(Fig. 7). Similarly, the results of local inversion estimates can also be interpolated between the holes, which is depicted in Fig. 8.

The logs are located at approximately 50 m apart, sounding hole 7 processed in the previous section is located at 300 m. The layers with different water saturations can be easily recognized on the derived maps and thus can help with site characterization. The good correlation between the two estimates confirm the applicability of the two-dimensional factor analysis for water saturation estimation from direct push logs.

Fig. 7 Water saturation estimated by the two-dimensional factor analysis in sounding holes 1–12

Fig. 8 Water saturation estimated by inverse modeling in sounding holes 1–12

5 Conclusions

The paper presents the results of a machine learning-based log analysis method for direct-push logging data. The suggested factor analysis based approach is shown to effectively reduce the dimension of direct push logging datasets and the newly extracted factor logs by the FA-PSO method can be used to estimate water saturation along arbitrary long sounding hole intervals. By incorporating more than one sounding hole into the procedure, even two-dimensional water saturation profiles can be derived by simultaneously processing DP logs from neighboring holes. The result of the presented method is also verified by independent inversion based estimates of water saturation. It is also shown that simulated annealing is capable to automatically select some of the hyperparameters of the utilized particle swarm optimization algorithm. Thus, the uncertainty of the optimization algorithm can be reduced, since the manual selection of the c_1 and c_2 parameters is no longer necessary. The suggested machine learning tool assures

reliable evaluation of unconsolidated and unsaturated near-surface formations being as target domain of several engineering and environmental geophysical problems.

Acknowledgements The research was carried out in the Project No. K-135323 supported by the National Research, Development and Innovation Office (NKFIH). The use of the dataset was permitted by Dezső Drahos from Loránd Eötvös University. The authors thanks for the continuous support of János Stickel from Elgoscarr Ltd.

Authors' contributions Armand Abordán: original draft, software, visualization. Norbert Péter Szabó: conceptualization, methodology, review and editing.

Funding Open access funding provided by University of Miskolc. The research was carried out in the Project No. K-135323 supported by the National Research, Development and Innovation Office (NKFIH).

Declarations

Conflict of interest The authors have no conflicts of interest to declare that are relevant to the content of this article.

Ethics approval Not applicable.

Consent to participate Not applicable.

Consent for publication Not applicable.

Availability of data and material Because of the data confidentiality, the experimental data is not published.

Code availability Because of the data confidentiality, the code is not published.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Abordán A, Szabó NP (2018) Particle swarm optimization assisted factor analysis for shale volume estimation in groundwater formations. *Geosci Eng* 6(9):87–97
- Ali A, Sheng-Chang C (2020) Characterization of well logs using K-mean cluster analysis. *J Petrol Explor Prod Technol* 10:2245–2256. <https://doi.org/10.1007/s13202-020-00895-4>
- Araya-Polo M, Dahlke T, Frogner C, Zhang C, Poggio T, Hohl D (2017) Automated fault detection without seismic processing. *Lead Edge* 36(3):208–214. <https://doi.org/10.1190/tle36030208.1>
- Balogh GP (2016) Interval inversion of engineering geophysical sounding logs. *Geosci Eng* 5(8):22–31
- Bartlett MS (1937) The statistical conception of mental factors. *Br J Psychol* 28:97–104. <https://doi.org/10.1111/j.2044-8295.1937.tb00863.x>
- Caté A, Perozzi L, Gloaguen E, Blouin M (2017) Machine learning as a tool for geologists. *Lead Edge* 36:215–219. <https://doi.org/10.1190/tle36030215.1>
- Chunduru RK, Sen MK, Stoffa PL (1997) Hybrid optimization methods for geophysical inversion. *Geophysics* 62:1196–1207. <https://doi.org/10.1190/1.1444220>
- Dietrich P, Leven C (2006) Direct push-technologies. In: Kirsch R (eds) *Groundwater geophysics*. Springer, Berlin. https://doi.org/10.1007/3-540-29387-6_11

- Dobróka M, Szabó NP, Tóth J, Vass P (2016) Interval inversion approach for an improved interpretation of well logs. *Geophysics* 81:D155–D167. <https://doi.org/10.1190/geo2015-0422.1>
- Drahos D (2005) Inversion of engineering geophysical penetration sounding logs measured along a profile. *Acta Geodetica Geophys Hungarica* 40:193–202. <https://doi.org/10.1556/AGeod.40.2005.2.6>
- Drams JS (2020) 70 years of machine learning in geoscience in review. *Adv Geophys*. <https://doi.org/10.1016/bs.agph.2020.08.002>
- Everett ME (2013) Near-surface applied geophysics. Cambridge University Press, Cambridge. <https://doi.org/10.1017/CBO9781139088435>
- Fejes I, Jóna E (1990) The engineering geophysical sounding method. Principles, instrumentation, and computerised interpretation. In: SH Ward (ed) *Geotechnical and environmental geophysics, Environmental and groundwater*, vol 2. SEG, pp 321–331, ISBN 978-0-931830-99-0.
- Geman S, Geman D (1984) Stochastic relaxation, Gibbs distributions, and Bayesian restoration of images. *IEEE Trans Pattern Anal Mach Intell PAMI-6*:721–741. <https://doi.org/10.1109/TPAMI.1984.4767596>
- Holland JH (1975) *Adaptation in natural and artificial systems*. University of Michigan Press
- Jöreskog KG (2007) Factor analysis and its extensions. In: Cudeck R, MacCallum RC (eds) *Factor analysis at 100, historical developments and future directions*. Lawrence Erlbaum Associates, pp 47–77
- Kaikkonen P, Sharma SP (2001) A comparison of performances of linearized and global nonlinear 2-D inversions of VLF and VLF-R electromagnetic data. *Geophysics* 66:462–475. <https://doi.org/10.1190/1.1444937>
- Kaiser HF (1958) The varimax criterion for analytical rotation in factor analysis. *Psychometrika* 23:187–200. <https://doi.org/10.1007/BF02289233>
- Kennedy J, Eberhart R (1995) Particle swarm optimization. *Proc IEEE Int Conf Neural Netw* 4:1942–1948. <https://doi.org/10.1109/ICNN.1995.488968>
- Kong Q, Trugman DT, Ross ZE, Bianco MJ, Meade BJ, Gerstoft P (2019) Machine learning in seismology: turning data into insights. *Seismol Res Lett* 90(1):3–14. <https://doi.org/10.1785/0220180259>
- Li G, Qiao Y, Zheng Y, Li Y, Wu W (2019) Semi-supervised learning based on generative adversarial network and its applied to lithology recognition. *IEEE Access* 7:67428–67437. <https://doi.org/10.1109/ACCESS.2019.2918366>
- Liu S, Liang M, Hu X (2018) Particle swarm optimization inversion of magnetic data: field examples from iron ore deposits in China. *Geophysics* 83(4):J43–J59. <https://doi.org/10.1190/geo2017-0456.1>
- Menke W (2012) *Geophysical data analysis: discrete inverse theory*, 3rd edn. Academic Press. <https://doi.org/10.1016/C2011-0-69765-0>
- Metropolis N, Rosenbluth MN, Rosenbluth AW, Teller AH, Teller E (1953) Equation of state calculations by fast computing machines. *J Chem Phys* 21:1087–1092. <https://doi.org/10.1063/1.1699114>
- Puskarczyk E, Jarzyna JA, Wawrzyniak-Guz K et al (2019) Improved recognition of rock formation on the basis of well logging and laboratory experiments results using factor analysis. *Acta Geophys* 67:1809–1822. <https://doi.org/10.1007/s11600-019-00337-8>
- Sen MK, Stoffa PL (2013) *Global optimization methods in geophysical inversion*: Cambridge University Press, Cambridge. <https://doi.org/10.1017/CBO9780511997570>
- Shi Y, Eberhart R (1998) A modified particle swarm optimizer. In: *The 1998 IEEE international conference on IEEE world congress on computational intelligence evolutionary computation proceedings*, pp 69–73. <https://doi.org/10.1109/ICEC.1998.699146>
- Soupios P, Akca I, Mpogiatzis P, Basokur AT, Papazachos C (2011) Applications of hybrid genetic algorithms in seismic tomography. *J Appl Geophys* 75(3):479–489. <https://doi.org/10.1016/j.jappgeo.2011.08.005>
- Szabó NP (2011) Shale volume estimation based on the factor analysis of well-logging data. *Acta Geophys* 59:935–953. <https://doi.org/10.2478/s11600-011-0034-0>
- Szabó NP (2018) A genetic meta-algorithm-assisted inversion approach: hydrogeological study for the determination of volumetric rock properties and matrix and fluid parameters in unsaturated formations. *Hydrogeol J* 26:1935–1946. <https://doi.org/10.1007/s10040-018-1749-7>
- Szabó NP, Dobróka M, Drahos D (2012) Factor analysis of engineering geophysical sounding data for water saturation estimation in shallow formations. *Geophysics* 77(3):WA35–WA44. <https://doi.org/10.1190/geo2011-0265.1>
- Szabó NP, Dobróka M (2018) Exploratory factor analysis of wireline logs using a float-encoded genetic algorithm. *Math Geosci* 50:317–335. <https://doi.org/10.1007/s11004-017-9714-x>
- Szabó NP, Balogh GP, Stickel J (2018) Most frequent value-based factor analysis of direct-push logging data. *Geophys Prospect* 66(3):530–548. <https://doi.org/10.1111/1365-2478.12573>

- Szabó NP, Dobróka M (2020) Interval inversion as innovative well log interpretation tool for evaluating organic-rich shale formations. *J Petrol Sci Eng.* <https://doi.org/10.1016/j.petrol.2019.106696>
- Vu MT, Jardani A (2021) Convolutional neural networks with SegNet architecture applied to three-dimensional tomography of subsurface electrical resistivity: CNN-3D-ERT. *Geophys J Int* 225(2):1319–1331. <https://doi.org/10.1093/gji/ggab024>
- Wang Z, Di H, Shafiq MA, Alaudah Y, AlRegib G (2018) Successful leveraging of image processing and machine learning in seismic structural interpretation: a review. *Lead Edge* 37(6):451–461. <https://doi.org/10.1190/tle37060451.1>
- Wang R, Yin C, Wang M, Wang G (2012) Simulated annealing for controlled-source audio-frequency magnetotelluric data inversion. *Geophysics* 77(2):E127–E133. <https://doi.org/10.1190/geo2011-0106.1>
- Xing Z, Mazzotti A (2019) Two-grid full-waveform Rayleigh-wave inversion via a genetic algorithm—Part 1: method and synthetic examples. *Geophysics* 84(5):R805–R814. <https://doi.org/10.1190/geo2018-0799.1>
- Zhang L-P, Yu H-J, Hu S-X (2005) Optimal choice of parameters for particle swarm optimization. *J Zhejiang Univ Sci* 6:528–534. <https://doi.org/10.1631/jzus.2005.A0528>
- Zhdanov MS (2015) *Inverse theory and applications in geophysics*. Elsevier, ISBN 978-0-444-62674-5. <https://doi.org/10.1016/C2012-0-03334-0>