Factor analysis of borehole logs for evaluating formation shaliness - A hydrogeophysical application for groundwater studies

Norbert Péter Szabó[1*], Mihály Dobróka[1,2], Endre Turai[1], Péter Szűcs[1,2]

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

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

*Corresponding author

E-mail: norbert.szabo.phd@gmail.com

Phone: +3646361936

Fax: +3646361936

Abstract

The calculation of groundwater reserves in shaly sand aquifers requires a reliable estimation of effective porosity and permeability. The amount of shaliness as a related quantity can be extracted from geophysical well log analysis. The conventionally used linear model connecting natural gamma-ray index to shale content often gives a rough estimate in shallow boreholes. To get a better result a non-linear model is suggested, which is derived from the factor analysis of well-logging data. An earlier study of hydrocarbon wells revealed an empirical relationship between the factor scores and shale volume independent of the well site. Borehole logs measured from three groundwater wells drilled in Hungary are analyzed to estimate the logs of factor variables, which are then correlated with shale volumes given from the method of Larionov. Shale volume logs estimated by the statistical procedure are in sufficiently close agreement with those derived from the Larionov's formula that confirms the validity of the non-linear approximation. The statistical results are accordance with laboratory measurements made on core samples. However, whereas the conventional borehole geophysical methods normally use a single well log as an input, factor analysis processes all available logs to provide groundwater exploration with highly reliable estimation results.

Keywords: shale volume, well-logging data, factor analysis, factor log, Larionov's formula

1. Introduction

Shale is a fine-grained sedimentary rock composed of a half-and-half mixture of clay minerals and silt. It can appear in three different forms, i.e. dispersed shale that distributes within the interstices of the rock, structural clays that are incorporated in the rock matrix, and shale laminae that build thin beds between porous-permeable formations. Petrophysical properties of clastic formations generally vary with the distribution and quantity of shale. According to empirical studies, the seismic velocity is proportional to clay volume (Han et al. 1986; Klimentos 1991). Since the effective porosity decreases with shale volume, well-logging data usually need to be corrected for shale effect (Thomas and Stieber 1975). The electrical resistivity measurement used to extract information on water saturation gives a minimum reading in shale, which is usually caused by the high surface conductivity of clay minerals. Resistivity models applied in well log analysis assume that free-water and dispersed clay particles conduct an electrical current like a mixture of electrolytes (De Witte 1955). The variation of grain size affect greatly the permeability of shallow formations, thus the permeability is strongly influenced by shale volume. Some authors suggest empirical formulae between permeability and shale volume based on field or laboratory measurements (Revil and Cathles 1999; Slater and Lesmes 2002).

The total volume of shale as one of the key-parameters can be calculated by deterministic, inverse or statistical modeling procedures. The most common deterministic approach is based on the individual analysis of well logs. Measurements that are the most sensitive to the variation of shale content such as spontaneous potential and natural gamma-ray intensity are substituted into explicit equations to calculate the amount of shale separately (Asquith and Krygowski 2004). More reliable estimation can be given by using more well logs simultaneously, e.g. the combination of porosity logs (Poupon and Gaymard 1970). Several near-surface geophysical and hydrogeological studies apply deterministic techniques to extract shale volume from well-logging data for the characterization of shallow geology (Paillet 1995; Fisher et al. 1998; Cripps and McCann 2000; Kvapil and Mares 2003; Doveton and Merriam 2004; Hsieh et al. 2007; Adeoti et al. 2009; Maliva et al. 2009).

Advanced shale volume estimation methods are based on the simultaneous processing of all suitable well logs. Well-logging data are usually integrated into a joint inversion procedure to evaluate petrophysical parameters such as porosity, water saturation, mineral content and shale volume. Probe response equations represent the link between the measurements and the petrophysical model, which are used to calculate theoretical data in the straightforward modeling phase of the inversion procedure. The optimal solution of the inverse problem can be developed by using an appropriate optimization algorithm that fits the calculated data to the measured data. Since the number of data types observed in a given depth is barely more than that of the unknowns and the noise level of the data are different, a marginally overdetermined inverse problem has to be solved in each depth by using the weighted least squares method (Menke 1984). The principles of the well-logging inverse problem are detailed in Mayer and Sibbit (1980), Alberty and Hashmy (1984), Ball et al. (1987), and a novel inversion methodology (called interval inversion) was suggested by the research team of the Department of Geophysics, University of Miskolc (Dobróka and Szabó 2005; Dobróka et al. 2009; Dobróka and Szabó 2011; Dobróka and Szabó 2012; Dobróka et al. 2012). Applications of inversion processing of well-logging data for the shallow region can be found, e.g. in Paillet and Crowder (1996), Beltrami et al. (1997), Moret et al. (2004), Drahos (2005), Jang and Kim (2008).

The third alternative for data processing is represented by statistical methods which describe the empirical relationships between the observed physical quantities and petrophysical parameters not directly measurable by well-logging. In this study, a multivariate statistical procedure is offered to find correlation between well-logging data and shale volume. Factor analysis is commonly used to reduce high-dimensional data sets to lower dimensions (Lawley and Maxwell 1962). Hempkins (1978) mentioned first that factor analysis seemed to be a powerful tool in formation evaluation. Principal component analysis as an option of solving the problem of factor analysis has been widely used in petrophysics and hydrogeology (Elek 1988; Elek 1990; Kazmierczuk and Jarzyna 2006; Tanos et al. 2011; Magyar et al. 2013). The feasibility of factor analysis has been demonstrated by the solution of various borehole geophysical problems (Rao and Pal 1980; Herron 1986; Buoro and Silva 1994; Grana et al. 2011; Szabó et al. 2012; Szabó 2012). A modern approach that is useful for considering the time dependence of factor variables is the dynamic factor analysis, which was used successfully in hydrogeological problems by Márkus et al. (1999) and Kovács et al. (2004).

An earlier study of deep wells revealed a non-linear relationship between one of the new variables derived by factor analysis and shale volume, which proved to be nearly independent of the measurement area (Szabó 2011). Szabó and Dobróka (2011) presented a field case for the factor analysis of a data set measured in a Hungarian thermal-water well, too. It is assumed that the same statistical methodology applied to geophysical data originated from shallow boreholes finds a similar regression relationship between the above quantities and the method can be applied also to shallow freshwater formations. In the present study, three field examples are shown to demonstrate that the correlation between a given factor and shale volume is strong and practically independent from the measurement area. The aim of the study is to establish an independent shale volume estimation method for hydrogeophysical applications that underlie the prediction of effective porosity and permeability of aquifers.

2. Conventional interpretation methods

2.1 Borehole logging in hydrogeophysics

A comprehensive summary of applying well-logging techniques to evaluate petrophysical parameters of aquifers was given by Tselentis (1985). The classification of measurements can be made by different principles. A practical approach was chosen, which is based on the concept of parameter sensitivities (Gyulai 1995; Dobróka and Szabó 2011). The parameter sensitivity in borehole geophysics informs us about the extent of influence on well-logging data, which is exerted by a given petrophysical parameter. Based on the theory of parameter sensitivities, three main classes can be distinguished that comprise well logs primarily sensitive to lithology, porosity and water saturation, separately. It can be mentioned that all types of well logs are sensitive to each petrophysical properties to some extent. For instance, the spontaneous potential (*SP*) log is highly sensitive to the variation of shale content, but the amount of influence is much higher than it is caused by the change in pore-fluid content. It means that the deflections of the *SP* curve can be easily transformed into shale volumes, but the suppression on the same curve caused by the presence of gas or other pore-fluids is not adequate to estimate the amount of saturation or permeability. As opposed to *SP* log, the neutron porosity (*NPHI*) log is also influenced by lithology, but it is the water-filled porosity which can be determined quantitatively from the measurement. Therewith, the precise interpretation of the neutron log requires the application of corrections on the log readings for the type of lithology and other environmental effects. From the point of view of corrections, nuclear magnetic resonance logging (*NMR*) is a highly favorable technique because it is marginally sensitive to lithology and provides direct (quantitative) information of effective porosity and permeability (Coates et al. 1999).

In groundwater prospecting lower cost logging programs are generally performed than used in hydrocarbon exploration. Therefore, the extraction of reliable petrophysical information from the observations requires advanced data processing techniques. In the present study, natural gamma ray intensity (*GR*) and spontaneous potential (*SP*) logs are applied as lithology logs. The former measures the natural radioactivity of formations, in cpm unit, caused by different amounts of potassium, thorium and uranium content. The latter records the values of electric potential between a surface and a down-hole electrode, in mV-s, excited by the ion movement between drilling mud and original pore fluid as well as the presence of shale. Both of them can be used to predict shale content quantitatively. For porosity determination nuclear logs such as density (*DEN*) and neutron porosity (*NPHI*) are used, which is sometimes complemented by acoustic traveltime (*AT*) or full-wave sonic (Variable Density Log - *VDL*) measurements. Bulk density measured by gamma-gamma probes, in g/cm^3 units, shows inverse proportionality to porosity. Neutron-neutron measurements are mainly sensitive to the hydrogen-index of formations, which infer the fluid-filled porosity in the absence of hydrogen atoms in the rock matrix. Several types of resistivity tools with different depth of investigation and vertical resolution can be applied to detect the invasion profile and to estimate water saturation in different zones around the borehole. Since freshwater has normally got higher resistivity than brine, traditional (non-focused) probes are generally suitable for groundwater exploration. While shallow resistivity (*RS*) tools measure the apparent resistivity, in ohm-m unit, of the zone invaded by mud, the deep resistivity (*RD*) instrument observes the same quantity in the original (non-invaded) formation. The resistivity readings are corrected to predict the true formation resistivity, which is of great importance in calculating water and gas (normally air) saturation in aquifers. The well-logging programs usually contain also some other types of technical measurements that are not used directly in the petrophysical characterization of rocks, but give important information on the technical conditions of the borehole wall and its environment, pressure, temperature distributions, the flow-rate and composition of original pore fluid along the borehole.

2.2 Deterministic methods for interpretation

Shale volume (V_{sh}) as the ratio of the quantity of shale to total volume of rock is most commonly determined by the analysis of the intensity of the natural gamma-ray log. Since a non-radioactive formation is indicated by a minimum (GR_{min}) in the profile and the radioactive shale by a maximum (GR_{max}) in the same profile, the natural gamma-ray index in a depth is

$$i_{GR} = \frac{GR_{log} - GR_{min}}{GR_{max} - GR_{min}}, \tag{1}$$

which can be used for the linear approximation of shale volume in the form of $V_{sh}=i_{GR}$ (Poupon and Gaymard 1970), where GR_{log} is the natural gamma-ray intensity measured in a given depth. Equation (1) gives only a rough estimate to the fractional volume of shale, because the selection of GR_{min} and GR_{max} highly depends on the available information about the local geology and, of course, the subjective decision of the log analyst. On the other hand, the evaluation assumes that radioactive non-clay minerals are not present in the rock, which can also mislead the interpretation of reservoirs with complex lithology.

According to field experiences, shale volume calculated by equation (1) is generally overestimated especially in young rocks. Lower values of shale volume can be estimated by using some non-linear relationships. For instance, Larionov (1969) introduced the following empirical formulae for sediments with different ages

$$V_{sh} = \begin{cases} 0.083\left(2^{3.7\, i_{GR}} - 1\right), & \text{Tertiary or younger rocks} \\ 0.33\left(2^{2\, i_{GR}} - 1\right), & \text{Pre-Tertiary rocks.} \end{cases} \tag{2}$$

Similar non-linear models were also introduced by Stieber (1970), Clavier et al. (1971), Bhuyan and Passey (1994), which all give more careful estimates than equation (1). Consider the results of different shale volume calculations for a field case (Fig. 1). Natural gamma-ray intensity data were collected from a shallow part of a South Hungarian deep well. Applying equations (1)-(2) and independent inverse modeling (next section), separately, shale volume was estimated in percent by different methods. It is concluded that much higher shale volumes than desired can be given by the linear approximation (Fig. 1a). As a result, estimated shale volume logs show significant deviations (Fig. 1b), which can greatly affect the calculation of other petrophysical parameters involved in the interpretation procedure.

2.3 Inversion methods for interpretation

Inverse modeling represents a different approach to estimate shale volume by processing all suitable well-logging data simultaneously. In shallow boreholes the natural gamma-ray intensity (*GR*), spontaneous potential (*SP*) and resistivity at different tool lengths (deep resistivity *RD* and shallow resistivity *RS*) are normally measured, which are occasionally expanded with some porosity measurements such as density (*DEN*), acoustic traveltime (*AT*) or neutron porosity (*NPHI*). Caliper (*CAL*) and temperature (*TE*) measurements are usually made for the purpose of well construction but in the lack of response equations they cannot be used directly in the inversion procedure. The main condition to solve an inverse problem is to have a known relationship between the observed data and the petrophysical model of the investigated structure. The following simplified probe response equations based on Poupon and Leveaux (1971), and Alberty and Hashmy (1984) can be applied to freshwater-saturated formations

$$GR = V_{sh} GR_{sh} + V_{sd} GR_{sd} , \tag{3}$$

$$SP = V_{sh}\left(SP_{sh} - SP_{sd}\right) + SP_{sd} , \tag{4}$$

$$\frac{1}{\sqrt{RD}} = \left[\frac{V_{sh}^{(1-0.5V_{sh})}}{\sqrt{R_{sh}}} + \frac{\left(\sqrt{\Phi}\right)^{m}}{\sqrt{aR_{w}}}\right]\left(\sqrt{S_{w}}\right)^{n} , \tag{5}$$

$$\frac{1}{\sqrt{RS}} = \left[\frac{V_{sh}^{(1-0.5V_{sh})}}{\sqrt{R_{sh}}} + \frac{\left(\sqrt{\Phi}\right)^{m}}{\sqrt{aR_{mf}}}\right]\left(\sqrt{S_{x0}}\right)^{n} , \tag{6}$$

$$NPHI = \Phi + V_{sh} NPHI_{sh} + V_{sd} NPHI_{sd} , \tag{7}$$

$$DEN = \Phi + V_{sh} DEN_{sh} + V_{sd} DEN_{sd} , \tag{8}$$

$$AT = \Phi AT_{w} + V_{sh} AT_{sh} + V_{sd} AT_{sd} , \tag{9}$$

$$\Phi + V_{sh} + V_{sd} = 1 , \tag{10}$$

where Φ denotes effective porosity, S_{x0} and S_w are water saturation in the flushed and undisturbed zones respectively, V_{sd} is the volume of sand. In the resistivity tool response equations R_{mf}, R_w, R_{sh} denote the resistivities of mud filtrate, pore-water and shale, respectively. Additional parameters with subscripts *sh*, *sd*, *w* refer to physical properties of shale, sand and water, respectively. Constants *m*, *n*, *a* represent the textural properties of rocks, that can be given from literature or alternatively by a special inversion method called interval inversion (Dobróka and Szabó 2011). Equation (10) is the material balance equation in the medium, which is used to constrain the search of model parameters in the inversion procedure.

Well-logging inversion methods can process various types of data measured in a certain depth to determine some petrophysical parameters to the same depth. Consider the column vector of the observed data in a given depth

$$\mathbf{d^{(m)}} = [GR, SP, RD, RS, DEN, NPHI, AT]^T, \tag{11}$$

where *T* is the symbol of transpose. Well-logging data in equation (11) are inverted to give an estimate for the vector of model parameters

$$\mathbf{m} = [\Phi, S_{x0}, S_w, V_{sh}, V_{sd}]^T . \tag{12}$$

Since V_{sd} can be derived from equation (10), the total number of data (7) is more than that of the unknowns (4), which constitutes an overdetermined inverse problem. Theoretical data are calculated by using equations (3)-(9), and the difference between observations and predictions is minimized in an iterative procedure. The objective function of the well-logging inverse problem is based on the weighted Euclidean norm of overall error (Mayer and Sibbit 1980)

$$E = \sum_{k=1}^{K} \left(\frac{d_k^{(o)} - d_k^{(c)}}{\sigma_k} \right)^2 = min , \tag{13}$$

where $d_k^{(o)}$ and $d_k^{(c)}$ denote the *k*-th observed and calculated data, respectively. As the uncertainty of well-logging data is different, it is advantageous to use weighting in the data space. The standard deviation σ_k of the *k*-th data variable depends on the probe type and borehole conditions that can be given from the literature (*K* is the number of applied logging instruments). The solution to equation (13) can be written as

$$\mathbf{m} = \left(\mathbf{J}^T \mathbf{W} \mathbf{J}\right)^{-1} \mathbf{J}^T \mathbf{W}\, \mathbf{d}^{(\mathbf{m})} , \qquad (14)$$

where **J** denotes the Jacobi's (sensitivity) matrix and $W_{qq} = \sigma_q^{-2}$ (q=1,2,…K) is the diagonal weighting matrix including a priori known data variances. As data and model covariance matrices can be related to each other, the inversion procedure can also provide the estimation errors of model parameters that characterize the accuracy and reliability of inversion estimates (Menke 1984). The result of inverse modeling in the thermal-water well (Section 2.2) can also be found in Fig. 1, where lithology (natural gamma-ray intensity, spontaneous potential, photoelectric absorption index), porosity (compensated neutron and density) and saturation sensitive (shallow and deep resistivity) logs were jointly inverted to quantify porosity, water saturation and shale volume. The inversion result emphasizes that deterministic procedures based on single log analysis are not universally valid and shale volumes are required to be estimated from different sources.

3. New statistical approach for interpretation

An alternative data processing approach based on multivariate statistical principles is suggested for improving shale volume estimates. Factor analysis is used to extract information about shale content from well-logging data. In the first step, all readings of available data types for the logging interval are gathered in one matrix

$$\mathbf{D} = \begin{pmatrix} GR_1 & SP_1 & RD_1 & RS_1 & DEN_1 & NPHI_1 & AT_1 & CAL_1 & TE_1 \\ GR_2 & SP_2 & RD_2 & RS_2 & DEN_2 & NPHI_2 & AT_2 & CAL_2 & TE_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ GR_j & SP_j & RD_j & RS_j & DEN_j & NPHI_j & AT_j & CAL_j & TE_j \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ GR_N & SP_N & RD_N & RS_N & DEN_N & NPHI_N & AT_N & CAL_N & TE_N \end{pmatrix}, \qquad (15)$$

where N is the total number of measuring points in the processed interval. An element of matrix **D** represents a datum collected by a particular probe in a given depth of the well. Such a combination of well logs given in equation (15) is only a theoretical possibility for groundwater applications, because some columns or submatrix might be missing in the lack of measurements. The minimum number of original variables (i.e. data) is normally four to extract new variables (i.e. factors), where natural gamma-ray intensity, spontaneous potential and some resistivity logs are always required for an appropriate analysis, because they are the most sensitive measurements to lithologic variations of formations.

Factor analysis reduces the N-by-K data matrix in equation (15) to a lower dimension by the following matrix decomposition

$$\mathbf{D} = \mathbf{F}\mathbf{L}^T + \mathbf{E}, \tag{16}$$

where **F** denotes the N-by-M matrix of factor scores, **L** is the K-by-M matrix of factor loadings and **E** is the N-by-K matrix of residuals. Dimension M denotes the number of factors, which is less than the number of original variables ($M<K$). Factor scores obtained in one column of matrix **F** represent a new well (factor) log and matrix **L** contains the weights of the data variables corresponding to the extracted factors. Factors are assumed to be linearly independent $\mathbf{F}^T\mathbf{F}/N = \mathbf{I}$ (**I** is the identity matrix), thus the correlation matrix of the standardized original variables is

$$\mathbf{R} = \frac{1}{N}\mathbf{D}^T\mathbf{D} = \mathbf{L}\mathbf{L}^T + \mathbf{\Psi}, \tag{17}$$

where $\mathbf{\Psi} = \mathbf{E}^T\mathbf{E}/N$ is the diagonal matrix of specific variances. Factor loadings can be derived by the approximation algorithm of Jöreskog (2007). At first the following matrix is calculated

$$\mathbf{S}^* = \left(\operatorname{diag}\mathbf{S}^{-1}\right)^{1/2}\mathbf{S}\left(\operatorname{diag}\mathbf{S}^{-1}\right)^{1/2}, \tag{18}$$

where **S** denotes the sample covariance matrix of the standardized data. Then the eigenvalues λ and eigenvectors **ω** of matrix $\mathbf{S}^*$ are computed, with which the matrix of factor loadings is given by

$$\mathbf{L} = \left(\operatorname{diag}\mathbf{S}^{-1}\right)^{-1/2}\mathbf{\Omega}_M\left(\mathbf{\Gamma}_M - \theta\mathbf{I}\right)^{1/2}\mathbf{U}, \tag{19}$$

where $\mathbf{\Gamma}_\mathbf{M}$ is the diagonal matrix of the first M number of sorted eigenvalues, $\mathbf{\Omega}_\mathbf{M}$ is a matrix of the first M number of eigenvectors (in its columns) and **U** is an arbitrary M-by-M orthogonal matrix. Parameter θ specifies the smallest number of factors when

$$\theta = \frac{1}{K - M}\left(\lambda_{M+1} + \lambda_{M+2} + \ldots + \lambda_K\right) < 1. \tag{20}$$

The factor scores are estimated by the maximum likelihood method using the following log-likelihood function

$$P = -\left(\mathbf{D} - \mathbf{F}\mathbf{L}^{T}\right)^{\mathrm{T}} \mathbf{\Psi}^{-1} \left(\mathbf{D} - \mathbf{F}\mathbf{L}^{T}\right) = \max \, . \tag{21}$$

A linear solution to equation (21) can be given by the Bartlett's method (1937), which provides factor scores as follows

$$\mathbf{F} = \left(\mathbf{L}^{T} \mathbf{\Psi}^{-1} \mathbf{L}\right)^{-1} \mathbf{L}^{T} \mathbf{\Psi}^{-1} \mathbf{D} \, . \tag{22}$$

Rotation of factor loadings makes it possible to interpret factors more easily. In this study, the varimax algorithm suggested by Kaiser (1958) is used to generate rotated factors, which are compared to shale volumes by regression analysis.

Factor analysis can be effectively used to extract shale volume and other petrophysical properties of rocks from the well logging data set. The workflow used in this study is based on the subsequent application of factor and regression analyses (Fig. 2). As a result, the empirical relationships existing between factors and petrophysical properties can be used in the comprehensive interpretation of well logs. For instance, petrophysical parameters given from factor analysis can be treated as known parameters during the procedure of inverse modeling (Section 2.3), thus, because of the increase of overdetermination of the inverse problem, the accuracy and reliability of the inversion results can be improved considerably. (The influence of data-to-unknowns ratio on the inversion results was detailed in the context of the interval inversion in Dobróka and Szabó (2011), Dobróka et al. (2012), Dobróka and Szabó (2012).) The feasibility of the statistical method depends on how much amount of data variance in the factors is culminated (eps in Fig. 2). We suggest that no less than 70% of variance must be explained by the first factor. To preserve the rest of the information, some of the succeeding factors must be calculated and maximum 5-10% of variance can be neglected. The degree of correlation between the factors and petrophysical properties is also of great importance (alpha in Fig. 2). The correlation coefficient between the given factor and petrophysical parameter should be at least 0.75 to accept the method feasible.

Szabó (2011) found a non-linear (exponential) relationship between the first factor (1st column of matrix **F**) and shale volume. The strong correlation can be explained by the high parameter sensitivities of well logs, namely a small difference in the value of shale content is followed by significant changes in the well-logging data. Each probe response function contains shale volume as it can be seen in Equations (3)-(10), thus, shale volume is responsible for a large part of data variance. Consider the following regression function for groundwater applications

$$V_{sh} = \alpha \mathrm{e}^{\beta F_1'} + \gamma \, , \tag{23}$$

where α, β, γ are local regression coefficients. For some deep wells the above empirical formula proved to be nearly independent of the measurement area. In equation (23) the first factor is scaled by the following procedure

$$F_1' = F_{1,min}' + \frac{F_{1,max}' - F_{1,min}'}{F_{1,max} - F_{1,min}} \left(F_1 - F_{1,min} \right), \tag{24}$$

where F_1 is the factor score estimated to a given depth, $F_{1,min}$ and $F_{1,max}$ are the smallest and largest values of the first factor log, $F_{1,min}^{'}$ and $F_{1,max}^{'}$ are the desired lower and upper limits of the scaled factor $F_1^{'}$ respectively. Regression tests on well-logging data originated from groundwater wells confirms the validity of equation (23), with a difference that new values of α, β, γ are required to be specified for the shallow region. The results of regression analysis on three data sets are shown in three case studies (Sections 4.1-4.3). For measuring the misfit between different shale volume estimation results the root mean squared error is calculated

$$RMSE = \sqrt{\frac{1}{N} \sum_{j=1}^{N} \left(V_{sh,j}^{(I)} - V_{sh,j}^{(II)} \right)^2} \, , \tag{25}$$

where superscripts (I) and (II) denote the estimation results of the first and second method, respectively. The dimensional unit of *RMSE* is per cent, but if the shaliness is given in fraction, it is dimensionless. The correlation relationships between the input data is characterized by the mean spread ($\bar{S}$), which gives the average of correlation coefficients excluding the self-correlations of data variables

$$\bar{S}(\mathbf{R}) = \sqrt{\frac{1}{K(K-1)} \sum_{j=1}^{K} \sum_{l=1}^{K} \left(R_{jl} - \delta_{jl} \right)^2} \, , \tag{26}$$

where δ denotes the Cronecker-delta symbol (δ_{il}=1 in case of i=j, and δ_{il}=0 otherwise). For characterizing the strength of the non-linear relationship between the relevant factor and shale volume, the rank correlation coefficient can be applied (Spearman 1904).

4. Application to groundwater exploration

The test sites are located in the north-eastern part of the Great Hungarian Plain (Fig. 3). Three shallow boreholes are investigated, two of them (Well-1 and Well-2) are situated on the bank of River Tisza and one (Well-3) is tens of kms away to the east. The investigation areas are part of the Pannonian Basin Province of Central Europe, which is a highly explored petroleum area, where many hydrocarbon systems have been discovered. The Pannonian Basin consists of a large extensional basin of Neogene age overlying Paleogene basins and a Mesozoic (or older) basement (Dolton 2006). Above the thick and large-extension Tertiary basin-fill sequence containing the oil and gas-bearing formations, there are various Pannonian aged postrift sediments including thermal-water resources. The succeeding Quaternary sediments are characterized by a high variety of paludal, fluvial and delta-plain deposits including thick gravel and sand deposits saturated with good quality fresh water, and confining clays. This shallowest region is the target of the present study.

4.1 Case study 1

The location of Well-1 is at the site of Tokaj Waterworks, which is situated at the junction of River Tisza and River Bodrog in North-East Hungary (Fig. 3). The aim of the geophysical survey composed of induced polarization (*IP*), vertical electrical sounding (*VES*) and well-logging measurements was to assess the local aquifer system for the purpose of protecting the quality of drinking-water supply. In the vicinity of Well-1, the estimated depth of the basement is approximately 250 m. The drillhole traversed young sediments overlaying Miocene volcanic rocks. The 150 m thick Pannonian (Late Miocene) water-bearing gravel and sand is covered by a dominantly coarse grained sand complex of Pleistocene age with clay interbeds, clayey-silty sands and then Holocene flood sediments thereupon.

The data set measured in Well-1 contains natural gamma-ray intensity (*GR*), spontaneous potential (*SP*), shallow resistivity (*RS*), deep resistivity (*RD*) and neutron porosity (*NPHI*) logs (Fig. 5). The gamma-ray image shows the shaly intervals with dark (brown) colors, while the sandy layers are characterized with the light (yellow) ones. The correlation matrix of well-logging data is shown in Table 1. The average of Pearson's correlation coefficients between the measured variables calculated by equation (26) is 0.58, which indicates a moderate linear relationship between the data types. The application of the maximum likelihood method requires that the distribution of the data must be near-Gaussian. The kurtosis and skewness of the given data set are -0.2 and 0.3, respectively, which practically fulfills the condition of normality.

Factor analysis is used to extract two new variables called Factor 1 and Factor 2. The former explains the 95% of total variance of the original variables. The rest of the information falls on to subsequent factors. A second factor is also calculated which is responsible for 3.3% of data variance. The rest of information (1.7%) is neglected. In equation (23) the exponential connection between the first scaled factor and shale volume is specified. For comparing the results of different wells, Factor 1 should be properly scaled. The new interval of factor scores is computed by equation (24), where $F'_{1,min}$=0 and $F'_{1,max}$=100 are chosen. (The new interval is the same that shale volumes fall into in percent). The estimated loadings of the first factor for Well-1 are shown in Table 2. The largest weights on the first factor are given by logs mainly sensitive to lithology (*GR* and *SP*). The role of porosity and saturation logs (*NPHI* and *RS*) is also important in the development of the new variable. The second factor correlates mostly with resistivity logs: -0.1 (*SP*), -0.23 (*GR*), -0.48 (*NPHI*), 0.72 (*RS*), 0.98 (*RD*). The exponential relationship between the first scaled factor and shale volume estimated from equation (2) is illustrated in Fig. 4a. The regression coefficients of equation (23) with their lower and upper limits are listed in Table 3. We can find an approximate value for parameter β, which is valid to the fourth decimal figure in the three boreholes, thus, it is fixed during the regression tests. Beside a constant value of β, the other regression coefficients were specified. The rank correlation coefficient of the nonlinear relationship is 0.96, which indicates a very strong correlation between the first scaled factor and shale volume computed from the natural gamma-ray intensity log based on equation (2). The factor logs and shale volume logs estimated by factor analysis and the Larionov's model are plotted in Fig. 5. The resultant logs show a close agreement between the independent interpretation results. The *RMSE* defined in equation (25) for measuring the difference between the shale volume logs is 7.2%. If we calculated shale volume from equation (1), the misfit between the results of factor analysis and the linear formula would increase up to 13.7%.

4.2 Case study 2

Well-2 was drilled in the area of Tiszalök Waterworks situated approximately 12 km south-west of Tokaj (Fig. 3). The purpose of geophysical survey and the local geological setting are similar as described in Case Study 1.

The following well logs are utilized such as natural gamma-ray intensity (*GR*), spontaneous potential (*SP*), shallow resistivity (*RS*) and caliper (*CAL*), which can be considered as a minimum set and sometimes a typical combination of low-cost measurements (Fig. 6). It can be mentioned that *CAL* log is normally treated as a technical measurement and cannot be explicitly used in inversion procedures. However, it is worth integrating into factor analysis, because it contains useful information about the lithology of clastic formations, i.e. cavernous clay appears with relatively large diameter and sand shows smaller hole size, because of the filter cake formed at the borehole wall by the infiltration process of drilling fluid. The mean spread of input data is 0.57, which indicates a moderate linear relationship between the data (the partial correlations are shown in Table 1). The kurtosis and skewness are 0.4 and 0.5, respectively.

Factor analysis results in one factor, which contains 99.3% of total variance of the four measured variables. The rest of information (0.7%) is neglected. The same scaling procedure is performed on the first factor as in Case Study 1. The factor loadings of the scaled factor in Well-2 are included in Table 2. In the lack of porosity logs the biggest weights go to *GR* and *SP* data, which corresponds to earlier results obtained in Case Study 1 and deep wells (Szabó 2011). The *RS* log has got also a great impact on the new variable and the *CAL* log shows also some weak correlation with the factor. Practically the same exponential relationship between the first scaled factor and shale volume calculated from equation (2) is obtained (Fig. 4b) as in Case Study 1. The misfit between the observed data points and the curve of the regression model is smaller than that of given in Case Study 1 (Fig. 4a), which is indicated also by an improved value of Spearman's correlation coefficient (0.99). The regression constants of equation (23) with their confidence intervals are listed in Table 3. The estimated factor and shale volume logs are plotted in Fig. 6. The *RMSE* value for the shale volume curves estimated by the statistical and deterministic (Larionov model) analysis is 1.9%, which shows a remarkable agreement between the independent results. (The model distance between the gamma-ray index based linear result and statistical calculation increases up to 11.6%.)

4.3 Case study 3

Well-3 is located in Baktalórántháza about 64 km east of Tiszalök and 51 km east south-east of Tokaj (Fig. 3). The aim of the geophysical survey was the investigation of the geological structure and exploration of thermal water. The local geology shows similar characteristics with the two previous surveying areas (Case Studies 1-2). The bottom of the well is at 1200 m where drilling information describes Lower Pannonian deposits as gravel, clayey sand, clayey silt, clayey marl and bituminous clay. The depth of the Pleistocene complex is approximately at 240 m, which dominantly consists of coarse grained gravel and sands at the bottom and finer fluvial sediments in the shallower region. At the top Holocene soil and aeolian sand can be found. In addition to the zone of the Pleistocene sequence of strata (that was investigated in Wells 1-2), further 250 m of logging interval is integrated into the statistical procedure.

The applied well logs represent natural gamma-ray intensity (*GR*), spontaneous potential (*SP*), shallow resistivity (*RS*), gamma-gamma (*GG*) and neutron-neutron (*NN*) data (Fig. 7). The well-logging data set is incomplete, because the interval of 89-99 m was not logged, but it causes no problems in the overall statistical analysis. The rest of the measured intervals can be processed together in one interpretation procedure. The average correlation between the measured data is 0.32, which indicates a weak linear relationship between the input logs (the correlation coefficients are shown in Table 1). The kurtosis and skewness are 0.8 and 0.3, respectively, which refer to a little more peaky data distribution than Gaussian.

The first factor represents the 99.6% of the total variance of the measured variables (the rest of 0.4% is neglected). The factor loadings for Well-3 are listed in Table 2. The biggest weights are represented by *GR*, *SP* and *RS* logs. The porosity logs (*NN*, *GG*) make a relatively small contribution to the extracted variable. Factor scores are scaled by the same procedure as in Case Studies 1-2. Figure 4c shows that the exponential function of equation (23) approximates well the relationship between the first factor and shale content estimated from equation (2). The coefficients of the regression function with their confidence intervals are in Table 3. The rank correlation coefficient is 0.99, which indicates almost entire correlation between the above quantities. This high value of correlation results from the very strong non-linear relationship and the large statistical sample (23,415 data). In Fig. 7 the factor log and the estimated shale volume logs are shown. The *RMSE* between the statistical and deterministic solution is 2.8%, which is ~5 times smaller than the model distance between the results of factor analysis and linear approximation (14.5%).

5. Discussion

The three parameter regression tests show consistent results for three different well sites. The synthesis of field results confirms that the coefficients α, β, γ do not change significantly in different measurement areas. In Table 3, the values of regression constants with their estimation errors can be compared for Wells-1-3. Since β represents almost the same value (~0.0146) in all drillholes, we can make a simplification by treating it as a fixed parameter (0.015). By accepting the average values of regression parameters (and mean estimation errors) given in Table 3, a local shale volume estimation formula is suggested in the form as

$$V_{sh} = 27.4e^{0.015\,F_1'} - 26.5\,. \tag{27}$$

The results given by equation (27) are ~3.3 times better than those of the linear approximation based on equation (1) in the three wells. The mean value of *RMSE* for the shale volume logs estimated by factor analysis and the Larionov's formula (applied to Tertiary or younger sediments) in the three boreholes is 4%, while it is 13.3% between the results of statistical and conventional gamma-ray index based methods. The results of factor analysis are close to the solutions given by the Larionov's formula. The biggest difference between the statistical and Larionov model can be found in Well-1, which can be explained with that in Well-1 the weight of *GR* log is relatively small and considerable loads are put on the rest of the logs. This causes some deflection of estimation results from the *GR* log based model (Fig. 8). On the other hand, the smallest number of data is processed by the statistical method in Well-1. Our experience shows that shale volumes calculated directly from natural gamma-ray index are approximately 20% higher than that of given by factor analysis for the interval of $20\%<V_{sh}<80\%$ independent of the measurement area.

The proposed exponential model in equation (27) gives more realistic and better estimates than the conventional linear approximation (Fig. 8). The results of factor analysis in Wells-1-3 are close to the model of Larionov, which confirms the validity of the suggested statistical method. To check the reliability of the regression formula, the shale volumes estimated from the geophysical well logs were compared to those of laboratory analysis made on core samples (grain-size distribution, comprehensive mineralogical and chemical analysis) in Well-3. The correlation relationship between shale volumes estimated by factor analysis and core measurements is strong (Fig. 9). The shale volume logs obtained by factor analysis and independent laboratory measurements are plotted in Track 3 of Fig. 10. Only those depth points are indicated by circles where core measurements were made (the red curve representing the results of factor analysis connects the same points). The *RMSE* between the results of core measurements and the calculations using the model of Larionov is 5.1%, while it is 4.5% between the results of laboratory tests and factor analysis. In this particular case, factor analysis showed slightly better estimation than the conventional Larionov's method (the relative improvement is 11%). If the linear approximation based on equation (1) is compared to the performance of factor analysis, the relative improvement for the statistical approach is ~440%

as the *RMSE* between the laboratory measurements and linear prediction is 19.9 %. The correction of linear results is highly recommended for a more accurate and reliable modeling of water-bearing formations.

6. Conclusion

It is shown that the presented shale indicator is applicable to estimate the shale content of water-bearing formations from the log of one special factor. Shale volume is normally given by deterministic or inversion procedures in each depth separately. On the contrary, the new statistical method jointly processes all available data within the logging interval allowing some shorter intervals of missing data to provide the vertical distribution of shale content. This procedure requires only a few seconds of CPU times using a quad-core processor based workstation. The method can easily be extended to multiwell applications as it was shown earlier by using engineering geophysical sounding data analogously (Szabó et al. 2012; Szabó 2012).

Factor analysis can be used to explore non-measurable petrophysical information from several kinds of well-logging data. The relatively large factor loadings of neutron porosity, resistivity and caliper logs suggest that the processing of other log types than natural gamma-ray and spontaneous potential is also advantageous to use for more reliable shale volume estimation. The new feature of integrating technical and other type of measurements into the interpretation process is different from inversion methodology, which requires a strict (explicit) mathematical connection between the data and model parameters.

The benefits of the suggested method related to the characterization of aquifers are manifold. It may confirm or bring more reliable information about the shale content in primary porosity aquifers. Derived quantities such as effective porosity, free and bound water saturation, hydraulic conductivity and storage capacity of aquifers can be re-interpreted even in large areas. The accuracy and reliability of estimated shale volumes are assured by using several types of well logs simultaneously in a joint interpretation problem. Shale volume information given by factor analysis can also be used to reduce the number of unknowns of the well-logging inverse problem. Moreover, it can also be applied to resolve the ambiguity existing between parameters of the geophysical model. In conclusion, the simultaneous application of the resultant shale volume logs and other available geophysical profiles measured by ground geophysical surveying methods can significantly improve the solution of hydrogeophysical interpretation problems.

Acknowledgement

The described work was carried out as part of the TÁMOP-4.2.2.A-11/1/KONV-2012-0049 project in the framework of the New Hungarian 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 Research Group of Geoengineering, the second author is grateful for the support of the Hungarian Academy of Sciences. The authors express their thanks also to László Bucsi Szabó from the Háromkő Deposit Company for providing field data and special knowledge about the well sites, Béla Latrán from Geokomplex Ltd. for the results of core analysis, and Stanley T. Paxton from USGS for his suggestions that highly improved the quality of the paper.

References

Adeoti L, Ayolabi E A, James P L (2009) An integrated approach to volume of shale analysis: Niger Delta example, Orire Field. World Applied Sciences Journal 7:448-452

Alberty M W, Hashmy K H (1984) Application of ULTRA to log analysis. SPWLA 25th Annual Logging Symposium, New Orleans, 1984-Z

Asquith G, Krygowski D (2004) Basic well log analysis, 2nd ed. AAPG, Tulsa

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

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

Beltrami H, Cheng L, Mareschal J C (1997) Simultaneous inversion of borehole temperature data for determination of ground surface temperature history. Geophysical Journal International 129:311-318

Bhuyan K, Passey Q R (1994) Clay estimation from GR and neutron-density porosity logs. SPWLA 35th Annual Logging Symposium, Tulsa, 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

Coates R, Xiao L, Prammer M G (1999) NMR Logging Principles and Applications. Halliburton Energy Services, Houston

Cripps A C, McCann D M (2000) The use of the natural gamma log in engineering geological investigations. Engineering Geology 55:313-324

De Witte L (1955) A study of electric log interpretation methods in shaly formations. Petroleum Transactions AIME 204:103-110

Dobróka M, Szabó P N (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 P N, 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 44:459-479

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

Doveton J H, Merriam D F (2004) Borehole petrophysical chemostratigraphy of Pennsylvanian black shales in the Kansas subsurface. Chemical Geology 206:249-258

Drahos D (2005) Inversion of engineering geophysical penetration sounding logs measured along a profile. Acta Geodetica et Geophysica 40:193-202

Elek I (1988) Some application of principal component analysis: Well-to-well correlation, zonation. Geobyte 3:2:46-53

Elek I (1990) Fast porosity estimation by principal component analysis. Geobyte 5:3:25-34.

Fisher A T, Barnhill M, Revenaugh J (1998) The relationship between hydrogeologic properties and sedimentary fades: An example from Pennsylvanian bedrock aquifers, Southwestern Indiana. Ground Water 36:901-912

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

Han D H, Nur A, Morgan D (1986) Effects of porosity and clay content on wave velocities in sandstones. Geophysics 51:2093-2107

Hempkins W B (1978) Multivariate statistical analysis in formation evaluation. SPE California Regional Meeting, San Francisco, 7144-MS

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

Hsieh B Z, Chilingar G V, Lu M T et al (2007) Estimation of groundwater aquifer formation-strength parameters from geophysical well logs: The southwestern coastal area of Yun-Lin, Taiwan. Energy Sources Part A 29:1-19

Jang H, Kim H J (2008) Permittivity inversion of borehole radar data measured in the vadose zone. SEG Annual Meeting, Las Vegas, 2008-1303

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

Kazmierczuk M, Jarzyna J (2006) Improvement of lithology and saturation determined from well logging using statistical methods, Acta Geophysica 54: 378-398

Klimentos T (1991) The effects of porosity-permeability-clay content on the velocity of compressional waves. Geophysics 56:1930-1939

Kovács J, Márkus L, Halupka G (2004) Dynamic factor analysis for quantifying aquifer vulnerability. Acta Geol Hung 47: 1-17

Kvapil J, Mares S (2003) Cone penetration logging for the direct detection of LNAPLs contamination. Near Surface Geophysics 1:157-160

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

Magyar N, Hatvani I G, Kovács Székely I, Herzig A, Dinka M, Kovács J (2013) Application of multivariate statistical methods in determining spatial changes in water quality in the Austrian part of Neusiedler See. Ecol Eng 55: 82-92

Maliva R G, Clayton E A, Missimer T M (2009) Application of advanced borehole geophysical logging to managed aquifer recharge investigations. Hydrogeology Journal 17:1547-1556

Mayer C, Sibbit A (1980) GLOBAL, a new approach to computer-processed log interpretation. SPE Annual Technical Conference and Exhibition, Dallas, 9341-MS

Márkus L, Berke O, Kovács J, Urfer W (1999) Analysis of spatial structure of latent effects governing hydrogeological phenomena. Environmetrics 10: 633-654.

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

Moret G J M, Clementz W P, Knollz M D et al (2004) VSP traveltime inversion: Near-surface issues. Geophysics 69:345-351

Paillet F L (1995) Integrating surface geophysics, well logs and hydraulic test data in the characterization of heterogeneous aquifers. Journal of Environmental and Engineering Geophysics 0:1-13

Paillet F L, Crowder R E (1996) A generalized approach for the interpretation of geophysical well logs in ground-water studies -Theory and application. Ground water 34:883-898

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

Poupon A, Leveaux J (1971) Evaluation of water saturation in shaly formations. The Log Analyst 12:3-8

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

Revil A, Cathles III L M (1999) Permeability of shaly sands. Water Resources Research 35:651-662

Slater L, Lesmes D P (2002) Electrical-hydraulic relationships observed for unconsolidated sediments. Water Resources Research 38:(31)1-13

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, Houston, 2961-MS

Szabó N P, Dobróka M (2011) Geostatistical approach for shale volume estimation in water-bearing formations. Near surface 2011, Leicester, P15

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 47:161-171

Tanos P, Kovács J, Kovács Székely I, Hatvani I G (2011) Exploratory data analysis on the Upper-Tisza section using single and multivariate data analysis methods. Cent Eur Geol 54: 345-356

Thomas E C, Stieber S J (1975) The distribution of shale in sandstones and its effect upon porosity. SPWLA 16th Annual Logging Symposium, New Orleans, 1975-T

Tselentis G A (1985) The processing of geophysical well logs by microcomputers as applied to the solution of hydrogeological problems. Journal of Hydrology 80:215-236

List of figure captions

Fig. 1 Shale volume estimation example using a South Hungarian well-logging data set. (a) Shale volume versus natural gamma-ray index crossplot including curves from different empirical models (see section: Deterministic Methods) and inversion results (see section: Inversion Methods). (b) Shale volume logs estimated from different interpretation methods

Fig. 2 Flowchart of the statistical methodology used for petrophysical parameter estimation

Fig. 3 Location map of the investigated well sites in Northeast Hungary

Fig. 4 Exponential relationship between the first scaled factor and shale volume estimated from the Larionov's formula in (a) Well-1, (b) Well-2, (c) Well-3

Fig. 5 Well logs measured from Well-1 as input for factor analysis (Tracks 1-4). *GR* is natural gamma-ray intensity, *SP* is spontaneous potential, *NPHI* is neutron porosity, *RS* and *RD* are shallow and deep resistivity, respectively. Well logs of the first scaled and second factor (Track 5). Shale volume logs estimated independently by factor analysis and the Larionov's formula (Track 6)

Fig. 6 Well logs measured from Well-2 as input for factor analysis (Tracks 1-4). *GR* is natural gamma-ray intensity, *SP* is spontaneous potential, *CAL* is caliper, *RS* is shallow resistivity. Well log of the first scaled factor (Track 5). Shale volume logs estimated independently by factor analysis and the Larionov's formula (Track 6)

Fig. 7 Well logs measured from Well-3 as input for factor analysis (Tracks 1-4). *GR* is natural gamma-ray intensity, *SP* is spontaneous potential, *GG* is gamma-gamma intensity, *NN* is neutron-neutron intensity, *RS* is shallow resistivity. Well log of the first scaled factor (Track 5). Shale volume logs estimated independently by factor analysis and the Larionov's formula (Track 6)

Fig. 8 Natural gamma-index vs. shale volume crossplot for comparing *GR* log based models with the estimation results of factor analysis in Wells-1-3

Fig. 9 Regression relationship between shale volume estimations made by laboratory and factor analyses (*R* denotes the Pearson's correlation coefficient)

Fig. 10 Gamma-ray log measured in Well-3 (Track 1) and results of factor analysis. Shale volume logs estimated by factor analysis and the Larionov's formula (Track 2). In Track 3 independent results of laboratory measurements and factor analysis are plotted, where shale volumes are represented in points where rock samples were cored out

Tables

Table 1 Correlation matrices calculated for borehole geophysical data sets originated from Wells-1-3.

Well-1	*SP*	*GR*	*NPHI*	*RS*	*RD*
SP	1	0.61	0.64	-0.38	-0.16
GR	0.61	1	0.69	-0.69	-0.32
NPHI	0.64	0.69	1	-0.68	-0.54
RS	-0.38	-0.69	-0.68	1	0.77
RD	-0.16	-0.32	-0.54	0.77	1
Well-2	*GR*	*SP*	*RS*	*CAL*	-
GR	1	-0.01	-0.35	-0.12	-
SP	-0.01	1	-0.82	-0.73	-
RS	-0.35	-0.82	1	0.77	-
CAL	-0.12	-0.73	0.77	1	-
Well-3	*RS*	*SP*	*GR*	*GG*	*NN*
RS	1	-0.01	-0.74	-0.33	-0.05
SP	-0.01	1	-0.12	-0.42	-0.14
GR	-0.74	-0.12	1	0.36	0.05
GG	-0.33	-0.42	0.36	1	0.02
NN	-0.05	-0.14	0.05	0.02	1

Table 2 Loadings of the first factor derived from borehole logging data originated from Wells-1-3 ("missing" represents not available log).

	GR	*SP*	*RD*	*RS*	*GG*	*NPHI*	*NN*	*CAL*
Well-1	0.92	0.65	-0.11	-0.56	missing	0.65	missing	missing
Well-2	0.96	0.80	missing	-0.84	missing	missing	missing	-0.30
Well-3	0.92	-0.12	missing	-0.81	0.40	missing	0.06	missing

Table 3 Estimated regression coefficients for shale volume versus first scaled factor relationship with 95% confidence bounds (beside the fixed value of β=0.015) in Wells-1-3.

	α_{min}	α	α_{max}	γ_{min}	γ	γ_{max}
Well-1	26.53	27.52	28.52	-29.97	-27.77	-25.58
Well-2	28.51	28.66	28.81	-27.94	-27.66	-27.37
Well-3	25.75	25.89	26.03	-24.28	-23.99	-23.69
Mean	26.93	27.36	27.79	-27.40	-26.48	-25.55

Figures

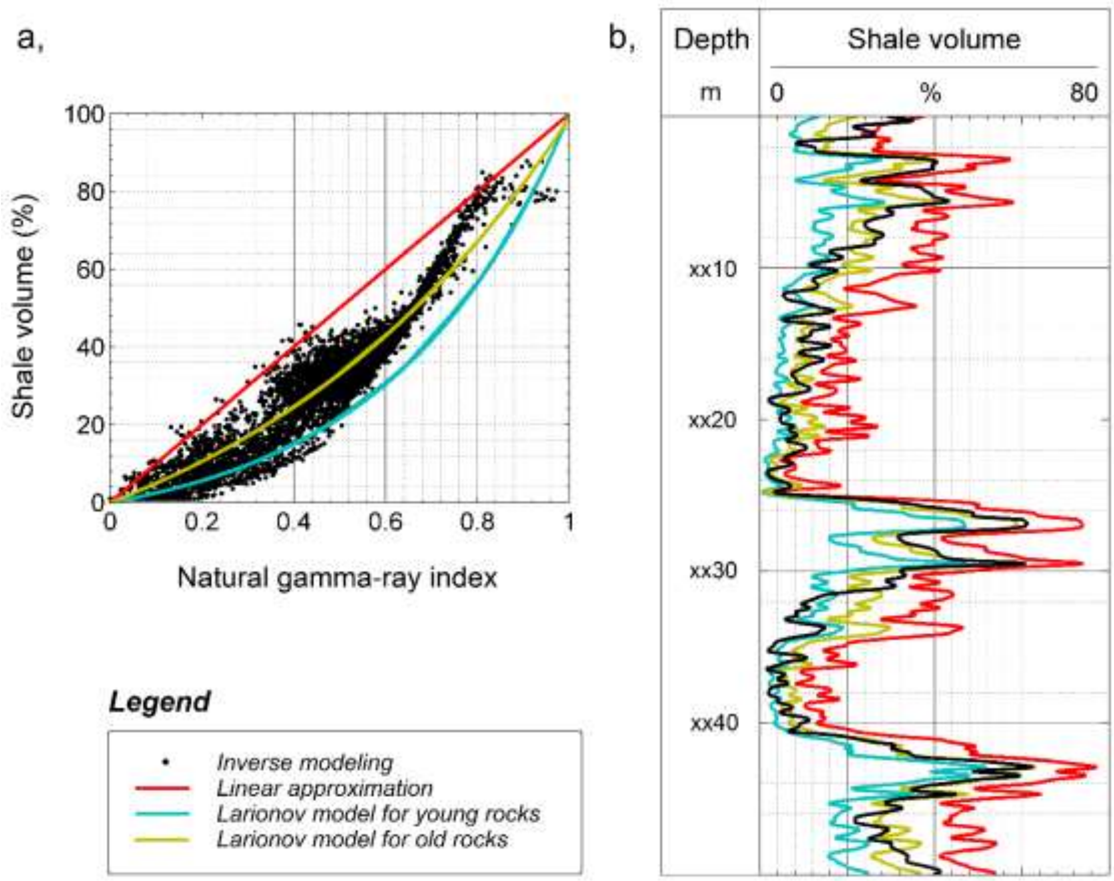


Fig 1.

Fig 1.

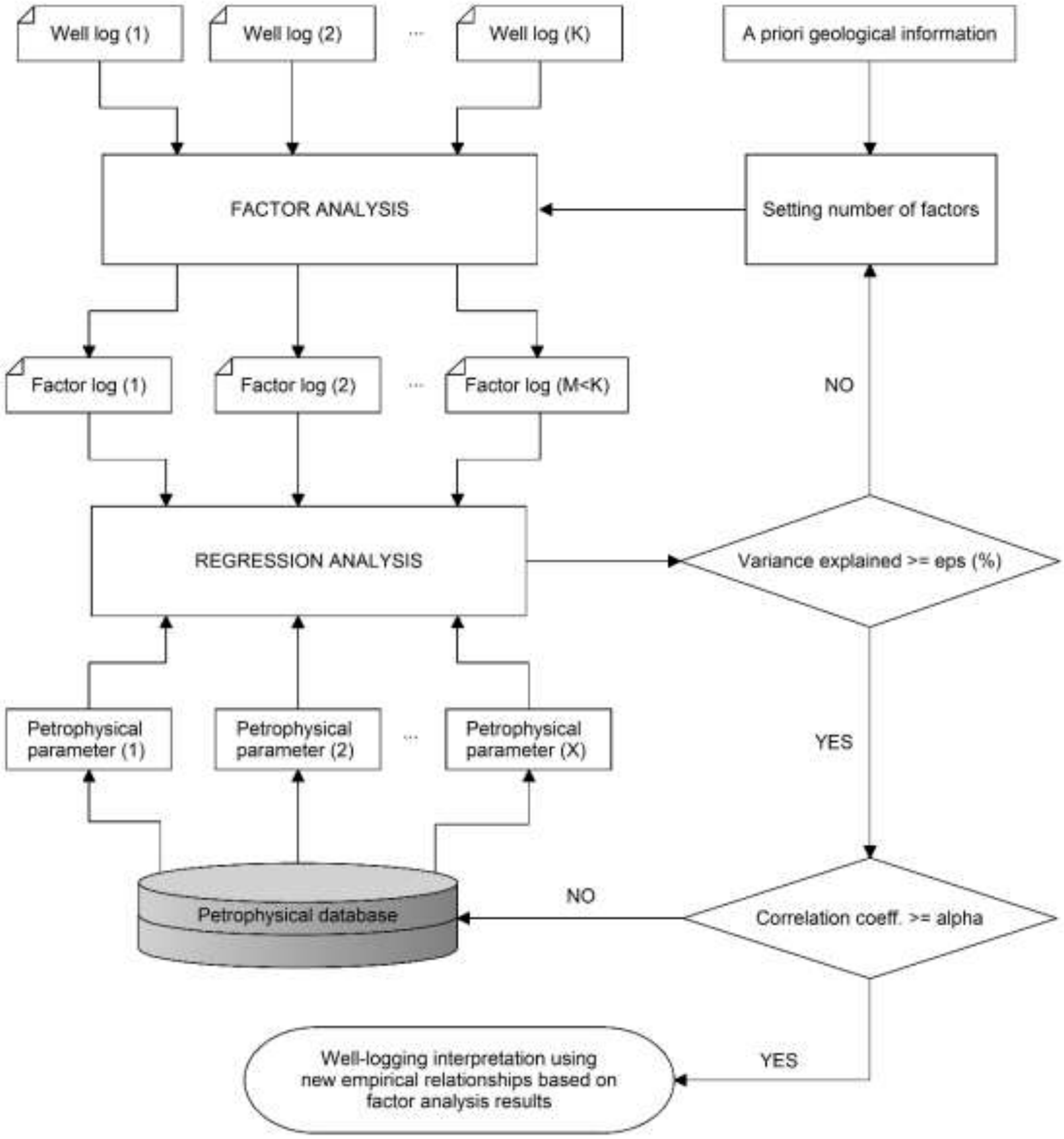

Well log (1)
Well log (2)
...
Well log (K)
A priori geological information
FACTOR ANALYSIS
Setting number of factors
Factor log (1)
Factor log (2)
...
Factor log (M<K)
NO
REGRESSION ANALYSIS
Variance explained >= eps (%)
Petrophysical parameter (1)
Petrophysical parameter (2)
...
Petrophysical parameter (X)
YES
Petrophysical database
NO
Correlation coeff. >= alpha
YES
Well-logging interpretation using new empirical relationships based on factor analysis results

Fig 2.


N
50 km
Europe
Slovakia
Ukraine
Romania
Hungary
Sajó
Bodrog
Tisza
Szamos
Miskolc
Well-1
Well-2
Well-3
Nyíregyháza
Debrecen
Legend
Country border
River
Town
Logged well

Fig 3.

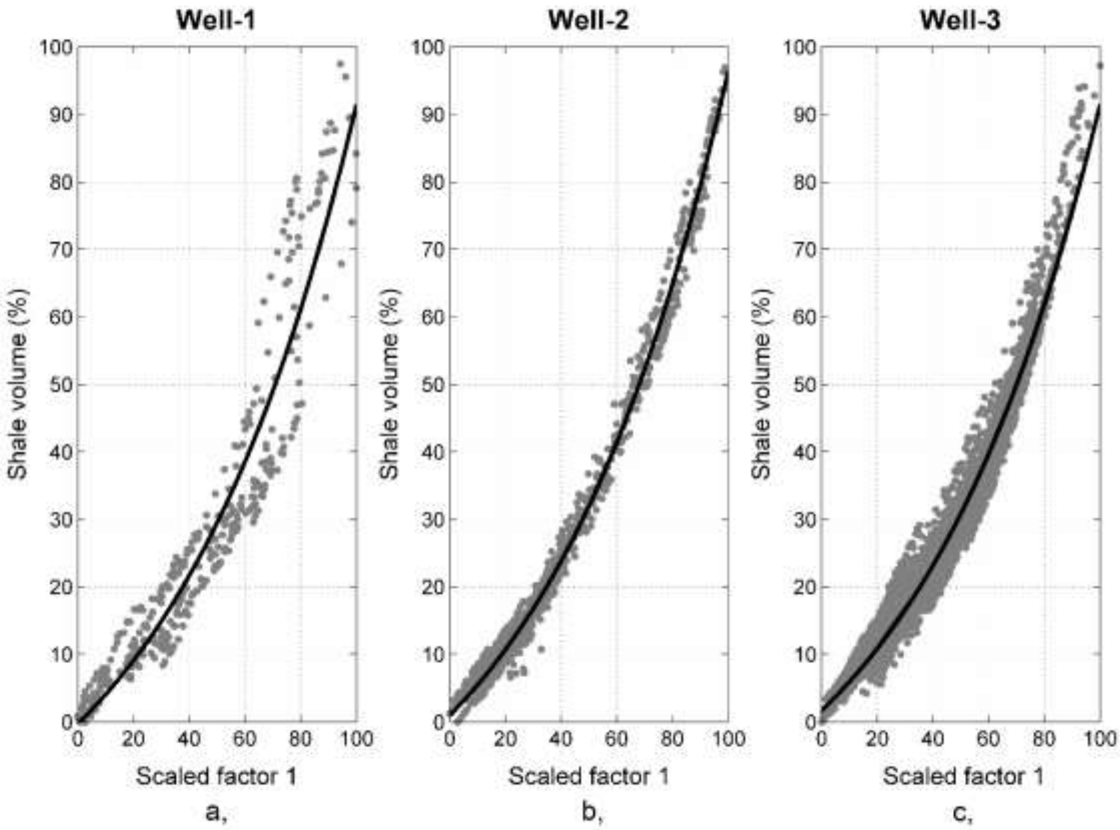

Well-1
Well-2
Well-3
Shale volume (%)
Scaled factor 1
0
10
20
30
40
50
60
70
80
90
100
0
20
40
60
80
100
a,
b,
c,

Fig 4.

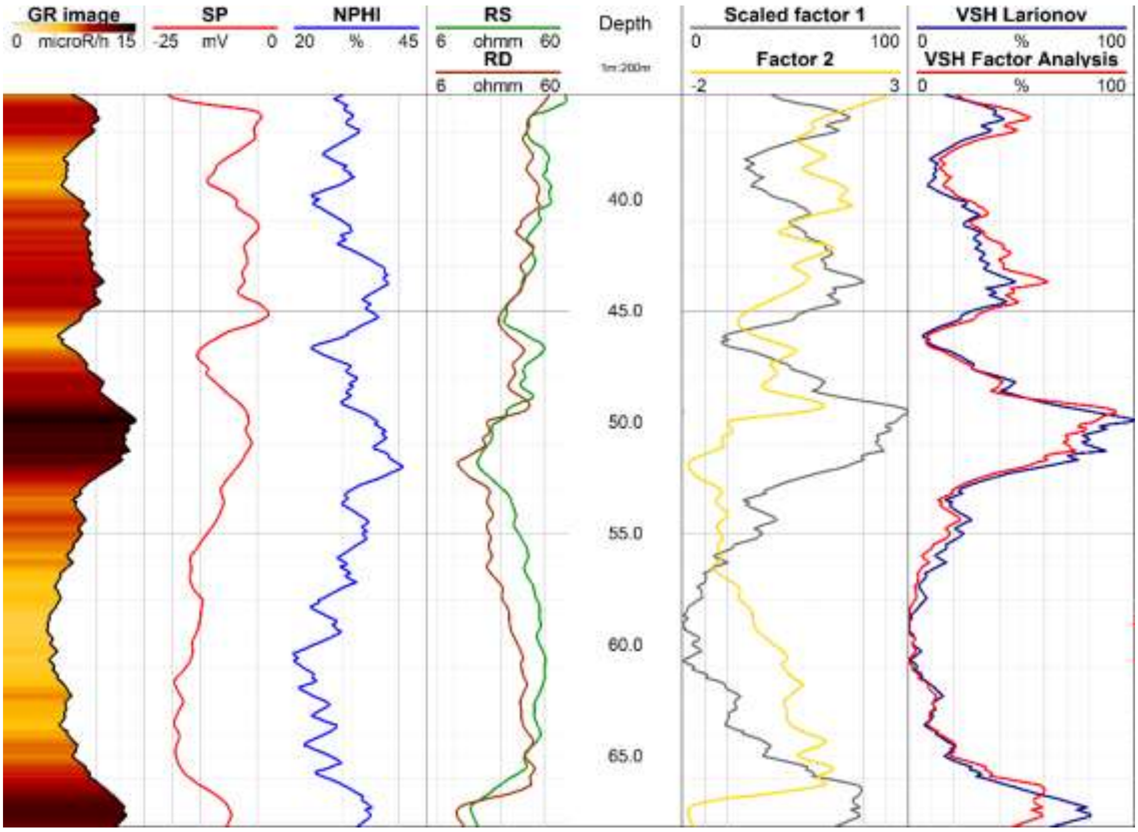
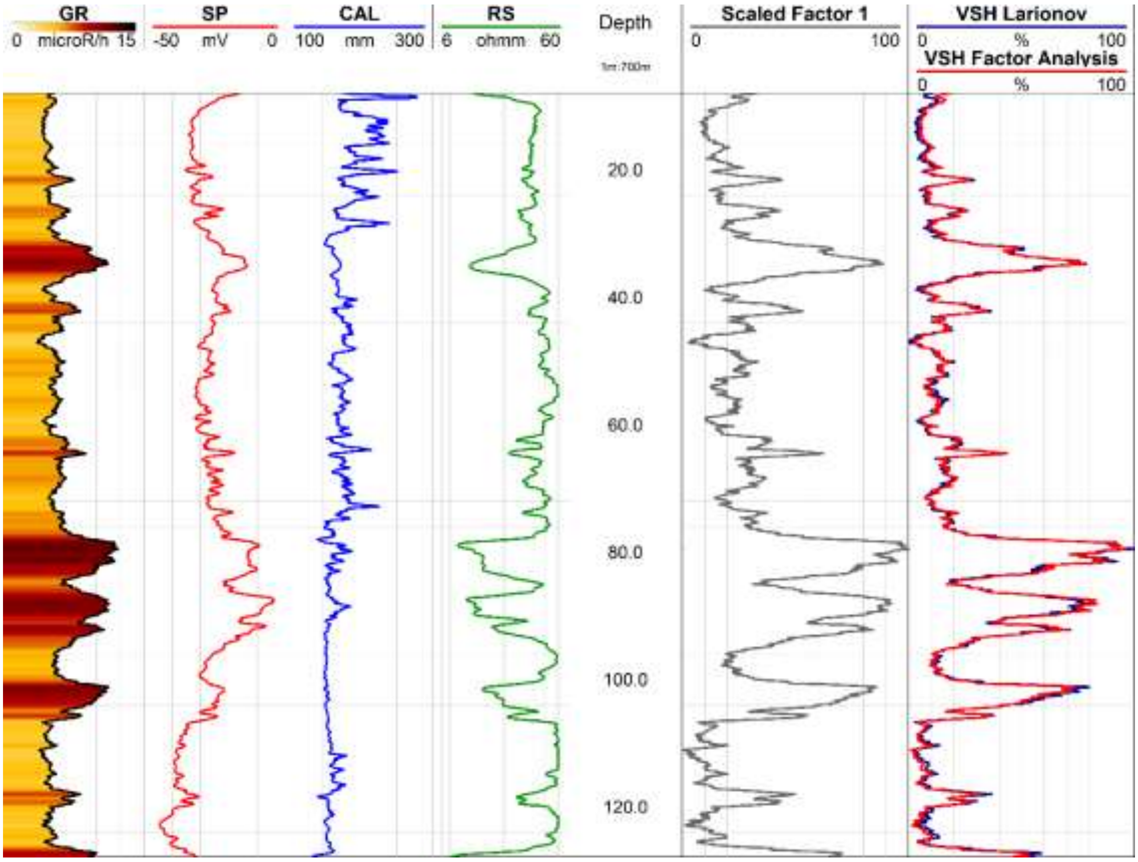

GR image
0 microR/h 15
SP
-25 mV 0
NPHI
20 % 45
RS
6 ohmm 60
RD
6 ohmm 60
Depth
40.0
45.0
50.0
55.0
60.0
65.0
Scaled factor 1
0 100
Factor 2
-2 3
VSH Larionov
0 % 100
VSH Factor Analysis
0 % 100

Fig 5.



GR
0 microR/h 15
SP
-50 mV 0
CAL
100 mm 300
RS
6 ohmm 60
Depth
Scaled Factor 1
0 100
VSH Larionov
0 % 100
VSH Factor Analysis
0 % 100
20.0
40.0
60.0
80.0
100.0
120.0

Fig 6.

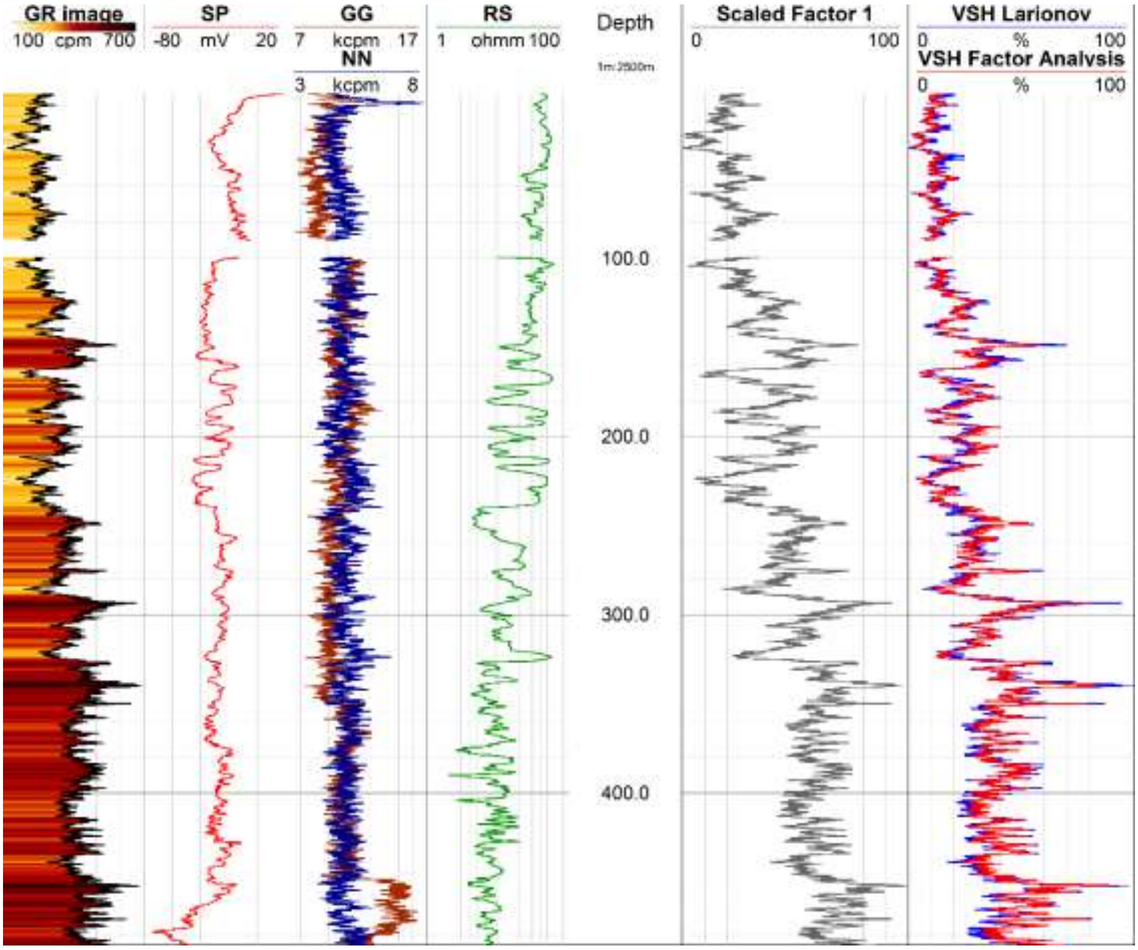

GR image
100 cpm 700
SP
-80 mV 20
GG
7 kcpm 17
NN
3 kcpm 8
RS
1 ohmm 100
Depth
100.0
200.0
300.0
400.0
Scaled Factor 1
0 100
VSH Larionov
0 % 100
VSH Factor Analysis
0 % 100

Fig 7.

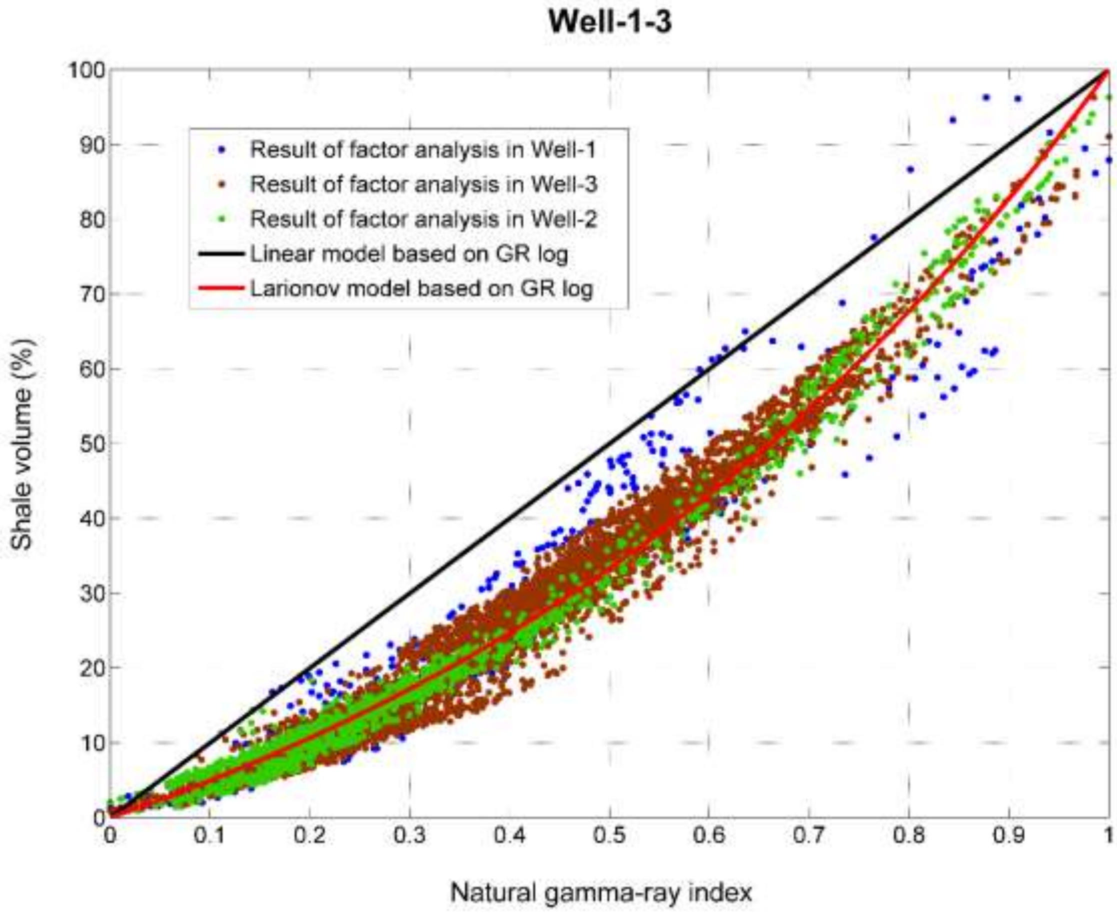

Well-1-3
Result of factor analysis in Well-1
Result of factor analysis in Well-3
Result of factor analysis in Well-2
Linear model based on GR log
Larionov model based on GR log
Shale volume (%)
Natural gamma-ray index
0
10
20
30
40
50
60
70
80
90
100
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

Fig 8.

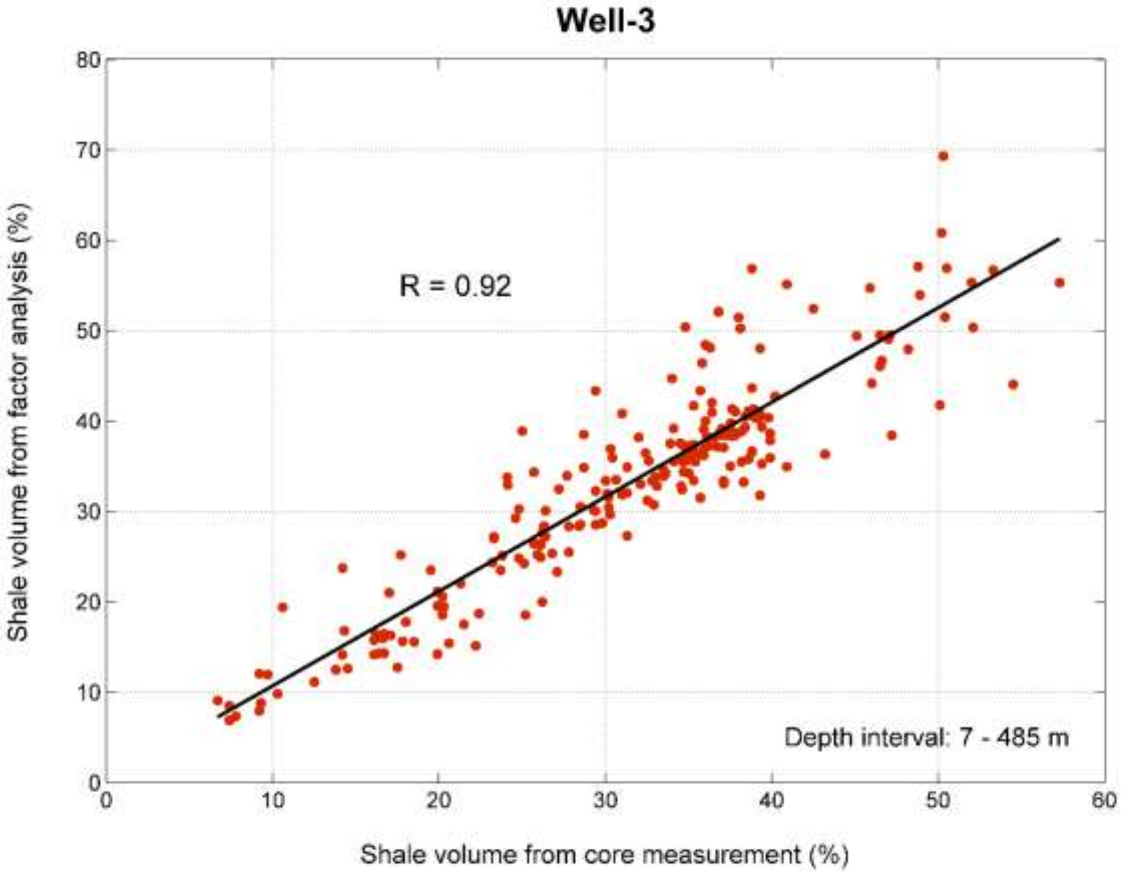

Well-3
R = 0.92
Depth interval: 7 - 485 m
Shale volume from factor analysis (%)
Shale volume from core measurement (%)
0
10
20
30
40
50
60
70
80

Fig 9.

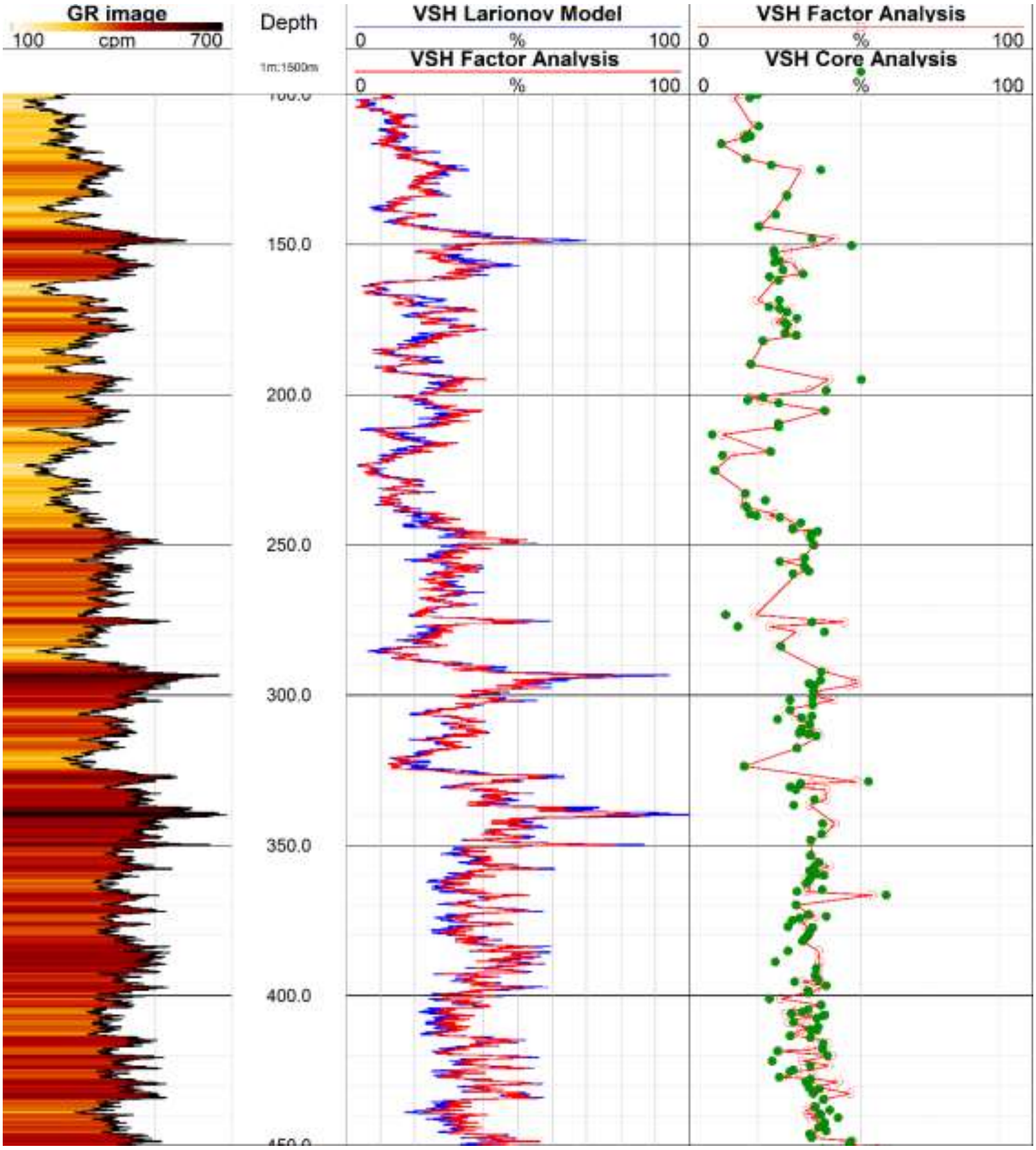

GR image
100
cpm
700
Depth
1m:1500m
VSH Larionov Model
0
%
100
VSH Factor Analysis
0
%
100
VSH Factor Analysis
0
%
100
VSH Core Analysis
0
%
100
150.0
200.0
250.0
300.0
350.0
400.0

Fig 10.