Factor analysis of engineering geophysical sounding data for water saturation estimation in shallow formations

Norbert Péter Szabó[1], Mihály Dobróka[1,2], and Dezső Drahos[3]

Running Head: Factor analysis: saturation estimation

[1]University of Miskolc, Department of Geophysics, Miskolc, Hungary. E-mail: norbert.szabo.phd@gmail.com

[2]MTA-ME Applied Geoscience Research Group, Miskolc, Hungary. E-mail: dobroka@uni-miskolc.hu.

[3]Eötvös Loránd University, Department of Geophysics and Space Science, Budapest, Hungary. E-mail: dezso.drahos@gmail.com.

Date of first submission: 28 July 2011.

Date of final submission: 28 January 2012.

ABSTRACT

The engineering geophysical sounding method has been applied for in-situ investigation of unconsolidated near-surface formations since the 1990’s. In this study, we offer an alternative to geophysical inversion methods for data processing. Factor analysis is applied to engineering geophysical sounding and cone penetration test data to extract information on water saturation of the shallow region. A linear correlation is indicated between water saturation and one of the new variables derived by factor analysis. We suggest a general formula for estimating water saturation and apply it on data measured in twelve shallow boreholes situated over a nuclear waste disposal site. We extend the statistical procedure to estimate the spatial distribution of water (air) saturation between the boreholes. We compare the statistical results to inversion estimations. The study shows that factor analysis gives a quick and reliable solution for the assessment of hydrogeological conditions in shallow prospecting sites.

INTRODUCTION

Detailed information about the shallow subsurface can be gained by surface geophysical methods (Sharma, 1997), which are routinely used, for example, in water prospecting, in delineating contaminated areas, and in planning waste disposal sites. When the subsoil is soft enough for penetration (i.e., there is no hard rock or concrete), cone penetration tests (CPT) can be effectively used to collect in-situ information about the subsoil. During the test a cone-shaped tip is pushed into the ground while mechanical parameters such as cone tip stress and sleeve friction are measured to evaluate geotechnical properties of soils such as soil type and density, stress conditions and shear strength. Several studies deal with the identification of soil types and stratigraphical profiling using CPT data (Douglas and Olsen, 1981; Robertson, 1990; Campanella and Robertson, 1991).

Recently CPT and surface geophysical methods are combined for more reliable site characterization (Wyatt et al., 1996; Clement et al., 1997; Cardimona et al., 1998; Chang and Basnett, 1999; Ghose and Goudswaard, 2004). Subsurface information can be further increased with better downhole measurement capabilities. For this purpose, the CPT tools are equipped with different measuring elements, such as geophones that enable shear-wave velocity measurements (Lech et al., 2008). A special type of CPT tool contains such probes attached to the penetration tube, which can measure the same physical parameters as open-hole logging instruments (Serra, 1984). Beside the different depths of investigation and measuring environments, a further difference between the two configurations is that probes applied in a borehole are separated from the rock environment by drilling mud, but in case of penetration soundings it is a steel tube that isolates the soil and the probe. In the latter case, data are transferred through the rods pushed into the ground. This measurement methodology and data processing technology is called engineering geophysical sounding (EGS), which was developed in the Loránd Eötvös Geophysical Institute of Hungary. The measurement principles were first published by Fejes and Jósa (1990), which was followed by several case studies and applications (Draskovits and Fejes, 1994; Draskovits et al., 1995, Magyar et al., 1995; Csapó et al., 2003; Drahos, 2005; Tillmann et al., 2008, Nyári et al., 2010).

Interpretation of EGS data acquired from nuclear and electrical sensors is performed by the same data processing methods used in well log analysis. Well-logging data are normally evaluated in an inversion procedure, when data acquired by different probes in a certain depth are processed together to determine the petrophysical properties of the formation only to that depth. The inversion method assumes a known relationship, called the probe-response function, between the measurement results and the petrophysical parameters. The solution of the inverse problem is determined by fitting theoretical data calculated by the response equations to the measured ones. The theory of well-logging inversion methods is detailed in Mayer and Sibbit (1980), Alberty and Hashmy (1984) and Ball et al. (1987); a novel inversion methodology was suggested by Szabó (2004), Dobróka and Szabó (2005), and Dobróka et al. (2009). By analogy with well logging, a proper inversion method for processing

EGS data was developed by Drahos (2005) to determine important physical parameters of soils (i.e., porosity, water/air saturation, shale volume), along with estimation errors. Local inversion, however, has some limitations. The number of relevant data measured by different probes at a point is only slightly more than the number of unknowns, which sets a limit to the accuracy and reliability of the parameter estimation. Separate inversion procedures are run point by point in depth to resolve the entire soil structure. Recently the EGS data are measured in several boreholes in the same area for the purpose of 2D or 3D soil mapping. The processing of the entire data set cannot be solved by a single inversion procedure.

We use a statistical tool as an alternative for interpretation of EGS data. Preliminary statistical research showed strong relationships between EGS measurements and some physical parameters of soils (Ocsenás, 1996). Here, we apply factor analysis to explore the possible correlation between the derived variables (factors) and petrophysical parameters of soils. The workflow of the proposed statistical procedure is in Figure 1. Lawley and Maxwell (1962) present the basic principles of factor analysis, which can enhance the essential information in large-scale multidimensional data sets, extract background variables not directly measurable or solve ambiguity problems (Duval, 1977; Silva and Hohmann, 1983; Pires and Harthill, 1989; Fraiha and Silva, 1994; Buoro and Silva, 1994; Asfahani et al., 2005; Kazmierczuk and Jarzyna, 2006; Grana et al., 2011).

In an earlier study, Szabó (2011) found strong correlation between the shale volume of deep-seated sedimentary rocks and one of the variables derived by factor analysis of well-logging data. Based on this empirical relationship, a general formula was introduced from which it can be inferred that shale volume caused the highest sensitivity to well-logging measurements. With EGS observations, the presence of water and structural or textural properties of soils have the greatest impact on probe responses. We demonstrate that there is a linear correlation between a derived factor and water (air) saturation in shallow structures. We suggest an independent water-saturation estimate based on this relationship, which can be used either in one borehole or more boreholes. In this study, EGS data originated from several boreholes situated along a profile are also processed in one (joint) statistical procedure. The 2D statistical interpretation of EGS data uses a larger statistical sample, which may improve the reliability of estimation results. The spatial distribution of water saturation obtained by 1D and 2D factor analysis is compared to estimates made by 1D inverse modeling. Some missing measurements in certain intervals along the borehole can be substituted effectively by the synthetic data generated by the factor analysis of rest of the data collected at the site.

THEORETICAL OVERVIEW

Applied petrophysical model

Soil consists of solid particles of different sizes, from very fine clay to sand and gravel, as well as other mineral constituents of small amounts, and pore space between the particles, which is occupied by variable amounts of water or gas (usually air). The solid part can be divided into several groups according to their average grain size. Generally two components are taken into account: the coarse grain component (sand) and the fine one (clay). Thus, the soil model consists of four components such as sand, clay, water and gas, of which volumetric ratios are V_s, V_{cl}, V_w, V_g, respectively, where

$$V_s + V_{cl} + V_w + V_g = 1\,. \tag{1}$$

The porosity Φ (pore volume/total soil volume) is the sum of V_w and V_g

$$\Phi = V_w + V_g\,. \tag{2}$$

The water saturation (S_w), which is the relative water content in the pore space is defined as

$$S_w = \frac{V_w}{\Phi}\,. \tag{3}$$

The gas saturation *S_g=1-S_w*. In the water saturated medium, the following response equations can be used for calculating EGS data (Drahos, 2005)

$$GR = V_{cl} GR_{cl} + V_s GR_s \,, \tag{4}$$

$$DEN = V_w \rho_w + V_{cl} \rho_{cl} + V_s \rho_s \,, \tag{5}$$

$$NPHI = V_w \Phi_{N,w} + V_{cl} \Phi_{N,cl} + V_s \Phi_{N,s} \,, \tag{6}$$

$$RES = \frac{a}{(\mathit{\Phi} + V_{cl})^m} \frac{R_p}{\left(\frac{V_w + V_{cl}}{V_w + V_{cl} + V_g} \right)^n} \,, \tag{7}$$

where *GR* denotes gamma ray intensity (cpm), *DEN* is density (g/cm^3), *NPHI* is neutron porosity (v/v) and *RES* is resistivity (ohm-m). The constituents have their own physical parameters, which are assumed to be constant. GR_{cl} and GR_s denote gamma ray intensity in clay and sand, respectively. The measured density is determined by different values in water (ρ_w), in clay (ρ_{cl}) and sand (ρ_s). The neutron porosities have also different values in water ($\mathit{\Phi}_{N,w}$), in clay ($\mathit{\Phi}_{N,cl}$) and sand ($\mathit{\Phi}_{N,s}$). In equations 5-6 the density of gas (ρ_g) and the neutron porosity of gas ($\mathit{\Phi}_{N,g}$=0) are neglected, and the porosity terms become $\mathit{\Phi}(S_w\rho_w+ S_g\rho_g)=V_w\rho_w$ and $\mathit{\Phi}(S_w\mathit{\Phi}_{N,w}+S_g\mathit{\Phi}_{N,g})=V_w\mathit{\Phi}_{N,w}$. The resistivity is calculated by the De Witte's approximation, where free water and clay particles dispersed in the pore space conduct an electrical current like a mixture of electrolytes (De Witte, 1955). The mixed electrolyte including air fills up the generalized pore space as $\Phi=V_w+V_g+V_{cl}$. This porosity term is combined with the Archie's equation for the case of unsaturated soils (Archie, 1942). In equation 7 the resistivity of the mixed electrolyte R_p is computed as

$$R_p=\left(\frac{q}{R_{cl}}+\frac{1-q}{R_w}\right)^{-1}, \tag{8}$$

where R_{cl} and R_w are clay and pore water resistivities, respectively, and $q=V_{cl}/(V_{cl}+V_w)$. Equation 7 contains three local specific constants which express the textural properties of formations: cementation exponent (*m*), saturation exponent (*n*), and tortuosity factor (*a*). Normally treated as constants, they can also be determined by using a special inversion method suggested by Dobróka and Szabó (2011).

Interpretation of EGS data gives estimates for parameters of the petrophysical model. In the inverse problem *GR, DEN, NPHI, RES* data are used to determine V_s, V_{cl}, V_w, V_g at each given depth. The evaluation is done by fitting the theoretical tool responses to measured data point-by-point. Drahos (2005) implemented the weighted sum of squares method (Menke, 1984) for solving the inverse problem, which used a priori known data variances for weighting of individual data sets. The advantage of this kind of inversion is that the unknown parameters of the soil structure can be determined with their estimation errors. Thus, soil layers can be recognized on the resultant logs and the measures of accuracy and reliability of inversion results are also quantified.

Factor analysis of EGS data

We formulate the problem by organizing CPT and EGS data into the following data matrix

$$\mathbf{D} = \begin{pmatrix} RCPT_1 & GR_1 & DEN_1 & NPHI_1 & RES_1 \\ RCPT_2 & GR_2 & DEN_2 & NPHI_2 & RES_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ RCPT_i & GR_i & DEN_i & NPHI_i & RES_i \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ RCPT_N & GR_N & DEN_N & NPHI_N & RES_N \end{pmatrix}, \qquad (9)$$

where *RCPT* denotes the cone tip stress (MPa) and the other data types are defined in equations 4-7. In equation 9, the D_{ik} element at *i*-th row and *k*-th column of **D** represents the datum collected by the *k*-th probe in the *i*-th depth of the borehole. The size of data-matrix **D** is *N*-by-*M*, where *N* is the total number of observation points along the logged interval and *M* is the number of measurement types (original variables).

Factor analysis decomposes **D** as

$$\mathbf{D} = \mathbf{F}\mathbf{L}^{\mathrm{T}} + \mathbf{E}, \qquad (10)$$

where **F** denotes the *N*-by-*a* matrix of factor scores, **L** is the *M*-by-*a* matrix of factor loadings and **E** is the *N*-by-*M* matrix of residuals (superscript *T* denotes matrix transpose). The number of factors (*a*) should be less than that of the original ones (*M*). The *j*-th column of **F** represents the values of the *j*-th factor computed for different depth points and **L** contains the weights of the original (data) variables on the derived factors. The correlation matrix **R** of the standardized data variables can be derived from equation 10 demanding that factors are linearly independent $\mathbf{F}^{\mathrm{T}}\mathbf{F}/N = \mathbf{I}$ (where **I** is the identity matrix)

$$\mathbf{R} = \mathbf{D}^{\mathrm{T}}\mathbf{D}/\mathrm{N} = \mathbf{L}\mathbf{L}^{\mathrm{T}} + \boldsymbol{\Psi}, \qquad (11)$$

where $\boldsymbol{\Psi} = \mathbf{E}^{\mathrm{T}}\mathbf{E}/N$ is the diagonal matrix of specific variances, independent of the common factors.

Jöreskog (2007) suggested a non-iterative approximate solution for the determination of factor loadings. Let **S** be the sample covariance matrix of the standardized data variables and consider the following derived quantity,

$$\mathbf{S}^{*} = \left(diag\mathbf{S}^{-1}\right)^{1/2}\mathbf{S}\left(diag\mathbf{S}^{-1}\right)^{1/2}. \qquad (12)$$

By calculating the eigenvalues ($\lambda_1,\lambda_2,\ldots,\lambda_M$) as well as the first *a* eigenvectors ($\omega_1,\omega_2,\ldots,\omega_a$) of $\mathbf{S}^{*}$ and sorting both quantities in descending order, a solution for the factor loadings is provided by

$$\mathbf{L}=\left(diag\mathbf{S}^{-1}\right)^{-1/2}\mathbf{\Omega}_a\left(\mathbf{\Gamma}_a-\theta\mathbf{I}\right)^{1/2}\mathbf{U}, \qquad (13)$$

where **U** is an arbitrary a-by-a orthogonal matrix, $\mathbf{\Gamma}_a$=diag($\lambda_1,\lambda_2,\ldots,\lambda_a$) and $\mathbf{\Omega}_a$=($\omega_1,\omega_2,\ldots,\omega_a$). The parameter θ, computed as

$$\theta=\frac{1}{M-a}\left(\lambda_{a+1}+\lambda_{a+2}+\ldots+\lambda_M\right), \qquad (14)$$

specifies the smallest number of factors when $\theta<1$.

The factor scores can be estimated by the maximum likelihood method, where the following log-likelihood function is optimized

$$P=-\left(\mathbf{D}-\mathbf{FL}^{\mathrm{T}}\right)^{\mathrm{T}}\mathbf{\Psi}^{-1}\left(\mathbf{D}-\mathbf{FL}^{\mathrm{T}}\right)=\max. \qquad (15)$$

The computation of factor scores is based on the condition $\partial\mathrm{P}/\partial\mathrm{F}=0$. An unbiased solution to equation 15 assuming linearity was suggested by Bartlett (1937), with factor scores estimated by

$$\mathbf{F}=\left(\mathbf{L}^{\mathrm{T}}\mathbf{\Psi}^{-1}\mathbf{L}\right)^{-1}\mathbf{L}^{\mathrm{T}}\mathbf{\Psi}^{-1}\mathbf{D}. \qquad (16)$$

For easier interpretation of factor variables, an orthogonal transformation of factor loadings is normally needed (Lawley and Maxwell, 1962). In this research, we used the varimax algorithm suggested by Kaiser (1958).

Estimation of water saturation

The purpose of our study is to find a relationship between the new variables (factors) obtained by factor analysis of EGS data and petrophysical parameters of shallow structures. The factor analysis gives an estimate for the factor scores. In equation 6 the columns of **F** represent new (uncorrelated) variables. Based on regression tests, we assume a linear correlation between the first factor (i.e., the first column of **F**) and water saturation defined in equation 3. Following Szabó (2011), we introduce the factor index i_F at a given depth as

$$i_{F_1} = \frac{F_1 - F_{1,min}}{F_{1,max} - F_{1,min}} , \tag{17}$$

where F_1 is the factor score computed at a depth point, while $F_{1,min}$ and $F_{1,max}$ are the minimum and maximum value of the first factor in the processed interval, respectively. We suggest a simple formula to compute water saturation directly from the factor scores as

$$S_w = i_{F_1} , \tag{18}$$

which is independent of the prospecting area, because equation 18 does not contain any additional local specific constant ($F_{1,min}$ and $F_{1,max}$ are derived from the factor analysis results). The above formula is based on the analogy to a frequently used shale volume estimation method using a simplified relationship between gamma ray index and shale content (Asquith and Krygowski, 2004). Our method also represents a simplified solution for computing water saturation from the factor index. If it is necessary, the gas saturation can also be computed from the factor scores by using $S_g = 1 - i_{F_1}$. It was shown in Szabó (2011) that the statistical method gave consistent results in different measurement areas.

For comparing water saturation estimations derived from factor analysis to that of a different method (e.g., inverse modeling), we introduce the model distance as a measure of fitting for the 1D case as

$$d^{(1D)} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(S_{w,i}^{(inv)} - S_{w,i}^{(F_1)}\right)^2}\ , \tag{19}$$

and for the 2D case as

$$d^{(2D)} = \sqrt{\frac{1}{QN^*}\sum_{q=1}^{Q}\sum_{i=1}^{N_q}\left(S_{w,qi}^{(inv)} - S_{w,qi}^{(F_1)}\right)^2}\ , \tag{20}$$

where Q is the number of boreholes and $N^*=N_1+N_2+\ldots+N_Q$ is the total number of measurement points to which water saturation values are estimated.

An assumption of the maximum likelihood method is that data in equation 9 are required to follow an M-dimensional normal distribution. We use the following statistics for checking the normality of data distribution. The skewness of data measured by the *k*-th probe is

$$\mu^{(k)} = \frac{\frac{1}{N}\sum_{i=1}^{N}\left(D_{ik} - \overline{D}_k\right)^3}{\left(\frac{1}{N}\sum_{i=1}^{N}\left(D_{ik} - \overline{D}_k\right)^2\right)^{3/2}}, \qquad (21)$$

where $\overline{D}_k$ is the mean of data measured by the *k*-th probe. For the case of $\mu \approx 0$ the probability density function of the *k*-th observed variable is symmetrical and data follow normal distribution. The kurtosis of the *k*-th data set is

$$\gamma^{(k)} = \frac{\frac{1}{N}\sum_{i=1}^{N}\left(D_{ik} - \overline{D}_k\right)^4}{\left(\frac{1}{N}\sum_{i=1}^{N}\left(D_{ik} - \overline{D}_k\right)^2\right)^{2}} - 3, \qquad (22)$$

which measures the peakedness of the probability density function. For the case of $\gamma \approx 0$ the data distribution is Gaussian type. Regression tests are used for the characterization of dependence between certain variables. For the linear case the Pearson’s correlation coefficient (*r*) can be applied for measuring the degree of correlation between two variables (Isaaks and Srivastava, 1989). In the paper, factor scores are also correlated with such data, where the relationship becomes nonlinear. For nonlinear relationships, the rank correlation coefficient is preferably used (Spearman, 1904).

FIELD RESULTS

The study area

Figure 2 shows the location of Bátaapáti (Üveghuta) Site, South-West Hungary. EGS and CPT surveys were made in twelve penetration holes (PH1-PH12) along a profile, where the first borehole (PH1) was drilled at x=0 m and the last one (PH12) at x=550 m. The geological structure was a loess complex deposited on a partly disintegrated granite basement. The thickness of the loess cover was approx. 40–60 m and the water level was mainly at the top of the granite. The aim of the measurement was to detect neotectonic events by tracing horizons connected with soils and explore local hydrogeological conditions. The measurement formed part of a complex geophysical survey conducted for establishing a repository for the disposal of low and intermediate level radioactive waste (Vértesy et al., 2004).

Preliminary inversion results

The EGS data set collected in penetration holes consisted of *GR*, *DEN*, *NPHI*, *RES* logs (equations 4-8). The petrophysical parameters of the formation constituents were selected on the basis of field experiences and a priori knowledge acquired by former geophysical measurements. The optimal constants appearing in equations 4-8 were chosen as follows: GR_s=1400 cpm, GR_{cl}=9900 cpm, ρ_s=2.5 g/cm^3, ρ_{cl}=2.0 g/cm^3, ρ_w=1 g/cm^3, $\Phi_{N,s}$=0, $\Phi_{N,cl}$=0.2, $\Phi_{N,w}$=1, R_w=12 ohm-m, R_{cl}=4 ohm-m, m=1.7, n=2 and a=1. Using these constants, a set of inverse modeling procedures was performed independently in PH1-PH12; detailed results can be found in Drahos (2005). In this study, we compare inversion estimates with results obtained from statistic analysis.

Application to 1D model

Since similar results were obtained in the twelve holes, we show the statistical processing of the data set measured in PH1 as a typical case. The logs of measured data (*RCPT, GR, DEN, NPHI, RES)* are in Figure 3, which provide the original variables for factor analysis. The maximal depth of penetration was 27.6 m, and data were sampled in every 0.1 m (totally 1,360 data). The average of Pearson's correlation coefficients between the measured variables was 0.10. The number of original variables was reduced to two uncorrelated factors. The number of factors was specified by equation 14 (it was also computed that these two factors had explained the 90% of the variance of original variables). In Table 1, the values of rotated factor loadings computed by equation 13 can be seen for each borehole, separately. It was inferred that logs sensitive mainly to the content of pore space such as *DEN, NPHI, RES* weighted the most on Factor 1. Considering the sign and magnitude of factor loadings, results were consistent in all boreholes. The factor scores were determined by equation 16, which were plotted in Figure 4.

Regression tests showed that the first factor correlates strongly (almost completely) with water saturation estimated by inverse modeling. It can be seen in Figure 5 that the correlation coefficients between Factor 1 and water saturation as well as its derived quantities (i.e., water and air content of the pore space) are high in every case. In Table 2 the correlation coefficients for the same quantities in the twelve boreholes were listed, which confirmed our statement. In Figure 6 the water saturation logs estimated separately by 1D inversion and factor analysis can be compared. The model misfit based on equation 19 was 5.7%, which shows that both independent solutions are essentially the same. The biggest difference was obtained in the first few meters, where a very loose soil was situated. The log of Factor 1 also informed us about the air volume of the pore space, which was approx. 75% by 25% irreducible water saturation in the limiting case.

Application to 2D model

We extended our factor analysis algorithm to a 2D model geometry. All EGS and CPT data from different boreholes are gathered in one data matrix as

$$\mathbf{D} = \begin{pmatrix} RCPT_{1} & GR_{1} & DEN_{1} & NPHI_{1} & RES_{1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ RCPT_{N_1} & GR_{N_1} & DEN_{N_1} & NPHI_{N_1} & RES_{N_1} \\ RCPT_{N_1+1} & GR_{N_1+1} & DEN_{N_1+1} & NPHI_{N_1+1} & RES_{N_1+1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ RCPT_{N_1+N_2} & GR_{N_1+N_2} & DEN_{N_1+N_2} & NPHI_{N_1+N_2} & RES_{N_1+N_2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ RCPT_{\Sigma N} & GR_{\Sigma N} & DEN_{\Sigma N} & NPHI_{\Sigma N} & RES_{\Sigma N} \end{pmatrix}, \tag{25}$$

where N_1,N_2, …,N_{12} are the numbers of measurement points in PH1-PH12, respectively, and ΣN denotes the number of measurement points in all boreholes. The total number of measured data is $5 \cdot \Sigma N$ with the five data variables in equation 25. We followed the same procedure for the decomposition of **D** in equation 25 as presented in equations 10-11. We adopted equations 12-16 to compute factor loadings and scores for the 2D model.

We processed 15,500 in-situ data simultaneously in the 2D statistical procedure. The empirical probability density functions showed that EGS data follow near-Gaussian distributions for an optimal maximum likelihood solution (Figure 7). It was confirmed by near-zero values of skewness and kurtosis defined in equations 21 and 22 (μ was 0.18 for *GR*, 0.2 for *DEN*, 0.08 for *NPHI* and 0.45 for *RES* data; γ was 0.07 for *GR*, -0.55 for *DEN*, -0.62 for *NPHI* and 0.77 for *RES* data). The *RCPT* data were farthest from a normal distribution.

The mean correlation between data variables was 0.09. We computed two factors as earlier. The CPU time of data processing was 20 s on a workstation with a quad-core processor. The rotated factor loadings in Table 3 are the same magnitude as in the 1D result.

In Figure 8 the regression relationship between the first factor and water saturation (water and gas content) is shown. Comparing 1D to 2D results, a small improvement in the correlation coefficient between Factor 1 and water saturation can be seen, because of the enlarged data set of the joint interpretation. We prepared a correlated section of water saturation values estimated by 1D inversion for the comparative study. In Figure 9, the spatial distribution of water saturation obtained by a set of 1D inversion procedures and 2D factor analysis were illustrated in separate sections, where five soil layers with different water/air content were distinguished by both methods. The distance between the two models was 5.6% (i.e., slightly better than in 1D case). It was concluded that the statistical method resolved the structure properly in a quick procedure.

Replacement of neutron measurement

We suggest a new technique, which is based on the strong correlation existing between Factor 1 and neutron porosity data (Tables 1 and 3). This relationship remains even when the neutron porosity is not on the list of input variables of factor analysis. Neutron measurements are relatively expensive therefore we tried to resolve the connection for making it possible to generate synthetic data directly from the factor scores.

In our experiment, we removed *NPHI* data from the factor analysis and computed only one factor using *RCPT, GR, DEN, RES* data measured from all boreholes. As a result, we found an exponential relationship between the resultant factor and *NPHI* data (Figure 10). The regression function depends on the area, where the environmental and probe type corrections on neutron data are useful to be taken into consideration. The rank correlation coefficient for this particular case was -0.75. We computed neutron porosity data for all boreholes by the regression function. The rms error between the measured and calculated neutron porosity data was 14%. In order to test the effect of the replacement of neutron data on the estimated water saturation, we reprocessed the data set measured in PH1, with a difference according to which measured neutron porosity data were exchanged to that of the synthetic ones computed from Factor 1. In Figure 11, the measured and calculated (synthetic) neutron porosity logs and the results of factor analysis can be seen. The distance between the models estimated separately by factor analysis and inverse modeling was 7.1%, which indicated an acceptable result. The experiment showed unequivocally that on depth intervals where neutron measurements were not available, the neutron porosity data were possible to be substituted by synthetic data generated by the factor analysis of the other data collected at the borehole.

CONCLUSIONS

A statistical method based on factor analysis gives an alternative for processing EGS data to extract information about water saturation of shallow structures. The results agree with inversion estimations. The statistical method uses not only EGS but also CPT data to increase the reliability of estimates. The method is non-iterative (a statistical transformation), and does not require response functions and local specific constants in the factor analysis phase. The CPU time of the process is also favourable and does not exceed that of the inversion procedures. The fundamental equation used for the determination of water saturation is equation 18. This formula does not depend on the actual geological setting and gives accurate and reliable estimation by assuming a linear approximation. The method is also applicable to estimate irreducible water saturation, which is a delicate parameter in borehole geophysical problems. Another advantage of the method is that neutron porosity data of some intervals where measurements are not available can be substituted by the same type of synthetic data generated by factor analysis. According to our opinion the accuracy of synthetic data may be improved in an iterative procedure based on the subsequent application of regression and factor analysis. This result is dependent on the specific conditions of the prospecting area and further studies covering different areas and soil conditions have to be made to confirm the generality of this approach. Other ideas of further development are the establishment of a robust algorithm for data sets following non-Gaussian statistics and searching for correlation between the resultant factors and other petrophysical properties in crosswell applications.

ACKNOWLEDGEMENTS

The described work was carried out as part of the TÁMOP-4.2.1.B-10/2/KONV-2010-0001 project in the framework of the New Hungary Development Plan. The realization of this project is supported by the European Union, co-financed by the European Social Fund. The authors are also grateful to Ivan Vasconcelos for his suggestions that appreciably improved the quality of the paper.

REFERENCES

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

Archie, G. E., 1942, The electrical resistivity as an aid in determining some reservoir characteristics: Transactions of the Mining and Metallurgical Engineers, 146, 54-62.

Asfahani, J., M. Aissa, and R. Al-Hent, 2005, Statistical factor analysis of aerial spectrometric data, Al-Awabed area, Syria: a useful guide for phosphate and uranium exploration: Applied Radiation and Isotopes, **62**, 649–661.

Asquith, G. B., and D. A. Krygowski, 2004, Basic Well Log Analysis, 2nd ed., AAPG Methods in Exploration Series, No. 16: AAPG, Tulsa.

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

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

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

Campanella, R. G., and P. K. Robertson, 1991, Use and interpretation of a research dilatometer: Canadian Geotechnical Journal, **28**, 113-126.

Cardimona, S. J., W. P. Clement, and K. Kadinsky-Cade, 1998, Seismic reflection and ground-penetrating radar imaging of a shallow aquifer: Geophysics, **63**, 1310–1317.

Chang, K.-R., and C. Basnett, 1999, Delineation of sinkhole boundary using Dutch cone soundings: Engineering Geology, **52**, 113–120.

Clement, W. P., S. Cardimona, A. L. Endres, and K. Kadinsky-Cade, 1997, Site characterization at the Groundwater Remediation Field Laboratory: The Leading Edge, **16**, 1617-1621.

Csapó, G., Z. Szabó, and L. Völgyesi, 2003, Changes of gravity influenced by water-level fluctuations based on measurements and model computation: Reports on Geodesy, Warsaw University of Technology, **64**, 143-153.

De Witte, L., 1955, A study of electric log interpretation methods in shaly formations: Transactions of American Institute of Mining, Metallurgical, and Petroleum Engineers, **204**, 103-110.

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

Dobróka, M., and N. P. Szabó, 2011, Interval inversion of well-logging data for objective determination of textural parameters: Acta Geophysica, **59**, 907-934, DOI: 10.2478/s11600-011-0027-z.

Douglas, B. J., and R. S. Olsen R.S., 1981, Soil classification using electric cone penetrometer: Symposium on Cone Penetration Testing and Experience, Geotechnical Engineering Division, ASCE, 209-227.

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

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

Draskovits, P., B. Magyar, and Á. Pattantyús, 1995, Geophysical methods in groundwater prospecting and environmental protection: Física de la Tierra, **7**, 53-86.

Duval, J. S., 1977, High sensitivity gamma-ray spectrometry - state of the art and trial application of factor analysis: Geophysics, **42**, 549-559.

Fejes, I., and E. Jósa, 1990, The engineering geophysical sounding method. Principles, instrumentation, and computerised interpretation, *in* S.H. Ward, ed., Geotechnical and Environmental Geophysics, Volume II: Environmental and Groundwater, SEG, 321–331.

Fraiha, S., and J. Silva, 1994, Factor analysis of ambiguity in geophysics: Geophysics, **59**, 1083-1091.

Ghose, R., and J. Goudswaard, 2004, Integrating S-wave seismic-reflection data and cone-penetration-test data using a multiangle multiscale approach: Geophysics, **69**, 440–459.

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

Isaaks, E. H., and R. M. Srivastava, 1989, An introduction to applied geostatistics: Oxford University Press Inc.

Jöreskog, K. G., 2007, Factor analysis and its extensions, *in* R. Cudeck, and R. C. MacCallum, eds., Factor analysis at 100, Historical developments and future directions: Lawrence Erlbaum Associates, Publishers, 47-77.

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

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

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

Lech, M., M. Bajda, and K. Markowska-Lech, 2008, The use of resistivity and seismic cone penetration tests for site characterization: Annals of Warsaw University of Life Sciences, Land Reclamation, **40**, 87–96.

Magyar, B., J. Stickel, L. Verő, and I. Pádár, 1995, Assessment and remediation of environmental damage in the abandoned Soviet military bases in Hungary: Proceedings of an international symposium, International Association of Hydro-logical Sciences, P233.

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

Menke, W., 1984, Geophysical data analysis: Discrete inverse theory: Academic Press Inc.

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

Ocsenás, P., 1996, The interpretation of engineering geophysical sounding data by statistical methods (in Hungarian): Magyar Geofizika, **37**, 247-272.

Pires, A. C. B., and N. Harthill, 1989, Statistical analysis of airborne gamma-ray data for geologic mapping purposes: Crixas-Itapaci area, Goias, Brazil: Geophysics, **54**, 1326-1332.

Robertson, P. K., 1990, Soil classification using the cone penetration test: Canadian Geotechnical Journal, **27**, 151-158.

Serra, O., 1984, Fundamentals of well-log interpretation: Elsevier.

Sharma, V., 1997, Environmental and engineering geophysics: Cambridge University Press.

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

Szabó, N. P., 2004, Global inversion of well log data: Geophysical Transactions, **44**, 313-329.

Szabó, N. P., 2011, Shale volume estimation based on the factor analysis of well-logging data: Acta Geophysica, **59**, 935-953, DOI: 10.2478/s11600-011-0034-0.

Silva, J. B. C., and G. W. Hohmann, 1983, Nonlinear magnetic inversion using a random search method: Geophysics, **48**, 1645-1658.

Tillmann, A., A. Englert, Zs. Nyári, I. Fejes, J. Vanderborght, and H. Vereecken, 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.

Vértesy, L., T. Fancsik, I. Fejes, Á. Gulyás, E. Hegedűs, A. Cs. Kovács, P. Kovács, J. Kiss, A. Madarasi, L. Sőrés, Z. Szabó, and Z. Tóth, 2004, Geophysical survey at the Bátaapáti (Üveghuta) Site: Annual Report of the Geological Institute of Hungary 2003, 239-256.

Wyatt, D. E., M. G. Waddell, and G. B. Sexton, 1996, Geophysics and shallow faults in unconsolidated sediments: Ground Water, **34**, 326-334.

LIST OF FIGURE CAPTIONS

Figure 1. The flowchart of the statistical procedure for exploring the relationship between the new variables obtained by factor analysis and soil properties.

Figure 2. Topographic map of Bátaapáti (Üveghuta) Site, South-West Hungary and the measurement geometry of engineering geophysical soundings.

Figure 3. Input logs measured in PH1 for factor analysis: *RCPT* is cone tip stress, *GR* is natural gamma-ray intensity, *DEN* is density, *NPHI* is neutron porosity and *RES* is formation resistivity.

Figure 4. The logs of Factor 1 and Factor 2 estimated by the factor analysis of EGS and CPT data measured in PH1.

Figure 5. Linear relationships between the factor scores estimated by factor analysis of EGS and CPT data measured in PH1 and water saturation (on the left), water content (in the middle), gas (air) content (on the right). Denotations are: r - Pearson's correlation coefficient, S_w - water saturation, V_w - water content and V_g - gas content.

Figure 6. Water saturation logs estimated in PH1 by inverse modeling (on the left) and factor analysis (on the right), respectively.

Figure 7. Histograms of engineering geophysical sounding data measured in PH1-PH12. Denotations are: *RCPT* - cone tip stress, *GR* -

natural gamma-ray intensity, *DEN* - density, *NPHI* - neutron porosity, *RES* - formation resistivity.

Figure 8. Linear relationships between the factor scores estimated by 2D factor analysis of EGS and CPT data measured in PH1-PH12 and water saturation (on the left), water content (in the middle), gas (air) content (on the right). Denotations are: r - Pearson's correlation coefficient, S_w - water saturation, V_w - water content and V_g - gas content.

Figure 9. Water saturation sections along the EGS profile estimated by a set of 1D inversion procedures (on the top) and 2D factor analysis (at the bottom).

Figure 10. The exponential relationship between Factor 1 estimated by the 2D factor analysis of EGS and CPT data observed in PH1-PH12 and measured neutron porosity (*NPHI*).

Figure 11. Synthetic neutron porosity log estimated by 2D factor analysis of EGS and CPT data in PH1-PH12 is compared to real neutron log (on the left). Water saturation log in PH1 estimated by 1D inverse modeling using observed neutron porosity data, and that of estimated by 1D factor analysis of CPT and EGS data including synthetic neutron porosity data (on the right).