Math Geosci
DOI 10.1007/s11004-017-9714-x

Exploratory Factor Analysis of Wireline Logs Using a Float-Encoded Genetic Algorithm

Norbert Péter Szabó[1,2] · **Mihály Dobróka**[1]

Received: 1 March 2017 / Accepted: 24 October 2017

Abstract In the paper, a novel inversion approach is used for the solution of the problem of factor analysis. The float-encoded genetic algorithm as a global optimization method is implemented to extract factor variables using open-hole logging data. The suggested statistical workflow is used to give a reliable estimate for not only the factors but also the related petrophysical properties in hydrocarbon formations. In the first step, the factor loadings and scores are estimated by Jöreskog's fast approximate method, which are gradually improved by the genetic algorithm. The forward problem is solved to calculate wireline logs directly from the factor scores. In each generation, the observed and calculated well logs are compared to update the factor population. During the genetic algorithm run, the average fitness of factor populations is maximized to give the best fit between the observed and theoretical data. By using the empirical relation between the first factor and formation shaliness, the shale volume is estimated along the borehole. Permeability as a derived quantity also correlates with the first factor, which allows its determination from an independent source. The estimation results agree well with those of independent deterministic modeling and core measurements. Case studies from Hungary and the USA demonstrate the feasibility of the global optimization based factor analysis, which provides a useful tool for improved reservoir characterization.

✉ Norbert Péter Szabó
norbert.szabo.phd@gmail.com

Mihály Dobróka
dobroka@uni-miskolc.hu

[1] Department of Geophysics, University of Miskolc, Egyetemváros, Miskolc 3515, Hungary

[2] MTA–ME Geoengineering Research Group, University of Miskolc, Egyetemváros, Miskolc 3515, Hungary

Journal: 11004 Article No.: 9714 TYPESET DISK LE CP Disp.:2017/11/2 Pages: 19 Layout: Small-X

Keywords Float-encoded genetic algorithm · Factor analysis · Shale volume · Permeability · Hungary · USA

1 Introduction

Soft computing methods have an emerging role in geosciences, especially in oilfield applications. Cranganu et al. (2015) present the latest developments in modern heuristics applied to hydrocarbon exploration problems such as uncertainty analysis, risk assessment, data fusion and mining, intelligent data analysis and interpretation, and knowledge discovery using a large amount of seismic, petrophysical, well logging and production data. State-of-the-art interpretation methods often use global optimization tools to find the best fit between the observations and predictions made by a deterministic, statistic or neural network-based modeling approach. Global optimization techniques, such as particle swarm optimization, simulated annealing and evolutionary algorithms, seek the global extreme of the objective function as a measure of data prediction error according to some criteria. A multidisciplinary selection of chapters including the theory, development, and applications of global optimization methods is presented in Michalski (2013). Geophysical inverse problems are conventionally solved by linearized optimization techniques (Menke 1984), which are quick but usually tend to trap in a local extreme of the objective function. Global optimization methods effectively avoid these localities and give a derivative-free and practically initial-model-independent solution. Despite these advantages, however, they have been found to be of limited use in industrial practice, because they require high computer processing time. They will become more common with the improvement of computer performance, especially in geophysical applications where the forward problem can be solved relatively quickly (Sen et al. 1993; Changchun and Hodges 2007; Bóna et al. 2009; Dobróka and Szabó 2011).

Genetic algorithm (GA) as a large class of evolutionary computation methods was first proposed by Holland (1975), which is based on the analogy between the optimization process and the natural selection of living organisms. The genetic search improves a population of artificial individuals in an iteration procedure. Model variables as possible solutions are represented by chromosomes, the genetic information of which are randomly exchanged during the procedure. In the classical GA, the model parameters are encoded using a binary coding scheme, which sets a limit to the resolution of the solution domain and the accuracy of the estimation results. Model parameters represented by real numbers makes a faster procedure and gives a higher resolution of the model space than binary algorithms (Michalewicz 1992). The float-encoded genetic algorithm (FGA) is known as one of the most efficient and adaptive global optimization methods, the fundamental theorem and geophysical aspects of which are detailed in Sen and Stoffa (2013). Applications to GA in ground geophysics and reservoir characterization are published in Boschetti et al. (1996), Dorrington and Link (2004), Alvarez et al. (2008), Akça and Basokur (2010), and Fang and Yang (2015). In well log analysis, an FGA-based inversion method called interval inversion is developed to automatically estimate not only the vertical distribution of porosity, shale

Journal: 11004 Article No.: 9714 TYPESET DISK LE CP Disp.:2017/11/2 Pages: 19 Layout: Small-X

volume, and hydrocarbon saturation but also the zone parameters and the positions of layer-boundaries (Dobróka and Szabó 2012; Dobróka et al. 2016).

Multivariate statistical methods are commonly used for lithology identification and facies analysis in hydrocarbon exploration. Factor analysis (FA) is applicable to reduce the dimensionality of statistical problems and extract non-measurable information from large-scale data sets (Lawley and Maxwell 1962). The statistical factors extracted from the measurements often correlate with the petrophysical properties of geological formations (Rao and Pal 1980; Puskarczyk et al. 2015). Szabó (2011) suggests the use of factor analysis for shale volume estimation in Hungarian unconsolidated gas reservoirs. Based on the same principles, a strong correlation between one of the factors and shale content is indicated in North American wells (Szabó and Dobróka 2013) and Syrian basaltic formations (Asfahani 2014). The classical GA is applicable to find hidden relations in binary data sets for the purpose of data compression and mining (Keprt and Snášel 2005), which can also be used as a preliminary data processing procedure to optimize the parameter structure of factor analysis (Yang and Bozdogan 2011).

In this paper, a highly adaptive method for the factor analysis of wireline logging data is presented. The proposed statistical approach developed by the combination of FGA and FA gives a reliable estimate to the factors and some related petrophysical quantities distributed along a borehole. The global optimization procedure improves the fit between the measured and calculated well logs and gives an estimate of the factor scores independently of their initial values. With suitably chosen genetic operators, the factor loadings and scores are estimated in a convergent iterative procedure. The basic method can be necessarily further improved using the L_1-norm or other weighted norms for fitness function to form a robust statistical procedure (Szabó and Dobróka 2017). Against the traditional methods of factor analysis (e.g., Bartlett's method), the new approach allows us to control the contributions of each datum to the solution by giving them individual weights. In this study, the shale volume and absolute permeability are directly estimated from the factor scores by the genetic algorithm-based factor analysis in Hungarian and North American wells. The permeability as a key-parameter in formation evaluation is not included in the probe response functions, thus, it cannot be determined by a traditional inversion procedure. Instead, it is usually derived from the inversion results using empirical formulae including porosity and irreducible water saturation. On the other hand, certain logs (e.g., caliper log) cannot be related explicitly to the petrophysical parameters. Thus, they cannot be utilized in the inversion procedure. In contrast, factor analysis makes use of the information of all well log types (including also the technical measurements) to give a reliable estimate of the petrophysical parameters in an independent well-log-analysis procedure.

2 Factor Analysis of Wireline Logs

2.1 Fast Approximate Algorithm

Observed wireline logs as input variables are simultaneously processed to derive a less number of statistical variables called factors, which are used to explore latent

Journal: 11004 Article No.: 9714 TYPESET DISK LE CP Disp.:2017/11/2 Pages: 19 Layout: Small-X

information not directly measurable by a logging tool. In the first step of factor analysis, the standardized well logging data are organized into an N-by-K matrix

$$\mathbf{D} = \begin{pmatrix} d_{11} & d_{12} & \cdots & d_{1k} & \cdots & d_{1K} \\ d_{21} & d_{22} & \cdots & d_{2k} & \cdots & d_{2K} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ d_{n1} & d_{n2} & \cdots & d_{nk} & \cdots & d_{nK} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ d_{N1} & d_{N2} & \cdots & d_{Nk} & \cdots & d_{NK} \end{pmatrix}, \tag{1}$$

where d_{nk} is the data recorded with the k-th logging instrument in the n-th depth (N is the total number of sampled depths, and K is that of the applied sondes). Data matrix in Eq. (1) is decomposed into two components

$$\mathbf{D} = \mathbf{F}\mathbf{L}^{\mathrm{T}} + \mathbf{E}, \tag{2}$$

where $\mathbf{F}$ is the N-by-M matrix of factor scores, $\mathbf{L}$ is the K-by-M matrix of factor loadings, and $\mathbf{E}$ is the matrix of residuals (M is the number of factors). Factor loading L_{km} practically measures the correlation between the k-th observed physical variable and the m-th factor, while factor scores given in the m-th column of matrix $\mathbf{F}$ constitute the well log of the m-th factor variable. The term FL^T in Eq. (2) can be regarded as the matrix of calculated data from the point of view of geophysical inversion. In this study, the greatest emphasis is placed on the first factor, which explains the largest part of variance of the observed data. In earlier studies, the first factor is identified as a lithology indicator, which carries information about the amount of shaliness in well logging applications (Szabó 2011). Since the factors are assumed linearly independent, the covariance matrix of standardized data can be directly expressed with the factor loadings

$$\mathbf{\Sigma} = N^{-1}\mathbf{D}^{\mathrm{T}}\mathbf{D} = \mathbf{L}\mathbf{L}^{\mathrm{T}} + \mathbf{\Psi}, \tag{3}$$

where $\mathbf{\Psi} = N^{-1}\mathbf{E}^{\mathrm{T}}\mathbf{E}$ is the diagonal matrix of specific variances, which does not explain the variances of measured variables. If the matrix $\mathbf{\Psi}$ is zero, Eq. (3) leads to the solution of principal component analysis. By knowing it, the factor loadings can be estimated by solving an eigenvalue problem. In the absence of specific variances, only an approximation can be made. In most cases, the factor loadings and scores are simultaneously estimated by the maximum likelihood method (Basilevsky 1994). The non-iterative method of Jöreskog (2007) gives an initial estimation of the factor loadings

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

where $\mathbf{\Gamma}$ is the diagonal matrix of the first M number of sorted eigenvalues (λ) 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. The factor loadings are usually

Journal: 11004 Article No.: 9714 TYPESET DISK LE CP Disp.:2017/11/2 Pages: 19 Layout: Small-X

rotated for a more efficient physical interpretation of factors. In this study, the varimax algorithm is applied to specify few data types to which the factors highly correlate (Kaiser 1958). Having estimated the factor loadings by Eq. (4), the matrix of factor scores are calculated by Bartlett's formula (1937)

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

The singular value decomposition of the reduced covariance matrix $\boldsymbol{\Sigma} - \boldsymbol{\Psi} = \mathbf{L}\mathbf{L}^{\mathrm{T}}$ can be applied to quantify the proportions of data variance explained by the factors. The total variance equals the trace of the singular value matrix, while the ratio of the m-th singular value and the trace gives the variance explained by the m-th factor. Jöreskog's method allows the estimation of the optimal number of factors, which normally depends on the applied well log suite and actual geological setting. In order to give a proper estimate, one must find the smallest number of factors for satisfying the inequality

$$\theta = (K - M)^{-1}\left(\lambda_{M+1} + \lambda_{\mathrm{M}+2} + \cdots + \lambda_{\mathrm{K}}\right) < 1. \tag{6}$$

Szabó and Dobróka (2017) study the impact of the selected number of factors on the result of factor analysis. The increase in the number of extracted factors improves the fit between the measured and calculated data, but simultaneously the variances and loadings of the factors (especially those of the first factor) significantly decrease. By increasing the number of factors, one neglects a relatively small amount of information, but the rest of information is shared more greatly by the factors. Thus, the correlation between the factors and petrophysical parameters is reduced. Both this experience and Eq. (6) suggest using the smallest possible number of factors with the condition that the misfit between the observations and predictions is acceptable. In order to give the best fit between the measured and calculated data, we introduce a global optimization method for the solution of factor analysis.

2.2 Genetic Algorithm Driven Factor Analysis

The GA search is based on an analogy to the process of natural selection, a mechanism that drives evolution in biology. In optimization problems, the model can be considered as an individual of an artificial population, the quality of which is characterized by a fitness value specifying its survival capability. The individuals with high fitness (or small data misfit) are more likely to survive, whereas those with low fitness tend to die out of the population. In the FGA procedure applied for seeking the absolute extreme of the fitness function, the model parameters are encoded as floating-point numbers, and real-valued operations are used to provide the highest resolution and optimal computer processing time. In this study, the FGA is implemented for the solution of the problem of factor analysis. At first, an estimate is given to the initial values of factor loadings and factor scores using Eqs. (4)–(5), which are then gradually refined by the highly effective FGA global optimization method. Experience shows that the factor loadings

Journal: 11004 Article No.: 9714 TYPESET DISK LE CP Disp.:2017/11/2 Pages: 19 Layout: Small-X

do not change significantly. Thus, they are assumed a priori known and fixed during the search of factor scores. It makes the procedure faster, but if necessary, the statistical algorithm allows the estimation of the factor loadings, too. Szabó and Balogh (2016) suggest a robust method of factor analysis for the simultaneous refinement of the factor scores and factor loadings by using the most frequent value method (Steiner 1991). The statistical method is based on the iterative reweighting of data prediction errors, which improves the accuracy of factor scores in case of non-Gaussian data sets including even a great number of outliers. The robust statistical method can be easily combined with the FGA to improve the results of factor analysis.

The classical model of factor analysis given in Eq. (2) is reformulated

$$\mathbf{d} = \tilde{\mathbf{L}}\mathbf{f} + \mathbf{e}, \tag{7}$$

where $\mathbf{d}$ denotes the $(K * N)$-by-1 vector of observed (standardized) data, $\tilde{\mathbf{L}}$ is the $(N * K)$-by-$(N * M)$ near-diagonal matrix of factor loadings, $\mathbf{f}$ is the $(M * N)$-by-1 vector of factor scores, and $\mathbf{e}$ is the $K * N$ length column vector of residuals. The column vector on the left side of the above equation includes the whole data set composed of K types of well-logging data measured in N number of adjacent depths. On the right side of the same equation, one can see that the factors are integrated in a column vector, in which all scores of the M number of factors are to be estimated in the same depth interval. The extended matrix of factor loadings includes all factor loadings measuring the correlations between the factors and the observed data referring to each depth. The chromosomes are built up from the factor scores, and the calculation of their fitness should be established. The fitness function is related to the vector of data deviations derived from Eq. (7)

$$F(\mathbf{f}) = -\left\|\mathbf{d} - \tilde{\mathbf{L}}\mathbf{f}\right\|_2^2 = \max, \tag{8}$$

which characterizes the goodness of the estimated factors. It is easily deduced from Eq. (8) that the theoretical data are calculated in terms of the factor scores by using the equation $\mathbf{d}^{(c)} = \tilde{\mathbf{L}}\mathbf{f}$, which corresponds to the solution of the forward problem. For checking the quality of fitting, one can calculate the distance between the measured and calculated (standardized) data in percent by multiplying the value of $-F$ by 100. One can also define the fitness function in a different form as

$$F(\mathbf{f}) = \left[\varepsilon^2 + \left\|\mathbf{d} - \tilde{\mathbf{L}}\mathbf{f}\right\|_2^2\right]^{-1} = \max, \tag{9}$$

where the positive constant ε^2 sets an upper limit of the value of fitness. In the FGA procedure, real genetic operators are suitably chosen to improve the fitness of the factor population. During the genetic process, the fittest individuals reproduce and survive to the next generation. The goal of the FGA is the increase of the average fitness of successive generations, which is achieved by the subsequent use of genetic operations, namely selection, crossover, mutation, and reproduction. A practical guide to the implementation of real genetic operators can be found in Houck et al. (1995).

Journal: 11004 Article No.: 9714 TYPESET DISK LE CP Disp.:2017/11/2 Pages: 19 Layout: Small-X

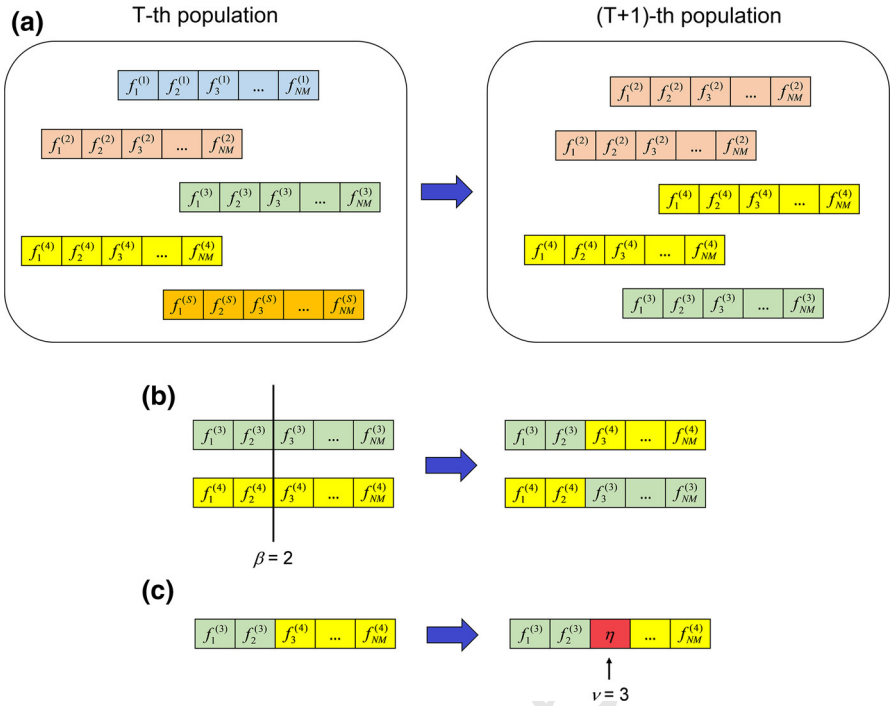


Fig. 1 Scheme of real-valued genetic operations applied in the FGA–FA procedure (**a**) selection (**b**) simple crossover (**c**) uniform mutation. Gene $f_{\mathrm{u}}^{(i)}$ denotes the u-th factor score of the i-th individual ($u = 1, 2, \ldots, NM$; $j = 1, 2, \ldots, S$), where N is the number of processed depths, M is the number of extracted factors, and S is the population size

Factor analysis is performed by the following evolutionary technique. In the first step, an initial population including a few tens of factor score vectors (**f**) is randomly generated. There are no restrictions for choosing the factor scores; only their upper and lower limits must be given. Several individuals are simultaneously tested during the optimization process, in which those with low fitness and having scores out of range are effectively rejected. In the first phase, the fittest individuals are selected for reproduction. Figure 1a shows that certain models of factors may be represented in the selected population several times (e.g., chromosomes nos. 2 and 4), while there are some that die (e.g., chromosome no. 1). The selection process is fitness proportionate, which allows the reselection of the fittest individuals. In this study, the selection of individuals is performed by the so-called normalized geometric ranking operator. At first individuals are sorted by their fitness value calculated using Eq. (8). The rank of the fittest model is 1, while that of the worst is S being the size of the population. The probability of selecting the i-th individual is

$$P_i = \frac{q}{1-(1-q)^S}(1-q)^{r_i-1}, \tag{10}$$

where r_i is the rank of the i-th individual, q is the probability of selecting the best individual. The cumulative probability of the ranked population is $C_i = \sum_{j=1}^{i} P_j$. If

Journal: 11004 Article No.: 9714 TYPESET DISK LE CP Disp.:2017/11/2 Pages: 19 Layout: Small-X

the condition $C_{i-1} < \alpha \leq C_i$ is fulfilled, the i-th individual is selected and copied into the new population (α is a random number from $U(0, 1)$). In the next step, a pair of individuals $\mathbf{f}^{(1)}$ and $\mathbf{f}^{(2)}$ is chosen from the selected population to exchange information between them. The simple crossover operator (Fig. 1b) cuts the chromosomes at crossover point β and swaps the factor scores located to the right of that

$$\mathbf{f}^{*(1)} = \begin{cases} f_u^{(1)}, & \text{if } \ u < \beta \\ f_u^{(2)}, & \text{otherwise} \end{cases}$$
$$\mathbf{f}^{*(2)} = \begin{cases} f_u^{(2)}, & \text{if } \ u < \beta \\ f_u^{(1)}, & \text{otherwise} \end{cases}, \tag{11}$$

where $\mathbf{f}^{*(1)}$ and $\mathbf{f}^{*(2)}$ are the updated individuals and index u runs through the total number of factor scores ($u = 1, 2, \ldots, \mathrm{NM}$). The operation of heuristic crossover extrapolates two individuals as follows

$$\mathbf{f}^{*(1)} = \mathbf{f}^{(1)} + \gamma \left(\mathbf{f}^{(1)} - \mathbf{f}^{(2)} \right)$$
$$\mathbf{f}^{*(2)} = \mathbf{f}^{(1)}, \tag{12}$$

where γ is a random number generated from $U(0, 1)$. During the application, it is assumed that the fitness of $\mathbf{f}^{(1)}$ is higher than that of $\mathbf{f}^{(2)}$. If any value of $\mathbf{f}^{*(1)}$ is out of bounds, a new value for γ is generated and Eq. (12) is recalculated. After a certain number of failures, the new values of factor scores are set as equal to the old ones. The third genetic operator is a uniform mutation (Fig. 1c). For the mutation process, individual $\mathbf{f}^{*(1)}$ is selected from the current population, and its v-th factor score is substituted with random number η generated from the possible range of factor scores

$$\mathbf{f}^{**(1)} = \begin{cases} \eta, & \text{if } \ v = h \\ f_v^{*(1)}, & \text{otherwise} \end{cases}, \tag{13}$$

where $\mathbf{f}^{**(1)}$ is the mutated individual. The genetic operations defined in Eqs. (10)–(13) are repeatedly applied in successive generations until a termination criterion is met. The stop criterion is usually the maximum number of generations or a specified threshold in the distance between the measured and calculated data. During the reproduction of individuals, the elitism can also be used, which copies the fittest individual of the previous generation to the new population whereas it removes the one with the smallest fitness. In the last generation, the individual with maximum fitness (including the optimal factor scores) is regarded as the result of factor analysis. The workflow of the above-described statistical procedure called FGA–FA is summarized in Fig. 2. In the last phase of the statistical procedure, the connections between the factors and petrophysical properties of hydrocarbon formations such as shale volume and permeability are explored by regression analyses. The strength of correlation between the above quantities is measured by the rank correlation coefficient (Spearman 1904).

Journal: 11004 Article No.: 9714 TYPESET DISK LE CP Disp.:2017/11/2 Pages: 19 Layout: Small-X

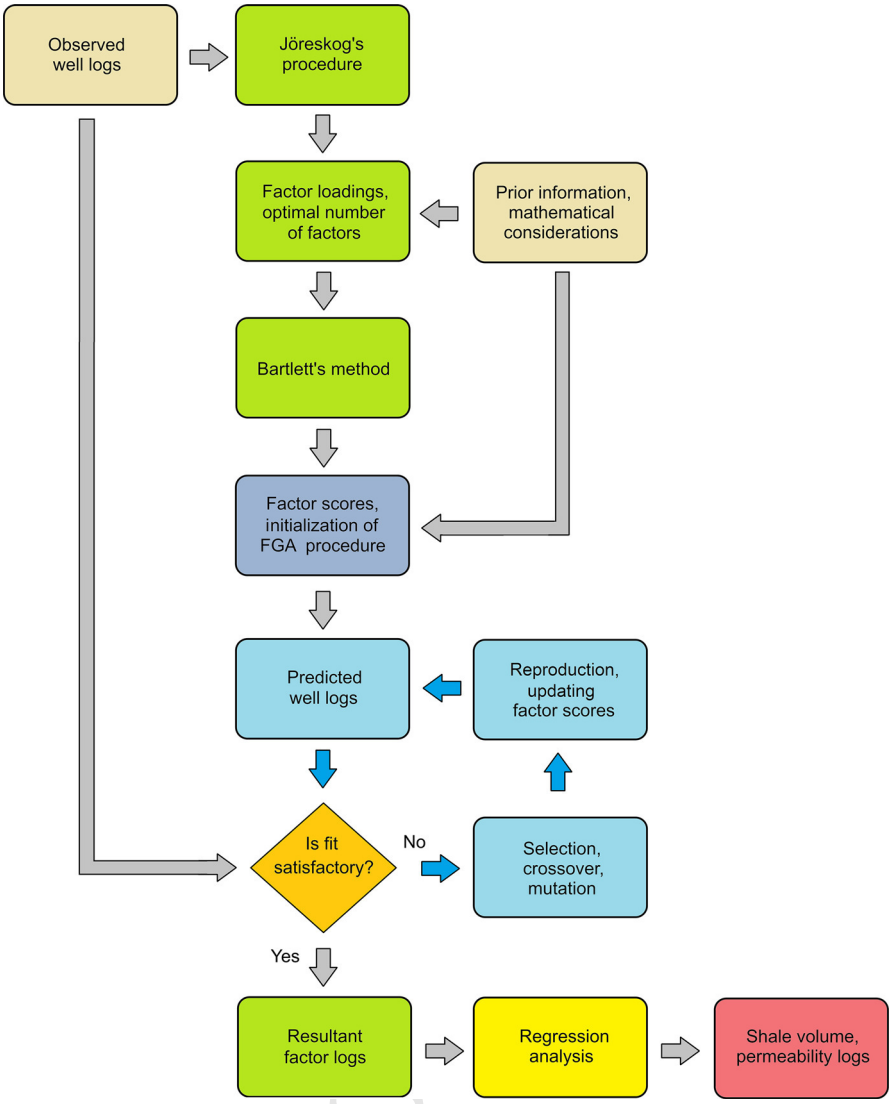


Fig. 2 Workflow of the genetic algorithm-based procedure of factor analysis applied to the estimation of petrophysical parameters

3 Test Computations

3.1 Case Study I

The FGA–FA method is first tested in a Hungarian hydrocarbon borehole. In Well-1, an unconsolidated gas-bearing formation of Pliocene age is investigated. Rock samples collected from the processed interval indicate high-porosity channel sands of good

Journal: 11004 Article No.: 9714 TYPESET DISK LE CP Disp.:2017/11/2 Pages: 19 Layout: Small-X

Table 1 Pearson's correlation matrix of wireline logs recorded in Well-1

	PHIN	GR	RD	DEN
PHIN	1	0.94	− 0.87	0.93
GR	0.94	1	− 0.85	0.96
RD	− 0.87	− 0.85	1	− 0.88
DEN	0.93	0.96	− 0.88	1

storage capacity interbedded by aleurite laminae and shaly layers. The natural gamma-ray intensity (*GR*), neutron-porosity (*PHIN*), density (*DEN*) and deep resistivity (*RD*) logs are used as input for factor analysis. The data are collected in 193 depths at 0.1 m intervals along the vertical well, where the total number of data is N=772. The average of the Pearson's correlation coefficients between the measured quantities is 0.9, which shows highly correlated well logs (Table 1). In this experiment, the information carried by the four well log types is concentrated into one factor, and it is studied which petrophysical properties of the reservoir are explained by the first factor. The initial values of factor loadings are calculated by Eq. (4), which are estimated as $L_{11}^{(PHIN)} = 0.96$, $L_{21}^{(GR)} = 0.98$, $L_{31}^{(RD)} = -0.89$, $L_{41}^{(DEN)} = 0.96$. They show a high correlation between the first factor and the processed well logs. The first factor is directly proportional to the readings of the nuclear logs, while it is inversely related to resistivity.

The first approximation for the factor scores is made by Eq. (5), which improved the FGA–FA process. The search domain of factor scores is set between the range of -2 and 2, which is specified by the preliminary results of the Jöreskog's method. In the initialization phase, the population size is set to 30. The fitness of individuals is calculated by Eq. (8), the values of which for the start population are plotted in Fig. 3. The FGA–FA procedure runs over 30,000 iterations, during which the genetic operators defined in Eqs. (10), (12), (13) are used to find the optimal values of factor scores. The control parameters of FGA are the probability of selecting the best individual ($q = 0.03$), crossover retry (100) and mutation probability ($p_m = 0.05$). An elitism-based reproduction is performed as the vector of factor scores with the maximum fitness is automatically copied into the next generation. The steady convergence of the FGA–FA procedure is illustrated in Fig. 4. The optimum is given at the maximal fitness of -7.2. At the end of the FGA–FA procedure, the vertical distribution of the first factor is estimated along the borehole. For further analysis, the first factor (F_{n1}) is suitably scaled

$$F'_{n1} = F'_{1,\min} + \frac{F'_{1,\max} - F'_{1,\min}}{F_{1,\max} - F_{1,\min}} \left(F_{n1} - F_{1,\min}\right), \tag{14}$$

where F'_{n1} is the score of the first scaled factor in the n-th depth, $F_{1,\min}$ and $F_{1,\max}$ are the extreme values of the first factor in the processed interval, respectively, $F'_{1,\min}$ and $F'_{1,\max}$ are those of the scaled factor ($n = 1, 2, \ldots, N$). In Well-1, the parameters of Eq. (14) are $F_{1,\min} = -1.64$, $F_{1,\max} = 1.89$, $F'_{1,\min} = 0$, $F'_{1,\max} = 1$.

Journal: 11004 Article No.: 9714 TYPESET DISK LE CP Disp.:2017/11/2 Pages: 19 Layout: Small-X

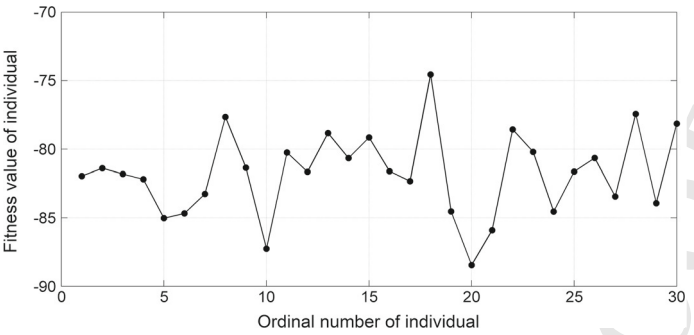
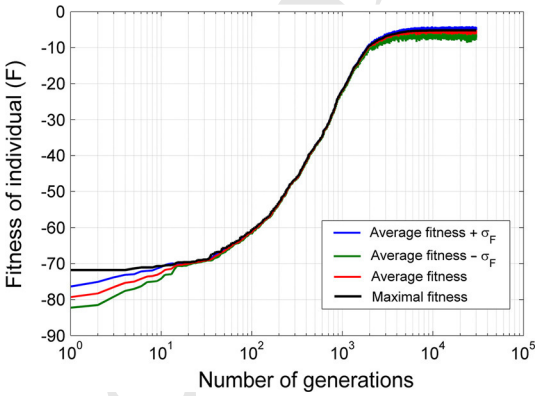


Fig. 3 Fitness values (F) of individuals (vectors of factor scores) generated in the initial population in Well-1

Fig. 4 Convergence plots of the FGA–FA procedure showing the fitness (F) of individuals (vectors of factor scores) versus the iteration steps in Well-1. The maximal fitness value calculated in the actual generation is represented by the black curve, the average fitness in the same generation is illustrated by the red curve, the average fitness plus or minus the standard deviation of fitness values (σ_F) is indicated by a blue and a green curve, respectively



Since all processed well logs are highly sensitive to reservoir shaliness, a strong correlation between the first factor and shale volume is found. Figure 5a shows the regression relation established between the first scaled factor (F_1') and the fractional volume of shale (V_{sh}) estimated by local (depth-by-depth) inversion of well-logging data (Dobróka et al. 2016). The regression function takes the form as

$$V_{sh} = ae^{bF_1'} + c, \tag{15}$$

where the regression coefficients are estimated with their 95% confidence bounds as $a = 0.08 \pm 0.02$, $b = -2.2 \pm 0.2$, $c = -0.01 \pm 0.01$. The rank correlation coefficient between the variables indicates a strong connection ($R = 0.97$). Permeability is derived from the well logs of porosity and irreducible water saturation. The former is estimated by the weighted least squares-based local inversion method, while the latter is calculated empirically as a function of the porosity-to-shale volume ratio available in Well-1 and neighboring boreholes. The reference values of absolute permeability (K given in mD unit) are calculated by the Timur formula (1968). The regression connection found between the first factor and the decimal logarithm of permeability is approximated by

Journal: 11004 Article No.: 9714 TYPESET DISK LE CP Disp.:2017/11/2 Pages: 19 Layout: Small-X

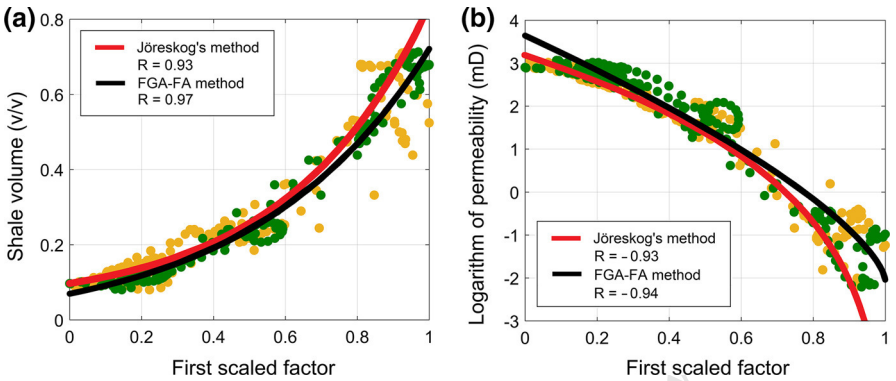


Fig. 5 Regression analysis made on the values of first scaled factor (F_1') and shale volume (V_{sh}) using the non-iterative Jöreskog's method (orange dots) and FGA–FA procedure (green dots) (**a**), and the first factor and the decimal logarithm of absolute permeability (K) (**b**) in Well-1. Rank correlation coefficient (R) indicates strong regression connection between the factor and petrophysical parameters

$$\lg(K) = a^* \left(1 - F_1'\right)^{b^*} + c^*, \tag{16}$$

where the regression coefficients are calculated as $a^* = 6.24 \pm 0.3, b^* = 0.48 \pm 0.05$, $c^* = -2.86 \pm 0.4$ (Fig. 5b). The rank correlation coefficient between the first factor and permeability shows a strong inverse relation ($R = -0.94$). The results of factor analysis are plotted in Fig. 6. Synthetic well logs calculated with the optimal values of factor scores show a good agreement with the observed well logs (tracks 1–4). The shale volume logs estimated separately by inversion, and the FGA–FA method shows high correlation (track 7), which is confirmed by the root mean square error (RMSE) of 2.3%. Permeability logs calculated by the two independent methods are also closely related (track 8), where the RMSE is 3.4%. The required CPU time of the FGA–FA procedure using a quad-core processor workstation is 55 s.

3.2 Case Study II

The FGA–FA method is tested in a North American borehole (Well-2), in which a low porosity and permeability (heavily cemented) oil-bearing sandstone formation of Late Permian age is investigated. Geophysical exploration is detailed in Gryc (1988), to which the well-logging data are provided by the USGS (1999). The spontaneous potential (*SP*), resistivity measured with a Laterolog-8 tool (*RLL8*), caliper (*CAL*), natural gamma-ray intensity (*GR*), neutron-porosity (*PHIN*) and acoustic transit-time (*AT*) logs are utilized for the analysis. The total number of processed depths (N) along the straight hole is 211, which are measured at 0.5 ft intervals (the total number of data is $N = 1266$). The overall strength of correlation between the input variables is moderate (Table 2), and the highest correlation is indicated between the lithology logs (*GR*, *SP*, *CAL*). The six well-log types are reduced to three independent factors by the FGA–FA procedure, which runs over 100,000 generations. The possible range of factor scores is set between -5 and 5. The same genetic operators and control

Journal: 11004 Article No.: 9714 TYPESET DISK LE CP Disp.:2017/11/2 Pages: 19 Layout: Small-X

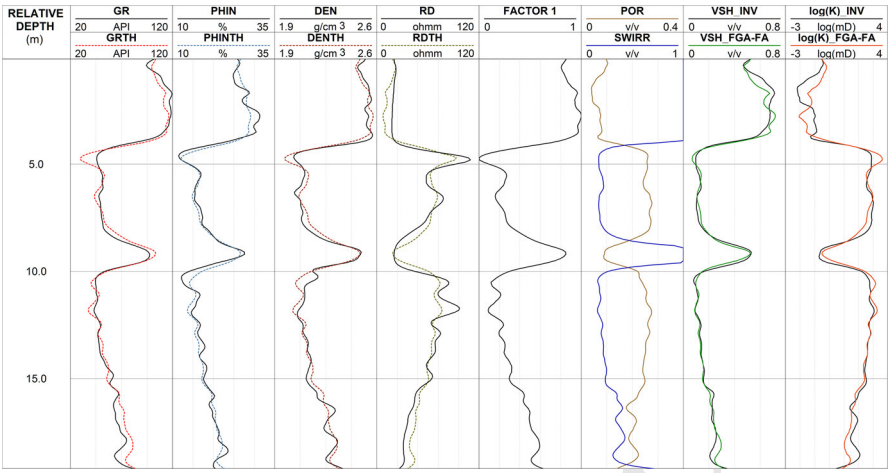


Fig. 6 Result of the FGA–FA procedure in Well-1. Observed well logs are: natural gamma-ray intensity (*GR*), neutron-porosity (*PHIN*), density (*DEN*), deep resistivity (*RD*). Estimated parameters are: theoretical data calculated from the factor scores (*TH*), effective porosity (*POR*), irreducible water saturation (*SWIRR*), first scaled factor (*FACTOR 1*), shale volume estimated by local inversion (*VSH_INV*) and factor analysis (*VSH_FGA–FA*), absolute permeability derived from local inversion (*K_INV*) and factor analysis (*K_FGA–FA*)

Table 2 Pearson's correlation matrix of well logs recorded in Well-2

	SP	RLL8	CAL	GR	PHIN	AT
SP	1	0.48	0.61	0.80	0.03	− 0.34
RLL8	0.48	1	0.59	0.22	0.16	− 0.54
CAL	0.61	0.59	1	0.35	0.06	− 0.18
GR	0.80	0.22	0.35	1	0.17	− 0.26
PHIN	0.03	0.16	0.06	0.17	1	− 0.36
AT	− 0.34	− 0.54	− 0.18	− 0.26	− 0.36	1

parameters are used as in Well-1. The optimum is obtained at the maximal fitness of −9.4. The rotated factor loadings are listed in Table 4, which measure the degrees of association between the extracted three factors and processed well logs. The first factor is strongly related to the *SP* and *GR* logs, which presents the first factor as a good lithology indicator. The second factor is mostly influenced by the resistivity and caliper log, while the third one is in inverse relation with the sonic log. The singular value decomposition of the covariance matrix $\mathbf{L}\mathbf{L}^{\mathrm{T}}$ shows that the first three factors explain the 57, 20, 16% parts of variances of the input data, respectively, while the rest of information is neglected (Table 3).

The first factor estimated by the FGA–FA procedure is used to calculate the shale volume. For checking the interpretation result, an independent estimation of shale volume is given by the method of Larionov (1969). The exponential regression relation between the first scaled factor and shale content in the tight oil formation is shown

Journal: 11004 Article No.: 9714 TYPESET DISK LE CP Disp.:2017/11/2 Pages: 19 Layout: Small-X

Table 3 Rotated factor loadings estimated in Well-2

	Factor 1	Factor 2	Factor 3
$L^{(SP)}$	0.84	−0.41	0.20
$L^{(RLL8)}$	0.13	0.64	0.59
$L^{(CAL)}$	0.31	−0.82	0.05
$L^{(GR)}$	0.92	−0.07	0.06
$L^{(NPHI)}$	0.06	−0.02	0.21
$L^{(AT)}$	−0.19	0.09	−0.78

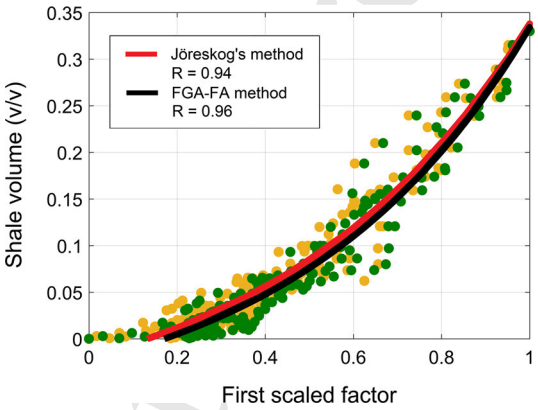


Fig. 7 Regression relation between the first scaled factor (F_1') and shale volume (V_{sh}) established by the non-iterative Jöreskog's method (orange dots) and FGA–FA procedure (green dots) in Well-2. Rank correlation coefficient (R) shows a strong regression relation between the studied quantities

in Fig. 7. The rank correlation coefficient between the first factor and shale volume indicates a strong connection (R=0.96), which is consistent with the results of the Hungarian experiment (Well-1). By assuming the model according to Eq. (15), the regression coefficients are estimated as $a = 0.07 \pm 0.02$, $b = -1.8 \pm 0.2$, $c = -0.09 \pm 0.02$. The result of factor analysis is plotted in Fig. 8. The fit between the observed and theoretical data is highly acceptable (tracks 1–6). The shale volume log estimated by the FGA–FA procedure correlates well to that given by Larionov's method (track 8). The RMSE calculated between the shale volume logs is 2.1%. The CPU time of the optimization procedure using a quad-core processor workstation is 14 min 46 s.

3.3 Case Study III

The FGA–FA method is tested in a Hungarian thermal-water well (Well-3) to validate the results of factor analysis with core data. In the processed interval, unconsolidated sediments composed of shale, sand, and gravel of Pleistocene age are deposited, which are fully saturated with water. The spontaneous potential (SP), natural gamma-ray intensity (GR), gamma–gamma intensity (GG), neutron-neutron (NN) and shallow resistivity (RS) logs are utilized as input for factor analysis. The average of Pearson's

Journal: 11004 Article No.: 9714 TYPESET DISK LE CP Disp.:2017/11/2 Pages: 19 Layout: Small-X

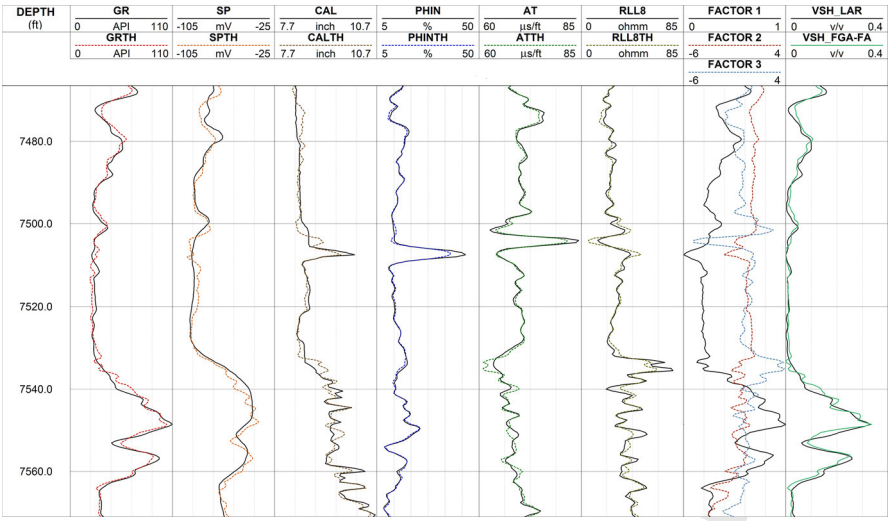


Fig. 8 Result of the FGA–FA procedure in Well-2. Measured well logs are: natural gamma-ray intensity (*GR*), spontaneous potential (*SP*), caliper (*CAL*), neutron-porosity (*PHIN*), acoustic traveltime (*AT*), shallow resistivity (*RLL8*). Estimated parameters are: theoretical data calculated from the factor scores (*TH*), first scaled factor (*FACTOR 1*), second and third factors (*FACTOR 2, FACTOR 3*), shale volume estimated by Larionov method (*VSH_LAR*) and factor analysis (*VSH_FGA–FA*)

Table 4 Pearson's correlation of wireline logs recorded in Well-3

	RS	SP	GR	GG	NN
RS	1	−0.75	−0.35	−0.49	0.52
SP	−0.75	1	0.65	0.47	−0.36
GR	−0.35	0.65	1	0.44	−0.28
GG	−0.49	0.47	0.44	1	−0.07
NN	0.52	−0.36	−0.28	−0.07	1

correlation coefficients is 0.4, which indicates a moderate correlation between the measured logs (Table 4).

One factor is extracted from five observed well logs, the scores of which are estimated at the depths of core sampling. The FGA–FA procedure runs over 10,000 generations, which improves 30 individuals. The possible range of factor scores is set between -3 and 3. The same genetic operators and control parameters are applied as in Wells-1–2. At the end of the procedure, the optimum is given at maximal fitness of -7.1. The resultant factor loadings are estimated as $L_{11}^{(RS)} = -0.83$, $L_{21}^{(SP)} = 0.86$, $L_{31}^{(GR)} = 0.63$, $L_{41}^{(GG)} = 0.56$, $L_{51}^{(NN)} = -0.48$. The first factor is strongly related to the *SP* and *RS* logs, which agrees with the results of Wells 1–2. The first factor is scaled between 0 and 1, which is related to the shale volume of the unconsolidated formations. Shale volumes available from the grain-size analysis of 24 core samples are used as a reference for regression analysis. Thus, the total number of observed well-logging data along the vertical well is N=120. The local regression

Journal: 11004 Article No.: 9714 TYPESET DISK LE CP Disp.:2017/11/2 Pages: 19 Layout: Small-X

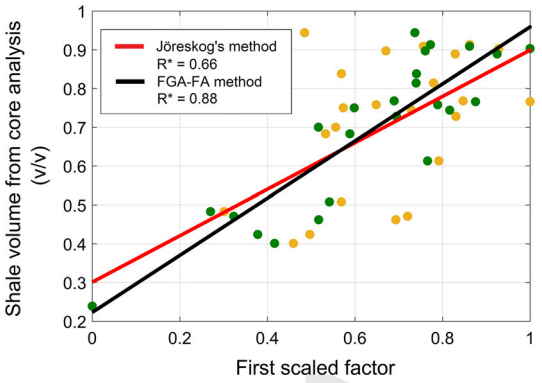


Fig. 9 Regression relation between the first scaled factor (F_1') and shale volume derived from grain-size analysis (V_{sh}) established by the non-iterative Jöreskog's method (orange dots) and FGA–FA procedure (green dots) in Well-3. Pearson's correlation coefficient (R^*) shows a strong regression relation between the studied quantities

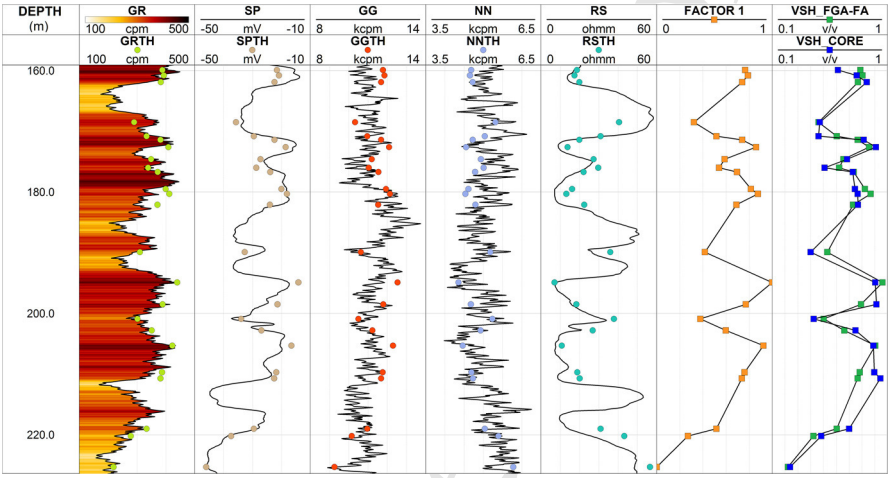


Fig. 10 Result of the FGA–FA procedure in Well-3. Measured well logs are: natural gamma-ray intensity (*GR*), spontaneous potential (*SP*), gamma–gamma intensity (*GG*), neutron-neutron (*NN*) and shallow resistivity (*RS*). Theoretical data calculated from the factor scores are indicated by colored dots (*TH*), first scaled factor is indicated by light orange squares (*FACTOR 1*), shale volume given by core measurements is indicated by blue squares (*VSH_CORE*) and shale volume predicted from factor analysis is indicated by green squares (*VSH_FGA–FA*)

relation between the first factor and shale volume is plotted in Fig. 9. The Pearson's correlation coefficient ($R^* = 0.88$) indicates a strong linear relation between the above quantities. The result of factor analysis is shown in Fig. 10. The *GR* image and the *SP* and *RS* logs show the cyclic variation of shales and sands along the processed interval. The color dots in tracks 1–5 represent the theoretical data calculated with the estimated factor scores at the depths of core sampling. The FGA–FA-derived shale volume log correlates well to the core data (track 7). The RMSE between the shale volumes estimated separately by factor analysis and laboratory measurements (represented by green and blue boxes) is 2.9%. The CPU time of the FGA–FA procedure using a quad-core processor workstation is 16 s.

Journal: 11004 Article No.: 9714 TYPESET DISK LE CP Disp.:2017/11/2 Pages: 19 Layout: Small-X

4 Conclusions

A global optimization approach for the factor analysis of wireline logging data is presented. The multivariate statistical method transforms the observed physical variables into factor logs, while it searches the absolute minimum of the misfit between the measured data and theoretical ones calculated directly from the factor scores. The genetic algorithm-based factor analysis predicts the essential petrophysical parameters also from the factors, which offers a new approach for improved formation evaluation. One can find the genetic algorithm-based factor analysis highly adaptive by the following reasons. The presented case studies show that the suggested method is feasible.

1. For analyzing well-logging data sets including different log types and vertical resolution,
2. By changing the correlation between the input data,
3. Both by equidistant measuring intervals and significant lack of data (e.g., core sampling),
4. By the same combination of genetic operators (and control parameters) in different wells.

The first factor estimated by the FGA–FA procedure is strongly related to shale volume and derived quantities in reservoir rocks, which still acts as a good shale indicator. The regression relation between the first factor and shale volume is consistent, where the functional coefficients are close to each other in different measurement areas. The factor scores can be directly used to solve the forward problem. Well logs, like the caliper log in Well-2, to which response function does not exist in the practice of inverse modeling, can be predicted by the proposed method of factor analysis. The FGA–FA algorithm can be further improved to give a robust solution for the factor scores. The processing of wireline logging data sets following non-Gaussian statistics requires the modification of the fitness function to be optimized. The weighted norm of data deviations is preferably used to measure the goodness of factors. For instance, the use of Steiner weights automatically calculated by the most frequent value method assures high statistical efficiency and excludes the outliers effectively from the solution. In future studies, the global optimization-based factor analysis technique will be applied for the lithological identification and petrophysical characterization of complex (unconventional) reservoirs, which may improve the reservoir model and the calculation of hydrocarbon reserves.

Acknowledgements This research was supported by the GINOP-2.3.2-15-2016-00010 “Development of enhanced engineering methods with the aim at utilization of subterranean energy resources” project in the framework of the Széchenyi 2020 Plan, funded by the European Union, cofinanced by the European Structural and Investment Funds. The first author thanks the support of the GINOP project. The research was partly supported by the National Research Development and Innovation Office (Project No. K109441), and as the leader of the project, the second author thanks the Office for its support. Both authors thank the Geokomplex Ltd. for providing well logs and grain-size data from Well-3.

Journal: 11004 Article No.: 9714 TYPESET DISK LE CP Disp.:2017/11/2 Pages: 19 Layout: Small-X

References

Akça I, Basokur AT (2010) Extraction of structure-based geoelectric models by hybrid genetic algorithms. Geophysics 75:F15–F22

Alvarez JPF, Martínez JLF, Pérez COM (2008) Feasibility analysis of the use of binary genetic algorithms as importance samplers application to a 1-D DC resistivity inverse problem. Math Geosci 40:375–408

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

Basilevsky A (1994) Statistical factor analysis and related methods: theory and applications. John Wiley & Sons, Hoboken, pp 367–381

Bóna A, Slawinski MA, Smith P (2009) Ray tracing by simulated annealing: bending method. Geophysics 74:T25–T32

Boschetti F, Dentith MC, List RD (1996) Inversion of seismic refraction data using genetic algorithms. Geophysics 61:1715–1727

Changchun Y, Hodges G (2007) Simulated annealing for airborne EM inversion. Geophysics 72:F189–F195

Cranganu C, Luchian H, Breaban ME (2015) Artificial intelligent approaches in petroleum geosciences. Springer International Publishing, Switzerland

Dobróka M, Szabó NP (2011) Interval inversion of well-logging data for objective determination of textural parameters. Acta Geophys 59:907–934

Dobróka M, Szabó NP (2012) Interval inversion of well-logging data for automatic determination of formation boundaries by using a float-encoded genetic algorithm. J Pet Sci Eng 86–87:144–152

Dobróka M, Szabó NP, Tóth J, Vass P (2016) Interval inversion approach for an improved interpretation of well logs. Geophysics 81:D163–D175

Dorrington KP, Link CA (2004) Genetic-algorithm/neural-network approach to seismic attribute selection for well-log prediction. Geophysics 69:212–221

Fang Z, Yang D (2015) Inversion of reservoir porosity, saturation, and permeability based on a robust hybrid genetic algorithm. Geophysics 80:R265–R280

Gryc G (1988) Geology and exploration of the National Petroleum Reserve in Alaska, 1974 to 1982. US Geol Surv Prof Pap 1399:1–940

Holland JH (1975) Adaptation in natural and artificial systems. University of Michigan Press, Ann Arbor

Houck CR, Joines J, Kay M (1995) A genetic algorithm for function optimization: a Matlab implementation. In: NCSU-IE technical report 95-09. North Carolina State University, Raleigh, pp 1–14

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, Mahwah, pp 47–77

Kaiser HF (1958) The varimax criterion for analytical rotation in factor analysis. Psychometrika 23:187–200

Keprt A, Snášel V (2005) Binary factor analysis with genetic algorithms. In: Abraham A, Dote Y, Furuhashi T, Köppen M, Ohuchi A, Ohsawa (eds) Soft computing as transdisciplinary science and technology. Proceedings of the fourth IEEE International workshop WSTST'05. Springer, Berlin, pp 1259–1268

Larionov VV (1969) Radiometry of boreholes. Nedra, Moscow (in Russian)

Lawley DN, Maxwell AE (1962) Factor analysis as a statistical method. The Statistician 12:209–229

Menke W (1984) Geophysical data analysis: discrete inverse theory. Academic Press, New York

Michalewicz Z (1992) Genetic algorithms + data structures = evolution programs. Springer, New York

Michalski A (2013) Global optimization: theory, developments and applications. In: Mathematics research developments. Computational mathematics and analysis series. Nova Science Publishers, New York

Puskarczyk E, Jarzyna J, Porebski Sz (2015) Application of multivariate statistical methods for characterizing heterolithic reservoirs based on wireline logs—example from the Carpathian Foredeep Basin (Middle Miocene, SE Poland). Geol Q 59:157–168

Rao BN, Pal PC (1980) Factor analysis for interpreting petrophysical data on Roro ultramafics, Singhbhum district, India. Geophys Prospect 28:112–118

Sen MK, Bhattacharya BB, Stoffa PL (1993) Nonlinear inversion of resistivity sounding data. Geophysics 58:496–507

Sen MK, Stoffa PL (2013) Global optimization methods in geophysical inversion. Cambridge University Press, Cambridge

Spearman C (1904) The proof and measurement of association between two things. Am J Psychol 15:72–101

Journal: 11004 Article No.: 9714 TYPESET DISK LE CP Disp.:2017/11/2 Pages: 19 Layout: Small-X

Steiner F (1991) The most frequent value: introduction to a modern conception of statistics. Academic Press, Budapest
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, Balogh GP (2016) Most frequent value based factor analysis of engineering geophysical sounding logs. In: 78th EAGE conference and exhibition, paper Tu SBT4 12. Vienna, pp 1–5
Szabó NP, Dobróka M (2017) Robust estimation of reservoir shaliness by iteratively reweighted factor analysis. Geophysics 82:D69–D83
Timur A (1968) An investigation of permeability, porosity, and residual water saturation relationship for sandstone reservoirs. Log Anal 9:8–14
USGS (1999) Selected data from eleven wildcat wells in the national petroleum reserve in Alaska. In: USGS open file report 99–015. https://pubs.usgs.gov/of/1999/ofr-99-0015/ofr-99-0015.html
Yang H, Bozdogan H (2011) Learning factor patterns in exploratory factor analysis using the genetic algorithm and information complexity as the fitness function. J Pattern Recognit Res 6:307–326

Journal: 11004 Article No.: 9714 TYPESET DISK LE CP Disp.:2017/11/2 Pages: 19 Layout: Small-X