SELECTING CONTROL PARAMETERS FOR THE PARTICLE SWARM OPTIMIZATION BASED FACTOR ANALYSIS

ARMAND ABORDÁN[1,2], NORBERT PÉTER SZABÓ[1,2]
[1]*Department of Geophysics, University of Miskolc, 3515 Miskolc-Egyetemváros, Hungary*
[2]*MTA-ME Geoengineering Research Group, University of Miskolc, 3515 Miskolc-Egyetemváros, Hungary*

1. INTRODUCTION

In this paper, shale volume is estimated from well-logging data by means of factor analysis. This multivariate statistical method has already been successfully applied in numerous cases in geosciences. Bücker et al. (2000) evaluated lithology by factor analysis from logging while drilling data. Odokuma-Alonge and Adekoya (2013) applied it on a geochemical dataset to interpret stream sediments. Szabó (2011) showed that by regression analysis the first extracted factor from a well-logging data set can be connected to the shale volume of an unconsolidated shaly-sand formation. Similar methodology was used for water saturation estimation in shallow formations based on engineering geophysical sounding measurements (Szabó et al. 2012). In this study, factor analysis is solved by a metaheuristic approach called particle swarm optimization (PSO) developed by Kennedy and Eberhart (1995) to improve its searching performance. As a metaheuristic, its output is highly effected by the control parameters that worth to be set as optimal as possible at the start of the searching mechanism. These parameters are usually chosen according to empirical suggestions from literature, which makes a problem ambiguous. In an attempt to generalize the presented PSO based factor analysis, some control parameters are set by simulated annealing (SA), a global optimization technique introduced by Metropolis et al. (1953). Then the suggested method is applied on a wireline logging dataset measured in a thermal water well to estimate shale volume. Results are also confirmed by core data.

2. SHALE VOLUME ESTIMATION BY FACTOR ANALYSIS

In the framework of factor analysis, first, the K number of measured well logs are standardized and collected into a matrix $\mathbf{D}$, where each column represents a different logging tool and there is N number of rows representing the measured depth points along the borehole. The basis of factor analysis is the following decomposition of matrix $\mathbf{D}$

$$\mathbf{D} = \mathbf{F}\mathbf{L}^T + \mathbf{E}, \tag{1}$$

where **F** denotes the *N*-by-*M* matrix of factor scores and *M* is the number of extracted factors. **L** represents the *K*-by-*M* matrix of factor loadings and **E** denotes the *N*-by-*K* error matrix. Based on Eq. (1), the measured well logs are derived as the linear combination of the extracted factors. The factor loadings quantify the correlation relationship between the measured data and the extracted factors. Most of the data variance is represented by the first factor log, which is the first column of the matrix **F**. The factor loadings are estimated by a non-iterative estimation method suggested by Jöreskog (2007)

$$\mathbf{L} = \left(diag\mathbf{S}^{-1}\right)^{-1/2} \mathbf{\Omega}\left(\mathbf{\Gamma} - \theta\mathbf{I}\right)^{1/2} \mathbf{U}, \tag{2}$$

where $\mathbf{\Gamma}$ denotes the diagonal matrix of the first *M* number of sorted eigenvalues of the sample covariance matrix **S**, $\mathbf{\Omega}$ is the matrix of the first *M* number of eigenvectors and **U** is an arbitrarily chosen *M*-by-*M* orthogonal matrix.

In this study, factor scores are estimated by the algorithm of PSO (Abordán and Szabó, 2018) for this reason, we have to rearrange the model of factor analysis defined in Eq. (1) to

$$\mathbf{d} = \tilde{\mathbf{L}}\mathbf{f} + \mathbf{e}, \tag{3}$$

where **d** denotes the *KN* length vector of measured (standardized) data, $\tilde{\mathbf{L}}$ is the *NK*-by-*NM* matrix of factor loadings, **f** is the *MN* length vector of factor scores and **e** is the *KN* length vector of residuals. At first, all data are put into a column vector, then the matrix $\tilde{\mathbf{L}}$ is estimated by Eq. (2) and then rotated with the varimax algorithm developed by Kaiser (1958) for easier interpretation. Then the vector of factor scores **f** is estimated by the PSO algorithm. For finding the optimal values of the factor scores, an energy function has to be defined, the minimization of which leads to the optimal solution. The objective function is based on the square of the L_2 norm as

$$E = \frac{1}{NK}\sum_{i=1}^{NK}\left(\mathbf{d}_i^{(m)} - \mathbf{d}_i^{(c)}\right)^2 = \min, \tag{4}$$

where $\mathbf{d}^{(m)}$ and $\mathbf{d}^{(c)}$ denote the measured and calculated (standardized) well-logging data vectors, respectively. In Eq. (3) $\tilde{\mathbf{L}}\mathbf{f}$ represents the calculated data and **D** denotes the measured data. This permits the estimation of the theoretical values of well logs, which can be considered as the solution of the forward problem. Factor loadings are fixed at the start of the procedure, and only the factor scores are updated to save CPU time. Then the factor scores are estimated by the algorithm of PSO. This metaheuristic is based on the social behavior of animals like birds or fish. This technique is especially effective for large search domains. The search mechanism utilizes a swarm of particles to find the optimal solution, which solution is found by using Eq. 5 and 6

$$\mathbf{x}_i(\mathbf{t}+1) = \mathbf{x}_i(\mathbf{t}) + \mathbf{v}_i(\mathbf{t}+1), \tag{5}$$

$$\mathbf{v}_i(\mathrm{t}+1) = w\mathbf{v}_i(\mathrm{t}) + r_1c_1(\mathbf{p}_i(\mathrm{t}) - \mathbf{x}_i(\mathrm{t})) + r_2c_2(\mathbf{g}(\mathrm{t}) - \mathbf{x}_i(\mathrm{t})), \tag{6}$$

where the position of the i-th particle is denoted by $\mathbf{x}_i$, $i=1,2,\ldots,S$ is the particle index, S is the size of the swarm, $t=1,\ldots,T$ is the current iteration step and T is the last iteration step. The velocity of the particle is $\mathbf{v}_i$ and its best position in the search space is denoted by $\mathbf{p}_i$ and $\mathbf{g}$ contains the very best position found by all the particles until the given iteration step. Control parameters c_1 and c_2 are set at the start, c_1 is the so-called cognitive scaling parameter, it defines the movement of particles regarding their own best position and c_2 is the social scaling parameter, which controls the movement of particles in the direction of the global best position found by the swarm. In this study, we treat these control parameters as unknowns during the process of factor analysis and determine them automatically by the improved data processing method. Constants r_1 and r_2 are uniformly distributed random numbers from the range of 0 and 1 and w is an inertia weight introduced by Shi and Eberhart (1998) to increase the performance of the algorithm.

Each particle represents a solution of the optimization problem, meaning that with an n-dimensional search space, each particle has n number of elements. In the current study, the unknowns are the factor scores $\mathbf{f}$ in Eq. (3). Once the optimum is found, the first extracted factor can be related to shale volume of the investigated interval by regression analysis.

3. Parameter selection for the Particle swarm optimization based factor analysis and results

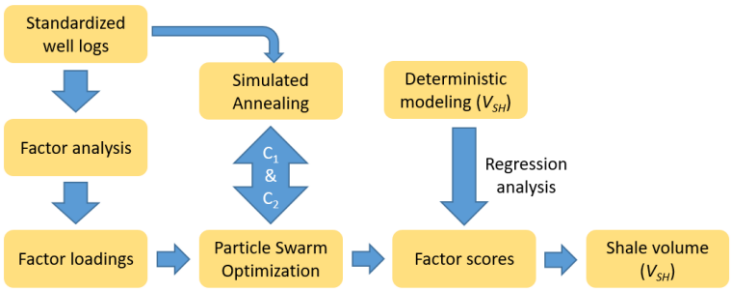

Figure 1: Workflow of the procedure

The utilized wireline logging dataset consists of natural gamma-ray intensity (GR), spontaneous potential (SP), shallow resistivity (RS), density (GG) and neutron-porosity (NN) logs measured in an East-Hungarian thermal water well. The investigated interval is from 448.5 m to 486.4 m, where shaly-sand layers are located. The workflow of the procedure can be seen in Figure 1. As a first step, factor loads are calculated for three factors by Eq. (2). For the first factor, loads are: $L^{(GR)}=0.75$, $L^{(SP)}=0.60$, $L^{(RS)}=-0.37$, $L^{(GG)}=0.14$, $L^{(NN)}=-0.13$. Then Eq. (4) is minimized by PSO. In this study, 250 particles are generated in the search space, each representing a solution for the factor scores $\mathbf{f}$, by extracting three factors, the number of unknowns to be estimated is (3×380 measured depth point) 1140. The inertia weight w is set

in each iteration step according to $w=w \cdot w_{damp}$, where w is set at the start of the algorithm to 3 and w_{damp} is a damping factor set to 0.99. Then the values of c_1 and c_2 are optimized by SA. The usually recommended choice is 2 for both parameters (Kennedy and Eberhart, 1995). To test the presented method, at the beginning of the SA procedure, both c_1 and c_2 are set as 1 and in every iteration step their value is slightly altered by adding a small b perturbation parameter to both values. Then with these control parameters, the PSO based factor analysis is run for a thousand iteration steps. If the energy difference in two subsequent iterations (ΔE) according to Eq. (4) is negative, then the new values of c_1 and c_2 are accepted and the procedure is continued. If the energy difference is greater than 0, then the probability of accepting the new control parameters is given by $P_a=\exp(-\Delta E/T)$, where T is the current temperature of the system. During the process, the temperature of the system is reduced according to $T^{(new)}=T^{(old)}/log(1+q)$, q denotes the number of previously computed iterations. The new control parameters are accepted only if a randomly generated number from the range of 0 and 1 is smaller than P_a. This is a fundamental feature of the SA algorithm; it prevents the search from being stuck in a local minimum.

These steps were repeated in 85 iteration steps to find the optimal values of c_1 and c_2 for the PSO based factor analysis. The convergence of the search can be seen in Figure 2.

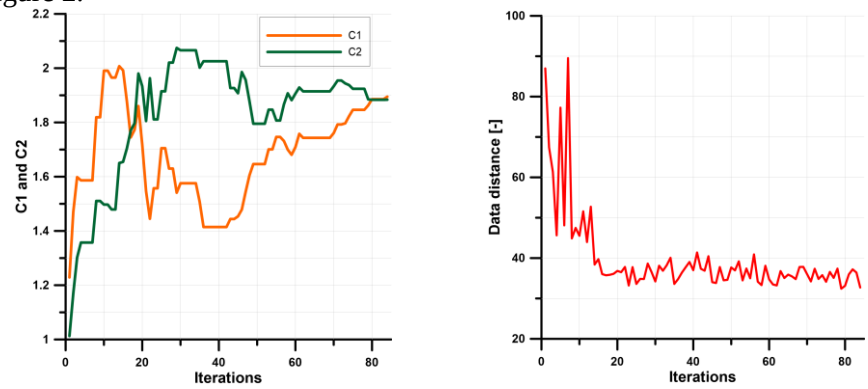

Figure 2: Convergence of data distance (on the right) by altering control parameters c_1 and c_2 (on the left)

In the last iteration step, the smallest data distance was found by using c_1=1,89 and c_2=1,88. Then the PSO based factor analysis was run for twenty thousand iteration steps with these parameters to find the optimal solution of the factor scores **f.** The convergence of data distance can be seen in Figure 3. After twenty thousand iterations the data distance was 17,4 [-]. By regression analysis between the first extracted factor (F_1) scaled in the range of 0 to 100 and the shale volume estimated by deterministic modeling resulted in a linear relationship (Figure 3) in the form of

$$V_{sh} = aF_1 + b\,, \tag{7}$$

where the regression coefficients with 95% confidence bounds are a=0.497 [a_{min}=0.472, a_{max}=0.523] and b=30.6 [b_{min}= 29.5, b_{max}= 31.71].

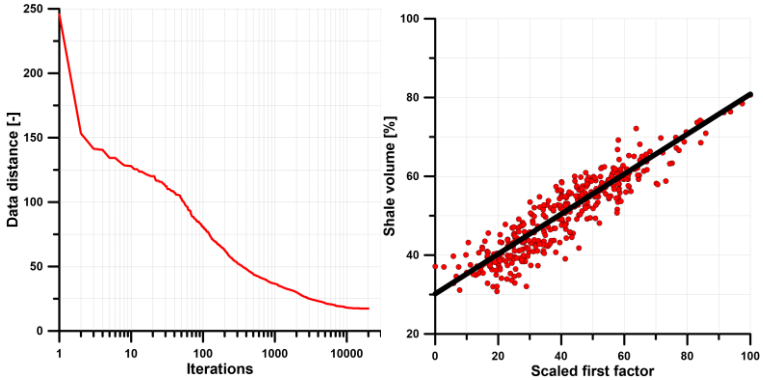

Figure 3: Convergence of the energy function (left) for FA-PSO and the relation between the first factor and shale volume (right)

The results of the procedure can be seen in Figure 4.

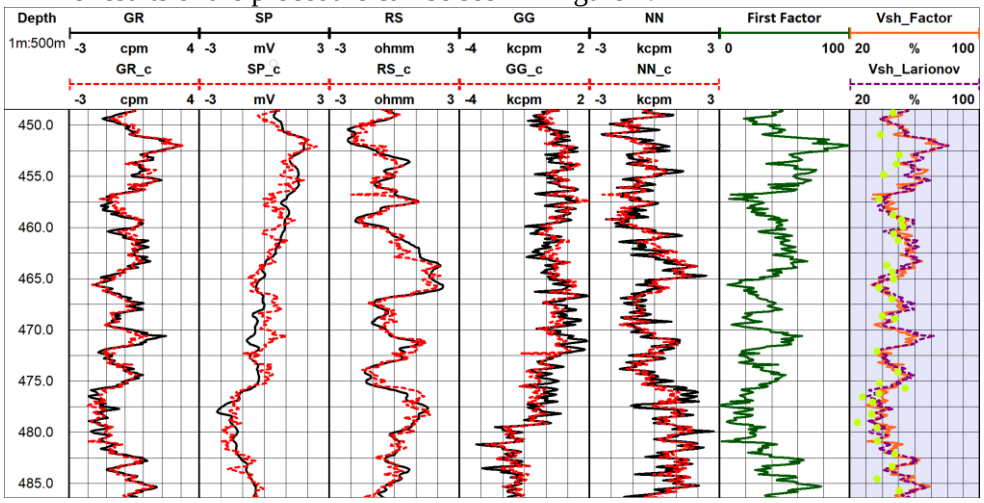

Figure 4: The input (black) and calculated well logs (red dashed line) in the first five track, extracted first factor log in the fifth, and the resultant shale volumes in the sixth track.

In the first five tracks, the standardized input well logs can be seen in black and the calculated logs from the factor model by red dashed line. The fit between the measurements and predictions is relatively good. In the sixth track, the scaled first factor log is shown in green and the last track contains the resultant shale volumes. Shale volume estimated by deterministic modeling (Larionov, 1969) is drawn by a purple dashed line and the one estimated by factor analysis is shown in orange. They match really well, which verifies the applicability of the method. The latter is also

confirmed by laboratory measurements for shale volumes, which are indicated by green dots in the same track.

4. CONCLUSIONS

In this paper, shale volume was estimated by an improved particle swarm optimization based factor analysis. For the generalization of the presented method, selection of c_1 and c_2 control parameters of PSO was done by an automated simulated annealing-based iterative procedure known commonly as a hyperparameter estimation approach in the terminology of machine learning. By regression analysis, a linear relationship was found between the first extracted factor and shale volume, which forms a basis of the estimation of the shaliness of the investigated formations from well logs. The results indicate the applicability of the method; however, further investigations are needed to increase the efficiency of the presented workflow.

5. ACKNOWLEDGEMENTS

The research was carried out within the GINOP-2.3.2-15-2016-00031 "Innovative solutions for sustainable groundwater resource management" project of the Faculty of Earth Science and Engineering of the University of Miskolc in the framework of the Széchenyi 2020 Plan, funded by the European Union, co-financed by the European Structural and Investment Funds.

6. REFERENCES

ABORDÁN A, SZABÓ N P. (2018) Particle swarm optimization assisted factor analysis for shale volume estimation in groundwater formations. Geosciences and Engineering: A publication of the University of Miskolc 6: (6) pp. 87-97.

BÜCKER C, SHIMELD J, HUNZE S, BRÜCKMANN W (2000) 2. Data Report: Logging while drilling data analysis of Leg 171A, a multivariate statistical approach, Proc. Ocean. Drill. Prog. Sci. Results., Scientific Results Vol. 171A.

JÖRESKOG K G (2007) Factor analysis and its extensions. In: Cudeck R, MacCallum RC (eds) Factor analysis at 100: historical developments and future directions. Erlbaum, Mahwah, pp 47–77.

KAISER H F (1958) The varimax criterion for analytical rotation in factor analysis: Psychometrika, 23, 187–200.

KENNEDY J, EBERHART R C (1995) Particle swarm optimization Proceedings of IEEE international conference on neural networks, 4:1942–1948.

LARIONOV V V (1969) Radiometry of boreholes (in Russian) Nedra, Moscow.

METROPOLIS N, ROSENBLUTH MN, TELLER AH, TELLER E (1953) Equation of State calculations by fast computing machines. J Chem Phys 21:1087–1092

ODOKUMA-ALONGE O, ADEKOYA J (2013) factor analysis of stream sediment geochemical data from Onyami drainage system, Southwestern Nigeria. Int J Geosci 4(3):656–661

Szabó N P. (2011) Shale volume estimation based on the factor analysis of well-logging data. Acta Geophys 59:935–953

Szabó N P., Dobróka M., Drahos D. 2012. Factor analysis of engineering geophysical sounding data for water saturation estimation in shallow formations. Geophysics **77**, (3), WA35–WA44.

SHI Y. & EBERHART R A.: modified particle swarm optimizer, Evolutionary Computation Proceedings. IEEE World Congress on Computational Intelligence., The 1998 IEEE International Conference on, 1998, 69–73.