INVASIVE WEED OPTIMIZATION DRIVEN FACTOR ANALYSIS FOR LITHOLOGICAL CHARACTERIZATION OF AQUIFERS

***Armand Abordán*[1,2]**
[1]PhD student, [2]Assistant research fellow
[1]Department of Geophysics, University of Miskolc, 3515 Miskolc-Egyetemváros, Hungary
[2]MTA-ME Geoengineering Research Group, University of Miskolc, 3515 Miskolc-Egyetemváros, Hungary

ABSTRACT

A globally optimized factor analysis is presented to estimate shale volume of groundwater formations. Wireline log data, including natural gamma-ray intensity (GR), spontaneous potential (SP), density (GG), neutron-neutron intensity (NN) and shallow resistivity (RS) logs are processed simultaneously to estimate the values of factor scores along a section of the investigated borehole. Then the so-called first factor log is related to the shale volume of formations by regression analysis. In the investigated borehole, a strong linear relationship is found between the first factor log and the shale volume. Invasive weed optimization is implemented into factor analysis to reduce the misfit between the measured and theoretical data calculated by the factor model and it also permits the estimation of theoretical values of well logging data. The method provides an independent estimate of shale volume, which can be later used in the further interpretation of data.

ACKNOWLEDGMENT

The described study was carried out as part of the EFOP-3.6.1-16-2016-00011 "Younger and Renewing University – Innovative Knowledge City – institutional development of the University of Miskolc aiming at intelligent specialisation" project implemented in the framework of the Szechenyi 2020 program. The realization of this project is supported by the European Union, co-financed by the European Social Fund.
The author is also grateful for the Geokomplex Kft. for the data used in this study.

INTRODUCTION

As the demand for freshwater is increasing, new approaches are being developed to more accurately model groundwater formations. As a statistical tool, factor analysis is capable to extract unobserved variables from measured data. Applied on well logging data, it can provide so-called factor logs that can be associated with petrophysical parameters by regression analysis [1]. Simone et al. suggested factor analysis for the investigation of ambiguity in geophysics [2]. Szabó et al. applied it on direct-push logging data for estimating water content in unconsolidated heterogeneous formations [3]. Xiang et al. used it for regional geochemical pattern recognition [4].

Global optimization methods are also widely used in earth sciences, such as the genetic algorithm or simulated annealing. It was shown that using the genetic algorithm, spontaneous potential data can be effectively inverted [5]. Yin and Hodges applied simulated annealing for the inversion of airborne electromagnetic data and it was shown that it greatly reduces the start model dependence of the inversion procedure [6].
In this paper, factor analysis is combined with invasive weed optimization (IWO), a global optimization approach that was introduces by Mehrabian and Lucas [7]. The method is suggested to more accurately calculate the factor scores and to improve the fit between the observed and calculated well logs.

INVASIVE WEED OPTIMIZATION CONTROLLED FACTOR ANALYSIS

IWO is inspired by the colonization of invasive weeds. In general, biological invasion is a phenomenon where groups of individuals migrate to new environments and challenge the native population. This phenomenon can be used as a framework for the design of optimization algorithms [8]. IWO has already been successfully applied to a variety of optimization problems [9][10][11].
For the formulation of the optimization algorithm the following steps need to be taken. First, we have to initialize a population of weeds, which are spread over the search space randomly, each representing a solution of the optimization problem. Then each weed of the population is allowed to produce seeds based on its own and based on the worst and best cost of the population to mimic the mechanism of natural selection. The number of produced seeds increases linearly, the weed with the worst cost produces the pre-defined minimal number of seeds and the one with the best cost value produces the pre-defined maximal number of seeds. The generated seeds are randomly distributed over the search space by normally distributed random numbers with mean equal to zero but with varying variance. This guarantees that the produced seeds will be generated near the parent weed. However, the standard deviation of the function is reduced according to Eq. 1 during the iteration process to ensure that the probability of generating a seed in a distant area decreases nonlinearly. This condition ensures that plants with better cost values are grouped and those with worse cost values are eliminated over time

$$\sigma_{iter} = \frac{(q_{\max} - q)^n}{q_{\max}^n}(\sigma_{initial} - \sigma_{final}) + \sigma_{final}, \quad (1)$$

where q_{max} is the maximum number of iterations, σ_{iter} is the current value of the standard deviation, q is the current iteration step, n is the nonlinear modulation index, $\sigma_{initial}$ is the pre-defined initial value of standard deviation and σ_{final} is the value of standard deviation at the end of the iteration process.
There is also a need for competition between plants to limit their number. After some iterations, the number of plants will reach a user-defined maximum number, $P_{max.}$ When P_{max} is reached each plant is allowed to produce seeds as detailed above and then the whole population is ranked based on their costs. Those with better cost values

survive and are allowed to reproduce in the next iteration step. This process continues until the user-defined maximum number of iterations is reached and the weed with the best cost value is accepted as the optimal solution. In this study, IWO is combined with factor analysis, which is a statistical tool for describing several measured quantities with fewer unobserved variables. In case of wireline logging, the measured well logs are the input variables, which are simultaneously processed to extract factors. Here, the extracted factors can be also seen as factor logs, which can be connected to petrophysical parameters through regression analysis [12]. In this paper, we derive the shale volume based on the first factor log in a Hungarian aquifer. First, wireline logging data are standardized and collected into the data matrix **D**, where every column represents the measured data of a given logging tool

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

where K represents the number of different logging tools and N is the number of measured depths in the given borehole. Factor analysis is realized by the following decomposition of matrix **D**

$$\mathbf{D}=\mathbf{F}\mathbf{L}^{\mathrm{T}}+\mathbf{E}, \tag{3}$$

where **F** is the N-by-M matrix of factor scores, **L** denotes the K-by-M matrix of factor loadings and **E** is the N-by-K error matrix. As it can be seen in Eq. 3, the observed variables are derived as linear combinations of extracted factors. The factor loadings describe the correlation between the observed variables and the factors. The largest part of data variance is described by the first column of matrix **F**, which is the so-called first factor log. The values of the factor loadings can be determined by a non-iterative approach suggested by Jöreskog [13]

$$\mathbf{L}=\left(diag\mathbf{S}^{-1}\right)^{-1/2}\boldsymbol{\Omega}\left(\boldsymbol{\Gamma}-\theta\mathbf{I}\right)^{1/2}\mathbf{U}, \tag{4}$$

where $\boldsymbol{\Gamma}$ is the diagonal matrix of the first M number of sorted eigenvalues of the sample covariance matrix **S**, $\boldsymbol{\Omega}$ denotes the matrix of the first M number of eigenvectors and **U** is an arbitrary M-by-M orthogonal matrix.
The factor scores generally are estimated by a maximum likelihood method suggested by Bartlett [14]

$$\mathbf{F}^{\mathrm{T}}=\left(\mathbf{L}^{\mathrm{T}}\boldsymbol{\Psi}^{-1}\mathbf{L}\right)^{-1}\mathbf{L}^{\mathrm{T}}\boldsymbol{\Psi}^{-1}\mathbf{D}^{\mathrm{T}}. \tag{5}$$

The traditional method of factor analysis is modified such that the factor scores are estimated by the IWO algorithm instead of Eq. 5. The developed method is called FA-IWO. In the first step of this optimization problem, the model of factor analysis defined in Eq. 3 is rearranged

$$\mathbf{d} = \tilde{\mathbf{L}}\mathbf{f} + \mathbf{e}, \tag{6}$$

where **d** denotes the *KN* length vector of standardized measured data, $\tilde{\mathbf{L}}$ represents the *NK*-by-*NM* matrix of factor loadings, **f** is the *MN* length vector of factor scores and **e** is the *KN* length residual vector. First, all data are collected into a column vector, then the matrix of factor loadings $\tilde{\mathbf{L}}$ is estimated by Eq. 4 and then rotated by the varimax algorithm suggested by Kaiser [15] for easier interpretation. The values of matrix $\tilde{\mathbf{L}}$ are then fixed for the whole procedure, while the vector of factor scores **f** is estimated by the IWO algorithm. To solve this inverse problem, an objective function is needed, which is then minimized to find the optimal solution. In this study, the objective function is defined based on the L_2 norm as

$$E = \frac{1}{NK}\sum_{i=1}^{NK}\left(\mathrm{d}_i^{(m)} - \mathrm{d}_i^{(c)}\right)^2 = \min, \tag{7}$$

where $\mathbf{d}^{(m)}$ and $\mathbf{d}^{(c)}$ are the standardized vectors of measured and calculated well logging data, respectively. In Eq. 6, the term $\tilde{\mathbf{L}}\mathbf{f}$ denotes the calculated data and **D** denotes the measured data. This multiplication of the factor loadings and factor scores permits the estimation of the theoretical values of well logs, which can be considered as the solution of the forward problem.

For finding the optimal values of matrix **f**, a population of weeds is generated with uniform distribution in the search space of the optimization problem. The size of the search space is estimated by solving Eq. 5. Then in each iteration step new seeds are generated and the whole population is evaluated based on Eq. 7 and the ones with the worst costs are eliminated to meet the criteria of population limit, $P_{max.}$ At the end of the iteration process, the member of the population with the best cost is accepted as the optimal solution of the problem.

Then the first factor extracted by the FA-IWO algorithm is directly used for the calculation of shale volume. It was previously published that for intervals up to 150 m, shale volume V_{sh} (in percent) correlates linearly to the first factor log (F_1) scaled into the range of 0 and 100 [12] [16]

$$V_{sh} = aF_1 + b, \tag{8}$$

where *a* is the slope, and *b* is the intercept of the regression line.

FIELD EXAMPLE

The suggested method is tested in a Hungarian thermal water exploratory well located in Baktalórántháza. The depth of the well is 1197 m and for this study data is taken from the interval of 100 m to 193 m. In the upper part of the well, upper Pleistocene aquifers are located with varying grain sizes. The horizontal porous layers are separated by shales. Between 100 m and 160 m, mainly sandy layers can be found, and below that thick coarse-grained beds with thickness of 10 to 15 m are deposited. The measured well logs include the natural gamma-ray intensity (GR), spontaneous potential (SP), density (GG), shallow resistivity (RS) and neutron-neutron intensity log (NN) logs, which serve as the input variables of the suggested method. At first, the values of factor loadings are estimated by Eq. 4 and can be seen in Table 1.

Table 1
Rotated factor loadings estimated by the FA-IWO method

Well logs	***Factor 1***	***Factor 2***	***Factor 3***
GR	0,7515	-0,1836	0,0127
SP	-0,0380	-0,7396	-0,0893
NN	0,0580	-0,1184	-0,5976
GG	-0,5930	-0,0948	0,3062
RS	-0,4955	0,6537	-0,0081

As seen above, three factors were extracted from the data set, the first of which is usually a lithological indicator in case of wireline log data [1]. In this case, where a groundwater formation is investigated, the natural gamma-ray intensity has the highest load (0,75) on the first resultant factor and both the density (-0,59) and shallow resistivity logs (-0,50) have moderate influence on it.
Having the factor loadings, in the next step, the factor scores can be estimated by the algorithm of IWO based on the objective function defined in Eq. 7. First, the control parameters of the algorithm need to be set, an initial population of 10 weeds is generated, and the maximal number of the population is limited at 25. The plant with the worst cost value is allowed to produce 1 seed per iteration step, and in a linearly increasing manner, the plant with the best cost produces 6 seeds. In Eq. 2 that controls the placement of the generated seeds, $\sigma_{initial}$ is set to 0.1 and σ_{final} is 0.001 and the *n*, nonlinear modulation index is 2. With these parameters the algorithm runs twenty thousand iterations and then the member of the population with the best cost is accepted as the optimal solution. Fig. 1 shows the decrease of the value of the objective function during the iteration process on the left and the value of the standard deviation according to Eq. 1 on the right. It can be seen that the algorithm is highly stable during the iteration process, and even starting from a very distant solution it was able to find the optimal solution. The objective function reached its minimum after about twenty thousand iterations under 5 minutes on a quad-core workstation.

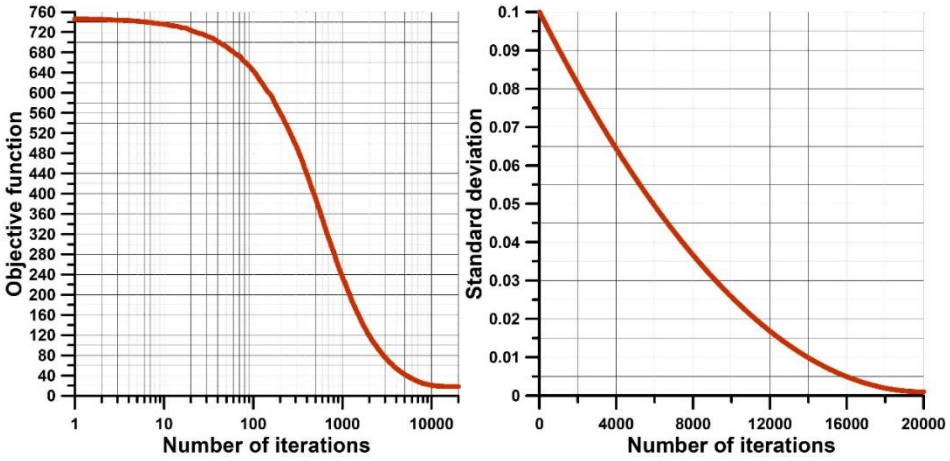

Fig. 1
Decrease of the objective function with the iteration steps on the left, value of standard deviation with the iteration steps on the right

Fig. 2 shows the regression relation between the scaled first factor and the shale volume estimated by the FA-IWO method using Eq. 8. The regression analysis revealed a strong linear connection between the first factor and the shale volume.

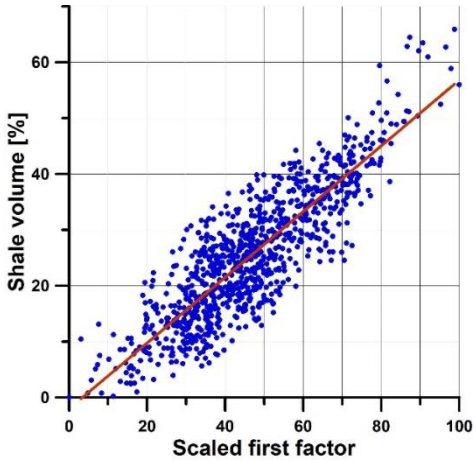

Fig. 2
Regression relation between the scaled first factor and the shale volume

In this example, the regression coefficients were found to be a=0,5879 [a_{min}=0.5646, a_{max}=0.6113] and b=–2,122 [b_{min}= -3.276, b_{max}= -0.9687] with 95% confidence bounds. The measured well logs and the ones calculated by the FA-IWO method is shown in Fig. 3. On tracks 1 to 5, the standardized input well logs (black solid line) and the calculated logs (red dotted line) are shown. The calculated logs are derived from Eq. 6, by multiplying the factor loadings with the factor scores. One can see that the misfit between the observed and calculated logs is quite small. Track 6 represents the first factor log scaled into the range of 0 to 100. On the last track, the shale volume determined by the suggested FA-IWO method (red dotted line) is compared to deterministic modeling [17] (black solid line) that was done based on the GR log and to core data (blue dots).

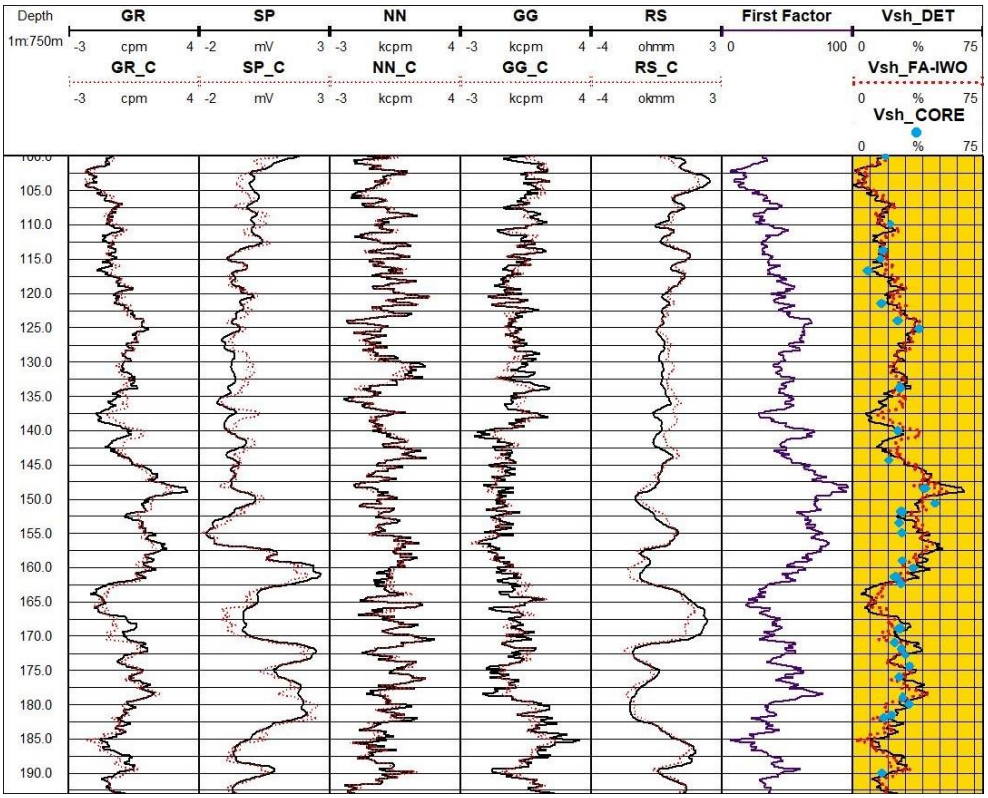

Fig. 3
The results of the FA-IWO algorithm

It can be seen that the derived shale volumes agree well with each other, which indicates that the FA-IWO method can be a useful tool for shale volume estimation in aquifers.

CONCLUSION

It is shown that the presented globally optimized factor analysis is capable of reliable shale volume estimation in ground water formations. The results of the presented FA-IWO method are compared and verified by gamma-ray log based deterministic modeling and core data. In the investigated section of the well, the study revealed a strong linear relation between the first extracted factor and the shale volume calculated by deterministic modeling. An added advantage of the implementation of the invasive weed optimization into factor analysis, that it permits the calculation of the theoretical values of well logs without any preliminary knowledge about the investigated area. The method delivers the results within 5 minutes and is also very reliable in the iteration process. It delivers an independent estimate for shale volume that can be later used during interpretation and also in further inversion procedures for increasing overdetermination and thus improving their estimations.

REFERENCES

[1] SZABÓ N, DOBRÓKA M.: **Exploratory Factor Analysis of Wireline Logs Using a Float-Encoded Genetic Algorithm**. Mathematical Geosciences 2018. 50:(3) pp. 317-335.

[2] SIMONE G. C. FRAIHA AND JOÃO B. C. SILVA.: **Factor analysis of ambiguity in geophysics**. Geophysics, 1994. 59(7), 1083-1091.
[3] SZABÓ N. P., BALOGH G. P., STICKEL J.: **Most frequent value-based factor analysis of direct-push logging data**. Geophysical Prospecting, 2018.Vol. 66, No. 3, pp. 530–548.
[4] SUN X., DENG J., GONG Q., WANG Q., YANG L., ZHAO Z.: **Kohonen neural network and factor analysis based approach to geochemical data pattern recognition.** Journal of Geochemical Exploration, 103 2009. 6–16.
[5] ABDELAZEEM M., GOBASHY M.: **Self-Potential Inversion Using Genetic Algorithm.** Journal of King Abdulaziz University, 2006. 17(1):83-101.
[6] YIN C. AND HODGES G.: **Simulated annealing for airborne EM inversion.** GEOPHYSICS, 2007. 72(4), F189-F195.
[7] MEHRABIAN, A. R., & LUCAS, C.: **A novel numerical optimization algorithm inspired from weed colonization**. Ecological Informatics, 2006. 1, 355–366.
[8] FALCO, I. D., CIOPPA, A. D., MAISTO, D., SCAFURI, U., & TARANTINO, E.: **Biological invasion–inspired migration in distributed evolutionary algorithms**. Information Sciences, 2012. 207, 50–65.
[9] GHOSH, A., DAS, S., CHOWDHURY, A., & GIRI, R.: **An ecologically inspired direct search method for solving optimal control problems with Bézier parameterization**. Engineering Applications of Artificial Intelligence, 2011. 24, 1195–1203.
[10] GIRI, R., CHOWDHURY, A., GHOSH, A., DAS, S., ABRAHAM, A., & SNASEL, V.: **A modified invasive weed optimization algorithm for training of feed-forward neural networks**. In IEEE International Conference on Systems, Man, and Cybernetics (IEEE SMC), Istanbul 2010. 3166–3173.
[11] MALLAHZADEH, A. R., & TAGHIKHANI, P.: **Shaped elevation pattern synthesis for reflector antenna**. Electromagnetics, 2013. 33, 40–50.
[12] SZABÓ N P.: **Shale volume estimation based on the factor analysis of well-logging data**. Acta Geophys., 2011. 59, 935-953.
[13] JÖRESKOG KG.: **Factor analysis and its extensions**. In: Cudeck R, MacCallum RC (eds) Factor analysis at 100: historical developments and future directions2007. Erlbaum, Mahwah, pp 47–77.
[14] BARTLETT, M S.: **The statistical conception of mental factors**. Br. J. Psychol., 1937. 28, 97-104.
[15] KAISER H F.: **The varimax criterion for analytical rotation in factor analysis**: Psychometrika, 1958. 23, 187–200.
[16] SZABÓ N P, DOBRÓKA M.: **Geostatistical Approach for Shale Volume Estimation in Water-bearing Formations**, Near Surface, 17th European Meeting of Environmental and Engineering Geophysics, Leicester, UK, 12-14 September 2011.
[17] LARIONOV V. V.: **Radiometry of boreholes** (in Russian). 1969. Nedra, Moscow.