Extending the application of a shale volume estimation formula derived from factor analysis of wireline logging data[1]

by Norbert P. Szabó[2] and Mihály Dobróka[2,3]

[1] Received; accepted.

[2] Department of Geophysics, University of Miskolc, 3515 Miskolc-Egyetemváros, Hungary, e-mail: norbert.szabo.phd@gmail.com; dobroka@uni-miskolc.hu

[3] MTA-ME Applied Geoscience Research Group, University of Miskolc, H-3515 Miskolc-Egyetemváros, Hungary

Corresponding Author:

Mihály Dobróka

Department of Geophysics

University of Miskolc

3515, Miskolc-Egyetemváros, Hungary

Phone: + 36 46 361936

fax: + 36 46 361936

e-mail: dobroka@uni-miskolc.hu

Abstract

A multivariate statistical procedure is developed for the estimation of shale volume in clastic sedimentary formations. The method offers an alternative to extract shale content from borehole geophysical measurements. Factor analysis of various well-logging data types generates a new well log that correlates with the shale content of shaly-sandy rocks. The mathematical relationship between shale volume and factor scores is represented by a non-linear equation, which seems to be applicable for data sets originating from different sedimentary basins. A comparative study is made between three different data sets originated from Hungary and the United States of America to check the validity of the proposed empirical formula. Shale volumes predicted from factor analysis are confirmed by estimates from independent deterministic and inverse modeling. Petrophysical information derived by factor analysis of logs recorded in deep wells can be used for a more accurate and reliable estimation of effective porosity and absolute permeability of reservoir rocks, for decreasing the estimation error of inversion estimates and for reducing the ambiguity in the solution of the well-logging inverse problem.

Keywords: shale volume, factor log, scaled factor, deterministic procedure, inverse modeling.

1. Introduction

Petrophysical properties of shaly sand formations can be extracted from borehole geophysical data either by deterministic or inversion procedures. The former ones substitute data to explicit equations in order to determine the petrophysical properties separately. There are several kinds of deterministic methods for the estimation of porosity, shale volume, mineral composition of rock matrix, water saturation, and other non-measurable properties of sedimentary formations (Serra 1984; Asquith and Krygowski 2004). Wireline logging data can also be evaluated by using an appropriate inversion method, which assumes a known relationship, called probe response function, between the observed data and the parameters of the petrophysical model. The solution of the inverse problem is given by fitting a theoretical data set calculated by the response equations to the measured data. As a result of the inversion procedure, the most probable set of model parameters with estimation errors is specified (Menke 1984). The theory of the inverse problem of borehole geophysics and its practical implementation are detailed in Mayer and Sibbit (1980), Alberty and Hashmy (1984), Ball et al (1987).

Factor analysis is a multivariate statistical method, which is generally used to reduce high-dimensional data sets to lower dimensions and to extract unobservable quantities hidden in the original measurements. The mathematical background of factor analysis is detailed in Lawley and Maxwell (1962). Various petrophysical applications of the statistical method have been published (Buoro and Silva 1994; Grana et al 2011; Herron 1986; Ma 2011; Rao and Pal 1980; Urbancic and Bailey 1988). Szabó et al (2012) used a linear factor model to process engineering geophysical sounding (EGS) data for the hydrogeological characterization of shallow formations by giving the spatial distribution of water saturation between twelve boreholes drilled at a Hungarian site. As the latest application, dry density of soil layers was

derived from factor analysis of EGS data with the purpose of providing geotechnical operations with in-situ information (Szabó 2012).

Shale volume is a key parameter in well log analysis, because it greatly affects all measurements and petrophysical interpretation results. For instance, in hydrocarbon prospecting a realistic calculation of reserves largely depends on the accurate prediction of shale content and other related quantities such as effective porosity and permeability. Factor analysis is applicable for the determination of shale volume from well-logging data. Szabó (2011) found strong correlation between one of the new variables (factor) and shale volume estimated by inversion processing. It was shown that shale volume was directly proportional to the first factor representing the largest amount of variance in the original (data) variables. Regression tests showed a non-linear connection between the above quantities, which proved to be valid for some different areas of the Pannonian Basin, Hungary.

In the paper, shale volume calculation based on factor analysis is tested in three boreholes drilled thousands of kilometers apart from each other. Data sets originated from Hungary and the United States of America are processed separately and the given results are compared to study the applicability of the non-linear formula proposed by Szabó (2011). The statistical research reveals that the empirical equation can be applied well to the selected areas representing very different geological ages and environments.

2. Estimation of shale volume

2.1 Deterministic and inverse modeling

Shale is treated in well-logging applications as a fine grained rock containing a sizable portion of clay minerals and silt (Ellis 2007). The volume fraction of dispersed shale occupying the pore space of permeable formations can be expressed from the material balance equation

$$V_{sh} = 1 - \Phi - \sum_{k=1}^{n} V_{ma,k} \ , \tag{1}$$

where Φ denotes effective porosity, $V_{ma,k}$ is the volume of the k-th matrix constituent, n is the number of mineral types forming the rock matrix. The deterministic approach for shale volume estimation is a relatively simple procedure, which is normally based on single well log analysis. Generally spontaneous potential (SP) or natural gamma-ray intensity (GR) data is substituted into an explicit equation to calculate the amount of shale in shaly-sand sequences. The most frequently used shale volume indicators are based on the calculation of gamma-ray index from GR log (Clavier et al 1971; Larionov 1969; Poupon and Gaymard 1970; Stieber 1970). Newer methods integrates more types of well logs for a more accurate estimation (Bhuyan and Passey 1994; Fertl and Frost 1980).

Inversion procedure is applicable to process all data types simultaneously to give an estimate for several petrophysical parameters. Consider **m** as the P-by-1 column vector of model parameters including shale volume and other petrophysical quantities. In the classical inversion methodology the model vector is defined locally in a certain measuring point. Data are collected in the same depth by borehole probes using different physical principles. Let $\mathbf{d^{(m)}}$ be the Q-by-1 column vector of measured data. The relationship between the model and data vector is basically non-linear. Theoretical data in vector $\mathbf{d^{(c)}}$ are calculated by using Q number of probe response equations from relationship $\mathbf{d^{(c)}=g(m)}$. Linearization based on Taylor-series expansion is frequently used for simplifying the above connection (Menke 1984)

$$\mathbf{d^{(c)}} = \mathbf{Gm} \ , \tag{2}$$

where **G** is the Jacobi's matrix including partial derivatives of data with respect to model parameters. Derivatives $\partial d_q^{(c)} / \partial m_p$ (q=1,2,…,Q; p=1,2,…,P) are calculated numerically in the

inversion procedure. The solution to the inverse problem is found by minimizing the L_2-norm of the difference between the measured and calculated data vectors. Since data have different orders of magnitude and measurement units an objective function measuring the weighted overall error is to be optimized

$$E = \sum_{q=1}^{Q} \left(\frac{d_q^{(m)} - d_q^{(c)}}{\sigma_q} \right)^2 = \min \tag{3}$$

where σ_q is the standard deviation of the q-th measured data that specifies the contribution of the given datum to the solution. The minimization of Equation (3) leads to a solution

$$\mathbf{m} = \left(\mathbf{G}^{\mathbf{T}} \mathbf{W} \mathbf{G} \right)^{-1} \mathbf{G}^{\mathbf{T}} \mathbf{W} \, \mathbf{d}^{(\mathbf{m})}, \tag{4}$$

where $W_{qq} = \sigma_q^{-2}$ ($q=1,2,\ldots Q$) is the q-th diagonal element of the weighting matrix **W**. Equation (4) is solved by an iterative process, in which the model is continuously refined. The quality of inversion results are checked by calculating a data misfit and parameter estimation errors (Menke 1984). The reliability is also possible to be quantified by using the matrix of Pearson's correlation coefficients, which measures the strength of linear relationships between the model parameters involved in the inversion procedure (Isaaks and Srivastava 1989).

The above detailed inversion method is generally accepted in the oil industry, which gives fast and good results in case of having a realistic initial model. However, as there are slightly more measured data than petrophysical unknowns in a given depth, the accuracy and reliability of inversion estimation performed point by point along the borehole are limited because of the marginal over-determination of the inverse problem. On the other hand, in case of poor starting models linearized inversion methods tend to trap in a local minimum of the objective function given in Equation (3). For getting a more robust inversion estimate of

inversion parameters a global optimization (Simulated Annealing or Genetic Algorithm) based inversion methodology for solving highly over-determined well-logging inverse problems was suggested by Dobróka and Szabó 2005, 2011, 2012 and Dobróka et al 2009, 2012.

2.2 Factor analysis

The first step in formulating the statistical problem is the construction of the *N*-by-*M* matrix of original data

$$\mathbf{D} = \begin{pmatrix} D_{11} & D_{12} & \cdots & D_{1M} \\ D_{21} & D_{22} & \cdots & D_{2M} \\ \vdots & \vdots & \vdots & \vdots \\ D_{i1} & D_{i2} & \cdots & D_{iM} \\ \vdots & \vdots & \vdots & \vdots \\ D_{N1} & D_{N2} & \cdots & D_{NM} \end{pmatrix}, \tag{5}$$

where each column represents one measurement type. The number of data variables is *M* and the total number of measuring (depth) points is *N*. The model of factor analysis is based on the following matrix decomposition

$$\mathbf{D} = \mathbf{F}\mathbf{L}^{\mathrm{T}} + \mathbf{E}, \tag{6}$$

where **F** denotes the *N*-by-*a* matrix of factor scores, **L** is the *M*-by-*a* matrix of factor loadings and **E** is the *N*-by-*M* matrix of residuals (superscript *T* denotes matrix transpose). The first term on the right side of Equation (6) can be detailed as follows

$$\mathbf{F}\mathbf{L}^{\mathrm{T}} = \begin{pmatrix} F_{11} & F_{12} & \cdots & F_{1a} \\ F_{21} & F_{22} & \cdots & F_{2a} \\ \vdots & \vdots & \vdots & \vdots \\ F_{N1} & F_{N2} & \cdots & F_{Na} \end{pmatrix} \begin{pmatrix} L_{11} & L_{12} & \cdots & L_{1M} \\ L_{21} & L_{22} & \cdots & L_{2M} \\ \vdots & \vdots & \vdots & \vdots \\ L_{a1} & L_{a2} & \cdots & L_{aM} \end{pmatrix} \quad (a < M), \tag{7}$$

where one column of matrix **F** represents the estimated values of a given factor in different measuring points (*a* is the number of factors). Factor scores of a column form a new well log that is called a factor log. The elements of matrix **L** represent the individual weights of the measured variables on the resultant factors. Assuming that factors are linearly independent $\mathbf{F}^{\mathrm{T}}\mathbf{F}/N = \mathbf{I}$ (where **I** is the identity matrix) the correlation matrix of the standardized original variables can be written as

$$\mathbf{R} = N^{-1}\mathbf{D}^{\mathrm{T}}\mathbf{D} = \mathbf{L}\mathbf{L}^{\mathrm{T}} + \mathbf{\Psi}, \tag{8}$$

where $\boldsymbol{\psi}=\mathbf{E}^{\mathbf{T}}\mathbf{E}/\boldsymbol{N}$ is the diagonal matrix of specific variances being independent of the common factors explaining the major part of variance of the original variables.

The factor loadings can be determined by a non-iterative approximate method suggested by Jöreskog (2007). Consider the sample covariance matrix **S** of the standardized data variables and the following derived quantity

$$\mathbf{S}^{*} = \left(diag\mathbf{S}^{-1}\right)^{1/2}\mathbf{S}\left(diag\mathbf{S}^{-1}\right)^{1/2}. \tag{9}$$

Calculating the eigenvalues λ and eigenvectors $\boldsymbol{\omega}$ of matrix $\mathbf{S}^{*}$ in Equation (9), the matrix of factor loadings can be computed as

$$\mathbf{L} = \left(diag\mathbf{S}^{-1}\right)^{-1/2}\mathbf{\Omega}_{a}\left(\mathbf{\Gamma}_{a} - \theta\mathbf{I}\right)^{1/2}\mathbf{U}, \tag{10}$$

where $\mathbf{\Gamma}_a$=diag($\lambda_1,\lambda_2,\ldots,\lambda_a$), $\mathbf{\Omega}_a=(\boldsymbol{\omega}_1,\boldsymbol{\omega}_2,\ldots,\boldsymbol{\omega}_a)$ and **U** is an arbitrary *a*-by-*a* orthogonal matrix. Constant θ is expressed by the rest of the eigenvalues not used in Equation (10)

$$\theta = \frac{1}{M - a}\left(\lambda_{a+1} + \lambda_{a+2} + \ldots + \lambda_{M}\right), \tag{11}$$

which specifies the smallest number of factors in case of θ<1. The factor scores can be estimated by the maximum likelihood method, which requires the optimization of the following log-likelihood function

$$P = -\left(\mathbf{D} - \mathbf{F}\mathbf{L}^{\mathrm{T}}\right)^{\mathrm{T}} \boldsymbol{\Psi}^{-1}\left(\mathbf{D} - \mathbf{F}\mathbf{L}^{\mathrm{T}}\right) = \max . \tag{12}$$

Fulfilling the condition $\partial P / \partial F = 0$ a linear solution can be given for factor scores (Bartlett 1937)

$$\mathbf{F} = \left(\mathbf{L}^{\mathrm{T}}\boldsymbol{\Psi}^{-1}\mathbf{L}\right)^{-1}\mathbf{L}^{\mathrm{T}}\boldsymbol{\Psi}^{-1}\mathbf{D} . \tag{13}$$

An orthogonal transformation of factor loadings is usually performed for the easier interpretation of uncorrelated factors. In this study, the varimax rotation algorithm was used that maximizes the sum of the variances of the squared factor loadings (Kaiser 1958).

Szabó (2011) concluded that there is a strong correlation between the first factor (1st column of matrix **F**) and shale volume estimated by inverse modeling. At first, the factor scores need to be rescaled for a comparative study

$$F_1' = F_{1,min}' + \frac{F_{1,max}' - F_{1,min}'}{F_{1,max} - F_{1,min}}\left(F_1 - F_{1,min}\right), \tag{14}$$

where $F_{1,min}$ and $F_{1,max}$ are the minimum and maximum values of factor scores derived from Equation (13) respectively, $F'_{1,min}$ and $F'_{1,max}$ are the desired lower and upper limit of the rescaled factor F_1' respectively. In the next step, the functional dependence between F'_1 and V_{sh} is determined. Regression analyses show the following non-linear relationship between the above quantities

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

where *a* and *b* are regression constants. This paper deals with an assumption that the choosing of coefficients *a* and *b* is nearly independent from the measurement area and Equation (15) can be treated as a general formula approximately valid in different sedimentary basins. To measure the quality of curve fitting based on Equation (15) we apply a special type of correlation calculus. The strength of non-linear relationship between two variables can be characterized by the rank correlation coefficient (Spearman 1904), which is applied to measure the goodness of regression results.

3. Geological settings

The first application of Equation (15) was showed by Szabó (2011). In this study, the earlier suggested empirical formula is extended to new prospecting areas. Three test sites representing different locations and geological environments were chosen for the comparative study. The short description of geology and the lists of processed well-logging data sets are presented below.

The first data set was collected from a hydrocarbon exploratory borehole drilled in the Great Hungarian Plain, East Hungary (Area-1). The investigation area is part of the Pannonian Basin Province of Central Europe, where several petroleum systems have been discovered and exploited. The Pannonian Basin consists of a large sedimentary basin of Neogene age overlying Paleogene basins and a Mesozoic or Paleozoic basement (Dolton 2006). The Pannonian sediments contain oil and gas-bearing formations, thermal and freshwater resources at the top. The sequence of the processing interval represents Upper Miocene (Pannonian) aged lacustrine deposits of the Lake Pannon, which form an unconsolidated sand complex interbedded with clay and silt layers. The reservoirs has got high and medium porosities and are saturated with water and gas. The processed well logs including caliper (CAL), spontaneous potential (SP), natural gamma-ray intensity (GR), compensated neutron

(CN), compensated density (DEN), acoustic traveltime (AT), microlaterolog (RMLL), shallow (RS) and deep resistivity (RD) measurements are plotted in Figure 1.

The second area was chosen from the literature (Asquith and Krygowski 2004). The investigated well was drilled in the Permian Basin of Texas, United States of America (Area-2). The Atoka Formation is a sequence of marine, tan to gray, silty sandstones and grayish black shales, which extends over a very large area from Arkansas to New Mexico and West Texas. The sediments of Area-2 represents different age and burial depth with respect to Pannonian sands. The common feature between the two formations is that both of them consist of unconsolidated sediments. The matrix of the Pennsylvanian aged Atoka sandstone was made up of loose and dominantly coarse mineral grains, which were poorly cemented. The sand was characterized by large porosity and permeability and the pore space was filled with water, gas and condensate. The data set consisted of caliper (CALI), spontaneous potential (SP), natural gamma-ray intensity (GR), neutron porosity (NPHI), density porosity (DPHI), deep induction (ILD), medium induction (ILM) and shallow resistivity (LL8) logs (Figure 2).

The third area was also chosen from the literature (Anna 2009). The well site can be found in the Powder River Basin of Wyoming, United States of America (Area-3). The hydrocarbon reservoir rocks shows similar age, lithologic characteristics and production history in the basin. The Minnelusa Formation is Pennsylvanian and Permian in age. The rocks of the reservoir zone as continental shelf sediments produce oil and gas, which are generally composed of sandstone, carbonate, and evaporites. The well logs represent a short zone with medium porosity and permeability, where the high resistivities indicates hard dolomites. The well logs applied to factor analysis were spontaneous potential (SP), natural gamma-ray

intensity (GR), acoustic (interval) traveltime (DT) and deep resistivity (RD). The data set can be seen in Figure 3.

4. Test computations

Szabó (2011) showed that practically the same regression coefficients for Equation (15) are valid in two hydrocarbon wells drilled ~200 km apart from each other in the Great Hungarian Plain, Hungary. Some new interpretation results of deep wells also suggest the use of $V_{sh} = 2.76e^{0.037F_1'}$ as a local formula for the Pannonian Basin (Area-1). To extend the validity of the above empirical relationship the three well-logging data sets presented in Section 3 are processed by factor analysis and the resultant shale volume logs are compared to that of independent shale volume calculations.

Factor analysis of the data sets was performed separately, where the number of factors was specified previously. The average of Pearson's correlation coefficients for the measured variables were 0.08, 0.10, 0.29 for Area-1, Area-2 and Area-3, respectively. The weakly correlated measurement variables were transformed into some uncorrelated factors. In East Hungary and West Texas the major part of the variance of the original data was explained by the first two factors (89% and 93% for Area-1 and Area-2, respectively), thus two factors were extracted. In Wyoming only one factor represented the main information of the original data set (99% of total variance). The estimated factor loadings for the three test sites are shown in Table 1. In each cases, the largest weights on the first factor were given by GR and SP logs. At the same time, other well logs sensitive to porosity and saturation also assisted in the development of the first factor. A significant difference was that resistivity measurements got very different values of factor loadings in the test sites. For making the results comparable the factors were transformed into the same scale. The new interval of factor scores was computed according to Equation (14). Since shale volume lies between 0% and 100% then

$F'_{1,min}$=0 and $F'_{1,max}$=100 were chosen. The exponential relationships between the (scaled) first factor and shale volume are illustrated in Figure 4. The regression analysis proved that Equation (15) including practically the same values of exponents *a* and *b* approximated well the relationships between the factor scores and shale volumes in the three test sites. While parameter *b* was fixed as 0.037, coefficient *a* was estimated 2.756 (a_{min}=2.744, a_{max}= 2.768), 2.763 (a_{min}=2.712, a_{max}=2.814), 2.705 (a_{min}=2.66, a_{max}=2.751) with 95% confidence bounds for Area-1, Area-2 and Area-3, respectively. The rank correlation coefficient was 0.98 in all cases (assuming linear connection the Pearson's correlation coefficients were calculated around 0.8).

Shale volume was calculated from the factor scores by using Equation (15) with the actual regression coefficients and the results were compared to estimates of independent well log analysis methods. In Area-1 factor analysis was checked by local (depth-by-depth) inverse modeling using a weighted least squares method [Eq. (4)]. In Area-2, the result of deterministic interpretation of several well logs was used for verification (Asquith and Krygowski 2004). In Area-3 the Larionov formula gave reference data for shale volume estimation. In Figure 4, shale volume logs estimated by factor analysis as well as traditional interpretation methods are shown for the measurement areas. It is concluded that there is a good agreement between the independent interpretation results. The RMS error between the resultant logs was 8.2%, 9.6%, 4.3% for Area-1, Area-2 and Area-3, respectively. The procedures took only a few seconds of CPU times using a quad-core processor based workstation.

5. Discussion

Shale volume estimation is frequently performed by using a shale indicator derived from the GR log. The linear model connecting gamma-ray index to shale volume is often overestimates

the fraction of shale in real formations (Poupon and Gaymard 1970). Field experiments showed that some non-linear fomulae gave more accurate results (Clavier et al 1971; Larionov 1969; Stieber 1970) that can be used independently from the location of the well. In Figure 6 the theoretical relationships between natural gamma-ray index and shale volume proposed by the above deterministic formulae with the estimation results of factor analysis can be compared in case of Area-1. The crossplot confirms the validity of Equation (15) with coefficients a=2.76 and b=0.037. It must be mentioned that the estimated constants are approximate values, and as a result of local regression analyses they have to be specified in different areas. Though, some data sets of other well-sites situated in North and South Hungary, Alaska and Ohio (USA) have already been successfully processed, additional measurements are required for further research. After processing them, it is expected that the statistical formula can be extended to even larger areas. The universality of the empirical formula is a strong assumption. In this stage of research its validity is proven in deep wells, clean and shaly water and hydrocarbon-bearing clastic reservoirs. In these formations all well logs are very sensitive to the amount of shaliness, that is why the first factor representing the most information of measurements highly correlates with shale volume. In multi-mineral rocks or complex reservoirs other matrix components may influence the same factor and the correlation is not so strong. According to our research Equation (15) holds also for shallow wells, but the coefficients are somewhat different.

A necessary feasibility condition for giving an optimal solution by the maximum likelihood method is that the input data set is to follow normal distribution. For measuring normality we calculated the skewness of individual well logs, which is the ratio of the third central moment to the cube of standard deviation of the data. Zero skewness indicates that the probability density function is symmetrical and data follow Gaussian distribution. In Areas-1-3 near zero skewnesses were calculated. To study the dependence of factor analysis on the nature of the

statistical distribution of data we generated several quasi measured data sets from that of Area-2. The new data sets (**D**') became antisymmetrical distributions, which were derived from the original (near-symmetrical) data (**D**) by using the transformation of $D'_{ij}=D_{ij}(1+R(\tau)/C_j)$, where $R(\tau)$ is a random number selected from Rayleigh distribution (scale parameter τ was set to 0.5) and C_j is a properly chosen constant depending on the range of the *j*-th well log to control the distance between the original and the new data set. The data distance was calculated as

$$DIST = \sqrt{\frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}\left(\frac{D_{ij}-D'_{ij}}{D_{ij}}\right)^2}\cdot 100(\%)\,. \qquad (16)$$

The new values of shale volume were calculated by the Larionov's formula. We performed factor analysis on the new data sets with different data distances. Regression tests were applied to estimate coefficient *a* beside the fixed value of parameter *b* (0.037). Figure 7 shows that the value of coefficient *a* slightly descreases with increasing skewness. It is concluded that the empirical formula proposed in this study provides a good approximation in case of near Gaussian or moderately skewed data distributions. However, in case of larger asymmetry of data requires a more robust form of the factor analysis algorithm.

6. Conclusions

It is inferred that factor analysis is applicable to extract the shale content as basic lithologic information of sedimentary rocks from well-logging data. The results of factor analysis of data sets acquired from three distant areas were compared and verified by independent estimates of deterministic and inverse modeling. In this stage of research an exponential relationship between the first factor and shale volume is assumed in sedimentary geological environments. This relationship proves to be straight and applicable both in water and

hydrocarbon reservoirs. With the implementation of the statistical method a more reliable estimation to shale volume and other derived petrophysical parameters, such as effective porosity, water and hydrocarbon saturation, absolute permeability, can be made in primer porosity reservoirs. These parameters given from different sources can be confirmed or updated by the results of factor analysis in the prospecting area. The estimation of shale volume is an important issue for the success of inverse modeling, too. By giving a reliable estimate for shale volume by an independent source, i.e. factor analysis, the number of unknowns can be decreased in the well-logging inverse problem. Releasing shale volume from the list of inversion unknowns, the over-determination of the inverse problem can be effectively increased. Consequently, the local (marginally over-determined) inversion procedure can give a more accurate estimate for the inversion model. The algorithm of factor analysis can also be implemented to multi-well applications (Szabó et al 2012). If data of several wellbores are integrated into one statistical procedure, it is possible to determine the 2D/3D spatial distribution of shale volume. Considering all of the above advantageous properties, it is expected that factor analysis will strengthen the potential of well-logging data processing in geological exploration.

Acknowledgments

The described work was carried out as part of the TÁMOP-4.2.1.B-10/2/KONV-2010-0001 project in the framework of the New Hungary Development Plan. The realization of this project is supported by the European Union, co-financed by the European Social Fund. The first author thanks to the support of the János Bolyai Research Fellowship of the Hungarian Academy of Sciences. As a member of the MTA-ME Applied Geoscience Research Group the second author is grateful for the support of the Hungarian Academy of Sciences.

References

Alberty M, Hashmy K (1984) Application of ULTRA to log analysis. SPWLA Symposium Transactions. Paper Z: 1-17.

Anna L O (2009) Geologic Assessment of Undiscovered Oil and Gas in the Powder River Basin Province, Wyoming and Montana. U.S. Geological Survey, Reston, Virginia.

Asquith G B, Krygowski D A (2004) Basic Well Log Analysis (2nd edition), AAPG Methods in Exploration Series 16. AAPG, Tulsa.

Ball S M, Chace D M, W H Fertl (1987) The Well Data System (WDS): An advanced formation evaluation concept in a microcomputer environment. Proceedings of SPE Eastern Regional Meeting. Paper 17034: 61-85.

Bartlett M S (1937) The statistical conception of mental factors. British Journal of Psychology 28: 97–104.

Bhuyan K. and Passey Q. R. 1994. Clay estimation from GR and neutron-density porosity logs. SPWLA 35th Annual Logging Symposium, Conference paper: 1994-DDD.

Buoro A B, Silva J B C (1994) Ambiguity analysis of well-log data. Geophysics 59: 336-344.

Clavier C, Hoyle W, Meunier D (1971) Quantitative interpretation of thermal neutron decay time logs: Part I. Fundamentals and techniques. Journal of Petroleum Technology 23: 743-755.

Dobróka M, Szabó N P (2005) Combined global/linear inversion of well-logging data in layer-wise homogeneous and inhomogeneous media. Acta Geodaetica et Geophysica Hungarica 40: 203-214.

Dobróka M, Szabó N P, Cardarelli E, Vass P (2009) 2D inversion of borehole logging data for simultaneous determination of rock interfaces and petrophysical parameters. Acta Geodaetica et Geophysica Hungarica 44: 459-482.

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

Dobróka M, Szabó N P, Turai E (2012) Interval inversion of borehole data for petrophysical characterization of complex reservoirs. Acta Geodaetica et Geophysica Hungarica 47: 172-184.

Dobróka M, Szabó N P (2012) Interval inversion of well-logging data for automatic determination of formation boundaries by using a float-encoded genetic algorithm. Journal of Petroleum Science and Engineering 86-87: 144-152.

Dolton G L (2006) Pannonian Basin Province, Central Europe (Province 4808) - Petroleum geology, total petroleum systems, and petroleum resource assessment. USGS Bulletin 2204-B: 1-47.

Ellis D V, Singer J M (2007) Well logging for earth scientists (2nd edition). Springer, Dordrecht.

Fertl W H, Frost Jr E (1980) Evaluation of Shaly Clastic Reservoir Rocks. Journal of Petroleum Technology 32: 1641-1646.

Grana G, Dvorkin J, Cibin P (2011) Factor analysis prediction of effective stress from measurable rock attributes and calibration data. First Break 29: 63-72.

Herron M M (1986) Mineralogy from geochemical well logging. Clays and Clay Minerals 34: 204-213.

Isaaks E H, Srivastava R M (1989) An introduction to applied geostatistics. Oxford University Press, Oxford.

Jöreskog K G (2007) Factor analysis and its extensions. In**:** Cudeck R, MacCallum R C (eds) Factor analysis at 100, Historical developments and future directions. Lawrence Erlbaum Associates, New Jersey.

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

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

Lawley D N, Maxwell A E (1962) Factor analysis as a statistical method. The Statistician 12: 209-229.

Ma Y Z (2011) Lithofacies Clustering Using Principal Component Analysis and Neural Network: Applications to Wireline Logs. Mathematical Geosciences 43: 401-419.

Mayer C, Sibbit A (1980) GLOBAL, a new approach to computer-processed log interpretation. Proceedings of 55th SPE Annual Fall Technical Conference and Exhibition. Paper 9341: 1-14.

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

Poupon A, Gaymard R (1970) The evaluation of clay content from logs. SPWLA 11th Annual Logging Symposium, Conference paper: 1970-G.

Rao B N, Pal P C (1980) Factor analysis for interpreting petrophysical data on Roro ultramafics, Singhbhum district, India. Geophysical Prospecting 28: 112-118.

Serra O (1984) Fundamentals of well-log interpretation: The acquisition of logging data. Elsevier, Amsterdam.

Spearman C (1904) The proof and measurement of association between two things. The American Journal of Psychology 15: 72-101.

Stieber S. J. 1970. Pulsed neutron capture log evaluation - Louisiana Gulf Coast. SPE Fall Meeting AIME, Conference paper: 2961-MS.

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

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

Szabó N P (2012) Dry density derived by factor analysis of engineering geophysical sounding measurements. Acta Geodaetica et Geophysica Hungarica 47: 161-171.

Urbancic T I, Bailey R C (1988) Statistical techniques applied to borehole geophysical data. Geophysical Prospecting 36: 752-771.

List of figure captions

Figure 1. Well logs recorded in Pannonian sand (Area-1) as input for the factor analysis. Measured quantities are: caliper CAL (inch), spontaneous potential SP (mV), natural gamma-ray GR (API), compensated neutron CN (%), density DEN (g/cm^3), acoustic traveltime AT (μs/ft), microlaterolog RMLL (ohmm), shallow resistivity RS (ohmm), deep resistivity RD (ohmm).

Figure 2. Well logs measured in Atoka sandstone (Area-2) as input for factor analysis. Observed quantities are: caliper CALI (inch), spontaneous potential SP (mV), natural gamma-ray GR (API), neutron porosity NPHI (v/v), density porosity DPHI (v/v), deep induction ILD (ohmm), medium induction ILM (ohmm), shallow resistivity LL8 (ohmm).

Figure 3. Well logs observed in Minnelusa formation (Area-3) as input for factor analysis. Measured variables are: spontaneous potential SP (mV), natural gamma-ray intensity GR (API), acoustic traveltime DT (μs/ft), deep resistivity RD (ohmm).

Figure 4. Regression relationships between the first scaled factor (F'_1) and shale volume (V_{sh}) for Area-1 (A), Area-2 (B), Area-3 (C).

Figure 5. Shale volume estimates from factor analysis represented by blue curve for Area-1 (A), red curve for Area-2 (B), orange curve for Area-3 (C). Independent petrophysical modeling results are illustrated with black curves.

Figure 6. Natural gamma-ray index vs. shale volume crossplot. Different natural gamma-ray intensity log based empirical equations are indicated with black curves. Estimation results of factor analysis are represented with grey dots.

Figure 7. Factor analysis results of positively skewed well-logging data sets. (A) Data distance vs. regression coefficient *a* (B), data distance vs. average skewness (C), first scaled factor vs. shale volume (C).

List of table captions

Table 1. Factor loadings estimated by the factor analysis of well-logging data measured from Pannonian sand (Area-1), Atoka sandstone (Area-2), Minnelusa formation (Area-3).

Tables

Table 1.

Area-1 Pannonian Sand (East Hungary)			Area-2 Atoka Sandstone (West Texas, USA)			Area-3 Minnelusa Formation (Wyoming, USA)	
Well log	**Factor 1**	**Factor 2**	**Well log**	**Factor 1**	**Factor 2**	**Well log**	**Factor 1**
CAL	0.46	-0.02	CALI	0.47	0.79	-	-
CN	0.91	0.25	NPHI	0.48	0.82	-	-
DEN	0.79	-0.60	DPHI	0.09	0.78	-	-
AT	0.12	0.79	-	-	-	AT	0.37
GR	0.94	-0.04	GR	0.76	0.56	GR	0.82
RD	-0.68	-0.06	ILD	-0.93	-0.31	RD	-0.26
RMLL	-0.72	0.57	LL8	-0.89	-0.26	-	-
RS	-0.18	-0.01	ILM	-0.75	-0.30	-	-
SP	-0.83	-0.15	SP	0.75	-0.08	SP	0.55

Figures

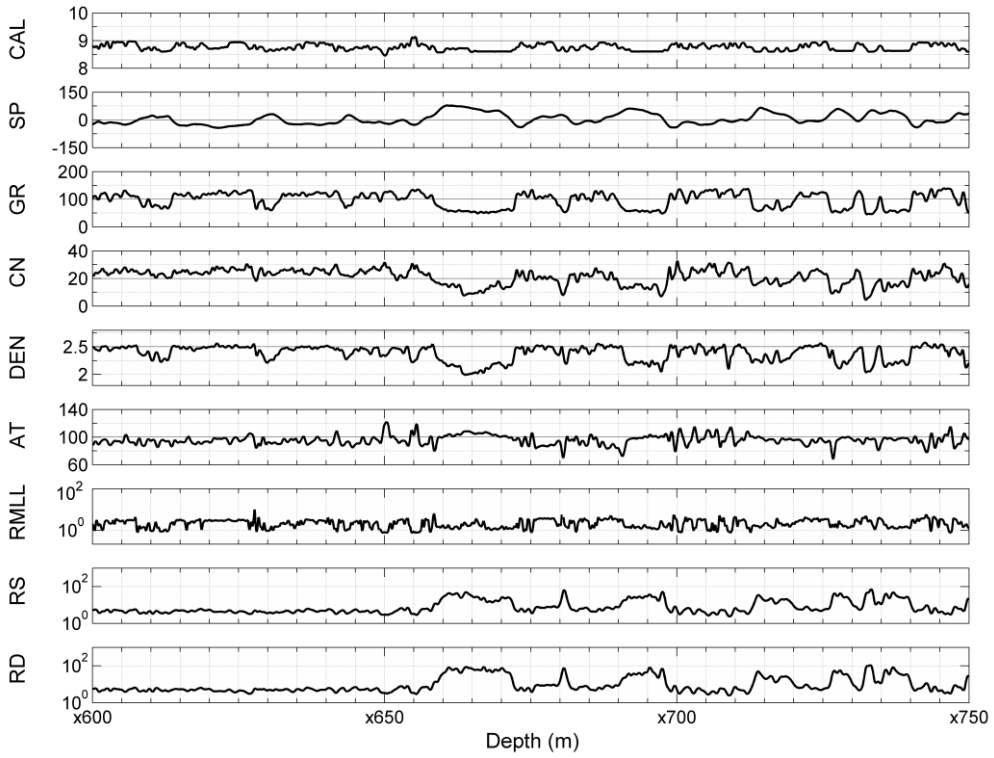

Fig. 1

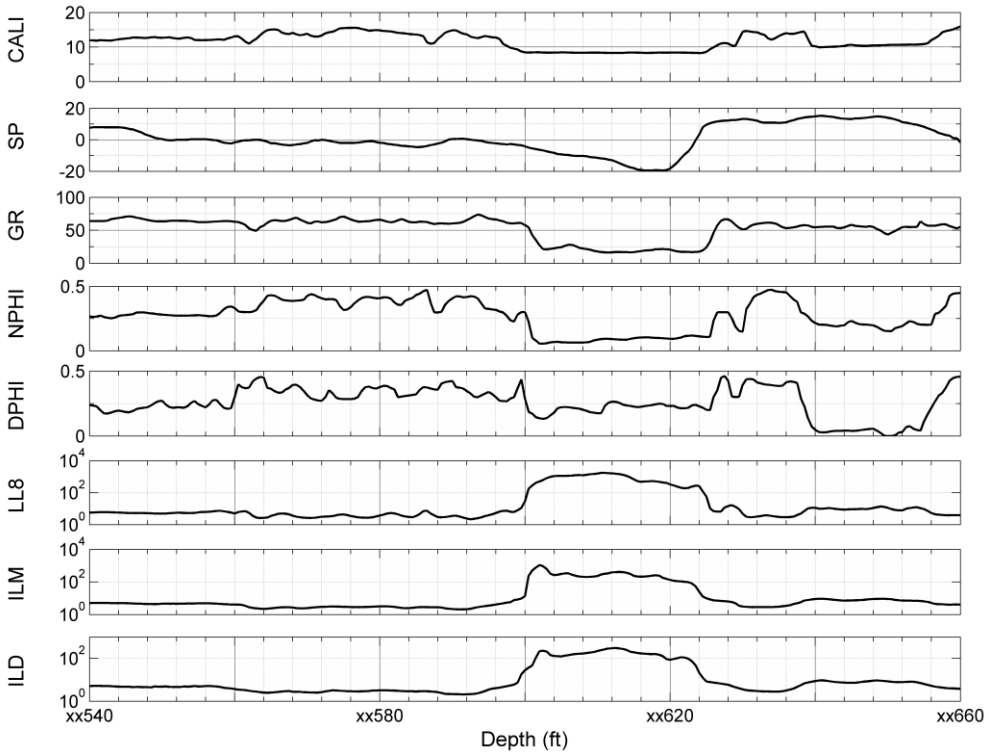
CALI
SP
GR
NPHI
DPHI
LL8
ILM
ILD
xx540
xx580
xx620
xx660
Depth (ft)

Fig. 2

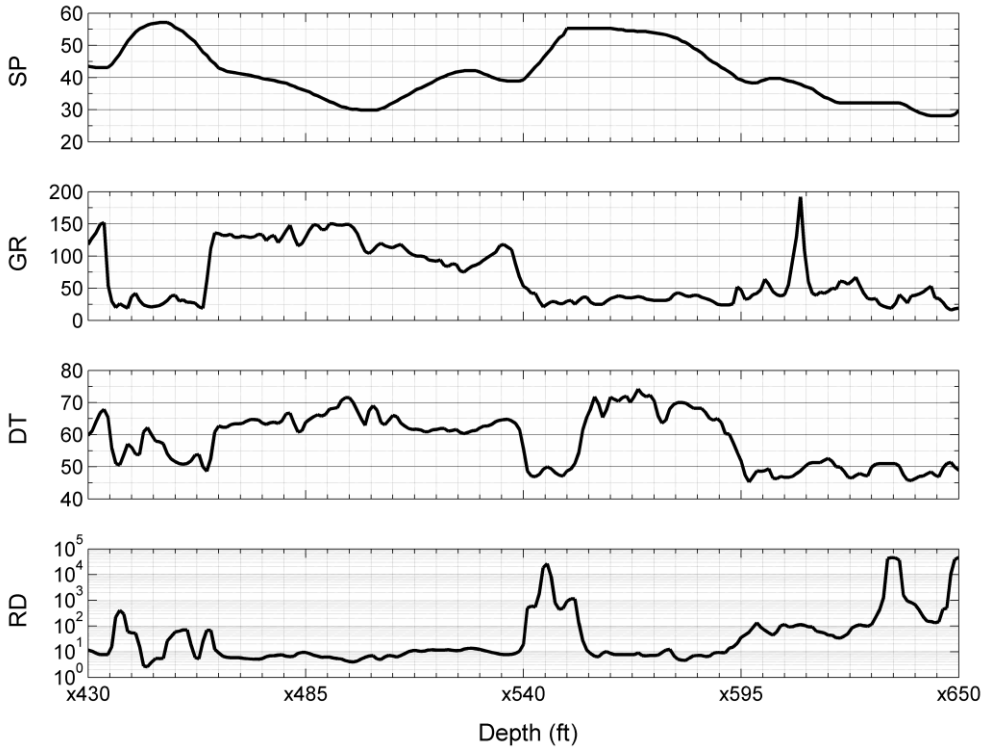
SP
60
50
40
30
20
GR
200
150
100
50
0
DT
80
70
60
50
40
RD
10^5
10^4
10^3
10^2
10^1
10^0
x430
x485
x540
x595
x650
Depth (ft)

Fig. 3

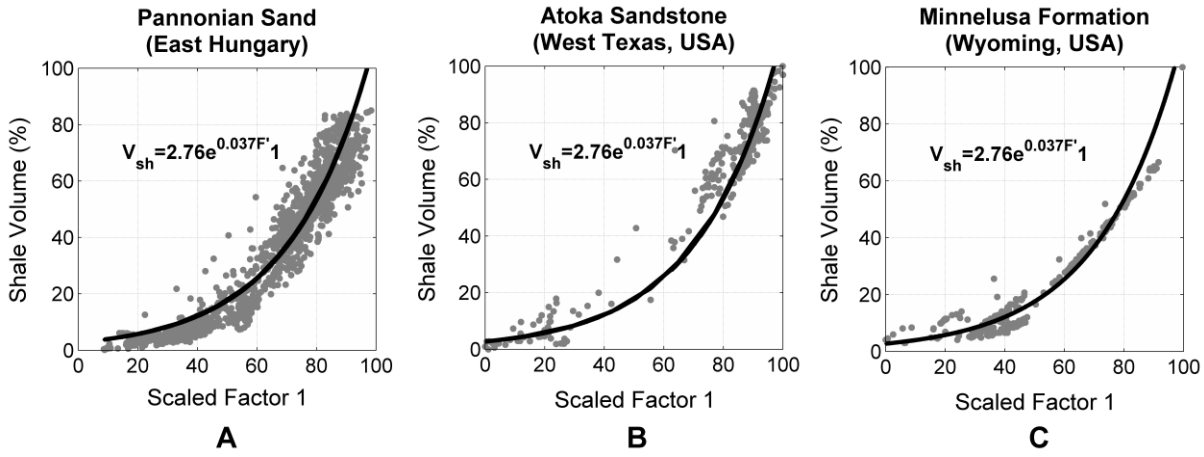
Pannonian Sand
(East Hungary)
Atoka Sandstone
(West Texas, USA)
Minnelusa Formation
(Wyoming, USA)
Shale Volume (%)
Scaled Factor 1
$V_{sh}=2.76e^{0.037F'1}$
A
B
C

Fig. 4

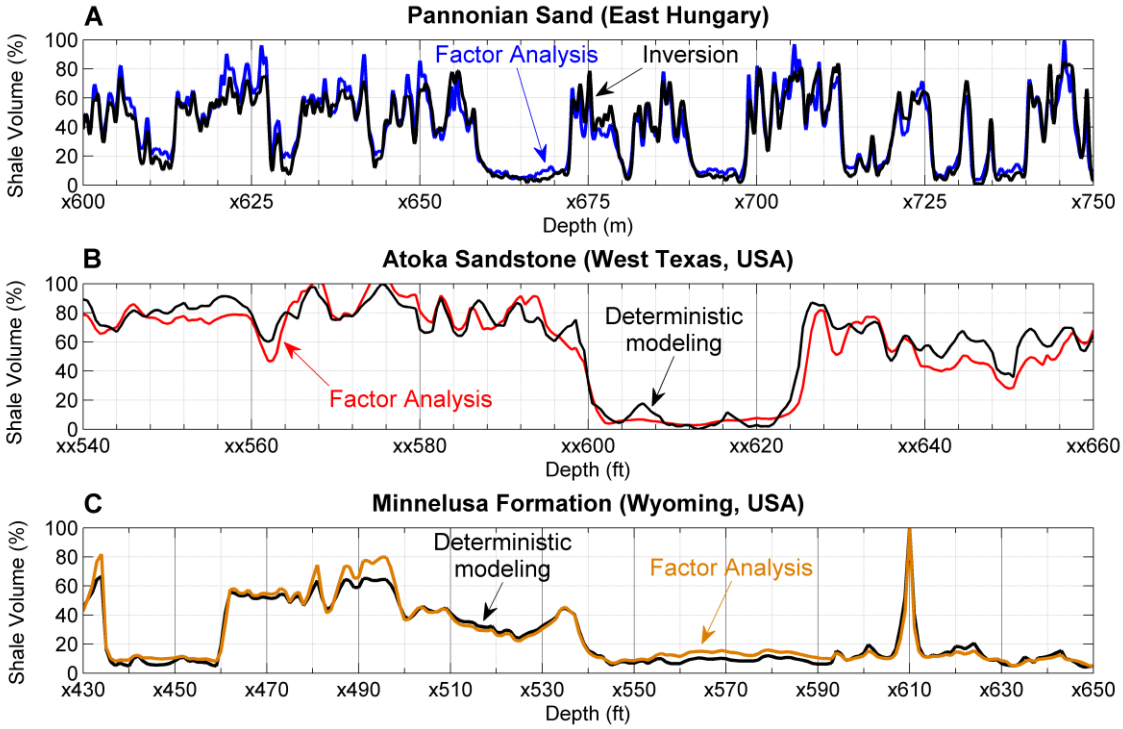
A
Pannonian Sand (East Hungary)
Factor Analysis
Inversion
Shale Volume (%)
100
80
60
40
20
0
x600
x625
x650
x675
x700
x725
x750
Depth (m)
B
Atoka Sandstone (West Texas, USA)
Deterministic modeling
Factor Analysis
Shale Volume (%)
xx540
xx560
xx580
xx600
xx620
xx640
xx660
Depth (ft)
C
Minnelusa Formation (Wyoming, USA)
Deterministic modeling
Factor Analysis
Shale Volume (%)
x430
x450
x470
x490
x510
x530
x550
x570
x590
x610
x630
x650
Depth (ft)

Fig. 5

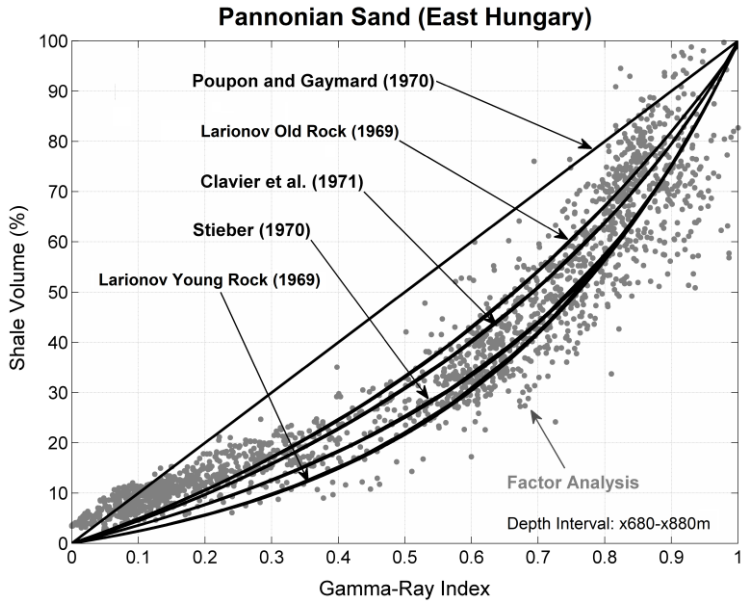
Pannonian Sand (East Hungary)
Poupon and Gaymard (1970)
Larionov Old Rock (1969)
Clavier et al. (1971)
Stieber (1970)
Larionov Young Rock (1969)
Factor Analysis
Depth Interval: x680-x880m
Shale Volume (%)
Gamma-Ray Index
0
10
20
30
40
50
60
70
80
90
100
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

Fig. 6

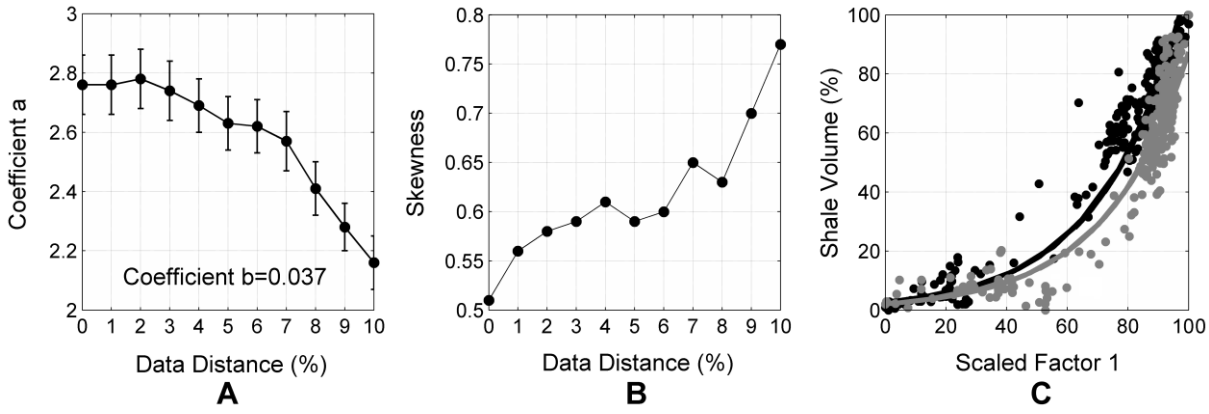
Coefficient a
Coefficient b=0.037
Data Distance (%)
A
Skewness
Data Distance (%)
B
Shale Volume (%)
Scaled Factor 1
C

Fig. 7