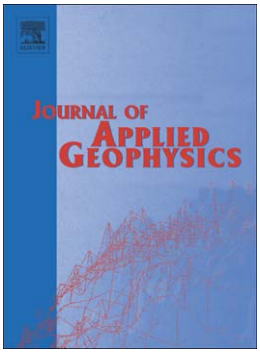

Integrating auxiliary data and geophysical techniques for the estimation of soil clay content using CHAID algorithm

Farideh Abbaszadeh Afshar, Shamsollah Ayoubi, Ali Asghar Besalatpour, Hossein Khademi, Annamaria Castrignano

PII: S0926-9851(16)30011-8
DOI: doi: 10.1016/j.jappgeo.2016.01.015
Reference: APPGEO 2901

To appear in: *Journal of Applied Geophysics*

Received date: 14 March 2015
Revised date: 6 January 2016
Accepted date: 15 January 2016

Please cite this article as: Afshar, Farideh Abbaszadeh, Ayoubi, Shamsollah, Besalatpour, Ali Asghar, Khademi, Hossein, Castrignano, Annamaria, Integrating auxiliary data and geophysical techniques for the estimation of soil clay content using CHAID algorithm, *Journal of Applied Geophysics* (2016), doi: 10.1016/j.jappgeo.2016.01.015

Integrating auxiliary data and geophysical techniques for the estimation of soil clay content using CHAID algorithm

Farideh Abbaszadeh Afshar[1], Shamsollah Ayoubi[1,*], Ali Asghar Besalatpour[2], Hossein Khademi[1], Annamaria Castrignano[3]

1- Department of Soil Science, College of Agriculture, Isfahan University of Technology, Isfahan, Iran
2 -Department of Soil Science, College of Agriculture, Vali-e-Asr University of Rafsanjan,Rafsanjan, Iran
3 -CRA — Research Unit for Cropping Systems in Dry Environments (SCA), Bari, Italy

[*]Corresponding author.
Shamsolalh Ayoubi, Professor of Soil science,
Tel: +98 31 33913470
Fax: +98 31 33913471.
E-mail address: ayoubi@cc.iut.ac.ir

Abstract

This study was conducted to estimate soil clay content in two depths using geophysical techniques (Ground Penetration Radar-GPR and Electromagnetic Induction-EMI) and ancillary variables (remote sensing and topographic data) in an arid region of the southeastern Iran. GPR measurements were performed throughout ten transects of 100 m length with the line spacing of 10 m, and the EMI measurements were done every

10 m on the same transect in six sites. Ten soil cores were sampled randomly in each site and soil samples were taken from the depth of 0-20 and 20-40 cm, and then the clay fraction of each of sixty soil samples was measured in the laboratory. Clay content was predicted using three different sets of properties including geophysical data, ancillary data, and a combination of both as inputs to multiple linear regressions (MLR) and decision tree-based algorithm of Chi-Squared Automatic Interaction Detection (CHAID) models. The results of the CHAID and MLR models with all combined data showed that geophysical data were the most important variables for the prediction of clay content in two depths in the study area. The proposed MLR model, using the combined data, could explain only 0.44 and 0.31 % of the total variability of clay content in 0-20 and 20-40 cm depths, respectively. Also, the coefficient of determination (R^2) values for the clay content prediction, using the constructed CHAID model with the combined data, was 0.82 and 0.76 in 0-20 and 20-40 cm depths, respectively. CHAID models, therefore, showed a greater potential in predicting soil clay content from geophysical and ancillary data, while traditional regression methods (i.e. the MLR models) did not perform as well. Overall, the results may encourage researchers in using georeferenced GPR and EMI data as ancillary variables and CHAID algorithm to improve the estimation of soil clay content.

Keywords: Clay content, Ground Penetration Radar, Electromagnetic Induction, Chi-Squared Automatic Interaction Detection (CHAID)

Abbreviations:

CHAID: Chi-Squared Automatic Interaction Detection, MLR: Multiple Linear Regressions, EMV: Electrical Conductivity in Vertical, EMH: Electrical Conductivity in Horizontal, TWI: Topographic Wetness Index, SPI: Stream Power Index, SR: Salinity Ratio, RVI: Ratio Vegetation Index, PVI: Perpendicular Vegetation Index, NDVI: Normalized Difference Vegetation Index, MrVBF: Multi-resolution Valley Bottom Flatness index, MrRTF: Multi

Resolution of Ridge Top Flatness Index, CI: Clay Index; CS: Catchment Area, CA: Catchment Slope and ELV: Elevation

1. Introduction

Assessing the spatial variations in soil clay content is quite vital especially when agronomists and farmers need this information for proper soil management activities. Amount of clay significantly affects soil water-holding capacity and hydraulic properties (Benedetto, 2010; De Benedetto et al., 2012) and considerably influences the cation exchange capacity (Fooladmand, 2008; Sharma et al., 2015), adsorption of the herbicide (Liu, 2008), soil nutrient regime (Zhao, 2013), and soil fertility and productivity (Davey, 1990).

In this regard, the determination and in situ monitoring of clay content at the field scale without disturbing the soil are keys in soil–vegetation systems studies. On the other hand, most soil surveys are costly, labor-intensive and time consuming. Therefore, alternative methods are being considered to complement conventional soil surveys for the estimation of soil properties. Such measurements can be performed using geophysical methods, like Ground Penetrating Radar (GPR) and Electromagnetic Induction (EMI). Both methods use electromagnetic principles, with the emission of radio waves at very high frequencies (10 to 1,000 MHz) (Annan et al., 1991), and can work quickly, accurately and continuously over long periods. They are also, not harmful to operators during use, thus non-destructive material can be analyzed by non-invasive approaches for mapping soil properties (De Benedetto et al., 2013). In particular, the high adsorptive capacity of clay minerals for water and exchangeable cations increases the dissipation of electromagnetic energy; therefore, geophysical techniques, such as

EMI and GPR, may help providing map and predict the clay content rapidly, inexpensively, and non-invasively (De Benedetto et al., 2012; Kalscheuer, et al., 2013; Benedetto and Tosti, 2013; Tosti et al., 2013; Sauvin, et al., 2014).

The GPR is widely employed as a tool to study the shallow subsurface in a broad range of applications and settings utilizing the transmission and reflection of high frequency electromagnetic waves (Jol, 2009). The performance of GPR depends on the electrical and magnetic properties of soils. High conductivities result in higher attenuation, which decreases the penetration depth of electromagnetic waves (Doolittle and Collins, 1995). Because most soils have no significant magnetic parts, the attenuation of electromagnetic waves in soils mainly depends on electrical conductivity. The electrical conductivity may be increased by clay content or ion concentration increments in the soil solution (Doolittle and Collins, 2004).

The EMI techniques use the electromagnetic energy to measure the apparent conductivity (ECa) of earthen materials. Apparent conductivity is the weighted average conductivity measured for a column of earthen materials to a specified observation depth. The EMI is sensitive to soil electrical conductivity, which is essentially affected by amount of water in soil, clay content and mineralogy, salinity, bulk density, organic matter and temperature (McNeill, 1980; Corwin and Lesch, 2003, 2005).

Recently, researchers considered developing new approaches for soil sensing based on fusing techniques with low time-consumption and high accuracy. Little literature on this new topic has been published so far (Taylor et al., 2006, 2010; De Benedetto et al., 2012). Therefore, identifying the use of geophysical techniques (i.e. EMI and GPR) and ancillary variables (remote sensing and topographic data) to estimate soil clay content can be considered as an innovative approach. With the increasing availability of new

sources of topographic and remote sensing based predictors, there is a growing interest to include such exhaustively sampled auxiliary data in the prediction of soil properties. The key point in the present study was to explore the feasibility of this combination by using the classification decision tree-based model on the Chi-Squared Automatic Interaction Detection (CHAID) method. These tree-based models are a type of algorithmic model that have already been widely used as a data-mining technique in medical, social, economic, and environmental sciences (Murthy, 1998; Tóth et al., 2012; Bichler et al., 2014). In the context of clay content estimation, however, very few studies have employed this robust and versatile data analysis method.

Therefore, the main objective of this study was applying and comparing two predictive approaches (including CHAID and MLR) for estimating soil clay content at two depths (0-20 and 20-40 cm) using three auxiliary data (including geophysics, topographic and remote sensing data) in an arid region of Bam district, Kerman province, in the southeastern Iran. In order to have an efficient/effective understanding of soil clay content, the efficacy of CHAID method was also compared with multiple linear regression (MLR) method for the clay content estimation as the benchmark for the comparison of their performances.

2. Material and Methods

2.1. Description of the study area

The study area was the Bam district located between 58°4′17″ to 58°28′8″ E longitudes and 28°52′51″ to 29°9′29″ N latitudes (Fig. 1), of Kerman province, Southeastern Iran. The area is surrounded by mountains (dominantly limestone and volcanic) from northwest toward southeast with major landforms including young

alluvial fans and pediment, clay flat and hills. Pasture species in the selected sites were mainly *Alhagi spp.* and *Artemisia sieberi*. The mean annual precipitation, temperature and potential evapotranspiration are respectively 64 mm, 23.8°C and 3000 mm with Aridic and Hyper thermic soil moisture and temperate regimes (Soil Survey Staff, 2010). The soils are classified into soil orders of Entisols and Aridisols and sub-great groups of Typic Haplosalids, Gypsic Haplosalids, Typic Haplocambids, Sodic Haplocalcids, Leptic Haplogypsids and Typic Torriorthents.

2.2. Ground penetrating radar (GPR) survey

The GPR as a non-invasive geophysical tool is specifically designed to penetrate into materials and provide proper images for shallow subsoils. The radar produces a high-frequency electromagnetic wave propagated through the sub-surface materials at the velocity determined by the soil dielectric permittivity. This amount of energy is reflected by an interface dependent upon the contrast in the relative dielectric permittivity of the two layers (Jol, 2009).

Soils with high electrical conductivity rapidly attenuate radar energy, which restricts penetration depth and severely limits the effectiveness of GPR. Factors influencing the electrical conductivity of soils include the amount and type of salts in solution and the clay content (Jol, 2009; De Benedetto et al., 2012).

The GPR measurement was conducted by the MALA GEOSCIENCE AB (Sweden), as commonly used in offset surveying with antennae 0.5 m apart with a center frequency of 250 MHz. Common offset surveys used a transmitting and receiving antenna held at a fixed distance from each other and moved incrementally along a survey line (Dominic et al., 1995; Mount et al., 2014). Common offset profiles are known as the most common

representation of GPR data because they resemble a geologic cross section where the depth scale is expressed as a two-way travel time. All GPR profiles were processed by ReflexW 7.0 software (Sandmeier Scientific Software, Karlsruhe, Germany). The common offset data was simplified to five components: (1) a dewow filter was used to remove the low frequency noise and DC signal component; (2) a time zero correction was considered to remove the lag between the triggering of the signal and the recording of the first arrival; (3) AGC Gain was used to enhance low amplitude reflections; (4) trapezoidal band pass filter with four frequency values for the removal of DC and high frequency noise components was applied to the data; and (5) the background removal was applied to make visible the useful information covered by temporally consistent noise and to emphasize hyperbolic signals indicating a point of anomaly.

Thereafter, a quadrature Hilbert-Transformation filter was used to calculate the instantaneous amplitude or envelop of the data, giving an estimation of the reflectivity and being proportional to the square root of the total energy of the signal at a given instant time (Claerbout, 1985). The envelope is expressed in the same measurement unit of signal amplitude in volts. Therefore, the envelope that could give an overview of the distribution of the different types of reflectors present in the subsoil (different depths) was used as the GPR output.

2.3. Electromagnetic induction (EMI) survey

EMI soil survey is based on the principle that a transmitter coil in contact with the soil surface produces a time-varying primary magnetic field in the subsoil. The eddy currents induced in the soil generate a secondary magnetic field recorded by a receiver coil in the EM unit. The apparent conductivity near the receiver is determined by the

ratio of the magnitude of the secondary magnetic field to one of the primary magnetic field (McNeill, 1980).

Bulk electrical conductivity (ECa) was measured using the electromagnetic induction instrument (EMI, Geonics EM38). This instrumentation could measure bulk electrical conductivity simultaneously in two orientations of polarization (horizontal and vertical) with a different depth response profile. In a homogeneous soil profile, the vertical (EMV) maximum sensitivity was at a depth of approximately 0.40 m and the signal penetrated to a depth of 1.5 m, whereas the horizontal (EMH) maximum sensitivity occurred at the surface and the signal primarily reflected the topsoil properties up to 0.75 m depth.

2.4. Soil survey

The GPR and EMI surveys were performed in six sites in the study area (Fig. 1b). In each site, the GPR surveys were run along 10 transects with 100 m length and the line spacing of approximately 10 m (Fig.2). The GPR data were collected with trace increments of 0.024 m and time increments of 0.05 ns. An EMI survey was carried out to cover the same area surveyed by the GPR survey, one measurement per 10 m length along each transect.

Ten soil cores were selected randomly in each site and soil samples were collected from depths of 0-20 cm and 20-40 cm (Fig. 2). Soil samples were air-dried and passed through a 2 mm sieve for particle size distribution analysis (Gee and Bauder, 1986).

2.5. Ancillary spatial variables

Two groups of ancillary variables including topographic attributes and remote sensing data were employed. Topography, as one of the major soil forming factors

controls various soil properties. Florinsky et al., (2002), Tajik et al., (2012) and Mehnatkesh et al., (2013) have reported that soil properties in the top soil layer are affected by topographic attributes. So, quantitative information on the topographic attributes has been applied in the form of digital elevation models. The topographic attributes were derived from the ASTER-GDEM with the cell size of 30 m × 30 m (METI and NASA, 2012) using the SAGA GIS software (Olaya, 2004). The derived terrain attributes obtained from the DEM included topographic wetness index (TWI), stream power index (SPI), catchment area (CA), catchment slope (CS), multi resolution of ridge top flatness index (MrRTF) and multi resolution valley bottom flatness index (MrVBF) (Gallant and Dowling, 2003). The derived topographic attributes are described in Table 1.

The physical factors (particle size and surface roughness) and components (surface mineralogy, organic matter content and moisture) control soil spectral reflectance (Irons et al., 1989). Therefore, remote sensing data could be used as auxiliary variables for predicting clay content at soil surface. One scene of the Landsat Enhanced Thematic Mapper (ETM) acquired in 2005 (U.S. Geology Survey, 2005) was used to extract remote sensing indices including the normalized difference vegetation index (NDVI; Boettinger et al., 2008), ratio vegetation index (RVI; Pearson and Miller, 1972), perpendicular vegetation index (PVI; Richardson and Wiegand, 1977), clay index (CI; Boettingeret al., 2008) and salinity ratio (SR; Metternicht and Zinck, 2003). A summary of the definition of the remote sensing data used in the present study is presented in Table 1. Processing of predictors was carried out using the SAGA GIS (Olaya, 2004).

2.6. Statistical analysis

Classical descriptive parameters of the experimental data, including mean, minimum, maximum, range, coefficient of variation (CV), skewness, and kurtosis, were determined using the statistical software SPSS (Statistical Package for the Social Sciences), v.16. The distribution of variables was also evaluated using the Kolmogorov Smirnov test (Massey, 1951). To interpret the interactions the correlations among the variables as well as clay content and geophysical and ancillary data (remote sensing and topographic data) were computed using SPSS software (Swan and Sandilands, 1995).

2.7. Multiple linear regressions (MLR)

MLR is one of the well-known statistical techniques that have long been used in many researches (Besalatpour et al., 2013; Ayoubi and Sahrawat, 2011). The basic linear regression model has the following form:

$$Y = \alpha + X^T \beta + \varepsilon \qquad (1)$$

where Y denotes the dependent variable, α is a constant called the intercept, $X= (X_1. . .X_n)$ is a vector of explanatory variables, $\beta=\{\beta_1, \ldots , \beta_n\}$ is the vector of regression coefficients (one for each explanatory variable), and ε represents random measured errors as well as any other variation not explained by the linear model. In this study, the stepwise regression procedure was used to develop the MLR models for estimating the clay content in both investigated depths using geophysics and ancillary data as explanatory variables. Factors for inclusion to the model were selected based on the probability of ≤ 0.05 (Freund and Littell, 2000). For developing the models, the data set (N= 60) was divided into two subsets of training and testing. The training subset was randomly chosen from 80% of the total set of the data (N= 46) and the remaining samples (20% of the data) were used as the testing set (N= 14).

2.8. Chi-squared automatic interaction detector (CHAID) method

The CHAID algorithm, originally proposed by Kass (1980) and further developed by Magidson (1993), is a well-known and widely used decision tree (DT) algorithm which constructs a tree using a recursive partitioning method. This method approximates the function for the target attribute by learning a DT from the previous examples. Each internal node in a DT specifies an attribute test, and each leaf represents the predicted target value. Chi-square analyses are used for splitting and merging operations in this algorithm. Accordingly, it takes two probabilities, whit the first one indicating the significance level for splitting the node, and the second one showing the significance level for merging the nodes. This algorithm deals with both qualitative (nominal or ordinal) and quantitative values (Demetgul, 2013).

The CHAID method can be used as an exploratory way for classifying categorical data. The purpose of the procedure is to split a set of objects in such a way that the subgroups differ significantly with respect to a designated criterion. The criterion matches the dependent variable, while the remaining attributes represent their predictors in the model. The segments derived by CHAID are mutually exclusive and exhaustive, implying that the segments do not overlap and each object of the sample is contained in exactly one segment. Therefore, the application of the method approves the classification of new objects by knowing the categories of the predictors (Magidson, 1993).

In CHAID trees, the homogeneity of the groups generated by the tree is evaluated by a Bonferroni corrected *p-value* obtained from the chi-square statistic applied to two-way classification tables with C classes and K splits for each tree node (Maroco et al., 2011):

$$X^2 = \sum_{c=1}^{c}\sum_{k=1}^{k}\frac{(n_{ck} - \hat{n}_{ck})^2}{\hat{n}_{ck}} \approx X^2(C-1)(K-1) \quad (2)$$

where n_{ck} refers to the observed frequencies of cell c_k and $\check{n}_{ck}$ is the expected frequencies under the null hypothesis of two-way homogeneity.

In the CHAID analysis employed here, the parameter epsilon for convergence and the maximum iteration for convergence were set at 0.001 and 100 (obtained by a trial and error procedure, as evaluated by the model performance), respectively. The parameter "epsilon for convergence" determined how much change had to occur for iterations to continue; if the change from the last iteration were smaller than the specified value, iterations would be stopped. The parameter "maximum iterations for convergence" also specified the maximum number of iterations before stopping, no matter if convergence had taken place or not. The Pearson method was used for the Chi-square of categorical target. The SPSS Clementine (IBM Com., Chicago, USA) software was also used to develop the CHAID models.

2.9. Model performance evaluation criteria

The performances of the developed models were evaluated using various standard statistical performance evaluation criteria calculated for the testing data sets of the models. The statistical measures were the mean estimation error (MEE), the root mean square error (RMSE), absolute error percentage (AEP), and the coefficient of determination (R^2) between the measured and predicted clay contents. The MEE, RMSE and AEP statistics were defined as:

$$MEE = \frac{1}{n}\sum_{i=1}^{n}\left[P(xi) - M(xi)\right] \quad (3)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[P(xi) - M(xi)\right]^2} \quad (4)$$

$$AEP = \frac{\sum_{i=1}^{n}\left|P(xi) - M(xi)\right|}{\sum_{i=1}^{n} M(xi)} \times 100 \quad (5)$$

where *P(xi)* denotes the predicted value of observation *i*, *M(xi)* is the measured value of observation *i*, and *n* is the total number of observations (Besalatpour et al., 2013).

3. Results and Discussion

3.1. Descriptive statistics

Descriptive statistics of the measured clay contents in the two investigated depths (0-20 and 20-40 cm) and the geophysical data are depicted in Table 2. The clay content in 0-20 cm depth exhibited a greater coefficient of variation (CV) than the one in 20-40 cm depth, indicating the high variability of clay in 0-20 cm depth. This might be due to the large variations of soil properties and more close to the soil surface. It could be presumably related to soil redistribution processes such as wind and water erosion in the study area. Ayoubi et al., (2012) reported that soil redistribution is responsible for almost 60% CV in clay content of the surface soil in the western hilly parts of Iran.

The clay contents in both depths were normally distributed as confirmed by the Kolmogorov Smirnov test and skewness values (Table 2). The highest and lowest CV values for the geophysical data were attributed to EMV (126.80%) and $GPR_{(40\text{-}60)}$ (amplitude for 250 MHz antenna frequency at 40-60 cm depth) (44.10%), respectively. The high variability in EMV might be related to the high variability of soil salinity in the

region. It is previously reported that ECa could be dramatically influenced by soil salinity (Lesch et al., 2005; Taghizadeh-Mehrjerdi et al., 2014).

To determine the relationships between clay contents of the depths 0-20 and 20-40 cm and the input data (i.e., geophysical and ancillary data), correlation analysis was carried out (Table 3). The results revealed significant negative relationships between the clay content and GPR data and positive relationships with EMI data; this was similar to the results obtained by De Benedetto et al., (2008, 2012) and Broge et al., (2004). Although the obtained correlation coefficients in this study were not as high as those reported by other researches, but very well indicating the relation between clay content and GPR/ EMI data. The quite small and generally insignificant correlation between EMI and GPR data (0.30, $p<0.01$ (i.e. the probability level of 0.01)) might be due to the differences in the size of vertical and horizontal supports and to the physical operational principle, as well as the high spatial variability in both horizontal and vertical directions.

The clay contents of both investigated depths showed a significantly positive correlation with the remote sensing indices (i.e., CI (0.38 $p<0.01$) and NDVI (0.33, $p<0.05$)) and a negative correlation with the ancillary data (i.e., SPI (-0.29, $p<0.05$) and TWI (-0.42, $p<0.01$)). The positive correlation between CI (as an indicator of clay content) and the clay content has already been reported by other researchers (Piccini et al., 2014). The positive relationship between NDVI and the clay content also confirmed the direct effects of soil texture on advancing plant vegetation via increasing water and nutrient availabilities. The clay content in the depth of 20-40 cm exhibited higher correlation coefficients with the most selected properties, as compared to the 0-20 cm depth. This might be due to the major changes that occurred near the soil surface,

leading to the enhancement of clay variation and subsequently reducing correlation coefficients of clay content in 0-20 cm with the geophysical and ancillary data.

3.2. Multiple linear regression (MLR) models

Resulting from the stepwise regression analysis, the NDVI, MrVBF and TWI were computed as affecting to the clay content variability at 0-20 cm depth, whereas the NDVI was solely selected as the predictor variable for the 20-40 cm depth among the ancillary variables (Table 4). Soil texture showed stronger absolute correlations with the Bands 4, 5 and 7 (as near-infrared bands) than the Bands 1–3 (visible bands) of ETM imagery, because the spectral signatures of the texture typically dominated the near-infrared spectra (Stenberg et al., 2010). Moreover, soil texture showed a straight effect on the short wave infrared reflectance, as the incoming radiation was scattered contrarily by the coarse particles, in comparison to finer materials. For instance, many researchers have reported that the spectral absorption band center of clay particles is around 2200 nm (in the range of the wavelengths of the Band 7) (Clark, 1999; Brown et al., 2006; Lagacherie et al., 2008).

The EMV and $GPR_{(40-60)}$ were introduced by the MLR model as the most influential factors when the geophysical data (GPR and EMI data) were used for the prediction of clay content in 20-40 cm depth and also, $GPR_{(0-15)}$ (amplitude for 250 MHz antenna frequency at 0-15 cm depth) remained as the only predictor variable for the clay content estimation in 0-20 cm depth. Doolittle and Collins (2004) stated that electrical conductivity might be increased by clay content or ion concentration in soil solution. Generally, time-lapse EMI measurements are significant for allowing the differentiation of the temporally stable contribution of static soil properties such as clay content from

the temporally dynamic contributions of water content and soil solution conductivity to the observed ECa. The $GPR_{(0-15)}$, NDVI, and MrVBF were the most important variables influencing the estimation of clay content in 0-20 cm depth using the combined data, while the $GPR_{(0-20)}$ (amplitude for 250 MHz antenna frequency at 20-40 cm depth) and NDVI were the most determinant factors for the prediction of clay content in 20-40 cm depth.

The obtained R^2 values for the measured and predicted clay content values via MLR models are presented in Table 5 and Fig. 3. The R^2 values for the clay content prediction, using the constructed MLR model with the ancillary data, were 0.17 and 0.12 for the 0-20 and 20-40 cm depths respectively, and 0.33 and 0.21 for the constructed MLR model using the geophysical data. The proposed MLR model of combined data, could explain only 0.44 and 0.31 % of the total variability of clay content in 0-20 and 20-40 cm depths, respectively. The coupling of ancillary and geophysical data improved the prediction accuracy, as confirmed by increasing R^2 values in the clay content prediction. The MLR of only ancillary data had the lowest R^2 among the proposed MLR models (Table 5).

Comparison of the model performances demonstrated that the utilization of a combination of ancillary and geophysical data as the input to the MLR model might give more accurate prediction results. This was evidenced by the obtained lower MEE, RMSE, and AEP and a higher R^2 value (Table 5 and Fig. 3). These results suggested the greater influence of combined data, rather than ancillary data and geophysical data alone, as the inputs of the models in clay content prediction by MLR approached. On the other hand, the proposed MLR models for 0-20 cm depth were generally more feasible than the MLR models in predicting clay content in the 20-40 cm depth when the

evaluation criteria were compared. However, the predictive capability of the constructed MLR model in 0-20 cm was higher than that of the MLR model in 20-40 cm depth.

According to the evaluation criteria (Table 5), it appeared that the conventional regression models were, to some extent, weak in predicting clay contents in the study region when the geophysical variables, ancillary data and the combined data are used. Besalatpour et al., (2013) and Ayoubi and Sahrawat (2011) also reported the low efficacy of MLR technique in similar studies.

3.3. CHAID models

Application of CHAID approach to discern the most important factors affecting the clay content variation resulted in different findings (Fig. 5). The components of TWI, MrVBF, SPI and CI were introduced as the effective combination of ancillary data affecting the clay content variation in 0-20 cm depth and CI in 20-40 cm. The TWI index was used to describe the effects of topography on the location and size of the saturated areas and to more accurately characterize the spatial variability of soil properties due to surface hydrology (Moore et al., 1993; Florinsky et al., 2002; Tajik et al., 2012). The MrVBF index was used to distinguish hillslope and valley bottoms, whereas the latter was characterized by lowness, flatness, and convergent water flow (Gallant and Dowling, 2003; Wang and Laffan, 2009), thereby significantly contributes to explain the variability of clay content in 0-20 cm depth (Fig. 5). The SPI, which measures the erosive power of flowing water (Wilson and Gallant, 2000) is one of the most important factors controlling clay content variability in this region. Also, this index could depict the area of both high slopes and the contributing areas over the landscape (Moore et al., 1993). The R^2 values for the clay content prediction, using the constructed

CHAID model with the ancillary data, were 0.79 and 0.86 for the 0-20 and 20-40 cm depths, respectively.

The $GPR_{(0-15)}$ (coefficient=0.413) and $GPR_{(40-60)}$ (coefficient=0.587) were the most important variables for the estimation of clay content in 0-20 cm depth using geophysical data, while the EMV and $GPR_{(30-45)}$ (amplitude for 250 MHz antenna frequency at 30-45 cm depth) were the most determinant factors for the prediction of the clay content in 20-40 cm depth. The proposed CHAID model, using geophysical data could explain only 0.57 and 0.62 % of the total variability of clay content in 0-20 and 20-40 cm depths, respectively (Fig. 5). Clay minerals display specific electrical properties as a consequence of physicochemical structure. Due to isomorphic substitution, clay minerals encompass a net negative charge. The net negative charge of a clay platelet is counterbalanced by an equivalent charge on cations, such as K^{+}, Na^{+}, Ca^{+2} and Mg^{2+}. These cations are concentrated in a diffuse double layer (DDL) that encloses clay minerals and provides an alternative pathway for electrical conduction. De Benedetto et al., (2008, 2010, 2012) reported significant advantages in using geophysical data for clay characterization along soil profiles.

A combination of geophysical data and ancillary properties as the input to the CHAID model showed that the GPR had the greatest influence on clay variation in the 0-20 cm depth (Fig. 5). These results confirmed the statistically significant relationships between clay content of the surface layers and GPR data. However, EMI data or ancillary variables had no significant effect in this respect. As shown in Fig. 4, the R^2 values for the clay content prediction, with constructed CHAID model with the combination of geophysical data and ancillary properties, were 0.82 and 0.76 for the 0-20 and 20-40 cm depths, respectively. De Benedetto et al., (2012), by utilizing a

stepwise regression model, reported that the EMI in vertical mode, antenna envelope at $GPR_{(15–30\ cm)}$ with 1600 MHz frequency and $GPR_{(30–60\ cm)}$ with 600 MHz frequency were the most influencing factors for the clay content prediction at 0–20 cm depth (R^2=0.89). In contrast, only one variable, the 600 MHz antenna envelope at $GPR_{(30–60\ cm)}$, was introduced as the significant factor for the clay content estimation at 20–40 cm; however, the obtained coefficient of determination was much lower (R^2=0.59), thereby indicating the higher attenuation of GPR signal at this soil depth. It is known that the performance of GPR is highly influenced by soil electrical conductivity. Particularly, soils with high electrical conductivity rapidly attenuate the radar energy. Consequently, it seems that the clay content is a key factor affecting the electrical conductivity of soils. Furthermore, the GPR sensor has been demonstrated to be a beneficial indicator of clay content in the shallower layer of the soil.

Constructing the CHAID model by combining geophysical data and ancillary properties to predict the clay content in 20-40 cm depth showed that the EMV (coefficient=0.124), EMH (coefficient=0.124), TWI (coefficient=0.124), MrRTF (coefficient=0.124), CI (coefficient=0.124) and RVI (coefficient=0.066) were the most important factors influencing the clay content variation (Fig. 5). Nevertheless, the constructed CHAID model, using a combination of the data, showed that the GPR data might not be important for the prediction of clay content in the lower depths (i.e., 20-40 cm, Fig. 5). Clay content has a high electrical conductivity strongly dependent on the EMI data but sand and silt are less conductive, hence rarely related to EMI data. Heil and Schmidhalter (2012) used the apparent soil electrical conductivity for characterizing soil texture variability at a highly variable site. Their results showed that the ECa was more closely related to clay and sand/gravel, whereas silt exhibited a stronger

dependency on the boundary depth. Furthermore, they showed that the ECa, in combination with the boundary depth between Tertiary and Quaternary sediments, elevation, aspect and the cultivation factors, provided a helpful and robust surveying technique to estimate soil texture for the Tertiary hill country in the southern Germany.

TWI described flow intensity and accumulation potential, and the significant correlation between TWI and clay content confirmed the TWI's impact on the distribution of clay particles within the landscape. The results of the CHAID models with the combined data set showed that the geophysics data were the most important variables for the estimation of clay in the study area (Fig. 5).

Comparison of the CHAID models performances demonstrated that the proposed CHAID models in 20-40 cm depth were, more appropriate than the CHAID models in predicting clay content in 0-20 depth when the evaluation criteria were compared. This was evidenced by the obtained lower MEE, RMSE, and AEP and a higher R^2 value (Table 5 and Fig. 4). However, the predictive capability of the constructed CHAID model in 20-40 cm was higher than that of the model in 0-20 cm depth.

3.4. Comparison of the MLR and CHAID models

Comparison of the two methods demonstrated that CHAID provided much more accurate predictions of clay content than MLR method, particularly when a combination of ancillary and geophysical data was employed to develop the model (Table 5). The statistical performance evaluation criteria revealed that the linear regression model was, to some extent, weak in predicting soil clay contents in the study region.

The obtained R^2 values for the MLR were much lower than the CHAID for all constructed models using the input data sets for both investigated depths (Table 5 and

Figs. 3 and 4). Therefore, when the aim is to study the relationship between the easily available characteristics and clay content, CHAID algorithm can be more favorable.

The constructed CHAID model, using geophysical data, could explain 57 and 62 % of the total variability of clay content for the 0- 20 and 20-40 cm depths respectively. However, the predictive capability of the constructed CHAID model, using geophysical data alone, was not higher than that of other input data (Table 5). The proposed CHAID model, using ancillary and geophysical data as the inputs to the model, resulted in the highest R^2 values for the measured and predicted clay in 0-20 cm among the other proposed CHAID models (Fig. 5).

According to the ME, RMSE, and AEP values (Table 5), all developed CHAID models showed a good performance during the reliability testing. The MEE, RMSE, and AEP values for the clay content based on MLR model with the ancillary data were 0.1, 6.32 and 37.42 in 0-20 cm depth and 2.53, 6.60 and 44.25 in 20-40 cm, respectively. A similar trend in clay content prediction using three constructed MLR models was also observed for all the data (Table 5). Tóth et al., (2012) developed some pedotransfer rules based on the CHAID classification tree and used available soil map information as the inputs. They concluded that the classification tree methods (regression tree and CHAID) were helpful in modeling the complex relationship between soil water retention and other soil properties of salt affected soils.

The scientific challenge is to acquire a higher-resolution model for topsoil and subsoil clay prediction, from sparse clay data, using geophysical data as the ancillary variables. However, further investigations are essential to know how the clay content and its variability in a soil could be predicted from geophysical data (such as GPR and EMI data), because such factors as the type and frequency of sensor could affect the results. It

could be speculated that the radar image might be efficiently used together with other geophysical sensors and easily available environmental variables such as topographic attributes and remote sensing data as the secondary information to improve the prediction of clay content using CHAID algorithm.

4. Conclusion

There is a great need for developing noninvasive technologies to obtain quantitative information regarding subsoil properties, such as clay content. In this study, geophysical and attribute data were used to estimate soil clay content using a decision tree based CHAID model and linear regression method in an arid region of Iran. The results showed that the geophysical data were the most important variables influencing the estimation of clay content. The CHAID technique showed a greater potential in predicting soil clay content from geophysical and ancillary data, whereas traditional regression methods (i.e. MLR models) did not perform well. These results may encourage researchers in using georeferenced GPR and EMI data as ancillary variables and CHAID algorithm to improve the estimation of soil clay content. However, further research should be conducted in this area and validated in the future, especially for various soils and different management practices. Finally, it is speculated that the introduced methods here will provide a novel tool (especially using geophysical data as the input to the model) for the quantitative estimation of soil clay content as an alternative to the existing conventional linear models. This can be very valuable for soil scientists who look for a particle size distribution prediction tool with the lowest error achievement and the highest efficiency.

Acknowledgment

Our special thanks are due to Professor Mohammad Ali Hajabbasi, visiting scientist in University of California, Berkeley, and Prof. Aghafakhr Mirlohi for their valuable comments and suggestions to improve of original manuscript.

References

Annan, A.P., Cosway, S.W., Redman, I.D. 1991. Water Table Detection with Ground Penetrating Radar. In Proceedings of the 61th Meeting of the Society of Exploration Geophysicists (SEG), p. 494-496.

Ayoubi, S., Ahmadi, M., Abdi, M.R., Abbaszadeh Afshar, F., 2012. Relationships of ^{137}Cs inventory with magnetic measures of calcareous soils of hilly region in Iran. J. Environ. Radioact. 112, 45-51.

Ayoubi, S., Sahrawat, K.L., 2011. Comparing multivariate regression and artificial neural network to predict barley production from soil characteristics in northern Iran. Arch. Agron. Soil Sci. 57, 549-565.

Benedetto, A., 2010. Water content evaluation in unsaturated soil using GPR signal analysis in the frequency domain. J. Appl. Geophys. 71, 26‑35.

Benedetto, F., Tosti, F., 2013. GPR spectral analysis for clay content evaluation by the frequency shift method. J. Appl. Geophys. 97, 89–96.

Besalatpour, A.A., Ayoubi, S., Hajabbasi, M.A., Mosaddeghi, M.R., Schulin, R., 2013. Estimating wet soil aggregate stability from easily available properties in a highly mountainous watershed.Catena. 111, 72-79.

Bichler, A., Neumaier, A., Hofmann, T., 2014. A tree-based statistical classification algorithm (CHAID) for identifying variables responsible for the occurrence of faecal indicator bacteria during waterworks operations. J. Hydrol. 519, 909–917.

Boettinger, J.L., Ramsey, R.D., Bodily, J.M., Cole, N.J., Kienast-Brown, S., Nield, S.J., Saunders, A.M., Stum, A.K., 2008. Landsat spectral data for digital soil mapping, In: Hartemink, A.E., McBratney, A.B., Mendonca-Santos, M.L. (Eds.), Digital Soil Mapping with Limited Data. Springer Science., Australia, pp. 193–203.

Broge, N.H., Thomsen, A.G., Greve, M.H., 2004. Prediction of Topsoil Organic Matter and Clay Content from Measurements of Spectral Reflectance and Electrical Conductivity.Arch. Agron. Soil Sci. 54(4), 232-240.

Brown, D.J., Shepherd, K.D., Walsh, M.G., Dewayne, M., Reinsch, T.G., 2006. Global soil characterization with VNIR diffuse reflectance spectroscopy. Geoderma. 132, 273–290.

Claerbout, J.F., 1985. Fundamentals of Geophysical Data Processing: with Applications to Petroleum Prospecting. Blackwell Scientific Publications, Palo Alto, California, USA.

Clark, R.N., 1999. Spectroscopy of rocks and minerals, and principles of spectroscopy, In: Rencz, A. (Eds), Manual of Remote Sensing., John Wiley and Sons., New York, pp. 3–58.

Corwin, D.L., Lesch, S.M., 2005. Apparent soil electrical conductivity measurements in agriculture.Comput Electron Agr. 46, 130–133.

Corwin, D.L.,Lesch, S.M., 2003. Application of soil electrical conductivity to precision agriculture: theory, principles, and guidelines. Agron J. 95, 455–471.

Davey, B.G., 1990. The chemical properties of soils, In: Campbell, K.O., Bowyer, J.W. (Eds.), Scientific Basis of Modern Agriculture. Sydney University Press., Sydney, Australia, pp. 54-78.

De Benedetto, D., Castrignanò, A., Sollitto, D., Campi, P., Modugno, F., 2008. Nonintrusive Mapping of Subsoil Properties in Agricultural Fields with GPR and EMI. Proceedings First global workshop on high resolution digital soil sensing and mapping, Sydney - Australia.

De Benedetto, D., Castrignano, A., Sollitto, D., Modugno, F., 2010. Spatial relationship between clay content and geophysical data. Clay Miner. 45, 197–207.

De Benedetto, D., Castrignano, A., Sollitto, D., Modugno, F., Buttafuoco, G., lo Papa, G., 2012. Integrating geophysical and geostatistical techniques to map the spatial variation of clay. Geoderma. 171-172, 53–63.

De Benedettoa, D., Castrignan, A., Quartoa, R., 2013. A geostatistical approach to estimate soil moisture as a function of geophysical data and soil attributes. Procedia Environ. Sci. 19, 436 – 445.

Demetgul, M., 2013. Fault diagnosis on production systems with support vector machine and decision trees algorithms. Int. J. Adv. Manuf. Technol. 67, 2183-2194.

Dominic, D.F., Egan, K., Carney, C., Wolfe, P.J, Boardman, M.R., 1995. Delineation of shallow stratigraphy using ground penetrating radar. J. Appl. Geophys. 33(1-3), 167-175.

Doolittle, J.A., Collins, M.E., 1995. Use of soil information to determine application of ground penetrating radar. J. Appl. Geophys. 33, 101–108.

Doolittle, J.A., Collins, M.E., 2004. Suitability of soils for GPR investigations, In: Daniels, D.J. (Eds.), Ground Penetrating Radar. Institution of Electrical Engineers, London, pp. 97–108.

Florinsky, I.V., Eilers, R.G., Manning, G.R., Fuller, L.G., 2002. Prediction of soil properties by digital terrain modeling. Environ. Model. Softw. 17, 295-311.

Fooladmand, H.R., 2008. Estimating cation exchange capacity using soil textural data and soil organic matter content: A case study for the south of Iran. Arch. Agron. Soil Sci. 54(4), 381–386.

Freund, R.J., Littell, R.C., 2000. SAS System for Regression.SASInst, Cary, NC.

Gallant, J.C., Dowling, T.I., 2003. A multiresolution index of valley bottom flatness for mapping depositional areas. Water Resour. Res. 39, 1347-1351.

Gee, G.W., Bauder, J.W., 1986. Particle size analysis, In: Klute, A. (Eds.), Methods of Soil Analysis. Part 1. Am. Soc. Agron, Madison, WI, p. 383-411.

Heil, K., Schmidhalter, U., 2012. Characterisation of soil texture variability using the apparent soil electrical conductivity at a highly variable site.Comput.Geosci. 39, 98–110.

Irons, J. R., Weismiller, R. A., Pterson, G.W., 1989. Soil reflectance. In Theory and Applications of Optical Remote Sensing; Asar, G., Ed.; John Wiley: New York, NY, USA, pp. 66–106.

Jol, H.M., 2009. Ground Penetrating Radar: Theory and Applications, first ed. Elsevier, Amsterdam.

Kalscheuer, T., Bastani, M., Donohue, S., Persson, L., Pfaffhuber, A.A., Reiser, F., Ren, Z., 2013.Delineation of a quick clay zone at Smørgrav, Norway, with electromagnetic methods under geotechnical constraints. J. Appl. Geophys. 92, 121–136.

Kass, G.V., 1980. An exploratory technique for investigating large quantities of categorical data. Appl. Stat. 29 (2), 119–127.

Lagacherie, P., Baret, F., Feret, J.B., Madeira Netto, J., Robbez- Masson, J.M., 2008. Estimation of soil clay and calcium carbonate using laboratory, field, and airborne hyperspectral measurements. Remote Sens. Environ. 112, 825–835.

Lesch, S.M., Corwin, D.L., Robinson, D.A., 2005. Apparent soil electrical conductivity mapping as an agricultural management tool in arid zone soils. Comput Electron Agr. 46, 351–378.

Liu, Z., He, Y., Xu, J., Huang, P., Jilani, G., 2008. The ratio of clay content to total organic carbon content is a useful parameter to predict adsorption of the herbicide butachlor in soils. Environ. Pollut. 152, 163-171.

Magidson, J., 1993. The use of the new ordinal algorithm in CHAID to target profitable segments. J. Database Market. 1, 29–48.

Maroco, J., Silva, D., Rodrigues, A., Guerreiro, M., Santana, I., de Mendon, A., 2011. Data mining methods in the prediction of Dementia, A real-data comparison of the accuracy, sensitivity and specificity of linear discriminant analysis, logistic regression, neural networks, support vector machines, classification trees and random forests. BMC Rese. Not. 4, 1-14.

Massey, F.J., 1951. The Kolmogorov-Smirnov test of goodness of fit. J Am. Stat. Assoc. 4, 70-71.

McNeill, J.D., 1980. Electromagnetic Terrain Conductivity Measurement at Low Induction Numbers. Geonics Limited, Technical Note TN 6, Geonics Ltd. Mississauga, Ontario, Canada.

Mehnatkesh, A., Ayoubi, S., Jalalian, J., Sahrawat, K.L., 2013. Relationships between soil depth and terrain attributes in a semi-arid hilly region in western Iran. J. Mount. Sci. 10, 163-172.

Metternicht, G.I., Zinck, J.A., 2003. Remote sensing of soil salinity: potentials and constraints. Remote Sens. Environ. 85, 1–20.

Moore, I.D.,Gessler, P.E., Nielsen, G.A., Peterson, G.A., 1993.Soil attribute prediction using terrain analysis. Soil Sci. Soc. Am. J. 57, 443–452.

Mount, G.J., Comas, X., Cunningham, K.J., 2014. Characterization of the porosity distribution in the upper part of the karst Biscayne aquifer using common offset ground penetrating radar, Everglades National Park, Florida. J. Hydrology. 515, 223-236.

Murthy, S.K., 1998. Automatic construction of decision trees from data: a multidisciplinary survey. Data Min. Know. Discov. 2 (4), 345–389.

Olaya, V.F., 2004. A Gentle Introduction to Saga GIS.The SAGA User Group e.V, Gottingen.

Pearson, R.L., Miller, L.D., 1972. Remote mapping of standing crop biomass for estimation of the productivity of the short-grass prairie, Pawnee National Grasslands, Colorado.Proceedings of the 8th International Symposiumon Remote Sensing of Environment. Environmental Research Institute of Michigan, Ann Arbor, pp. 1357–1381.

Piccini, C., Marchetti, A., Francaviglia, R., 2014. Estimation of soil organic matter by geostatistical methods: Use of auxiliary information in agricultural and environmental assessment. Ecol. Indic. 36, 301–314.

REFLEX Software. 2012. Sandmeier Scientific Software, Karlsruhe, Germany.

Richardson, A.J.,Wiegand, C.L., 1977. Distinguishing vegetation from soil background information. Photogramm. Eng. Rem. S. 43, 1541–1552.

Sauvin, G., Lecomte, I., Bazind, S., Hansen, L., Vanneste, M., L'Heureux, G.S., 2014. On the integrated use of geophysics for quick-clay mapping: The Hvittingfoss case study, Norway. J. Appl. Geophys. 106, 1–13.

Sharma, A., Weindorf, D.C., Wang, D., Chakraborty, S., 2015.Characterizing soils via portable X-ray fluorescence spectrometer: 4. Cation exchange capacity (CEC). Geoderma. 239–240, 130–134.

Soil Survey Staff. 2010. Keys to Soil Taxonomy, eleventh ed. United States Department of Agriculture, Washington.

Stenberg, B.,ViscarraRossel, R.A., Mouazen, A.M.,Wetterlind, J., 2010.Visible and near infrared spectroscopy in soil science. Adv. Agron. 107, 163–215.

Swan, A.R.H., Sandilands, M., 1995.Introduction to Geological Data Analysis, first ed. Blackwell Science, USA.

Taghizadeh-Mehrjerdi, R., Minasny, B., Sarmadian, F., Malone, B.P., 2014. Digital mapping of soil salinity in Ardakan region, central Iran. Geoderma. 112, 45-51.

Tajik, S., Ayoubi, S., Nourbakhsh, F., 2012. Prediction of Soil Enzymes Activity by Digital Terrain Analysis: Comparing Artificial Neural Network and Multiple Linear Regression Models. Environ. Eng. Sci. 29, 798- 806.

Taylor, J.A., McBratney, A.B., ViscarraRossel, R., Minasny, B., Taylor, H.J., Whelan, B.M., Short, M., 2006. Development of a Multisensor Platform for Proximal Soil Sensing.Proceedings of 18th World Congress of Soil Science, Philadelphia, Pennsylvania, USA.

Taylor, J.A., Short, M., McBratney, A.B., Wilson J., 2010. Comparing the ability of multiple soil sensors to predict soil properties in a Scottish potato production system, In: Viscarra Rossel, R.A., McBratney, A.B., Minasny, B. (Eds.), Proximal Soil Sensing: Progress in Soil Science, Springer Science +Business Media B.V., Netherlands, pp. 387–396.

The Ministry of Economy, Trade and Industry of Japan (METI) & theNational Aeronautics and Space Administration (NASA) 2012. AsterGlobal Digital Elevation Model (Aster GDEM). NASA Official [WWWdocument]. URL http://www.gdem.aster.ersdac.or.jp

Tosti, F., Patriarca, C., Slob, E., Benedetto, A., Lambot, S., 2013. Clay content evaluation in soils through GPR signalprocessing. J. Appl. Geophys.97, 69–80.

Tóth, B., Makó, A., Guadagnini, A., Tóth, G., 2012. Water Retention of Salt-Affected Soils: Quantitative Estimation Using Soil Survey Information. Arid Land Res Manage. 26, 103-121.

Wang, D., Laffan, S.W., 2009. Characterization of valleys from DEMs. 18th World IMACS/ MODSIM Congress, Cairns, Australia.

Wilson, J.P., Gallant, J.C., 2000. Terrain analysis. Wiley and Sons, New York.

Zhao, Z., Ashraf, M.I., Keys, K.S., Meng, F.R., 2013. Prediction of soil nutrient regime based on a model of DEM-generated clay content for the province of Nova Scotia, Canada. Can. J. Soil Sci. 93, 193-203.

Figure Captions

Fig. 1.Location of the study area (a) and GPR and EMI surveys in the selected sites (b), Bam district, Kerman province, southeastern Iran.

Fig. 2.Sampling scheme the GPR survey, soil sampling points (triangle), transects of the measurements with GPR (blue arrow), and sampling with EM (circles).

Fig. 3. Relationships between the predicted and measured clay content values in the two investigated depths for the testing data sets of the constructed MLR models using different data sets. a) $Clay_{(0\text{-}20\ cm)}$- topography and remote sensing data b) $Clay_{(20\text{-}40\ cm)}$- topography and remote sensing data, c) $Clay_{(0\text{-}20\ cm)}$- GPR and EM data, d) $Clay_{(20\text{-}40\ cm)}$- GPR and EM data, e) $Clay_{(0\text{-}20\ cm)}$- combined data, f) $Clay_{(20\text{-}40\ cm)}$- combined data

Fig. 4.Relationships between the predicted and measured clay content values for the testing data sets of the constructed CHAID models using different input data sets. a) $Clay_{(0\text{-}20\ cm)}$- topography and remote sensing data b) $Clay_{(20\text{-}40\ cm)}$- topography and remote sensing data, c) $Clay_{(0\text{-}20\ cm)}$- GPR and EM data, d) $Clay_{(20\text{-}40\ cm)}$- GPR and EM data, e) $Clay_{(0\text{-}20\ cm)}$- combined data, f) $Clay_{(20\text{-}40\ cm)}$- combined data.

Fig. 5.Factors affecting soil clay content variations in the study area resulted from the CHAID analysis. a) $Clay_{(0\text{-}20\ cm)}$ - topography and remote sensing data, b) $Clay_{(20\text{-}40\ cm)}$ - topography and remote sensing data, c) $Clay_{(0\text{-}20\ cm)}$ - geophysical data, d) $Clay_{(20\text{-}40\ cm)}$ - geophysical data, e) $Clay_{(0\text{-}20\ cm)}$ - combined data, f) $Clay_{(20\text{-}40\ cm)}$ - combined data

(EMV: Electrical conductivity in vertical; EMH: Electrical conductivity in horizontal; $GPR_{(0\text{-}15)}$:Amplitude in 0-15 cm; $GPR_{(15\text{-}30)}$:Amplitude in 15-30 cm; $GPR_{(30\text{-}45)}$:Amplitude in 30-45 cm; $GPR_{(0\text{-}20)}$:Amplitude in 0-20 cm; $GPR_{(20\text{-}40)}$:Amplitude in 20-40 cm; $GPR_{(40\text{-}0)}$:Amplitude in 40-60 cm; $GPR_{(10)}$:Amplitude in10 cm; $GPR_{(20)}$:Amplitude in20 cm; $GPR_{(30)}$:Amplitude in 30 cm; $GPR_{(40)}$:Amplitude in40 cm; TWI: Topographic Wetness Index; SPI: Stream Power Index; SR: Salinity Ratio; RVI: Ratio Vegetation Index PVI: Perpendicular Vegetation Index; NDVI: Normalized Difference Vegetation Index; MrVBF: Multi-resolution Valley Bottom Flatness index; MrRTF: multi resolution of ridge top flatness index; CI: Clay Index; CS: Catchment Area; CA: Catchment Slope)

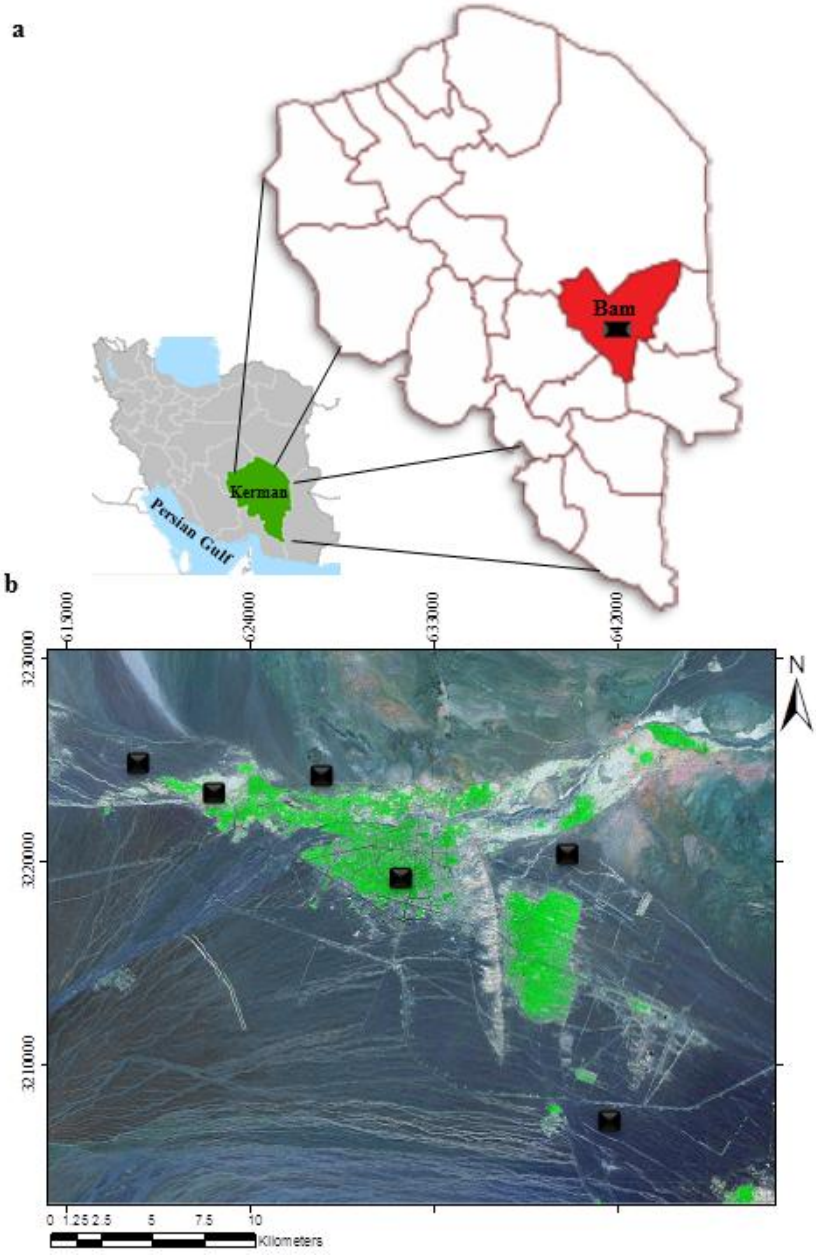

Fig. 1. Location of the study area (a) and GPR and EMI surveys in the selected sites (b), Bam district, Kerman province, southeastern Iran.

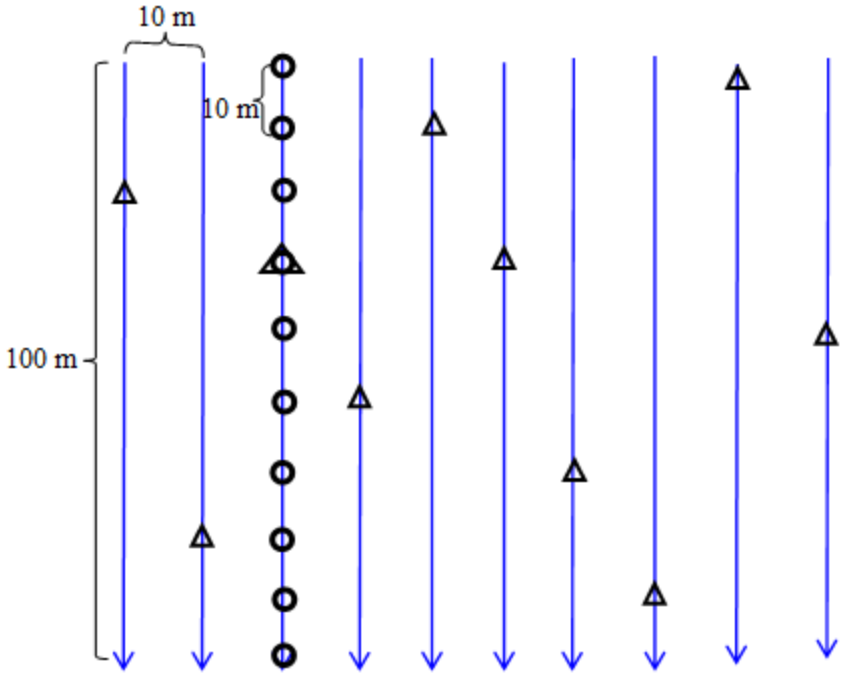

Fig. 2. Sampling schematic GPR survey, soil sampling points (triangle), transects of the measurements with GPR (blue arrow), and sampling with EM (circles).

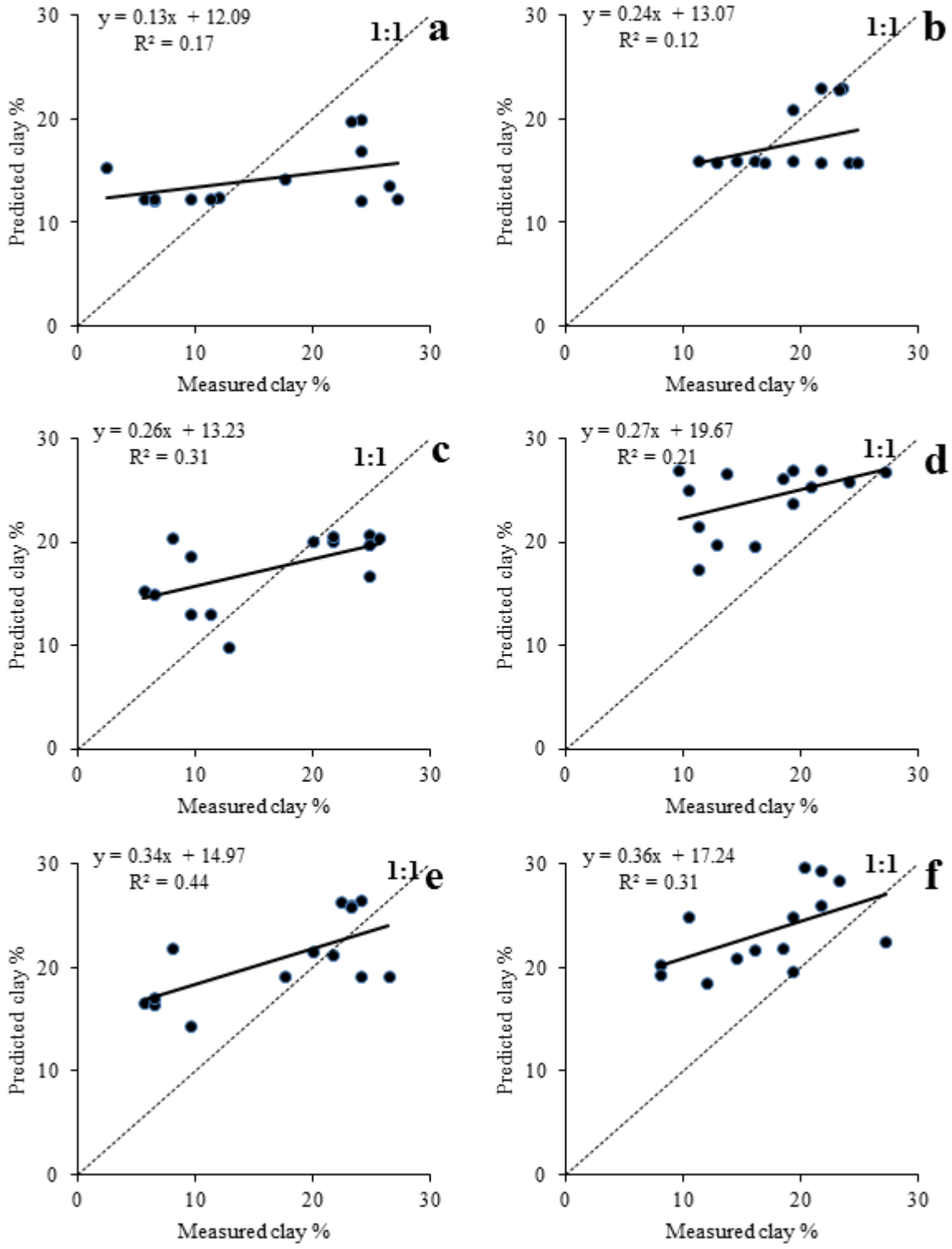

Fig. 3. Relationships between the predicted and measured clay content values in the two investigated depths for the testing data sets of the constructed MLR models using different data sets. a) $Clay_{(0\text{-}20\ cm)}$- topography and remote sensing data b) $Clay_{(20\text{-}40\ cm)}$- topography and remote sensing data, c) $Clay_{(0\text{-}20\ cm)}$- GPR and EM data, d) $Clay_{(20\text{-}40\ cm)}$- GPR and EM data, e) $Clay_{(0\text{-}20\ cm)}$- combined data, f) $Clay_{(20\text{-}40\ cm)}$- combined data

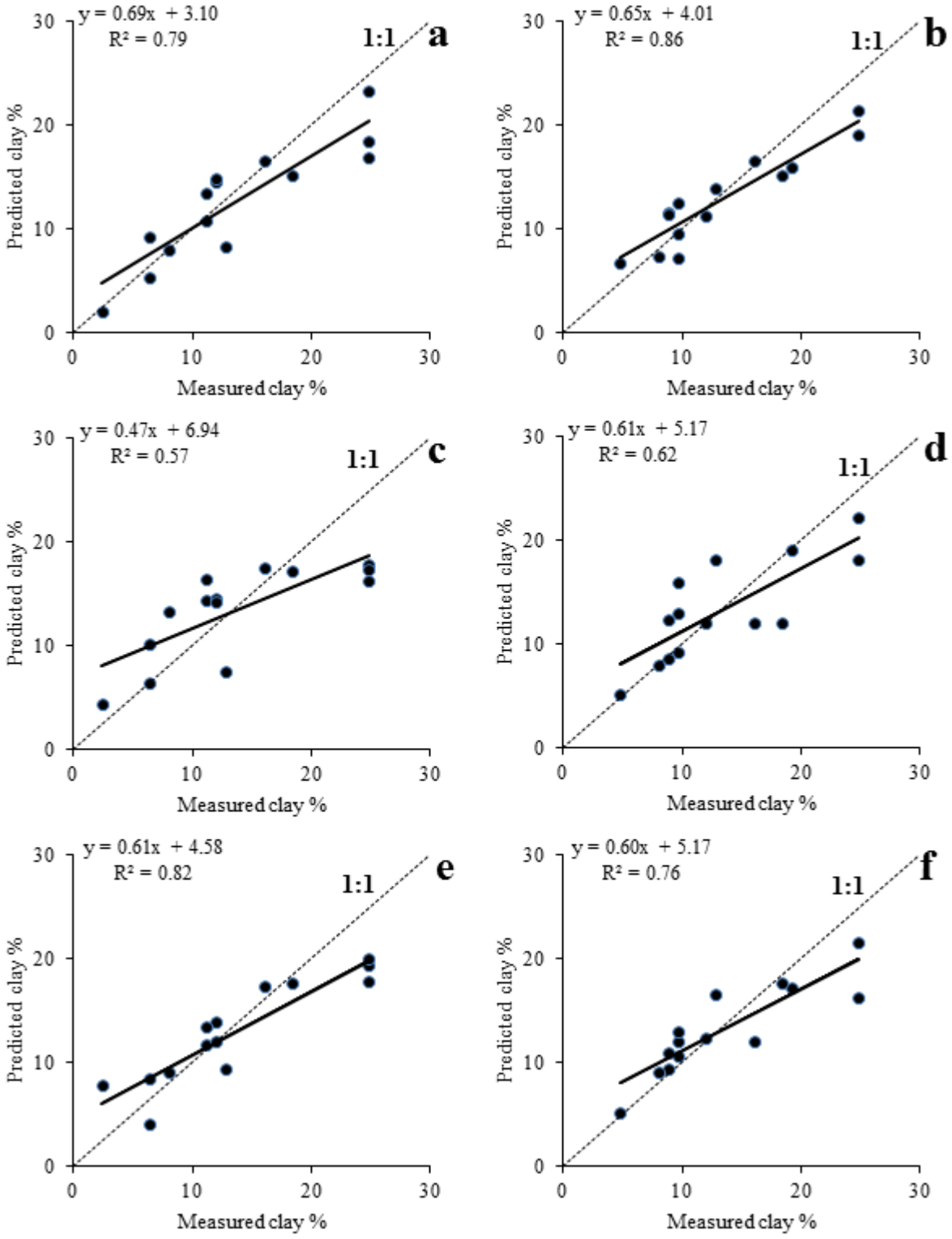

Fig. 4. Relationships between the predicted and measured clay content values for the testing data sets of the constructed CHAID models using different input data sets. a) $Clay_{(0\text{-}20\ cm)}$- topography and remote sensing data b) $Clay_{(20\text{-}40\ cm)}$- topography and remote sensing data, c) $Clay_{(0\text{-}20\ cm)}$- GPR and EM data, d) $Clay_{(20\text{-}40\ cm)}$- GPR and EM data, e) $Clay_{(0\text{-}20\ cm)}$- combined data, f) $Clay_{(20\text{-}40\ cm)}$- combined data.

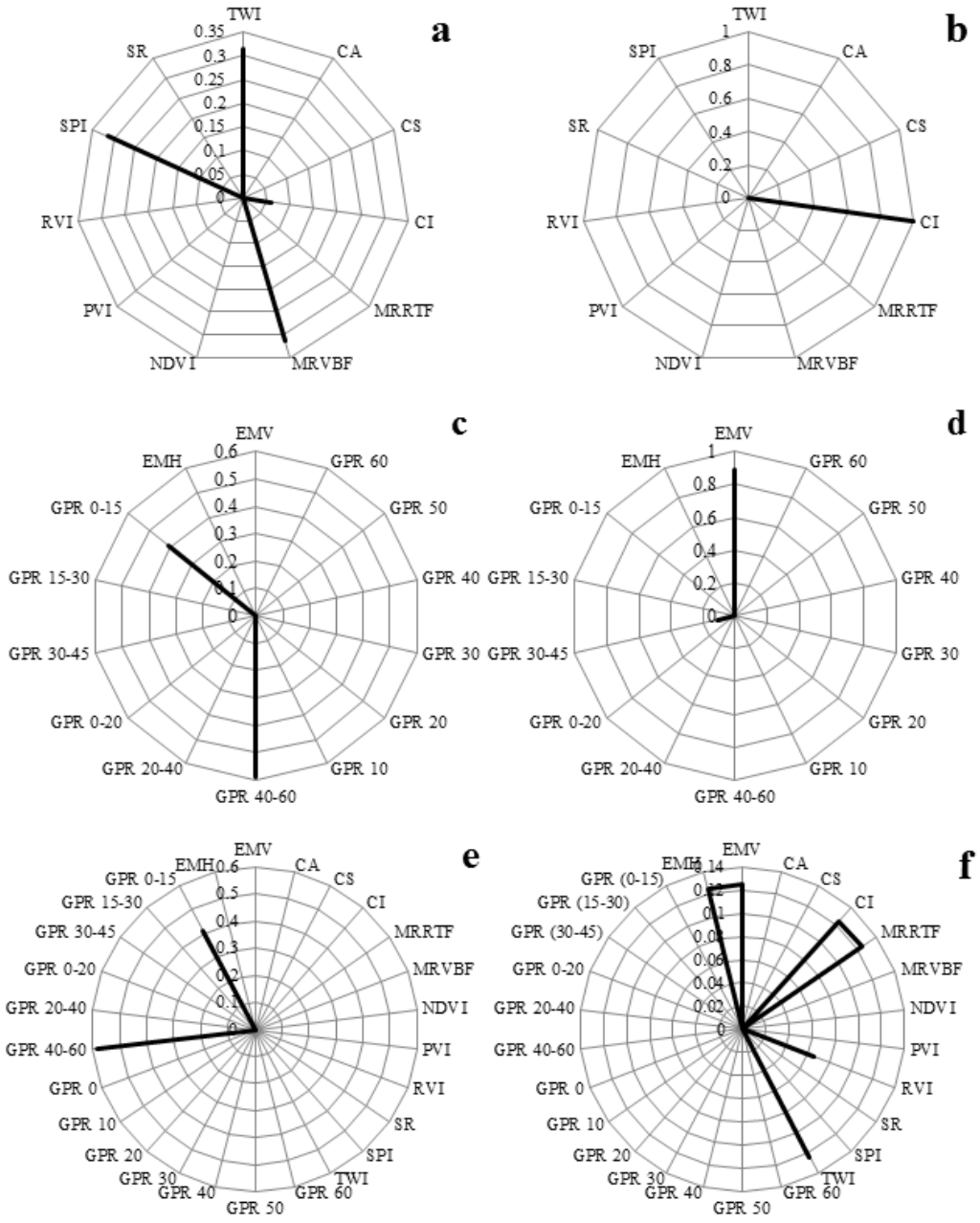

Fig. 5.Factors affecting soil clay content variations in the study area resulted from the CHAID analysis. a) $Clay_{(0\text{-}20\ cm)}$ - topography and remote sensing data, b) $Clay_{(20\text{-}40\ cm)}$ - topography and remote sensing data, c) $Clay_{(0\text{-}20\ cm)}$ - geophysical data, d) $Clay_{(20\text{-}40\ cm)}$ - geophysical data, e) $Clay_{(0\text{-}20\ cm)}$ - combined data, f) $Clay_{(20\text{-}40\ cm)}$ - combined data

:Amplitude in $_{)15\text{-}30(}$:Amplitude in 0-15 cm; $GPR_{)0\text{-}15(}$(EMV: Electrical conductivity in vertical; EMH: Electrical conductivity in horizontal; GPR :Amplitude in 40-60 $_{)40\text{-}0(}$:Amplitude in 20-40 cm; $GPR_{)20\text{-}40(}$:Amplitude in 0-20 cm; $GPR_{)0\text{-}20(}$:Amplitude in 30-45 cm; $GPR_{)30\text{-}45(}$15-30 cm; GPR :Amplitude in40 cm; TWI: Topographic $_{)40(}$:Amplitude in 30 cm; $GPR_{)30(}$:Amplitude in20 cm; $GPR_{)20(}$:Amplitude in10 cm; $GPR_{)10(}$cm; GPR Wetness Index; SPI: Stream Power Index; SR: Salinity Ratio; RVI: Ratio Vegetation Index PVI: Perpendicular Vegetation Index; NDVI: Normalized Difference Vegetation Index; MrVBF: Multi-resolution Valley Bottom Flatness index; MrRTF: multi resolution of ridge top flatness index; CI: Clay Index; CS: Catchment Area; CA: Catchment Slope)

Table 1. Land surface attributes used for spatial prediction of soil clay content

Auxiliary data	Land surface parameters	Definition	Reference/source
Terrain attributes	Multi-resolution Ridge-top Flatness Index (MrRTF)	Measure of flatness and lowness	Gallant and Dowling (2003)
	Multi-resolution Valley Bottom Flatness Index (MrVBF)	Measure of flatness and lowness	Gallant and Dowling (2003)
	Topographic wetness index (TWI)	Ln (Area/tan(slope))	SAGA GIS
	Catchment slope (CS)	Average gradient above flow path	SAGA GIS
	Catchment area (CA)	Top-down processing of flow accumulation	SAGA GIS
	Stream Power index (SPI)	CA * tan(Slope)	SAGA GIS
Remote sensing data	Normalized Difference Vegetation Index (NDVI)	(B4 − B3)/(B4 + B3)	Boettinger et al. (2008)
	Clay index (CI)	B5/B7	Boettinger et al. (2008)
	Salinity ratio (SR)	(B3 − B4)/(B2 + B4)	Metternicht and Zinck (2003)
	Ratio Vegetation Index (RVI)	B4/B3	Pearson and Miller (1972)
	Perpendicular Vegetation Index (PVI)	-sin a (B4)* cos a (B3)	Richardson andWiegand (1977)

Table 2. Descriptive statistics of the soil clay content and geophysical properties used for developing the prediction models

Variables	Unit	Min	Max	Mean	Skewness	Kurtosis	K-S	CV (%)
$Clay_{(0\text{-}20\ cm)}$	%	2.4	27.0	16.1	-0.07	-1.05	0.20*	45.41
$Clay_{(20\text{-}40\ cm)}$	%	4.8	37.0	15.9	0.69	0.74	0.20*	40.90
EMV	mS/m	0.0	137.0	28.7	1.48	1.10	0.00	126.80
EMH	mS/m	2.00	88.0	30.3	1.46	1.24	0.00	68.12
$GPR_{(0\text{-}15)}$	Volt	866.4	11909.4	3826.9	1.43	2.60	0.01	59.17
$GPR_{(15\text{-}30)}$	Volt	1083.0	23334.0	5955.4	2.08	6.88	0.00	64.07
$GPR_{(30\text{-}45)}$	Volt	1815.3	17117.3	6307.3	1.20	1.44	0.00	53.12
$GPR_{(0\text{-}20)}$	Volt	789.7	17736.7	4429.5	1.99	5.79	0.01	69.11
$GPR_{(20\text{-}40)}$	Volt	1000.3	21297.6	6115.7	1.86	5.05	0.03	58.75
$GPR_{(40\text{-}60)}$	Volt	2643.3	16507.6	7176.6	0.81	0.25	0.04	44.10
$GPR_{(10)}$	Volt	1012.0	18837.0	4667.3	1.97	5.68	0.00	69.89
$GPR_{(20)}$	Volt	338.0	23685.0	5870.0	2.09	6.95	0.00	66.46
$GPR_{(30)}$	Volt	431.0	22267.0	6111.7	1.86	5.36	0.00	61.78
$GPR_{(40)}$	Volt	613.0	16654.0	6289.2	1.06	1.05	0.02	54.63

K-S: Kolmogorov-Smirnov test; CV: coefficient of variation; EMV: Electrical conductivity in vertical; EMH: Electrical conductivity in horizontal; $GPR_{(0\text{-}15)}$:Amplitude in 0-15 cm; $GPR_{(15\text{-}30)}$:Amplitude in 15-30 cm; $GPR_{(30\text{-}45)}$:Amplitude in 30-45 cm; $GPR_{(0\text{-}20)}$:Amplitude in 0-20 cm; $GPR_{(20\text{-}40)}$:Amplitude in 20-40 cm; $GPR_{(40\text{-}0)}$:Amplitude in 40-60 cm; $GPR_{(10)}$:Amplitude in10 cm; $GPR_{(20)}$:Amplitude in20 cm; $GPR_{(30)}$:Amplitude in 30 cm; $GPR_{(40)}$:Amplitude in40 cm;
*: Significant at 95% probability level

Table 3. Correlation coefficients between the clay content and the geophysical, topography, and remote sensing properties.

Variables	$Clay_{(0\text{-}20\ cm)}$	$Clay_{(20\text{-}40\ cm)}$	Variables	$Clay_{(0\text{-}20\ cm)}$	$Clay_{(20\text{-}40\ cm)}$
EMV	0.43**	0.48**	TWI	-0.42**	-0.30*
EMH	0.33**	0.31*	SPI	-0.16	-0.29*
$GPR_{(0\text{-}15)}$	-0.41**	-0.21	SR	-0.32**	-0.33**
$GPR_{(15\text{-}30)}$	-0.23*	-0.31*	RVI	0.32**	0.31*
$GPR_{(30\text{-}45)}$	-0.27*	-0.28*	PVI	0.39**	0.17
$GPR_{(0\text{-}20)}$	-0.26*	-0.35**	NDVI	0.33**	0.33**
$GPR_{(20\text{-}40)}$	-0.29*	-0.31*	MrVBF	-0.36**	0.04
$GPR_{(40\text{-}60)}$	-0.27*	-0.28*	MrRTF	-0.09	0.11
$GPR_{(10)}$	-0.26*	-0.28*	CI	0.38**	0.31**
$GPR_{(20)}$	-0.24*	-0.32**	CS	0.28*	0.01
$GPR_{(30)}$	-0.23*	-0.26*	CA	-0.28*	-0.10
$GPR_{(40)}$	-0.25*	-0.26*	ELV	0.20	0.10

EMV: Electrical conductivity in vertical; EMH: Electrical conductivity in horizontal; $GPR_{(0\text{-}15)}$:Amplitude in 0-15 cm; $GPR_{(15\text{-}30)}$:Amplitude in 15-30 cm; $GPR_{(30\text{-}45)}$:Amplitude in 30-45 cm; $GPR_{(0\text{-}20)}$:Amplitude in 0-20 cm; $GPR_{(20\text{-}40)}$:Amplitude in 20-40 cm; $GPR_{(40\text{-}0)}$:Amplitude in 40-60 cm; $GPR_{(10)}$:Amplitude in10 cm; $GPR_{(20)}$:Amplitude in20 cm; $GPR_{(30)}$:Amplitude in 30 cm; $GPR_{(40)}$:Amplitude in40 cm; TWI: Topographic Wetness Index; SPI: Stream Power Index; SR: Salinity Ratio; RVI: Ratio Vegetation Index PVI: Perpendicular Vegetation Index; NDVI: Normalized Difference Vegetation Index; MrVBF: Multi-resolution Valley Bottom Flatness index; MrRTF: multi resolution of ridge top flatness index; CI: Clay Index; CS: Catchment Area; CA: Catchment Slope, ELV: Elevation

"*" and "**" are significant at 95 and 99 % probability level.

Table 4. Factors affecting soil clay content variations in the study area resulted from the stepwise linear regression analysis.

Target variable	Input data	selected variables
$Clay_{(0\text{-}20\ cm)}$	Topography and remote sensing data	NDVI, MrVBF, TWI
$Clay_{(20\text{-}40\ cm)}$	Topography and remote sensing data	NDVI
$Clay_{(0\text{-}20\ cm)}$	GPR and EMI data	$GPR_{(0\text{-}15)}$
$Clay_{(20\text{-}40\ cm)}$	GPR and EMI data	EMV, $GPR_{(40\text{-}60)}$
$Clay_{(0\text{-}20\ cm)}$	Combined data	$GPR_{(0\text{-}15)}$, NDVI, MrVBF
$Clay_{(20\text{-}40\ cm)}$	Combined data	$GPR_{(0\text{-}20)}$, NDVI

EMV: Electrical conductivity in vertical; $GPR_{(0\text{-}15)}$:Amplitude in 0-15 cm; $GPR_{(0\text{-}20)}$:Amplitude in 0-20 cm; $GPR_{(40\text{-}0)}$:Amplitude in 40-60 cm; TWI: Topographic Wetness Index; NDVI: Normalized Difference Vegetation Index; MrVBF: Multi-resolution Valley Bottom Flatness index; CI: Clay Index.

Table 5. Goodness-of-fit of the proposed CHAID and MLR models for the prediction of clay content (%) in the two study depths.

Model	Variable	Input data	Evaluation criterion			
			ME	RSME	AEP	R^2
MLR	$Clay_{(0\text{-}20\ cm)}$	Topography and remote sensing data	0.10	6.32	37.42	0.17
	$Clay_{(20\text{-}40\ cm)}$	Topography and remote sensing data	2.53	6.60	44.25	0.12
	$Clay_{(0\text{-}20\ cm)}$	GPR and EMI data	3.64	8.13	50.58	0.33
	$Clay_{(20\text{-}40\ cm)}$	GPR and EMI data	11.63	13.89	91.29	0.21
	$Clay_{(0\text{-}20\ cm)}$	Combined data	4.93	8.80	55.68	0.44
	$Clay_{(20\text{-}40\ cm)}$	Combined data	8.86	11.42	73.90	0.31
CHAID	$Clay_{(0\text{-}20\ cm)}$	Topography and remote sensing data	0.65	3.43	19.10	0.79
	$Clay_{(20\text{-}40\ cm)}$	Topography and remote sensing data	0.54	2.67	16.67	0.86
	$Clay_{(0\text{-}20\ cm)}$	GPR and EMI data	0.27	4.66	28.86	0.57
	$Clay_{(20\text{-}40\ cm)}$	GPR and EMI data	0.11	3.76	20.94	0.62
	$Clay_{(0\text{-}20\ cm)}$	Combined data	0.67	3.41	19.93	0.82
	$Clay_{(20\text{-}40\ cm)}$	Combined data	0.22	3.17	17.85	0.76

MEE: mean estimation error, RMSE: root mean square error, AEP: absolute error percentage, R^2: coefficient of determination, CHAID: Chi-Squared Automatic Interaction Detection, MLR: multiple linear regression.

Highlights

- Geophysical data were significant variables for predicting clay content
- CHAID models showed greater potential in predicting soil clay as compared to MLR.
- Geophysical data provide a novel tool for quantitative estimation of clay content