Most frequent value based factor analysis of direct-push logging data

Norbert Péter Szabó[1,2*], Gergely Pál Balogh[1], János Stickel[3]

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

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

[3]Elgoscar-2000 Ltd., H-1095 Budapest, Soroksári Street 164, Hungary

[*]norbert.szabo.phd@gmail.com

ABSTRACT

An improved iteratively re-weighted factor analysis procedure is presented to interpret engineering geophysical sounding logs in shallow unsaturated sediments. We simultaneously process cone resistance, electric resistivity and nuclear data acquired by direct-push tools to give robust estimates of factor variables and water content in unconsolidated heterogeneous formations. The statistical procedure is based on the iterative re-weighting of the deviations between the measured and calculated data using the Most Frequent Value method famous for its robustness and high statistical efficiency. The iterative approach improves the result of factor analysis for not normally distributed data and extremely noisy measurements. By detecting a strong regression relation between one of the extracted factors and the fractional volume of water, we establish an independent method for water content estimation along the penetration hole. We verify the estimated values of water volume by using a highly overdetermined quality checked interval inversion procedure. The multidimensional extension of the statistical method allows the estimation of water content distribution along both the vertical and horizontal coordinates. Numerical tests using engineering geophysical sounding data measured in a Hungarian loessy-sandy formation demonstrate the feasibility of the Most Frequent Value based factor analysis, which can be efficiently used for a more reliable hydrogeophysical characterization of the unsaturated zone.

Keywords: Inversion, Log analysis, Noise rejection, Modelling.

1 INTRODUCTION

Cone penetration tests (CPT) can be effectively used for the in-situ investigation of shallow sediments such as clay, silt, sand, gravel and other loose formations. In soft rocks, the penetration of the tube can reach some tens of meters, which allows the classification of soil types and estimation of their petrophysical properties. Lunne, Robertson and Powell (1997) provide an extensive overview of CPT instruments and tests including geotechnical applications. Traditional CPT tools used for measuring geotechnical parameters of soils are usually complemented by geophysical instruments for a more reliable site investigation. Analogously to well-logging methods, several physical parameters are observed in penetration holes with an advantage that only a steel tube isolates the probe from the soil, there is no invasion of drilling fluid into the formation, and the measured data are transferred to the surface unit through the rods pushed into the ground. Schulmeister *et al.* (2003) show the downhole electric resistivity method as an efficient tool in stratigraphic characterization of unconsolidated sediments, differentiation between sand and clay intervals, estimation of shale content and hydraulic conductivity underlying the process of groundwater flow and solute transport. In the same study, hydrostratigraphic facies mapping was done by the lateral correlation of electric logs. Direct-push techniques including the state-of-the-art methods and hydrogeophysical applications are detailed in Kirsch (2006).

A Hungarian direct-push technology called Engineering Geophysical Sounding (EGS) allows the measurement not only of soil resistivity, but cone resistance, sleeve friction, natural gamma-ray intensity, neutron-porosity and bulk density (Fejes and Jósa 1990). The EGS measurement technique illustrated in Fig. 1 can be efficiently applied in solving environmental, hydrogeological and civil engineering problems. Several studies have shown its successful application in soil mechanical studies and treatment, environmental risk assessment, planning recultivation programs, surveying of dams, studying of water resources, landfill characterization, mapping groundwater contamination, mining damage assessment and delineation of hydrocarbon contamination. Draskovits and Fejes (1994) presented two case studies in which shallow water-bearing formations and their overburden were investigated by the combination of surface geoelectric and EGS methods. For gravel terraces, the

combination of direct current resistivity maps and EGS logs allowed the characterization of the hydrogeological value of groundwater formations and the protecting capacity of the overburden. The interpretation results can be improved by using different surface geoelectric arrays (Szalai *et al.* 2013; Szalai *et al.* 2015). The EGS method is recently used for solving hydrocarbon contamination problems using special instruments sensitive to the effects of induced polarization and UV fluorescence. For a more detailed interpretation of induced polarization data, a series expansion based inversion methodology was introduced by Turai and Dobróka (2011) to which several case studies were added by Turai (2011).

In practice, EGS data processing mostly incorporates deterministic methods adapted from oilfield well-log analysis (Tillman *et al.* 2008; Nyári *et al.* 2010). Drahos and Galsa (2007) developed a finite difference method to calculate the response of the electric tool for inversion applications. Drahos (2005) developed a method for volumetric compositional analysis of EGS logs based on weighted least-squares inversion. As a new alternative, Szabó, Dobróka and Drahos (2012) suggested a multivariate statistical approach for the estimation of water saturation in shallow formations. The soil moisture in unsaturated formations is an important parameter in hydrogeological studies, to which intensive research is made by involving innovative technologies. One example is the application of the nuclear magnetic resonance (NMR) measurement in shallow penetration holes. The NMR tool gives additional information on the fractional volumes of mobile and bound water, the results of which are consistent with neutron measurements (Walsh *et al.* 2013).

Multivariate statistical methods are widely used in geophysical data processing, especially in the simultaneous analysis of well logs. Factor analysis is applicable to reduce the dimensionality of statistical problems and extract not directly measurable information from multivariate data sets (Lawley and Maxwell 1962). The statistical factors as new variables extracted from the observed data can be correlated to petrophysical properties of rocks. Grana, Dvorkin and Cibin (2011) applied the method of factor analysis to effective stress prediction from seismic attributes, which gave a new possibility for estimating the abnormal pore-pressure of reservoir rocks. Principal Component Analysis (PCA) as a practical approximation method for solving the problem of factor analysis has been widely used in formation evaluation. Puskarczyk, Jarzyna and Porebski (2015) used the PCA to

reduce well-logging data sets and differentiate thin layers of sands and mudstones in middle Miocene gas-reservoirs in Poland. Niculescu, Andrei and Ciuperca (2016) applied the same technique to separate lithostratigraphic units and delineate gas-reservoirs in the Moldavian Platform. A quantitative use of factor analysis was suggested by Szabó (2011) to calculate the shale volume directly from the factor scores in hydrocarbon-bearing clastic rocks. Asfahani (2014) applied this approach successfully in the basaltic area of Southern Syria. The hydraulic conductivity of groundwater formations was explored by factor analysis of hydrogeophysical logs (Szabó 2015), the results of which was validated by core measurements and pumping tests. Szabó (2016) solved the problem of factor analysis by using a float-encoded genetic algorithm-based inversion approach, which gave the best fit between the measured and calculated logs in estimating the factors and related petrophysical properties of hydrocarbon reservoirs.

Jöreskog (2007) suggested a non-iterative factor analysis technique, which finds the solution fast regardless of the scaling of observed data. The maximum-likelihood method frequently used for solving the problem of factor analysis simplifies to the least-squares method for Gaussian distributed input variables. The drawback of the Jöreskog's approximate algorithm is that it gives optimal results only for Gaussian distributed data. In consequence, it works as a relatively noise-sensitive data processing procedure in the field. Since EGS data rarely follow Gaussian statistics, the classical method of factor analysis must be improved to give a robust solution. We adapted the algorithm of Iteratively Reweighted Factor Analysis (IRFA) suggested by Szabó and Dobróka (2017), which has proved to be a useful tool in the evaluation of multimineral hydrocarbon reservoirs. The IRFA method updates the factor scores by iteratively re-weighting the difference between the measured and calculated data. The statistical method uses a weighting process, which is analogous to ground geophysical inversion applications using the same strategy (Drahos 2008; Gyulai, Baracza and Szabó 2014). In this study, we combine the IRFA technique with the Most Frequent Value (MFV) method suggested by Steiner (1991). By the MFV method, optimal weights can be automatically calculated for the observed data to improve the result of statistical estimation. Dobróka *et al.* (1991) used the Steiner weights given by the MFV method for the establishment of a joint inversion algorithm to interpret seismic and geoelectric data collected in underground mine. The same weights were used in the

development of robust seismic tomography methods (Dobróka and Szegedi 2014) and a series-expansion based Fourier transformation method that showed high noise rejection capability (Szegedi and Dobróka 2014). In Gyulai *et al.* (2017), the Steiner weights were used for the automatic separation of dip- and strike direction of apparent resistivity data, measured over a 3D thermal water structure as part of a 2.5D geoelectric inversion procedure. Other geophysical and hydrogeological applications of the MFV method can be found in Steiner (1997) and Szűcs, Civan and Virág (2006).

We employ the MFV method for the factor analysis of penetration logs to generate optimal weights for each component of the deviation vector. By this manner, the factor logs can be calculated more accurately, and abrupt changes caused by outliers included in the data set may be prevented. We call the further developed iterative factor analysis procedure as MFV-IRFA. We calculate the water content of shallow sediments by making use of the strong correlation between the first statistical factor and water volume. The statistical results are compared to those of a joint inversion method called interval inversion, which was originally developed for the processing of oilfield well logs (Dobróka *et al.* 2016). The highly overdetermined interval inversion procedure gives significantly more accurate estimation results than the local (depth-by-depth) inversion methods. We perform regression analysis to determine the functional relation between the first factor estimated by the MFV-IRFA method and water volume, which allows the calculation of water saturation distribution in the borehole or between neighbouring boreholes. The noise rejection capability and other advantages of the MFV-IRFA method is shown in a Hungarian case study.

2 THEORY AND METHODS

2.1 Forward modelling

The petrophysical parameters of subsoils are normally extracted from EGS data using well-log analysis techniques originally developed for deep boreholes (Serra 1984). A deterministic approach for estimating the 3D resistivity distribution from the parameters of EGS measurements was suggested by

Nyári *et al.* (2010), in which the Archie's (1942) and De Witte (1955) models were compared. Nuclear data collected by EGS tools provide information about the density, water content and porosity, which can be used in resistivity calculations for modelling the transport of water and contaminants. EGS data processing techniques also incorporate inversion-based and multivariate statistical methods. For the evaluation of volumetric parameters in unsaturated sediments, Drahos (2005) introduced a local inversion technique using a weighted least squares optimization algorithm. By following this idea, we establish the petrophysical model with the assumption that the rock matrix is composed of coarse and fine grain components and the pore-space is occupied by freshwater and gas (normally air). In the framework of local inversion, the fractional volumes of water (V_w), gas (V_g), clay (V_{cl}) and sand (V_s) with their estimation errors are estimated at each depth along the penetration hole. From the inversion results, one can derive the hydraulic conductivity (Nyári *et al.* 2010) and other geotechnical parameters, e.g., dry density (Szabó 2012). The above volumetric parameters are extracted by using the following EGS data typically measured in penetration tests: natural gamma-ray intensity, *GR* (cpm), bulk density, ρ_b (g/cm^3), neutron-porosity, Φ_N (V/V) and resistivity, *R* (ohm-m). The classical CPT logs such as cone resistance *RCPT* (MPa) cannot be used in inverse modelling, because there is not any response function to connect the measured data with the petrophysical model. The *RCPT* informs about the drillability of soils, while *GR* is sensitive to clay content and lithology, ρ_b and Φ_N respond to porosity and *R* is used for water saturation estimation. As opposed to inverse modelling, all of the above log types can be used as input for factor analysis.

We calculate the EGS logs in the forward modelling process by assuming a known petrophysical model. The following response functions can be used to relate the observed quantities with the model parameters for unsaturated clastic sediments (Drahos 2005)

$$GR = V_{cl} GR_{cl} + V_s GR_s, \tag{1}$$

$$\rho_b = V_w \rho_w + V_{cl}\, \rho_{cl} + V_s \rho_s, \tag{2}$$

$$\Phi_N = V_w \Phi_{N,w} + V_{cl} \Phi_{N,cl} + V_s \Phi_{N,s}\,, \tag{3}$$

$$R = a\left(V_w + V_g + V_{cl}\right)^{-m} \left(\frac{V_{cl}/\left(V_w + V_{cl}\right)}{R_{cl}} + \frac{1 - \left[V_{cl}/\left(V_w + V_{cl}\right)\right]}{R_w} \right)^{-1} \left(\frac{V_w + V_{cl}}{V_w + V_g + V_{cl}} \right)^{-n}, \tag{4}$$

where the physical constants of rock constituents and pore-filling fluids are indicated with *cl* (clay), *s* (sand), *w* (water), *g* (gas). Symbols *m*, *a*, *n* represent the Archie's (textural) parameters such as cementation exponent, tortuosity factor and saturation exponent, respectively. The detailed list of functional constants are in Table 1. In equations (2)–(3), the density and neutron-porosity of gas are set to zero. Response function (4) applicable to calculate the resistivity was suggested by De Witte (1955). In fact, the fluid and matrix properties are not always constant as they may vary in the heterogeneous sediment. However, these zone parameters are treated as constant to avoid an ambiguous underdetermined inverse problem. In this study, the local inverse problem has three unknowns (V_w, V_{cl}, V_s) and four data types (GR, ρ_b, Φ_N, R). The gas saturation is derived from the inversion results by using the material balance equation $V_g = 1 - V_w - V_{cl} - V_s$. The inverse problem is overdetermined, which has a unique solution. On the other hand, rock samples can be collected easily from the shallow holes and several physical parameters can be a priori given using reliable laboratory information. In this study, we use equations (1)–(4) for solving a highly overdetermined inverse problem called interval inversion to give an estimate to the petrophysical model, which is used to validate the results of factor analysis.

2.2 Robust method of factor analysis

Factor analysis offers a new alternative for the evaluation of unsaturated sediments. We extract the water volume as an important parameter of the applied petrophysical model by the IRFA method suggested by Szabó and Dobróka (2017). In this study, the IRFA method is improved by using the Steiner weights to reduce the noise sensitivity of the procedure and give a robust solution. In the first step, we organize the standardized EGS data into an *N*-by-*K* matrix (**D**), where *N* is the total number of sampled depths and *K* is the number of applied direct-push tools. Factor analysis reduces the *K*

dimensional problem to a lower dimensional one by extracting *M* number of new variables (factors) from the data set. The model of factor analysis is

$$\mathbf{D} = \mathbf{F}\mathbf{L}^{\mathrm{T}} + \mathbf{E}, \tag{5}$$

where **F** is the *N*-by-*M* matrix of factor scores, $\mathbf{L}^{\mathrm{T}}$ is the *M*-by-*K* transpose matrix of factor loadings and **E** is the matrix of residuals. In equation (5), the observed variables are developed as a linear combination of the statistical factors. Factor loadings practically quantify the strength of correlation between the log types and factors. The scores in a given column of matrix **F** build up the well log of the relevant factor. For instance, the first column defines the first factor explaining the largest part of variance of the observed data. If the factors are linearly independent, the covariance matrix of observed data can be expressed with the factor loadings

$$\mathbf{\Sigma} = N^{-1}\mathbf{D}^{\mathrm{T}}\mathbf{D} = \mathbf{L}\mathbf{L}^{\mathrm{T}} + \mathbf{\Psi}, \tag{6}$$

where $\mathbf{\Psi}$ is the *K*-by-*K* matrix of specific variances representing the portion of data variances not explained by the common factors. In several cases, the matrix of specific variances is set to be zero and the PCA is used to calculate the factors. If the specific variances are known, factor loadings can be estimated by solving an eigenvalue problem. For the lack of specific variances, only approximation can be made and the factor loadings are simultaneously estimated with the specific variances using the maximum likelihood method. In our methodology, any of the above approaches can be used for giving an initial estimate to the factors, which are then refined by the IRFA method. We apply the non-iterative approach of Jöreskog (2007), which gives a quick solution and an objective estimate to the number of factors.

Jöreskog's method gives optimal results for Gaussian distributed data. Since EGS data sets rarely follow normal distribution, we apply the IRFA procedure. Equation (5) is properly modified as

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

where **d** denotes the *KN*-element column vector of standardized observed data, $\tilde{\mathbf{L}}$ is the *NK*-by-*NM* matrix of factor loadings, **f** is the *MN*-length column vector of factor scores, **e** is the *KN*-element column vector of residuals. The actual values of factor scores and loadings are refined in an iterative

algorithm (Szabó and Dobróka 2017), which uses the combination of the damped least squares and the weighted least squares methods (Menke 1984)

$$\mathbf{L}^{\mathrm{T}(q)} = \left(\mathbf{F}^{\mathrm{T}(q-1)}\mathbf{F}^{(q-1)} + \alpha^2\mathbf{I}\right)^{-1}\mathbf{F}^{\mathrm{T}(q-1)}\mathbf{D}\,, \tag{8}$$

$$\mathbf{f}^{(q)} = \left(\tilde{\mathbf{L}}^{\mathrm{T}(q-1)}\mathbf{W}\tilde{\mathbf{L}}^{(q-1)}\right)^{-1}\tilde{\mathbf{L}}^{\mathrm{T}(q-1)}\mathbf{W}\mathbf{d}\,, \tag{9}$$

where α is a properly chosen damping factor and q is the index running through the number of iterations. The recursive process is based on the iterative re-weighting of deviation between the measured (**d**) and calculated data ($\tilde{\mathbf{L}}\mathbf{f}$). In each iteration, the larger the distance is between the measured and predicted data, the less weight is given to the relevant datum. Szabó and Dobróka (2017) applied the Cauchy weights for solving the IRFA problem. The scale parameter of the Cauchy weight function is to be arbitrary chosen, which may have considerable impact on the solution.

We improve the Cauchy-IRFA method by using a fully automated weighting procedure for optimizing the values of weight coefficients. The Most Frequent Value (MFV) method is known as a robust estimator with high statistical efficiency (Steiner 1991). The MFV is defined as the weighted average of the *N*-element statistical sample

$$\mathrm{MFV} = \frac{\sum_{i=1}^{N}\left[\dfrac{\varepsilon^2}{\varepsilon^2 + (x_i - \mathrm{MFV})^2}\right]x_i}{\sum_{i=1}^{N}\left[\dfrac{\varepsilon^2}{\varepsilon^2 + (x_i - \mathrm{MFV})^2}\right]}\,, \tag{10}$$

where x_i denotes the i-th data and ε is the dihesion given as the scale parameter of the weight function represented by the fraction in square brackets. It is noticeable that the dihesion controls the relative importance of the data in the weighting process. If the value of ε is high, all data get approximately the same weight. For small values of ε, only data in the near vicinity of the MFV affect the estimation considerably. Since it appears on both sides of equation (10), the MFV is improved by an iterative algorithm. In the q-th iteration, the dihesion is calculated analogously to equation (A-6) using the actual value of MFV

$$\varepsilon_q = \left(\frac{3\sum_{i=1}^{N} \frac{\left(x_i - \mathrm{MFV}_{q-1}\right)^2}{\left[\varepsilon_{q-1}^2 + \left(x_i - \mathrm{MFV}_{q-1}\right)^2\right]^2}}{\sum_{i=1}^{N} \frac{1}{\left[\varepsilon_{q-1}^2 + \left(x_i - \mathrm{MFV}_{q-1}\right)^2\right]^2}} \right)^{1/2} . \quad (11)$$

By using the updated value of ε, the MFV is refined by equation (10). The development of convergence usually requires some tens of iterations. The iteration process runs until a stop criterion (e.g., the difference between the old and new value of MFV is under a given threshold or maximum iteration number) is met. Compared to Cauchy weighting the advantage of the MFV method is that the scale parameter ε is automatically calculated during the iterative procedure, which allows the finding of optimal weight coefficients for the actual data set. In the IRFA procedure, we choose the weight function given in equation (9) analogously to equation (10). The elements of the *NK*-by-*NK* diagonal weight matrix are proportional to the deviations between the (standardized) measured and calculated EGS data

$$W_{ii} = \frac{\varepsilon^2}{\varepsilon^2 + \left(e_i\right)^2} \quad \left(i = 1,2,...,KN\right). \quad (12)$$

The modified iterative factor analysis procedure is named as MFV-IRFA, in which optimal weights are calculated automatically for each component of the deviation vector to estimate the factors more accurately.

Statistical factors are generally rotated for a more efficient physical interpretation. The reduced covariance matrix of observed data $\mathbf{\Sigma}^* = \mathbf{L}^*\mathbf{L}^{*\mathrm{T}}$ can be computed in several ways using a *K*-by-*K* orthogonal matrix **V**, where $\mathbf{L}^* = \mathbf{LV}$. Orthogonal transformation performed on the factor loadings, as geometric rotation, results in an equivalent solution to the factors. In this study, the varimax algorithm suggested by Kaiser (1958) is used to simplify the structure of factor loadings by maximizing the sum of the variances of the squared factor loadings. By this method, any given factor is influenced by only a few observed variables, while the remaining variables have near-zero loadings on the same factor. As a result, there will be few log types to which the resultant factor strongly correlates. The first factor, which explains the largest part of variance of the observed data, highly correlates to the water

content of shallow formations (Szabó *et al.* 2012). We assume that the relation between the relevant factor and water volume as the product of water saturation and porosity is also strong. To test this relation, the water volume derived independently from the interval inversion of EGS logs is correlated to the first factor estimated by the MFV-IRFA procedure.

2.3 Interval inversion method

In practice, fast inversion methods are normally used to predict the petrophysical parameters using a depth-by-depth approach (Drahos 2005). As having barely more EGS data than unknown model parameters at a given depth, we solve a set of marginally overdetermined inverse problems sensitive to data noises and limited in estimation accuracy. Dobróka *et al.* (2016) suggested an interval inversion approach to improve the quality of inversion results. The interval inversion method is used to process the data set over a longer depth interval to predict the vertical distributions of petrophysical parameters in a joint inversion procedure. The depth-dependent model parameters are discretized using series expansion and the inverse problem is solved for much smaller number of expansion coefficients than data. The resulting highly overdetermined inverse problem leads to significantly more accurate solution than local inversion methods and, if necessary, gives an estimate to the layer-thicknesses and zone parameters within the inversion procedure (Dobróka and Szabó 2012).

We approximate the depth-dependent model parameters, such as fractional volumes of clay, sand and water, by the following series expansion

$$m_l(z) = \sum_{j=1}^{J^{(l)}} B_j^{(l)} P_{j-1}(z), \tag{13}$$

where m_l is the l-th petrophysical parameter, B_j is the j-th expansion coefficient, P_j is the j-th degree Legendre polynomial and $J^{(l)}$ is the number of expansion coefficients suitably chosen for describing the l-th model parameter. Legendre polynomials in equation (13) are used as basis function, which are known quantities depending only on the depth coordinates. The orthonormal polynomials can be favourably used in inversion applications for estimating slightly correlated model parameters. We collect the EGS data of different types measured from an arbitrary depth interval in the column vector

$\mathbf{d}^{(obs)}$. To calculate the theoretical logs to the same interval, we solve the forward problem by using equations (1)–(4). The objective function of the inverse problem is

$$E = \left\| \mathbf{e}^* \right\|_2^2 + \lambda^2 \left\| \mathbf{B} \right\|_2^2 = \min , \tag{14}$$

where **e*** denotes the deviation vector including the normalized differences between the observed and calculated data, **B** is the vector of expansion coefficients and λ is a regularization parameter necessary for the numerical stabilization of the inversion procedure. As indicated by the above formulation, the inverse problem is solved for the series expansion coefficients

$$\mathbf{B} = \mathbf{G}^{-g}\mathbf{d}^{(obs)}, \tag{15}$$

where $\mathbf{G}^{-g}$ is the generalized inverse matrix of the damped least squares method (Marquardt 1959). The vertical distribution of petrophysical parameters can be derived directly from the inversion results using equation (13). According to Menke (1984), the covariance matrix of the model parameters estimated by a linearized inversion method is directly proportional to the data covariance matrix including the variances of observed data. The accuracy of volumetric parameters derived from equation (13) requires the propagation of errors taken into consideration. At first the covariance matrix of series expansion coefficients are calculated by Menke's formula, which is then related to the covariance matrix of the petrophysical parameters (Dobróka *et al.* 2016). The quality-checked inversion result, including the water content, serves as a reference model for regression analysis, in which the statistical factors are correlated to the petrophysical parameters of unsaturated formations.

3 APPLIED STATISTICAL WORKFLOW

We perform the formation evaluation by using the workflow given in Fig. 2. At the beginning, we have *K* number of measured logs as input, which are simultaneously processed to derive less number of statistical variables. The number of factors are specified by the Jöreskog's algorithm. The initial values of factor loadings and scores estimated by the same algorithm are simultaneously refined in the

MFV-IRFA procedure. The weight function related to the difference between the observed and predicted logs is automatically re-calculated using the MFV method in each iteration. Technically, the value of dihesion is updated in an inner loop of iterations during the MFV-IRFA procedure, while the factor scores and loadings are estimated with the actual weights using equations (8)–(9). The output of the MFV-IRFA procedure is a set of factor logs showing the vertical variation of factor scores along the penetration hole. The factors may carry information on not directly measurable petrophysical properties of rocks. To extract this information from the EGS data set, we correlate factors to different petrophysical parameters given from independent sources, e.g., core data, pumping tests or independent well-log-analysis techniques like the interval inversion procedure. Partial regression analyses may reveal the connection between the factors and petrophysical properties. In this study, we demonstrate the strong connection between the first factor and water content, which has been tested in different areas in Hungary. Based on the consistent relation between the two quantities, we suggest the use of a regression formula to extract the water volume directly from the EGS data set. Water content estimated by the MFV-IRFA method can be used as an initial model for inverse modelling, which can be refined by the repeated use of interval inversion and factor analysis. On the other hand, by treating the estimated petrophysical quantities as known parameter, we can increase the overdetermination (data-to-unknowns ratio) of the inverse problem as well as the estimation accuracy of other inversion unknowns.

4 RESULTS AND INTERPRETATION

We use the MFV-IRFA method for the processing of EGS data originated from Bátaapáti, South West Hungary. In the test area, detailed ground geophysical surveys were previously made for establishing a nuclear waste repository beneath the sedimentary formations (Vértesy *et al.* 2004). The shallow structure is composed of a loessy-sandy sequence deposited on a partially weathered granite basement. The thickness of the loess cover has an average thickness of 50 m and the water level is mainly at the top of the granite. Engineering geophysical soundings were carried out in the upper 20–25 m of the unconsolidated, partially water-saturated formation. In this study, we extract the statistical factors

using EGS data measured in single holes (sections 4.1–4.2) and several neighbouring holes (section 4.3). Then, the first factor is related to water content estimated independently from the interval inversion of the EGS logs.

4.1 One-dimensional application

We first simultaneously process the GR, ρ_b, Φ_N, R logs using the interval inversion method to estimate the vertical distribution of water, clay and sand volumes in Hole−4. The air volume is calculated from the material balance equation. The input logs and their confidence intervals are plotted in Fig. 3. The accuracy of observed data is given after Drahos (2005). The standard deviation of the EGS data are assumed as σ_1=0.22 kcpm, σ_2=0.07 g/cm^3, σ_3=0.04 V/V, σ_4=2.1 ohm-m. We have totally 944 data by a sampling distance of 0.1 m. We discretize the model parameters (V_w, V_{cl}, V_s) using equation (13). The series expansion is performed for the whole length of Hole−4, which allows the estimation of petrophysical parameters to the same interval. The degree of Legendre polynomials for all model parameters is set to 40, and the depth coordinates are properly scaled to the range of −1 and 1. The number of expansion coefficients is optimized by preliminary tests minimizing the correlation of model parameters (Dobróka *et al.* 2016). By doing this, a trade-off must be taken between the vertical resolution and the stability of the inversion procedure. The total number of expansion coefficients is 123, which are estimated by the interval inversion procedure run over 15 iterations. The data-to-unknown ratio is approximately seven, which is significantly higher than that of the local inverse problem (where it is 4/3). We set the initial values of the zeroth-order expansion coefficients to 0.5 for clay volume and 0.2 for water and sand volume, respectively. The expansion coefficients of higher-order Legendre polynomials are equal to zero. Theoretical EGS logs are calculated for the entire logging interval in each iteration by equations (1)–(4). We find that the inversion procedure is still stable without using regularization, i.e., $\lambda \approx 0$ is used in equation (14). In the first step, the data distance (D_d) as the average root-mean-square error (RMSE) between the observed and predicted data is 36 %, which decreases down to 6.9 % at the end of the inversion procedure. The estimation error of the

zeroth-order series expansion coefficients is 1–2 %. The average of correlation coefficients between the inversion unknowns is 0.08, which refers to practically uncorrelated model parameters. This remarkable value indicates a stable inversion procedure. The distribution of volumetric parameters are derived from the estimated model vector using equation (13). The input logs and the results of interval inversion are plotted in Fig. 3. The estimation errors of volumetric parameters (σ_5, σ_6, σ_7) are 2–6 V/V, while the mean correlation between the volumetric parameters is approximately 0.48. Both quantities show stable and reliable inversion results.

The vertical variation of factors is determined by the MFV-IRFA of the *RCPT*, *GR*, ρ_b, Φ_N, *R* logs. Two uncorrelated factors are calculated by the procedure. The initial values of factor loadings are given by the method of Jöreskog (2007), which are updated simultaneously with the factor loadings over 15 iterations. Singular value decomposition of the reduced covariance matrix $\mathbf{\Sigma}^* = \mathbf{L}\mathbf{L}^{\mathrm{T}}$ shows that the first factor explains 76 % part of the total variance, while the 24 % part of that is given by the second factor. The estimated factor loadings are listed in Table 2, which shows that the strongest correlation is between the first factor and EGS log types sensitive to water saturation (Φ_N and *R*).

The main purpose of our study is to find correlation between the factors and petrophysical parameters of the shallow structure. The regression tests show a strong exponential relation between the first factor and water volume estimated by the interval inversion method (Fig. 4a), which is described by the empirical function $V_w = 0.29\,\mathrm{e}^{(0.11 F_1)} - 0.09$. The linear connection between the water volumes given separately by factor analysis and interval inversion is also strong. Both methods shows consistent estimation results (Fig. 4b). The result of factor analysis and the soil composition estimated by interval inversion are illustrated in Fig. 5. The RMSE as a measure of misfit between the water volume logs estimated by the statistical and inversion methods is 1.45 %, which indicates a close agreement between the interpretation results.

4.2 Application to data set with outliers

The MFV-IRFA procedure is used for the processing of non-Gaussian distributed data. The measurements may be contaminated with outliers, for example, when we drill a hard formation, the probe readings sometimes lie an abnormal distance from those of other points. In this test, we process the *RCPT*, *GR*, ρ_b, Φ_N, *R* logs measured in Hole–12. The non-Gaussian nature of the data set is characterized by non-zero values of empirical kurtosis and skewness. The value of the former is 17, while that of the latter is 2. To test the outlier-sensitivity of factor analysis, we make a comparative study between the Jöreskog's method, which we have chosen to call Traditional Factor Analysis (TFA), and the robust MFV-IRFA method. In the regression phase of the study, we use the inversion-derived water content as reference value.

We calculate two factors by using the TFA and MFV-IRFA methods, respectively. The result of the TFA procedure is used as an initial model for the MFV-IRFA procedure, in which the factor loadings and scores are updated jointly over 15 iterations. In each step of the iterative procedure, the Steiner weights in equations (9) and (12) are re-calculated in further 30 steps. In the inner loop of iterations, the dihesion is automatically decreased and changed differently for each EGS log (Fig. 6a). The value of ε is continuously decreased with the number of iterations, which makes the big deviations contribute less to the solution. For a given value of ε, the larger the distance is between the measured and calculated data, the smaller the amplitude of the weight coefficient (Fig. 6b). The figure shows that the large deviations give a negligible contribution to the solution. The probability distribution function of the optimal weights given in the last iteration step are shown in Fig. 6c, while that of the prediction errors are in Fig. 6d. The MFV-IRFA procedure is convergent and stable. The development of convergence is shown in Fig. 7, where the relative distance between the measured and calculated data (and the weighting factors) gradually decreases with the number of iterations. At the end of the iterative procedure, the rotated loadings of the first factor is estimated as $L^{(RCPT)}$= –0.01, $L^{(GR)}$=0.17, $L^{(\rho_b)}$=0.86, $L^{(\Phi_N)}$=0.73, $L^{(R)}$=–0.88. The first factor, which explains the 74 % part of the data variance, has still strong connection with the penetration logs highly sensitive to saturation. The noise rejection capability of TFA and MFV-IRFA methods can be compared in Fig. 8. The extreme values of factor scores in Fig. 8a show that the TFA method is unable to suppress the outliers in the calculation of

factors. In contrast to the outlier-sensitive TFA procedure, the outliers are efficiently removed by the MFV-IRFA procedure (Fig. 8b), which shows the robustness of the iterative method.

We correlate the first factor to water saturation calculated as the ratio of the inversion-derived water content and porosity. (We solve a set of local inverse problems using a weighted least squares technique, because the effect of outliers can be studied in more detail). The first factor versus water saturation relation is strong for both methods, with a difference that the MFV-IRFA solution is more accurate as it lacks for the outliers (Fig. 9b). Both regression functions follow the exponential model $S_w = \alpha e^{(\beta F_1)} + \gamma$, the coefficients of which are close to each other. For the MFV-IRFA procedure, the regression coefficients are calculated at 95 % confidence level as α=0.77±0.14, β=0.22±0.04, γ=–0.12±0.11.

The results of the comparative study is summarized in Fig. 10. The well logs of the two factors are given in the fifth track. The factor logs estimated by the TFA method include erroneous peaks caused by the outliers (e.g., in the consolidated layer at the depth of around 24–27 m), while much smoother estimate to the factor variables are given by the MFV-IRFA procedure. Tracks 6–7 also show that the water saturation log derived by the TFA method contains abrupt changes in the places of outliers, while these peaks are completely absent when using the MFV-IRFA method. The RMSE measured between the results of factor analysis and local (weighted) inversion is 7.4 % for the TFA method, while it is 4.6 % for the MFV-IRFA procedure. They show a close agreement between the independent well-log-analysis results. It is also concluded that the application of MFV-IRFA method makes a significant improvement in the estimation accuracy of water saturation.

The single borehole application of the MFV-IRFA method is studied also by synthetic modelling experiments. We calculate EGS logs from an exactly known petrophysical model (last track in Fig. 11) using equations (1)–(4). The water table in the clayey sand formation is found at the depth of around 7.5 m, under which the pore-space is fully saturated with fresh-water and the water volume changes only with porosity. By adding different amount of random noise to the synthetic data (5 % Gaussian and 5 % Gaussian plus eight times higher random noise added to every randomly chosen eighth data, respectively), we calculate two factors by the TFA and MFV-IRFA methods, respectively.

In both tests, the absolute value of factor loadings of the neutron-porosity and resistivity logs is not smaller than 0.83. The results confirm the exponential relation between the first factor and the exactly known values of the water content. In Fig. 11, one can observe the good noise suppression capability of the MFV-IRFA method.

4.3 Two-dimensional application

We further develop the MFV-IRFA algorithm for multidimensional applications, which allows the simultaneous processing of EGS data acquired from several neighbouring drill-holes. Let $\mathbf{d}^{(h)}$ denote the observed data vector, defined in equation (7), in the *h*-th hole (*h*=1,2,…,*H*). After gathering all data to a column vector, the model of factor analysis takes the form

$$\begin{pmatrix} \mathbf{d}^{(1)} \\ \vdots \\ \mathbf{d}^{(h)} \\ \vdots \\ \mathbf{d}^{(\mathrm{H})} \end{pmatrix} = \begin{pmatrix} \tilde{\mathbf{L}}^{(1)} & 0 & 0 & 0 & 0 \\ 0 & \ddots & 0 & 0 & 0 \\ 0 & 0 & \tilde{\mathbf{L}}^{(h)} & 0 & 0 \\ 0 & 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & 0 & \tilde{\mathbf{L}}^{(\mathrm{H})} \end{pmatrix} \begin{pmatrix} \mathbf{f}^{(1)} \\ \vdots \\ \mathbf{f}^{(h)} \\ \vdots \\ \mathbf{f}^{(\mathrm{H})} \end{pmatrix} + \begin{pmatrix} \mathbf{e}^{(1)} \\ \vdots \\ \mathbf{e}^{(h)} \\ \vdots \\ \mathbf{e}^{(\mathrm{H})} \end{pmatrix}, \tag{16}$$

where $\tilde{\mathbf{L}}^{(h)}$ is the matrix of factor loadings and $\mathbf{f}^{(h)}$ is the vector of factor scores estimated along the *h*-th borehole. Since we have N_h number of depth points in the *h*-th hole, the total number of processed depths is $N^* = N_1 + N_2 + \ldots + N_H$. By applying the same number of probes in all holes, the size of the weight matrix in equation (12) modifies to KN^*-by-KN^*. In the multidimensional algorithm of MFV-IRFA, the matrix of factor loadings and the column vector of factor scores in equation (16) are analogously determined by equations (8)–(9). The CPU time of the modified procedure can be highly reduced by employing special algorithms developed for sparse matrices, e.g., by means of the MATLAB software package. The estimated factor scores are necessarily interpolated between the holes to derive the factor maps.

We test the 2D MFV-IRFA procedure in twelve penetration holes drilled along a profile in Bátaapáti test site. The holes are located approximately 50 meters away from each other on the northwest-southeast oriented profile, the total length of which is 550 m. Penetration holes studied in

sections 4.1–4.2 (Hole–4 and Hole–12) are located at the horizontal distances 150 m and 550 m along the survey line, respectively. We analyse the *RCPT*, *GR*, ρ_b, Φ_N, *R* logs, where the total number of processed data is 15,500. Two factors are calculated by the 2D MFV-FA procedure. The first factor explains the 72 % part of total variance of original data. The initial values of factor loadings estimated by the method of Jöreskog (2007) are updated simultaneously with the factor loadings over 20 iterations. The Steiner weights are re-calculated separately for each log in further 50 iterations. The magnitude and sign of the rotated loadings of the first factor are consistent with the results of the one-dimensional cases: $L^{(RCPT)}$=–0.08, $L^{(GR)}$=0.03, $L^{(\rho_b)}$=0.85, $L^{(\Phi_N)}$=0.77, $L^{(R)}$=–0.88. The second factor shows the highest correlation with the *RCPT* log. The first factor is strongly correlated to water saturation (Fig. 12a), where the functional relation is slightly exponential. Water content is linearly proportional to the same factor (Fig. 12b). Earlier studies showed that the first factor could be related also to some quantities measured by the EGS tools. Szabó *et al.* (2012) made an experiment to simulate the neutron log to unmeasured intervals using the results of factor analysis. These calculations were based on the high correlation between the first factor and the neutron log. It was experienced that these connections also existed when the relevant log was removed from the procedure of factor analysis. In this study, we find a strong correlation between the first factor and neutron-porosity (Fig. 12c) and resistivity (Fig. 12d) for the 2D case. The crossplot in Fig. 12e confirms the reliability of the factor analysis-based water content estimation. The high correlation coefficient shows a strong linear proportionality between the inversion-based and statistical estimation results. The correlation coefficients of the above regression relations are inversely related to the scattering of data points, which depend on not only the data noise but also the lateral variation in the lithology of the shallow formation.

The 2D MFV-IRFA procedure can be favourably used for the fast automatic processing of large statistical samples including outliers. To investigate the horizontal and vertical distribution of water content, we first calculate the 2D sections of factor variables. The factor scores are interpolated by using a standard kriging algorithm to give an estimate for the spatial distribution of factors between the penetration holes (Fig. 13). By using the regression relation between the first factor and water content, we derive the 2D section of water content directly from the first factor (Fig. 14). The result of

1D inverse modelling (Fig. 14a) show a close agreement with that of the 2D MFV-IRFA procedure (Fig. 14b). The Pearson's correlation coefficient calculated between the water content sections is 0.86, which shows consistent results.

5 CONCLUSIONS

We suggest a robust algorithm for the factor analysis of engineering geophysical sounding data to improve the estimation of water content in shallow formations. The case study and synthetic modelling experiments demonstrate that the Most Frequent Value-based factor analysis procedure gives highly acceptable results for non-Gaussian distributed data sets. The advantage of the method is that it gives continuous in-situ information along the borehole and makes a significant improvement in the estimation of water saturation for outlying measurement data. Optimal weights are calculated in a fully automated weighting process, while the effect of extreme noises are efficiently suppressed. According to our experience, at least 40 % (or even better) relative improvement of estimation accuracy can be achieved compared to traditional factor analysis.

We find a strong correlation between the first factor and water content in shallow sediments, which is consistent in different measurement areas in Hungary. Other derived quantities, e.g., air saturation and dry density can be extracted from the factors scores more reliably to get information that is more detailed on the petrophysical characteristics of near-surface layers. The study of subsequent factors may also be of importance in further studies. By the interval inversion-assisted factor analysis, additional unknowns can be estimated such as effective porosity and hydraulic conductivity. In cross-hole applications, penetration logs originated from several holes can be simultaneously processed using the extended version of the MFV-IRFA method, which allows the fast calculation of water saturation and derived quantities between the holes. Physical parameters such as neutron-porosity or resistivity are sometimes not measured in some depth intervals or boreholes. Synthetic logs of the same quantities generated by factor analysis of other observed EGS data can be effectively used for the replacement of missing observations. In an ongoing research, we make efforts to estimate the textural parameters included in the probe response functions using an interval inversion

approach. The development of the inversion method supported by robust factor analysis can be an important step forward for further improvements in a more accurate and reliable interpretation of engineering geophysical sounding data. The statistical workflow introduced in the paper can be employed as a powerful data processing tool for hydrogeophysical, environmental and civil engineering applications for a more reliable assessment of unsaturated sedimentary structures.

ACKNOWLEDGEMENTS

The research was carried out within the GINOP-2.3.2-15-2016-00031 "Innovative solutions for sustainable groundwater resource management" project of the Faculty of Earth Science and Engineering of the University of Miskolc in the framework of the Széchenyi 2020 Plan, funded by the European Union, co-financed by the European Structural and Investment Funds. The authors thank to the support of Professor Mihály Dobróka, University of Miskolc and Associate professor Dezső Drahos, Roland Eötvös University of Budapest for their scientific advices and cooperation.

REFERENCES

Archie G.E. 1942. The electrical resistivity log as an aid in determining some reservoir characteristics. Petroleum Transactions of the AIME **146**, 54–62.

Asfahani J. 2014. Statistical factor analysis technique for characterizing basalt through interpreting nuclear and electrical well logging data (case study from Southern Syria). Applied Radiation and Isotopes **84**, 33–39.

De Witte L. 1955. A study of electric log interpretation methods in shaly formations. Petroleum Transactions of the AIME **204**, 103–110.

Dobróka M., Gyulai Á., Ormos T., Csókás J. and Dresen L. 1991. Joint inversion of seismic and geoelectric data recorded in an underground coal mine. Geophysical Prospecting **39**, 643–665.

Dobróka M. and 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.

Dobróka M. and Szegedi H. 2014. On the generalization of seismic tomography algorithms. American Journal of Computational Mathematics **4**, 37–46.

Dobróka M., Szabó N.P., Tóth J. and Vass P. 2016. Interval inversion approach for an improved interpretation of well logs. Geophysics **81**, (2), D155–D167.

Drahos D. 2005. Inversion of engineering geophysical penetration sounding logs measured along a profile. Acta Geodaetica et Geophysica **40**, 193–202.

Drahos D. 2008. Determining the objective function for geophysical joint inversion. Geophysical Transactions **45**, 105–121.

Drahos D. and Galsa A. 2007. Finite element modelling of penetration electric sonde (in Hungarian). Magyar Geofizika **48**, 22–30.

Draskovits P. and Fejes I. 1994. Geophysical methods in drinkwater protection of near-surface reservoirs. Journal of Applied Geophysics **31**, 53–63.

Fejes I. and Jósa E. 1990. The engineering geophysical sounding method. Principles, instrumentation, and computerised interpretation. In: Geotechnical and environmental geophysics, Environmental and groundwater, Vol. 2 (ed. S.H. Ward), pp. 321–331, SEG, ISBN 978-0-931830-99-0.

Grana G., Dvorkin J. and Cibin P. 2011. Factor analysis prediction of effective stress from measurable rock attributes and calibration data. First Break **29**(7), 63–72.

Gyulai Á., Baracza M.K. and Szabó N.P. 2014. On the application of combined geoelectric weighted inversion in environmental exploration. Environmental Earth Sciences **71**, 383–392.

Gyulai Á., Szűcs P., Turai E., Baracza M.K. and Fejes Z. 2017. Geoelectric characterization of thermal water aquifers using 2.5D inversion of VES measurements. Surveys in Geophysics **38**, 503–526.

Jöreskog K.G. 2007. Factor analysis and its extensions. In: Factor analysis at 100, historical developments and future directions, (eds. Cudeck R. and MacCallum R.C.), pp. 47–77. Lawrence Erlbaum Associates, ISBN 0805862129.

Kaiser H.F. 1958. The varimax criterion for analytic rotation in factor analysis. Psychometrika **23**, 187–200.

Kirsch R. 2006. Groundwater geophysics: A tool for hydrogeology. Springer, ISBN 978-3-540-29383-5.

Lawley D.N. and Maxwell A.E. 1962. Factor analysis as a statistical method. The Statistician **12**, 209–229.

Lunne T., Robertson P.K. and Powell J.J.M. 1997. Cone-penetration testing in geotechnical practice. Taylor & Francis, ISBN 9780419237501.

Marquardt D.W. 1959. Solution of non-linear chemical engineering models. Chemical Engineering Progress **55**, 65–70.

Menke W. 1984. Geophysical data analysis: Discrete inverse theory. Academic Press, ISBN 978-0-12-490920-5.

Niculescu B.M., Andrei G. and Ciuperca C. 2016. Improved formation evaluation through principal component analysis. 78th EAGE Conference & Exhibition 2016, Vienna, Austria, Expanded Abstracts, Paper Tu STZ2 15.

Nyári Z., Kanli A.I., Stickel J. and Tillmann A. 2010. The use of non-conventional CPTe data in determination of 3-D electrical resistivity distribution. Journal of Applied Geophysics **70**, 255–265.

Puskarczyk E., Jarzyna J. and Porebski S. 2015. Application of multivariate statistical methods for characterizing heterolithic reservoirs based on wireline logs – example from the Carpathian Foredeep Basin (Middle Miocene, SE Poland). Geological Quarterly **59**, 157–168.

Schulmeister M.K., Butler J.J., Healey J.M., Zheng L., Wysocki D.A. and McCall G.W. 2003. Direct-push electrical conductivity logging for high-resolution hydrostratigraphic characterization. Ground Water Monitoring & Remediation **23**, 52–62.

Serra O. 1984. Fundamentals of well-log interpretation 1. The acquisition of logging data. Developments in Petroleum Science vol. 15. Elsevier. ISBN 978-0-444-42132-6.

Steiner F. 1988. Most frequent value procedures (A short monograph). Geophysical Transactions **34**(2–3), 139–260.

Steiner F. 1991. The most frequent value: Introduction to a modern conception of statistics. Akadémiai Kiadó, ISBN 9789630556873.

Steiner F. 1997. Optimum methods in statistics. Akadémiai Kiadó, ISBN 978-9630574396.

Szabó N.P. 2011. Shale volume estimation based on the factor analysis of well-logging data. Acta Geophysica **59**, 935–953.

Szabó N.P. 2012. Dry density derived by factor analysis of engineering geophysical sounding measurements. Acta Geodaetica et Geophysica **47**, 161–171.

Szabó N.P., Dobróka M. and Drahos D. 2012. Factor analysis of engineering geophysical sounding data for water saturation estimation in shallow formations. Geophysics **77**, (3), WA35–WA44.

Szabó N.P. 2015. Hydraulic conductivity explored by factor analysis of borehole geophysical data. Hydrogeology Journal **23**, 869−882.

Szabó N.P. 2016. Hydrocarbon formation evaluation using an efficient genetic algorithm-based factor analysis method. 15th European Conference on the Mathematics of Oil Recovery, Amsterdam, The Netherlands, Paper Mo P071.

Szabó N.P. and Dobróka M. 2017. Robust estimation of reservoir shaliness by iteratively reweighted factor analysis. Geophysics **82**(2), D69–D83.

Szalai S., Koppan A., Szokoli K. and Szarka L. 2013. Geoelectric imaging properties of traditional arrays and of the optimized Stummer configuration. Near Surface Geophysics **11**, 51–62.

Szalai S., Lemperger I., Metwaly M., Kis A., Wesztergom V., Szokoli K. and Novák A. 2015. Increasing the effectiveness of electrical resistivity tomography using γ_{11n} configurations. Geophysical Prospecting **63**, 508–524.

Szegedi H. and Dobróka M. 2014. On the use of Steiner's weights in inversion-based Fourier transformation: robustification of a previously published algorithm. Acta Geodaetica et Geophysica **49**, 95–104.

Szűcs P., Civan F. and Virág M. 2006. Applicability of the most frequent value method in groundwater modeling. Hydrogeology Journal **14**, 31–43.

Tillmann A., Englert A., Nyári Zs., Fejes I., Vanderborght J. and Vereecken H. 2008. Characterization of subsoil heterogeneity, estimation of grain size distribution and hydraulic conductivity at the Krauthausen test site using Cone Penetration Test. Journal of Contaminant Hydrology **95**, 57–75.

Turai E. and Dobróka M. 2011. Data processing method developments using TAU-transformation of Time-Domain IP data I. Theoretical basis. Acta Geodaetica et Geophysica **46**, 283–290.

Turai E. 2011. Data processing method developments using TAU-transformation of Time-Domain IP data II. Interpretation results of field measured data. Acta Geodaetica et Geophysica **46**, 391–400.

Vértesy L., Fancsik T., Fejes I., Gulyás Á., Hegedűs E., Kovács A.Cs., Kovács P., Kiss J., Madarasi A., Sőrés L., Szabó Z. and Tóth Z. 2004. Ground-based geophysical surveys at the Bátaapáti (Üveghuta) Site and in its vicinity. In: Annual Report of the Geological Institute of Hungary 2003, 239–256.

Walsh D., Turner P., Grunewald E., Zhang H., Butler J.J., Reboulet E., Knobbe S., Christy T., Lane J.W., Johnson C.D., Munday T. and Fitzpatrick A. 2013. A small-diameter NMR logging tool for groundwater investigations. Groundwater **51**, 914–926.

APPENDIX: DERIVATION OF DIHESION IN EQUATION (11)

The most frequent value (MFV) in equation (10) is the symmetry point of the probability distribution function $f(x)$. For large sample sizes, the corresponding integral formula is

$$\mathrm{MFV} = \frac{\int_{-\infty}^{\infty} \frac{\varepsilon^2 x}{\varepsilon^2 + (x - \mathrm{MFV})^2} f(x)\mathrm{d}x}{\int_{-\infty}^{\infty} \frac{\varepsilon^2}{\varepsilon^2 + (x - \mathrm{MFV})^2} f(x)\mathrm{d}x}, \tag{A-1}$$

where x denotes the sampled variable and ε is the dihesion. The weigh function in the above expression is

$$\varphi(x) = \frac{\varepsilon^2}{\varepsilon^2 + (x - \mathrm{MFV})^2}, \tag{A-2}$$

which has a maximum value in the centre of gathering and it decreases to approximately zero for outlying values of the sample. The dihesion controls the relative importance of the observed data in the weighting process. For large values of ε, the sample is equally weighted, while for small values of ε, only the values in the immediate vicinity of the MFV affect the estimation considerably. We need to define a suitable measure for the number of data playing significant role in computing the MFV. Steiner (1988) applied the sum of weights given by (A-2) to calculate the number of effective data $\xi_{eff}(\varepsilon)$. The optimal value of dihesion is found at the maximum of the expression $\xi_{eff}(\varepsilon)/\varepsilon$. Hence, the objective function to be optimized for large sample size is

$$\Omega = \int_{-\infty}^{\infty} \frac{\varepsilon^{3/2}}{\varepsilon^2 + (x - \mathrm{MFV})^2} f(x)\mathrm{d}x = \max . \tag{A-3}$$

Consider a simplified form of equation (A-3) by fixing the MFV as zero. We derive the function Ω with respect to ε

$$\int_{-\infty}^{\infty} \left\{ \frac{(3/2)\varepsilon^{1/2}\left(\varepsilon^2 + x^2\right) - 2\varepsilon^{5/2}}{\left[\varepsilon^2 + x^2\right]^2} \right\} f(x)\mathrm{d}x = 0 . \tag{A-4}$$

After reorganizing the above equation, we obtain

$$3\int_{-\infty}^{\infty} \frac{x^2}{\left[\varepsilon^2 + x^2\right]^2} f(x)\mathrm{d}x = \varepsilon^2 \int_{-\infty}^{\infty} \frac{1}{\left[\varepsilon^2 + x^2\right]^2} f(x)\mathrm{d}x . \tag{A-5}$$

The square root of dihesion is derived from equation (A-5)

$$\varepsilon = \left(\frac{3\int_{-\infty}^{\infty} \frac{x^2}{\left[\varepsilon^2 + x^2\right]^2} f(x)\mathrm{d}x}{\int_{-\infty}^{\infty} \frac{1}{\left[\varepsilon^2 + x^2\right]^2} f(x)\mathrm{d}x} \right)^{1/2} . \tag{A-6}$$

By substituting the original expression $(x - \mathrm{MFV})^2$ instead of x^2 into equation (A-6), an equivalent formula for equation (11) is given, which is used for practical computations. The optimal value of dihesion is automatically estimated in an iterative process by the subsequent use of equations (10)–(11).

TABLES

Table 1 Zone parameters used in equations (1)–(4) for calculating engineering geophysical sounding logs in Bátaapáti test site.

Zone parameter	Notation	Texture	Clay	Sand	Water	Unit
Gamma-ray intensity	GR	–	11.6	1.45	0	cpm
Density	ρ_b	–	2.10	2.60	1.0	g/cm^3
Neutron-porosity	Φ_N	–	0.23	0	1.0	V/V
Resistivity	R	–	6.50	–	9.0	ohm-m
Cementation exponent	m	1.68	–	–	–	–
Tortuosity factor	a	1.0	–	–	–	–
Saturation exponent	n	2.0	–	–	–	–

Table 2 Rotated factor loadings estimated by the MFV-IRFA procedure in Hole−4.

EGS log	Factor loading	First factor	Second factor
Cone resistance	$L^{(RCPT)}$	–0.41	0.30
Natural gamma-ray	$L^{(GR)}$	0.31	–0.32
Density	$L^{(\rho_b)}$	0.88	–0.09
Neutron-porosity	$L^{(\Phi_N)}$	0.93	–0.08
Resistivity	$L^{(R)}$	–0.97	0.15

FIGURES

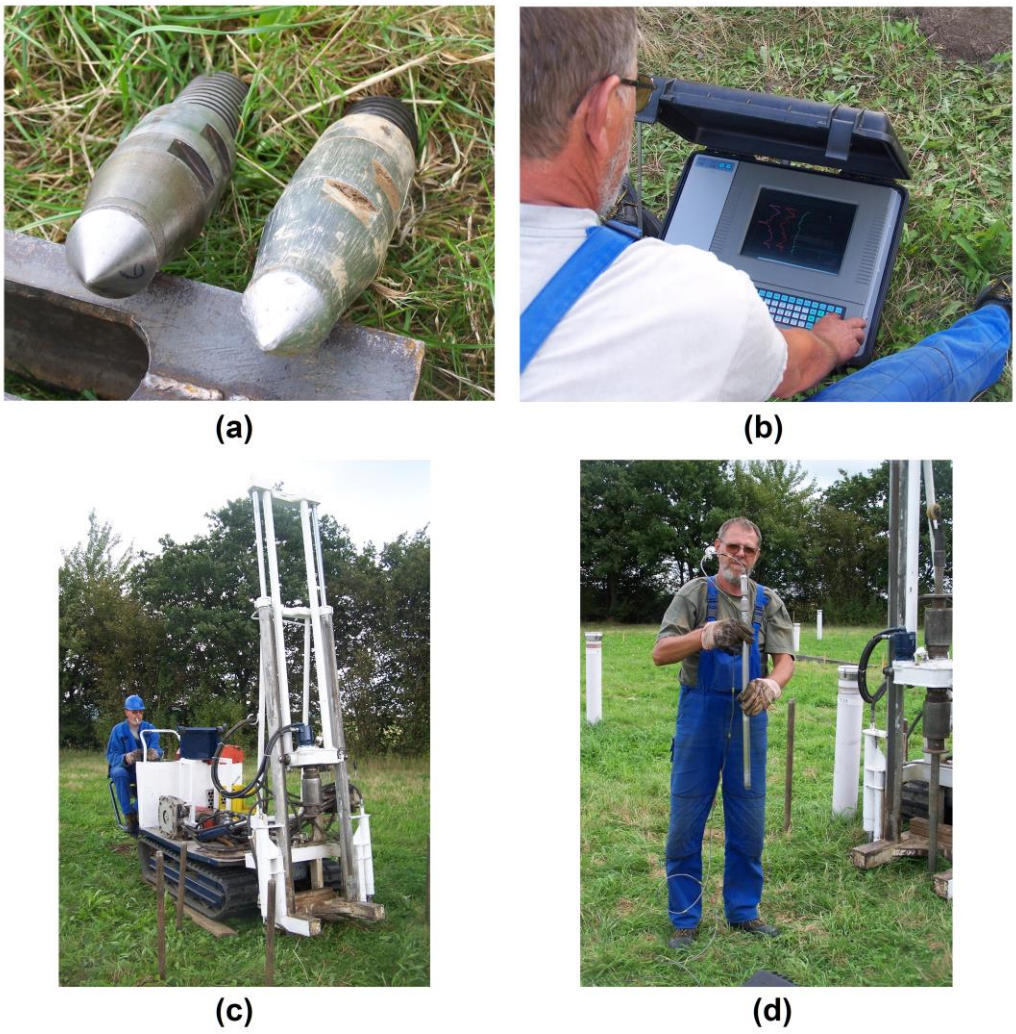

Figure 1 (a) New cone head (on the left) and cone head after sounding in boulder stone (on the right) in Jülich, Germany; (b) surface unit model E1009; (c) engineering geophysical sounding equipment ready for measurement; (d) neutron probe used for estimating porosity and water content.

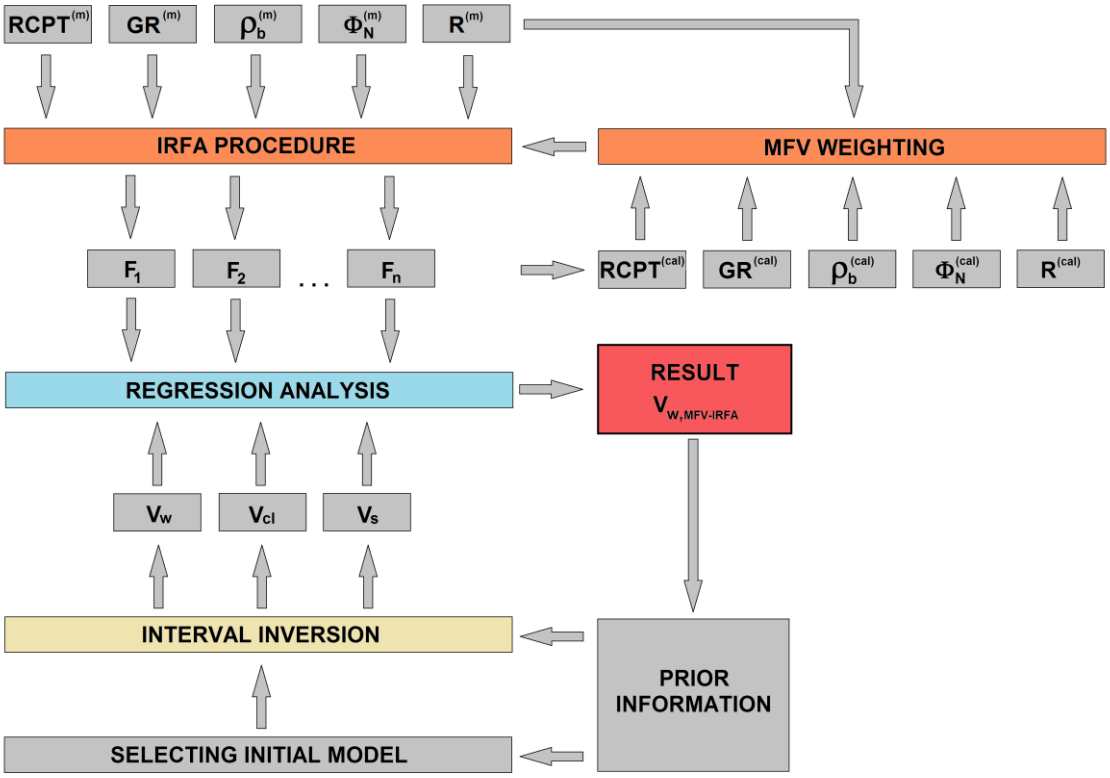

Figure 2 Flowchart of the interval inversion-assisted iteratively re-weighted factor analysis procedure used for estimating petrophysical parameters from engineering geophysical sounding logs.

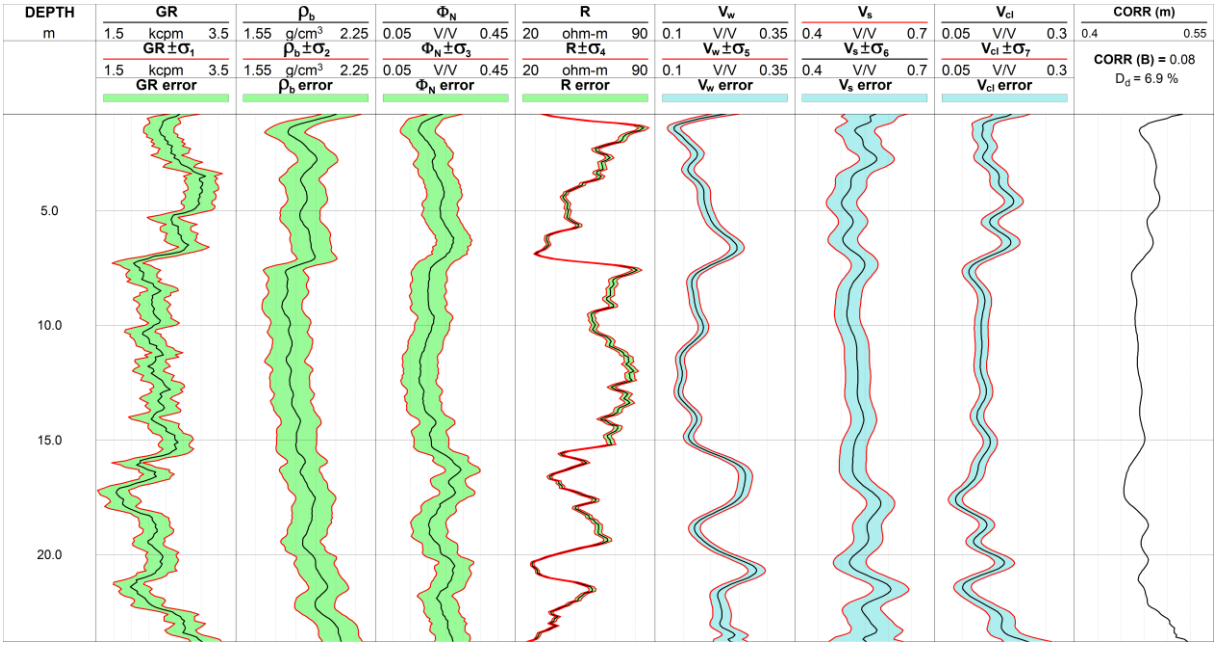

Figure 3 Penetration logs measured in Hole-4, Bátaapáti, South-West Hungary: natural gamma-ray intensity, *GR*, density, ρ_b, neutron-porosity, Φ_N, electric resistivity, *R*; result of interval inversion: water content, V_w, clay volume, V_{cl}, sand volume, V_s, average correlation of model parameters, CORR(**B**), average correlation of petrophysical parameters, CORR(**m**).

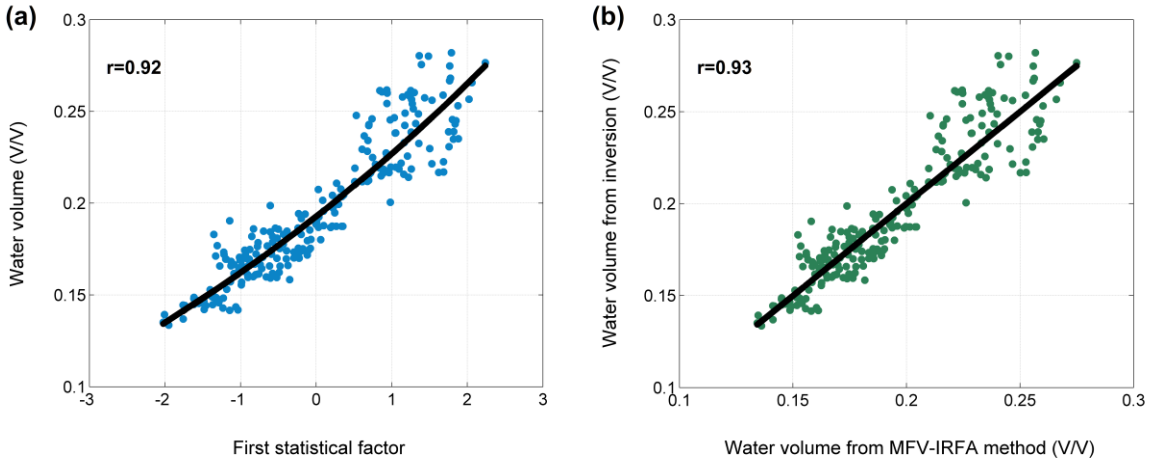

Figure 4 (a) Regression relation between the first factor extracted by MFV-based iteratively re-weighted factor analysis and water saturation estimated by interval inversion procedure in Hole−4; (b) crossplot of water content values estimated separately by factor analysis and interval inversion; *r* denotes the Pearson's correlation coefficient.

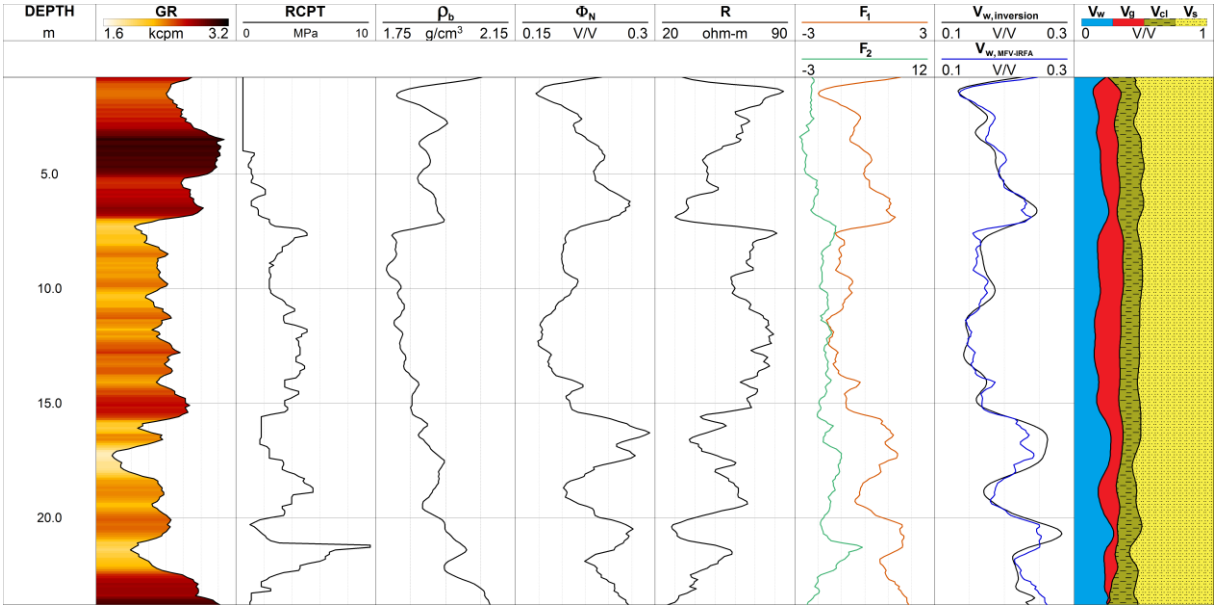

Figure 5 Penetration logs measured in Hole-4, Bátaapáti, South-West Hungary: cone resistance, *RCPT*, natural gamma-ray intensity, *GR*, density, ρ_b, neutron-porosity, Φ_N, electric resistivity, *R*; result of MFV-based iteratively re-weighted factor analysis: first factor, F_1, second factor, F_2, water content, $V_{w,MFV\text{-}IRFA}$; the result of interval inversion: water content, V_w and $V_{w,inversion}$, clay volume, V_{cl}, sand volume, V_s, gas volume, V_g.

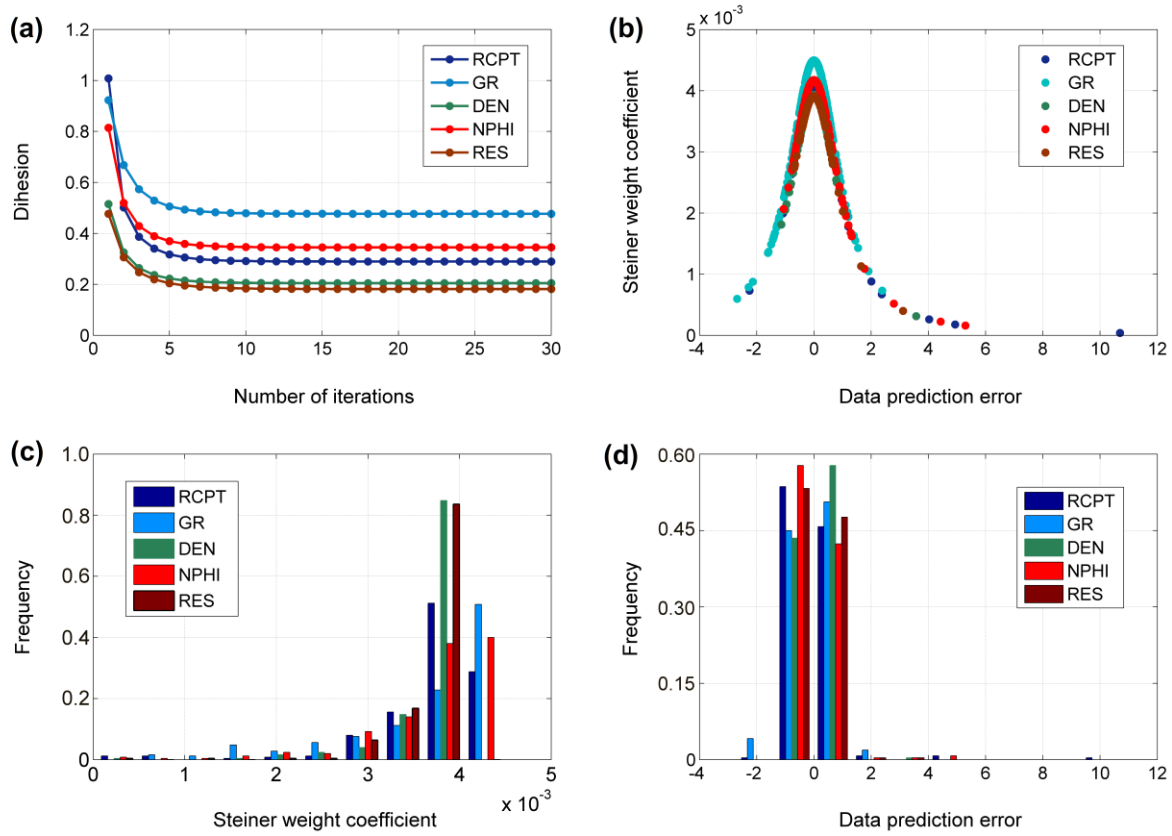

Figure 6 MFV-based weighting process used in the factor analysis of penetration logs observed in Hole–12, *RCPT* is cone resistance, *GR* is natural gamma-ray intensity, *DEN* is density, *NPHI* is neutron-porosity, *RES* is resistivity; (a) dihesion is optimized for each log separately; (b) weight coefficients in the function of the deviation between observed and predicted data; (c) frequency (pieces normalized to the number of sampled depths) plot of Steiner weights; (d) frequency plot of data deviations.

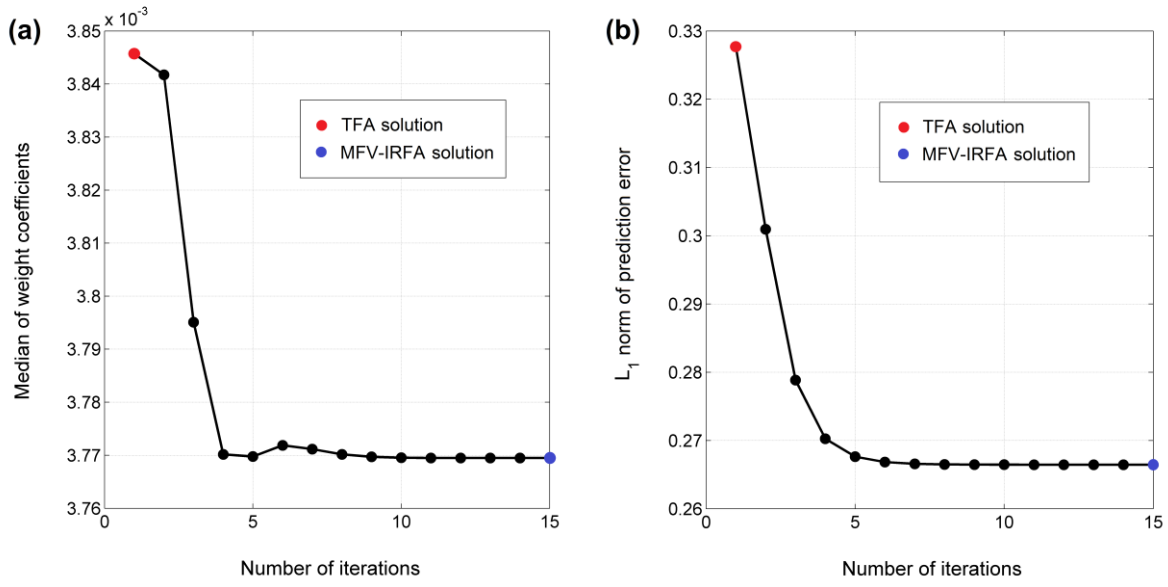

Figure 7 MFV-based iterative factor analysis of penetration logs observed in Hole–12; (a) decrease of Steiner weights during the iterative process; (b) improvement of the overall deviation between the measured and calculated (standardized) EGS data during the iteration process.

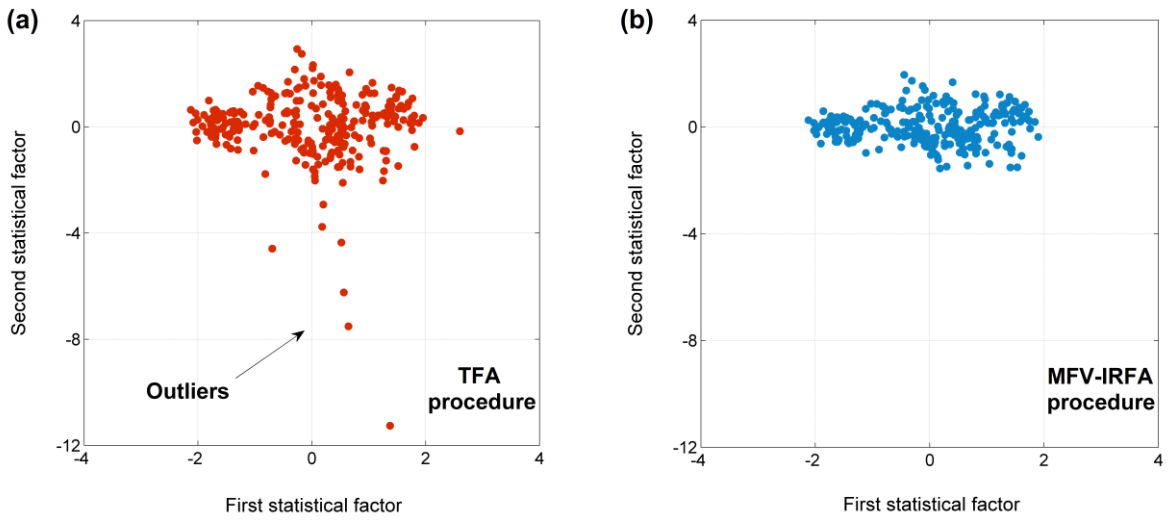

Figure 8 Crossplot of the first and second factor extracted from EGS data observed in Hole–12; (a) result of traditional factor analysis; (c) result of the robust MFV-based iteratively re-weighted factor analysis.

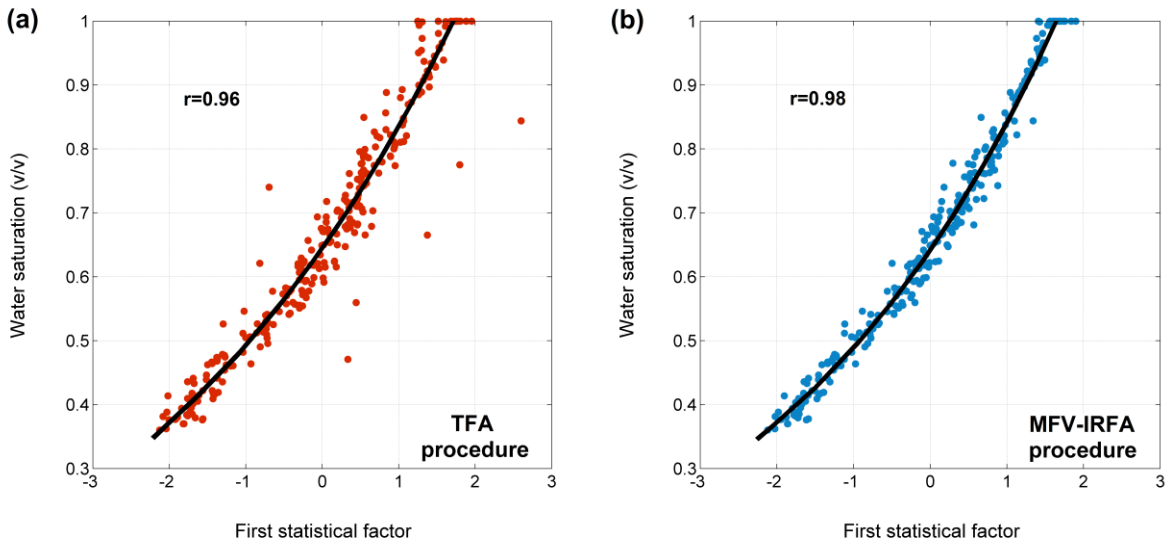

Figure 9 Regression relation between the first factor and water saturation in Hole–12; (a) result of traditional factor analysis; (b) result of MFV-based iteratively re-weighted factor analysis; *r* denotes the Pearson's correlation coefficient.

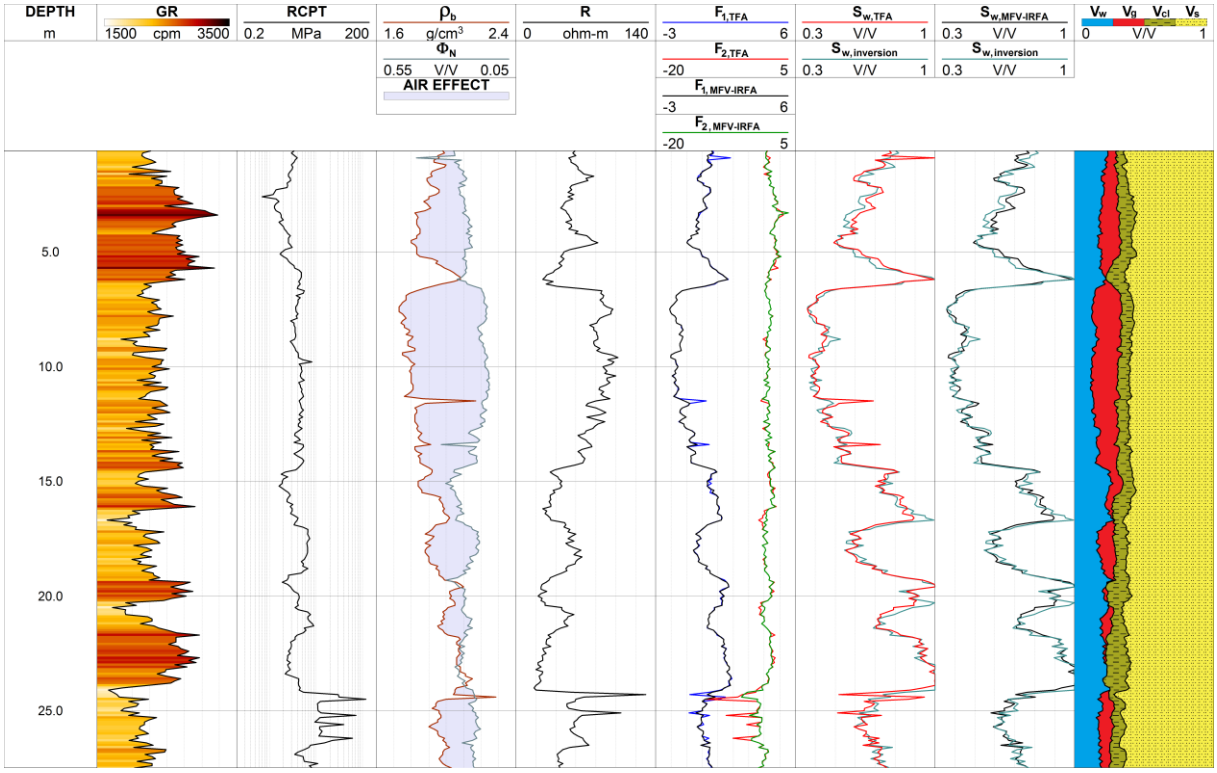

Figure 10 Penetration logs measured in Hole-12, Bátaapáti, South-West Hungary: cone resistance, *RCPT*, natural gamma-ray intensity, *GR*, density, ρ_b, neutron-porosity, Φ_N, electric resistivity, *R*; result of factor analysis: first factor estimated by traditional factor analysis, $F_{1,TFA}$, and MFV-based factor analysis, $F_{1,MFV\text{-}IRFA}$, second factor estimated by traditional factor analysis, $F_{2,TFA}$, and MFV-based factor analysis, $F_{2,MFV\text{-}IRFA}$, water saturation estimated by traditional factor analysis, $S_{w,TFA}$, and MFV-based factor analysis, $S_{w,MFV\text{-}IRFA}$; result of local inversion: water saturation, $S_{w,inversion}$, water content, V_w or $V_{w,inversion}$, clay volume, V_{cl}, sand volume, V_s, gas volume, V_g.

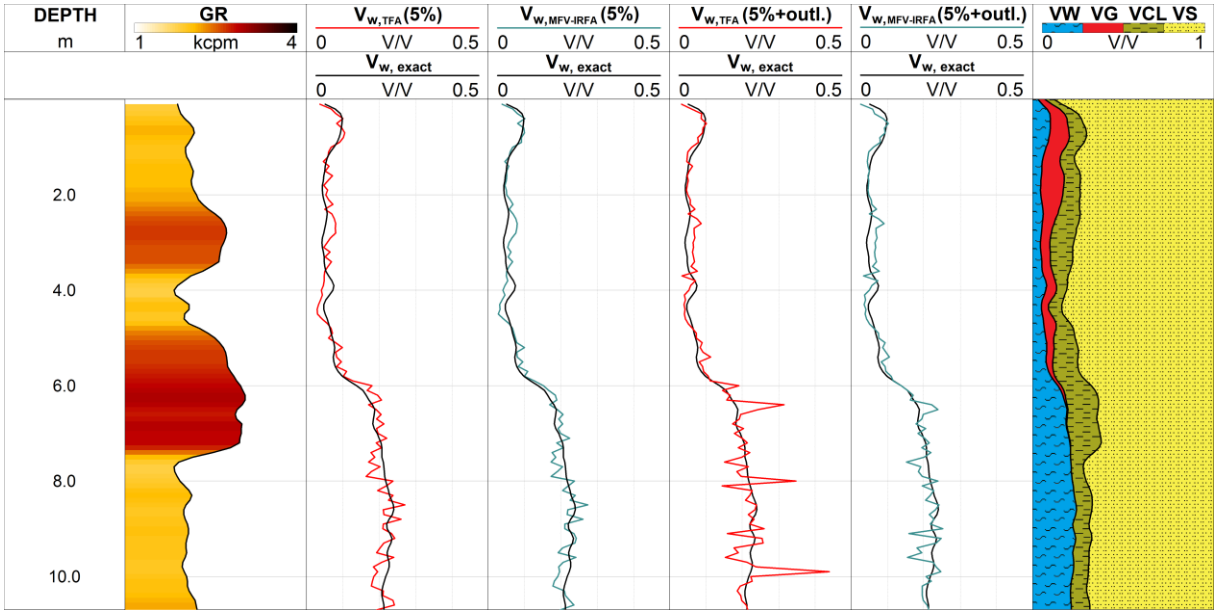

Figure 11 Results of synthetic statistical tests: natural gamma-ray intensity log, *GR*, water volume estimated by traditional factor analysis using 5% Gaussian noise, $V_{w,TFA}$ (5 %) and MFV-based factor analysis, $V_{w,MFV-IRFA}$ (5 %), water volume estimated by traditional factor analysis using 5% Gaussian noise and outliers, $V_{w,TFA}$ (5 %+outl.) and MFV-based factor analysis, $V_{w,MFV-IRFA}$ (5 %+outl.), exactly known water volume, $V_{w,exact}$ and V_w, clay volume, V_{cl}, sand volume, V_s, gas volume, V_g.

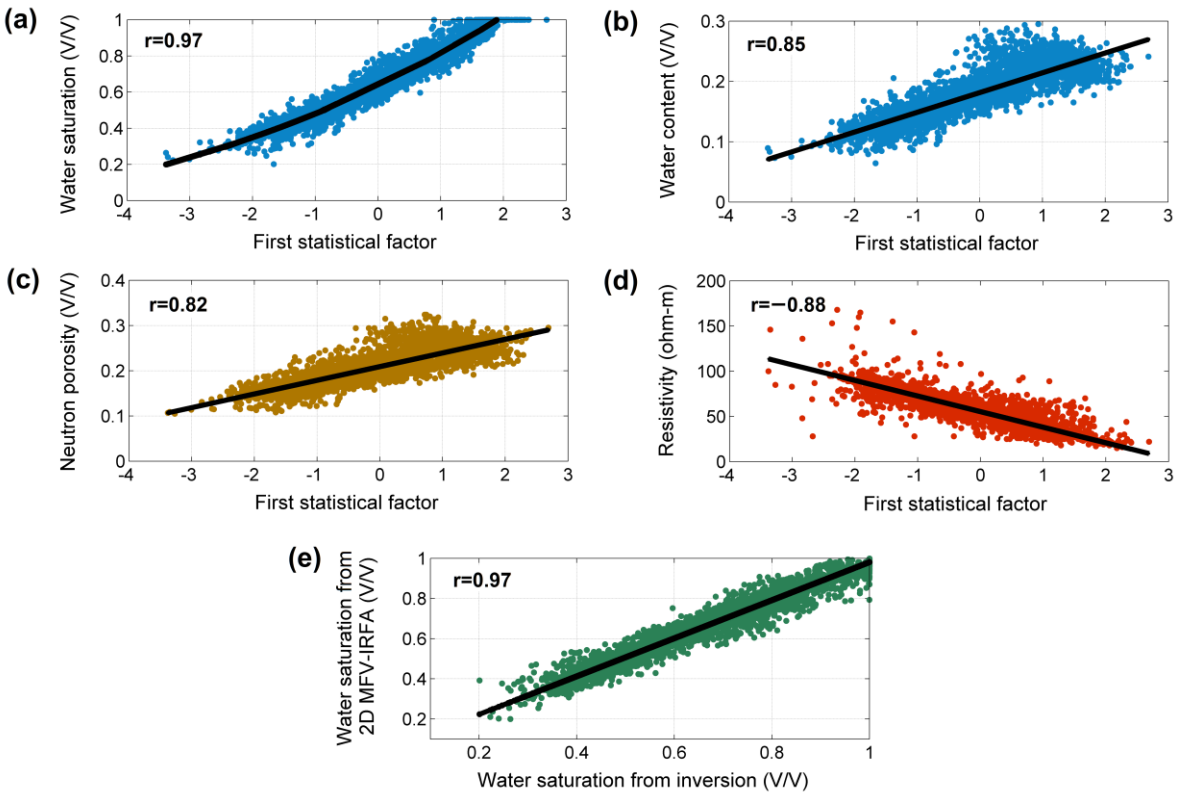

Figure 12 Regression relations between the first factor estimated by 2D MFV-based iterative factor analysis and petrophysical parameters in Holes 1–12, Bátaapáti, South-West Hungary; (a) first factor vs. water saturation (b) vs. water volume (c) vs. observed neutron-porosity (d) vs. observed resistivity; (e) water saturation estimated separately by local inversion and factor analysis; *r* denotes the Pearson's correlation coefficient.

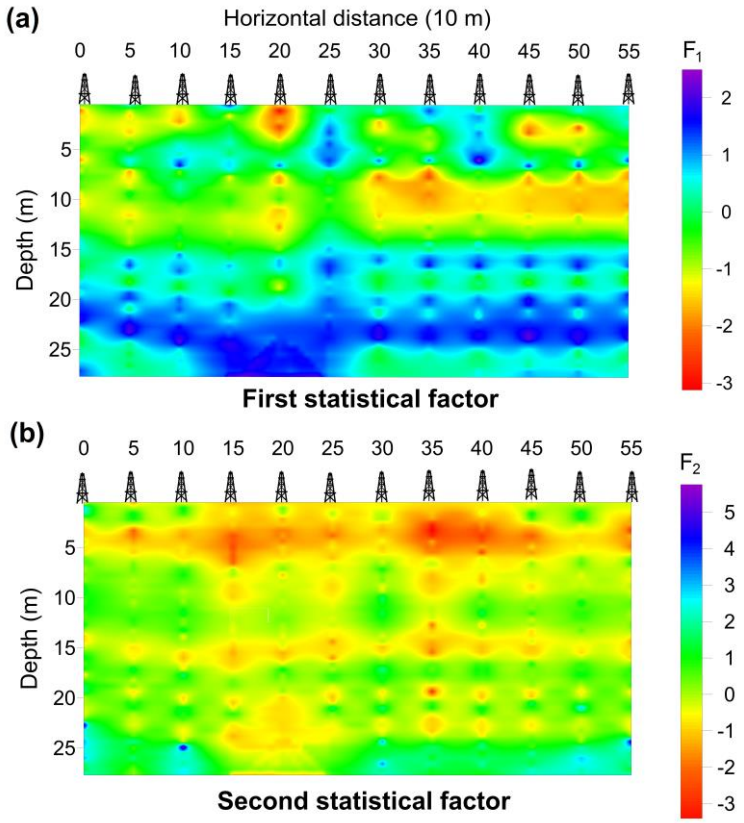

Figure 13 Factor maps estimated by 2D MFV-based iterative factor analysis of penetration logs measured in Holes 1–12, Bátaapáti, South-West Hungary; (a) cross sections of the first statistical factor, F_1 (b) second statistical factor, F_2; borehole symbols indicate the locations of penetration holes drilled along the profile.

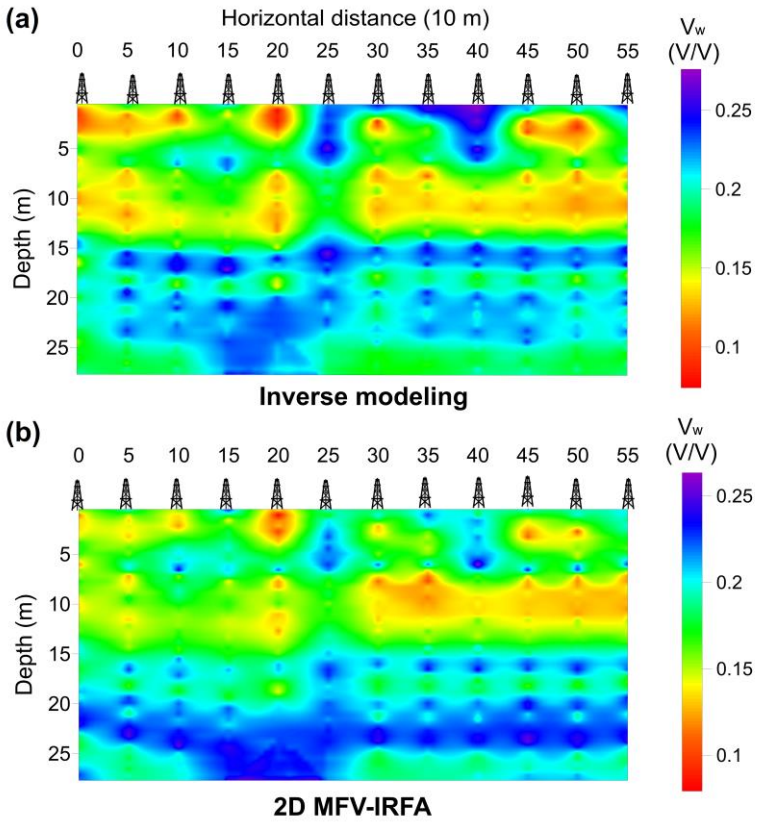

Figure 14 Water volume (V_w) maps estimated by 2D MFV-based iterative factor analysis of penetration logs measured in Holes 1–12, Bátaapáti, South-West Hungary; (a) result of a set of 1D local inverse modelling (b) 2D MFV-based iterative factor analysis; borehole symbols indicate the locations of penetration holes drilled along the profile.