

Metropolis algorithm driven factor analysis for lithological characterization of shallow marine sediments

A. Abordán^{1,2} · N. P. Szabó^{1,2}

Received: 21 July 2017 / Accepted: 14 December 2017 / Published online: 6 January 2018
© Akadémiai Kiadó 2018

Abstract Factor analysis of well logging data can be effectively applied to calculate shale volume in hydrocarbon formations. A global optimization approach is developed to improve the result of traditional factor analysis by reducing the misfit between the observed well logs and theoretical data calculated by the factor model. Formation shaliness is directly calculated from the factor scores by a nonlinear regression relation, which is consistent in the studied area in Alaska, USA. The added advantage of the implementation of the Simulated Annealing method is the estimation of the theoretical values of nuclear, sonic, electrical as well as caliper well-logging data. The results of globally optimized factor analysis are compared and verified by independent estimates of self-potential log-based deterministic modeling. The suggested method is tested in two different shaly-sand formations in the North Aleutian Basin of Alaska and the comparative study shows that the assumed nonlinear connection between the factor scores and shale volume is applicable with the same regression constants in different burial depths. The study shows that factor analysis solved by the random search technique provides an independent in situ estimate to shale content along arbitrary depth intervals of a borehole, which may improve the geological model of the hydrocarbon structure in the investigated area.

Keywords Factor analysis · Simulated Annealing · Shale volume · Global optimization · Metropolis criterion

✉ A. Abordán
gfaa@uni-miskolc.hu

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

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

1 Introduction

Multivariate statistical methods act as a powerful tool of lithology determination in hydrocarbon exploration. Factor analysis is capable of extracting unobserved, latent petrophysical information from well log data. By correlating the calculated factors to petrophysical parameters, new approaches are being developed for the better modeling of hydrocarbon reserves. Bücken et al. (2000) suggested the use of factor analysis for a fast and objective evaluation of lithology from logging while drilling data. Odokuma-Alonge and Adekoya (2013) applied the same statistical method for the interpretation of stream sediments using geochemical data. Coimbra et al. (2017) used Principal Component Analysis as a related method to calibrate and contrast stratigraphic patterns in bulk elemental abundances. Szabó (2011) used factor analysis for shale volume calculation in unconsolidated shaly-sand hydrocarbon formations. Based on this study, Szabó and Dobróka (2013) found a strong exponential relation between a statistical factor explaining the most part of variance of the input well-logging data and the shale content for different North-American and Hungarian wells. The applicability of factor analysis was also indicated by Asfahani (2014) for the classification of Syrian basaltic formations. Szabó et al. (2014) confirmed for the results of shale volume estimation by core data.

The traditional method of factor analysis normally solves a maximum likelihood problem to extract the statistical factors (Beauducel et al. 2016). In this study, we offer an alternative approach, which treats the factor analysis as an inverse problem. In the frame of the inversion-based factor analysis, the observed well logs are approximated by calculated data directly computed from the factors, and the misfit between them is continuously optimized in an iterative process. For the computer-based minimization process, one of the most favorable methods are the global optimization methods. Sen and Stoffa (1995) provided a detailed description on the applicability of global optimization techniques used in geophysics including Simulated Annealing (SA). Soupios et al. (2011) utilized the Genetic Algorithms as a similar method to SA for seismic travel time inversion. Shaw and Srivastava (2007) successfully applied Particle Swarm Optimization as a third group of global optimization methods for inverting direct current, induced polarization and magnetotelluric data. The Simulated Annealing algorithm is an adaptation of the Metropolis algorithm (Metropolis et al. 1953) that is based on the annealing of solids in a heat bath. Analogously to the metallurgical process, SA is often used to solve multivariate optimization problems. Yin and Hodges (2007) suggested the use of SA for the inversion of electromagnetic data to eliminate the strongly start model dependent nature of traditional linearized inversion algorithms, e.g. the Marquardt–Levenberg method (Marquardt 1963). Szűcs et al. (2006) proved that in the characterization of groundwater formations, the geophysical inverse problem could be effectively solved by the Very Fast Simulated Annealing method. In this study, factor analysis is combined with SA to calculate the statistical factors more accurately and improve the fit between the measured and calculated well logs. By choosing suitable control parameters, SA allows the estimation of the factor variables in a convergent iterative procedure.

2 Simulated Annealing assisted factor analysis

SA is a random search technique used for approximating the global optimum of an objective function. The SA algorithm is based on the analogy between the simulation of the annealing of solids and the problem of solving multivariate optimization problems (van

Laarhoven and Aarts 1987). Annealing denotes a physical process in which a solid is heated up by increasing the temperature to a maximum value. At this stage, all particles of the solid randomly arrange themselves in the liquid phase. Then, a slow cooling is carried out by slowly lowering the temperature. In this way, all particles organize themselves in the low energy ground state of a corresponding lattice, provided the maximum temperature is sufficiently high and the cooling is sufficiently slow. If the cooling schedule is too rapid, then the solid is not allowed to reach thermal equilibrium at a given temperature, and defects can be frozen into the solid and a metastable amorphous structure is reached (in our point of view a local minimum) rather than the lowest possible energy crystalline lattice structure (global minimum). Since the number of atoms are in the order of 10^{23} , there are many possibilities for reaching an imperfect structure, in other words, a local minimum, but there is only one global minimum (perfect crystalline lattice structure). The analogy with geophysical inversion problems is obvious as we have a great number of measured and calculated data and we look for the absolute minimum of an objective function related to the data misfit with many local minima.

Factor analysis is generally used to describe several measured quantities with potentially fewer unobserved variables. In well log analysis, the measured well logs represent the input variables, which are simultaneously processed to extract the statistical factors. The scores of a given factor plotted against depth called a factor log can be associated with petrophysical parameters by regression analysis. In this study, we derive the shale volume from the factor logs in different sedimentary formations deposited along the Alaska Peninsula, USA.

In the first step, open-hole wireline logging data are standardized and put into the data matrix \mathbf{D} , in such a way that each column contains data of different logging tools

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

where K is the number of different types of well logs and N is the total number of observed depths. The basis of factor analysis is the following decomposition of matrix \mathbf{D}

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

where \mathbf{F} denotes the N -by- M matrix of factor scores, \mathbf{L} represents the K -by- M matrix of factor loadings and \mathbf{E} denotes the N -by- K residual matrix. Based on Eq. (2), the observed variables are derived as the linear combination of the factors. The factor loadings quantify the correlation relation between the observed variables and the extracted factors. The largest part of the total data variance is represented by the first column of the matrix \mathbf{F} , which is called the first factor log. The factor loadings can be estimated by the non-iterative approach of Jöreskog (2007)

$$\mathbf{L} = (\text{diag}\mathbf{S}^{-1})^{-1/2}\mathbf{\Omega}(\mathbf{\Gamma} - \theta\mathbf{I})^{1/2}\mathbf{U}, \quad (3)$$

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

We alter the traditional method (Bartlett 1937) of factor analysis by using the SA algorithm to calculate the factor scores. The combined method has been chosen to call as FA-SA. In the first step of this optimization problem, the model of factor analysis defined in Eq. (2) is rearranged

$$\mathbf{d} = \tilde{\mathbf{L}}\mathbf{f} + \mathbf{e}, \quad (4)$$

where \mathbf{d} denotes the KN length vector of observed (standardized) data, $\tilde{\mathbf{L}}$ is the NK -by- NM matrix of factor loadings, \mathbf{f} is the NM length vector of factor scores and \mathbf{e} is the KN length vector of residuals. At first, all data are put into a column vector, from where the matrix $\tilde{\mathbf{L}}$ is estimated by Eq. (3) and then rotated with the varimax algorithm developed by Kaiser (1958). This orthogonal rotation simplifies the interpretation of factors by giving to each factor a few large loadings and many small loadings. Then the vector of factor scores \mathbf{f} is estimated by the FA-SA algorithm. To solve the inverse problem and estimate the factor scores, an objective function needs to be defined, the minimization of which finds the optimal solution. We choose the objective function, named energy function in the terminology of optimization theory, based on the L_2 norm as

$$E = \frac{1}{NK} \sum_{i=1}^{NK} \left(\mathbf{d}_i^{(m)} - \mathbf{d}_i^{(c)} \right)^2 = \min, \quad (5)$$

where $\mathbf{d}^{(m)}$ and $\mathbf{d}^{(c)}$ denote the measured and calculated (standardized) well-logging data vectors, respectively. In the modified model of factor analysis, the term of $\tilde{\mathbf{L}}\mathbf{f}$ represents the calculated data and \mathbf{D} denotes the measured data. The former multiplication allows the estimation of the theoretical values of well logs, which can be considered as the solution of the forward problem. During the iterative procedure, we keep the values of factor loadings fixed to minimize the CPU time, and only the factor scores are updated. In each iteration, a randomly generated number (b) is added to any of the factor scores in vector \mathbf{f}

$$\mathbf{f}_j^{(new)} = \mathbf{f}_j^{(old)} + b \quad (j = 1, \dots, NM), \quad (6)$$

where parameter b is smaller than the maximal perturbation (b_{\max}) that has to be defined in the initialization of the FA-SA algorithm. In the current procedure, we select the initial values of factor scores as zero. If the energy difference of factor models—estimated in two subsequent iterations (ΔE) according to Eq. (5)—is negative (i.e. better fit between the observed and calculated data), the new model is accepted and the process is continued with the new energy state. However, in the reverse case (if $\Delta E > 0$), the probability of acceptance is given by the formula $P_a = \exp(-\Delta E/T)$, where T is the current temperature of the artificial system with no physical meaning. The new factor model is accepted only when a randomly generated number from the range of 0 and 1 is smaller than P_a . This acceptance rule for new energy states is referred to as the Metropolis criterion (Metropolis et al. 1953). This is a fundamental part of the FA-SA algorithm as it prevents the search from being stuck in a local minimum of function E in Eq. (5). During the annealing process, the temperature of the system is reduced iteratively according to Geman and Geman (1984)

$$T^{(new)} = T^{(old)} / \ln(1 + q), \quad (7)$$

where q denotes the number of iterations already computed. The maximal perturbation term (b_{\max}) is also reduced according to $b_{\max} = b_{\max} \cdot \varepsilon$, where ε is an arbitrary chosen

constant from the interval of 0 and 1. These steps are repeated in each iteration until the pre-defined maximal number of iteration steps is reached and then the values of factor scores in the last iteration step are accepted as the solution. The factors estimated by the FA-SA algorithm are directly used to reveal hidden petrophysical information from the well-logging data set. Szabó and Dobróka (2013) showed that the shale volume (in percent) correlates strongly to the first factor log (F_1) scaled into the range of 0 and 100

$$V_{sh} = 2.76e^{0.037F_1}, \quad (8)$$

the validity of which is verified in Alaska coastal sediments in Sect. 3. According to Eq. (8), the first factor has been identified as a lithologic indicator. Shale volume can also be independently calculated by a series expansion based inversion method (Dobróka et al. 2016). Szabó and Dobróka (2017) confirmed the validity of Eq. (8) by a series expansion based interval inversion procedure.

3 Field examples

3.1 Milky River Formation

The FA-SA method is tested in two hydrocarbon exploratory wells drilled in Alaska, USA shown in Fig. 1. First, we investigate Well-1 that penetrated the Milky River Formation, which is mainly built up of conglomerates, sandstone and mudstone and was formed in a shallow marine environment in the Pliocene age. It has high porosity and high permeability. As the input of factor analysis, we utilized the natural gamma-ray intensity (GR), density (RHOB), caliper (CAL), deep induction resistivity (ILD), spontaneous potential (SP) and neutron-porosity (NPHIS) logs. At the beginning of the statistical procedure, the factor loadings are calculated by Jöreskog's non-iterative approach using Eq. (3). Table 1 contains the resultant factor loadings representing the impact of the different well logs on the extracted factors for Well-1.

As Table 1 shows, the deep resistivity and the spontaneous potential logs have the highest loads on the first factor and unusually the load of the natural gamma-ray intensity is only -0.5213 . From the first factor considered as lithological indicator, we would expect the natural gamma-ray intensity to have higher load, but in this formation, the abundance of lithic detritus causes a mineralogical overlap between sandstones and mudstones. Therefore, the natural gamma-ray tool which response is mainly due to some aspect of mineralogy recorded only little or no difference between sandstone and mudstone and this causes the relatively low factor loading of the natural gamma-ray intensity on the first factor.

Then the factor scores are estimated by the FA-SA algorithm. In the step of initialization, we define the objective function, also called the energy function according to Eq. (5) and set the initial temperature (T_0) to 0.00015, cooling schedule according to Eq. (7), maximal parameter perturbation (b_{max}) to 0.5, perturbation reduction parameter (ϵ) to 0.98 and the maximal number of iteration steps. Figure 2a shows the decrease of the difference between the measured and calculated well logs by the iterations steps and Fig. 2b presents the temperature decrease of the system. The difference between the measured and calculated reached a minimum at about eighty thousand iterations, which took approximately 5 min on a quad-core based workstation. The continuous decrease of the value of the objective function indicates the highly stable nature of the FA-SA method.

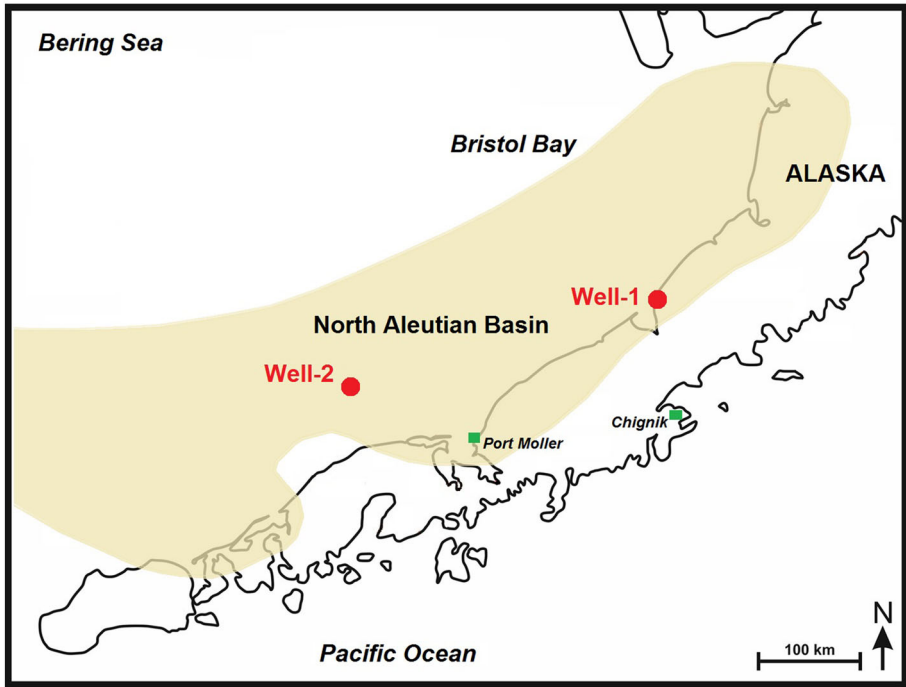


Fig. 1 The location of the investigated wells

Table 1 Rotated factor loadings estimated in Well-1

Well logs	Factor 1	Factor 2	Factor 3
CAL	− 0.3892	− 0.1484	− 0.5221
GR	− 0.5213	− 0.0205	0.4602
SP	0.9538	−0.0707	0.0320
RHOB	− 0.0883	0.7609	0.0144
ILD	0.9386	− 0.0780	0.0134
NPHIS	0.0969	− 0.7397	− 0.0612

Figure 3 presents the relation between the first factor and the shale volume estimated by the FA-SA method utilizing Eq. (8). The regression coefficients of the suggested exponential relation agree well with earlier studies (Szabó and Dobróka 2013). The same coefficients in different formations shows consistent results and confirm the applicability of the method.

Shale volume derived from the FA-SA method is shown in Fig. 4. On tracks 1–6, the standardized (input) well logs (black solid line) and the calculated logs (red dotted line) can be seen. Track 7 is the scaled first factor log, while next to it on the right the shale volume calculated from the first factor log (red solid line) is compared to that of calculated by deterministic modeling using the SP log (purple dotted line).

The fit between the observed and calculated well logs is quite good. Theoretical well logs represented by red dashed lines were directly calculated as the multiplication of the rotated factor loadings and factor scores. The shale volume calculated by the FA-SA

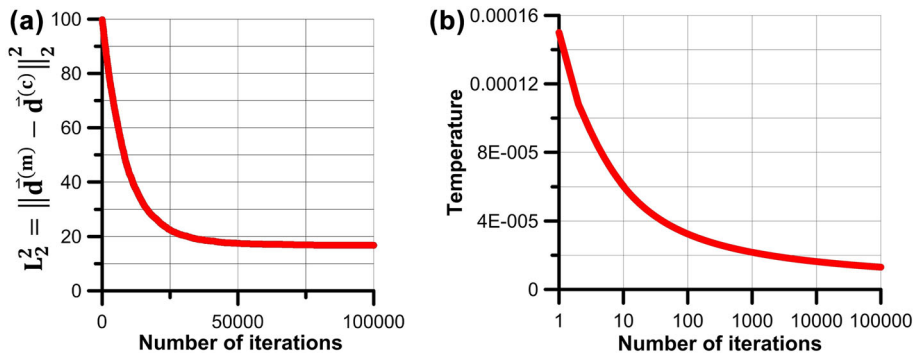


Fig. 2 a Development of convergence of the FA-SA procedure for Well-1, b temperature of the system during the procedure

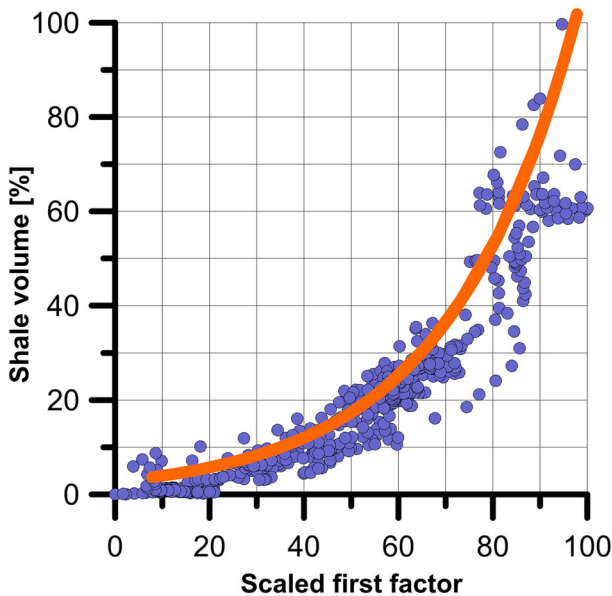


Fig. 3 Regression relation between the scaled first factor and the shale volume for Well-1

method also agrees well with the results of deterministic analysis. These results indicate the applicability of the FA-SA method for shale volume estimation in shallow marine sedimentary formations.

3.2 Bear Lake Formation

In case of Well-2, the penetrated horizon is the Bear Lake Formation, which is of Miocene age, and consists mainly of sandstones, conglomerates and thin mudstones. It formed in neritic to tidal flat environments and it can be characterized by high porosity and moderate permeability. The inputs of the factor analysis were the caliper (CAL), interval transit time (DT), natural gamma-ray intensity (GR), deep induction resistivity (ILD), neutron-porosity

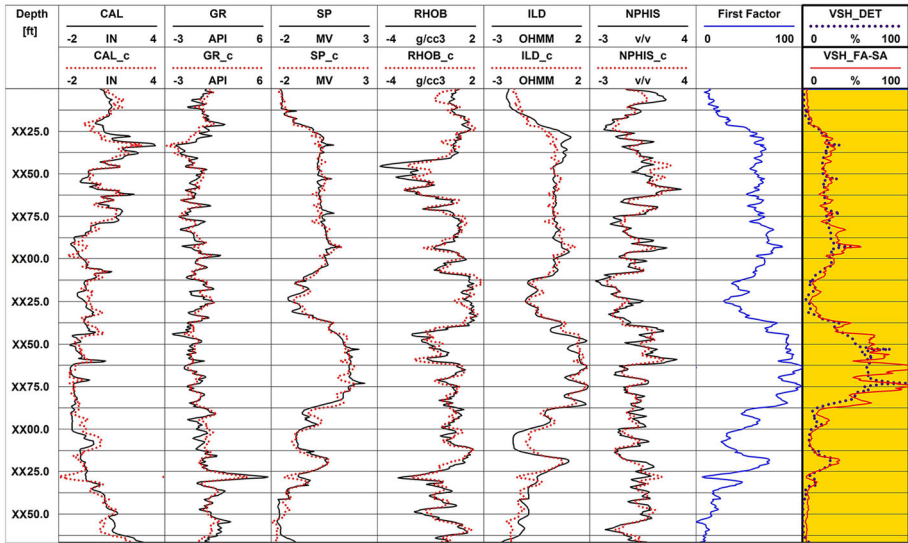


Fig. 4 The results of FA-SA for Well-1. (Color figure online)

(NPHIS), density (RHOB) and spontaneous potential (SP) logs. The FA-SA procedure run by the same control parameters as in Well-1. Table 2 contains the factor loadings related to three factors in case of Well-2.

In this case, not just the deep resistivity and the spontaneous potential logs have high loads on the first factor, but both the gamma-ray intensity and the caliper logs, too. We implemented the caliper log into the procedure, because washouts and the thickening of the mudcake might have strong relation in some cases to lithology. As we can see in this example, the caliper log affects highly the lithology-sensitive first factor. Figure 5 shows the decrease of the value of energy function by the iteration steps. It reached a minimum at about ninety thousand iterations in approximately 4 min. In this case, the FA-SA method again proves to be very stable in the iteration process. The cooling schedule was identical as in Fig. 2b which indicates the highly adaptive nature of this approach.

The regression function of the first factor and the shale volume estimated by the FA-SA method is illustrated in Fig. 6, which refers to a strong relation also for Well-2.

The interpretation results of the FA-SA method applied to Well-2 is shown in Fig. 7. The first seven columns from the left represents the input well logs (black solid line) and the calculated logs (red dotted line), the next is the scaled first factor log (blue solid line),

Table 2 Rotated factor loadings calculated in Well-2

Well logs	Factor 1	Factor 2	Factor 3
CAL	− 0.6222	− 0.6081	− 0.1344
GR	0.6529	0.5218	0.2648
SP	0.9338	− 0.0429	0.0081
RHOB	− 0.1935	0.5788	0.4545
ILD	0.8967	− 0.0588	− 0.1758
NPHIS	0.2424	− 0.8167	− 0.1616
DT	− 0.2733	− 0.7565	0.0344

Fig. 5 The decrease of the model energy by the iteration steps for Well-2

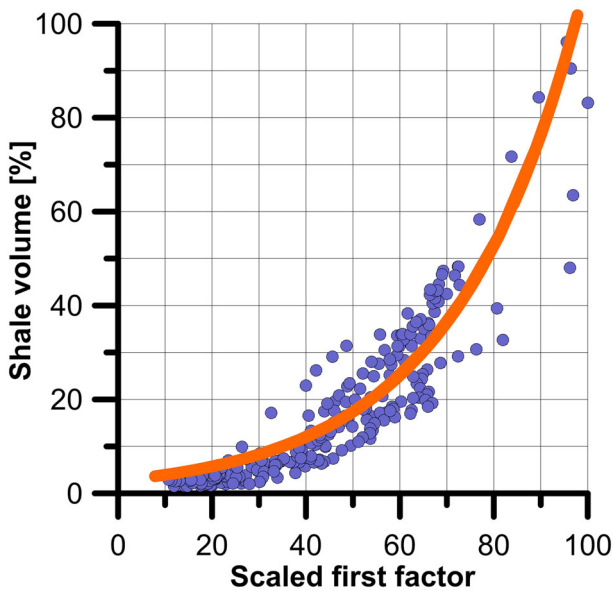
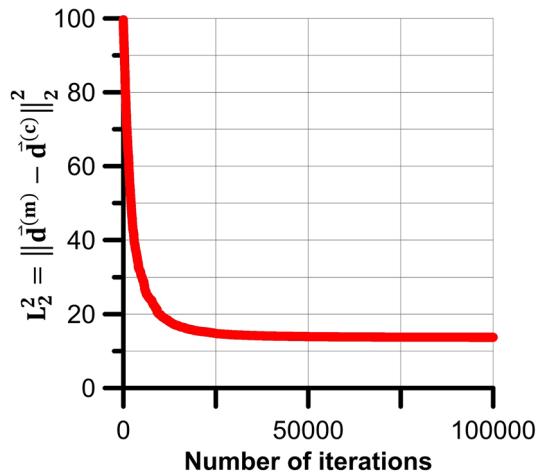


Fig. 6 Regression relation between the first factor and the shale volume for Well-2

and on the very right the shale volume calculated from the first factor log (red solid line) can be compared to the shale volume calculated by deterministic modeling (purple dotted line). Here the calculated logs again fit the measured data acceptably well. In conclusion, the shale volume calculated by the FA-SA method is consistent with the shale volume calculated by deterministic modeling.

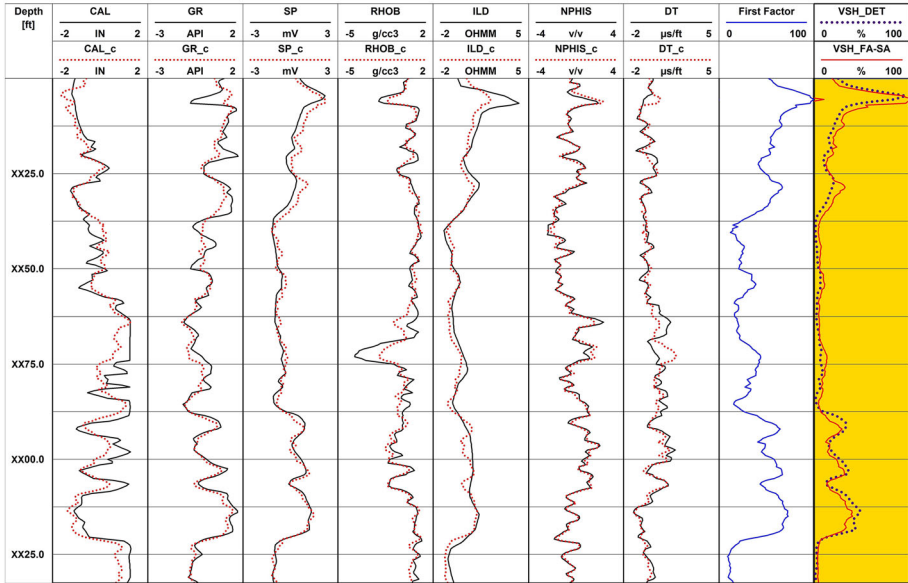


Fig. 7 The results of FA-SA for Well-2. (Color figure online)

4 Discussion and conclusions

The study presents the results of the newly developed method (FA-SA) of factor analysis assisted by the Simulated Annealing technique. With a global optimization approach, this method transforms the measured well logs into factor logs in a reliable way. One finds that the first factor log calculated by the FA-SA method strongly correlates with the independently calculated shale volume of different coastal formations. However, it should be noted, that as it is a general method, it should work in most lithologies. Based on the factor loadings, in both examples the second factor log is influenced by porosity sensitive logs such as RHOB, NPHIS and DT, two of which, RHOB and DT, are also sensitive to elastic properties, which might be the basis of further studies. The implementation of the FA-SA allows also for the estimation of the theoretical values of well logs, which neglects the preliminary knowledge of zone parameters and other petrophysical information. It also proves to be very stable in the iteration process and delivers the results within some minutes for sections of few hundred meters. Therefore, the statistical method can serve as an alternative tool for calculating shale volume in the oilfield. With the further improvement of the mathematical algorithm of the FA-SA method, we intend to give a robust solution, so that it can also be utilized for highly noisy data sets with arbitrary (e.g. not normally distributed) statistical distributions.

Acknowledgements The research was supported by the National Research, Development and Innovation Office—NKFIH, Project No. K109441. The authors would like to thank The Alaska Oil and Gas Conservation Commission (AOGCC) for the permission to use their digital well log data that was obtained from their website, via the link: <http://doa.alaska.gov/ogc/DigLog/diglogindex.html>.

References

- Asfahani J (2014) Statistical factor analysis technique for characterizing basalt through interpreting nuclear and electrical well logging data (case study from Southern Syria). *Appl Radiat Isot* 84:33–39
- Bartlett MS (1937) The statistical conception of mental factors. *Br J Psychol* 28:97–104
- Beauducel A, Harms C, Hilger N (2016) Reliability estimates for three factor score estimators. *Int J Stat Probab* 5(6):94
- Bücker C, Shimeld J, Hunze S, Brückmann W (2000) Data report: logging while drilling data analysis of Leg 171A, a multivariate statistical approach. *Proc Ocean Drill Program Sci Results Sci Results* 171A:1
- Coimbra R, Horikx M, Huck S, Heimhofer U, Immenhauser A, Rocha F, Dinis J, Duarte LV (2017) Statistical evaluation of elemental concentrations in shallow-marine deposits (Cretaceous, Lusitanian Basin). *Mar Pet Geol* 86:1029–1046
- Dobróka M, Szabó NP, Tóth J, Vass P (2016) Interval inversion approach for an improved interpretation of well logs. *Geophysics* 81(2):D155–D167
- Geman S, Geman D (1984) Stochastic relaxation, Gibbs distributions, and Bayesian restoration of images. *IEEE Trans Pattern Anal Mach Intell PAMI-6*:721–741
- Jöreskog KG (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 HF (1958) The varimax criterion for analytical rotation in factor analysis. *Psychometrika* 23:187–200
- Marquardt DW (1963) An algorithm for least-squares estimation of nonlinear parameters. *J Soc Ind Appl Math* 11(2):431–441
- 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
- Sen MK, Stoffa PL (1995) *Global optimization methods in geophysical inversion*, vol 4, 1st edn. Elsevier, Amsterdam
- Shaw R, Srivastava S (2007) Particle swarm optimization: a new tool to invert geophysical data. *Geophysics* 72(2):F75–F83
- Soupiosa P, Akcab I, Mpogiatzisc P, Basokurb AT, Papazacho C (2011) Applications of hybrid genetic algorithms in seismic tomography. *J Appl Geophys* 75(3):479–489
- Szabó NP (2011) Shale volume estimation based on the factor analysis of well-logging data. *Acta Geophys* 59:935–953
- Szabó NP, Dobróka M (2013) Extending the application of a shale volume estimation formula derived from factor analysis of wireline logging data. *Math Geosci* 45:837–850
- Szabó NP, Dobróka M (2017) Robust estimation of reservoir shaliness by iteratively reweighted factor analysis. *Geophysics* 82(2):D69–D83
- Szabó NP, Dobróka M, Turai E, Szűcs P (2014) Factor analysis of borehole logs for evaluating formation shaliness: a hydrogeophysical application for groundwater studies. *Hydrogeol J* 22:511–526
- Szűcs P, Civan F, Virág M (2006) Applicability of the most frequent value method in groundwater modeling. *Hydrogeol J* 14:31–43
- van Laarhoven PJ, Aarts EH (1987) *Simulated annealing: theory and applications*, volume 37 of the series mathematics applied, pp 7–15
- Yin C, Hodges G (2007) Simulated annealing for airborne EM inversion. *Geophysics* 72(4):F189–F195