Digital soil mapping in a selected arid landscape in southeastern Iran

By

Azam Jafarisirizi

February 2012

Table of Contents

Acknowledgements

I am deeply thankful to my God, for providing the opportunity of study. I wish to thank the members of my committee, Dr. Shamsollah Ayoubi, Prof. Dr. Hossein Khademi, and Prof. Dr. Peter Finke, for their guidance and support throughout my research and encouragement during the task of completing my thesis. I am deeply thankful to my advisors, Prof. Dr. Ahmad Jalalian, Dr. Nourair Toumanian, and Dr. Mohammad Hadi Farpour, for their advice, patience, and support throughout my research. I also express gratitude to my committee chair, Dr. Mostafa Karimian Eghbal, Dr. Mehran Salehi, and Prof. Dr. Mohammad Ali Hajabbasi, for having the time to read and provide important comments to my thesis.

To my mom and dad, thank you for your support and words of encouragement. I thank my husband, Sayyed Shahabaldin Nooraldini, for his support throughout the thesis process.

I thank Johan Van de Wauw for helping in using R software and also thank Ann Zwertvaegher for moral support in Belgium.

Lastly, I would like to extend my sincere gratitude to all those in the Department of Soil Science at Isfahan University of Technology and the Department of Geology and Geography at Ghent University in providing me the opportunity to pursue the PhD's program.

List of Tables

List of Figures

Chapter 1
Research necessity and overall objectives

Geographic information science (GIS) and technology have great potential to improve the efficiency and quality of methods used to gather spatial soil information (McBratney and Odeh, 1997). Technological advances in GIS and remote sensing have created a tremendous potential for improvement in soil resource inventory (McKensie et al., 2000, Bui and Moran, 2001). Information on the distribution of soil properties over the landscape is required for a variety of hydrological, ecological and land management applications. In detailed hydroecological and other environmental modeling applications, considering variability of soil properties over an area is needed to approximate the resolution of other environmental parameters gathered from remote sensing and digital terrain analysis (Band et al. 1991, 1993). Unfortunately, traditional soil maps seldom provide information about the spatial distribution of soil properties at the desired resolution (both at spatial and attribute levels) and this soil information is very difficult to directly obtain over large areas as soils show inherently high and continuous spatial variation, and are often obscured by a vegetation canopy.

Traditional soil survey methods for mapping soil distribution are outdated because they were formulated prior to the introduction of computer based geographic information systems and remote sensing techniques. Computer based approaches to digital soil modelling combine remotely sensed data, terrain analysis data, field-collected data, vegetation, climate and

lithology distribution and expert knowledge to infer soil characteristics at unvisited sites. Digital soil mapping (DSM) is the computer-assisted production of digital maps of soil type and soil properties (Schull et al., 2005). It typically implies the use of mathematical and statistical models that combine information from soil observations with information contained in correlated variables and remote sensing images (Figure 1). Such methods hold promise to reduce survey cost and increase objectivity, while producing soil landscape information more appropriate for the systems science applications.

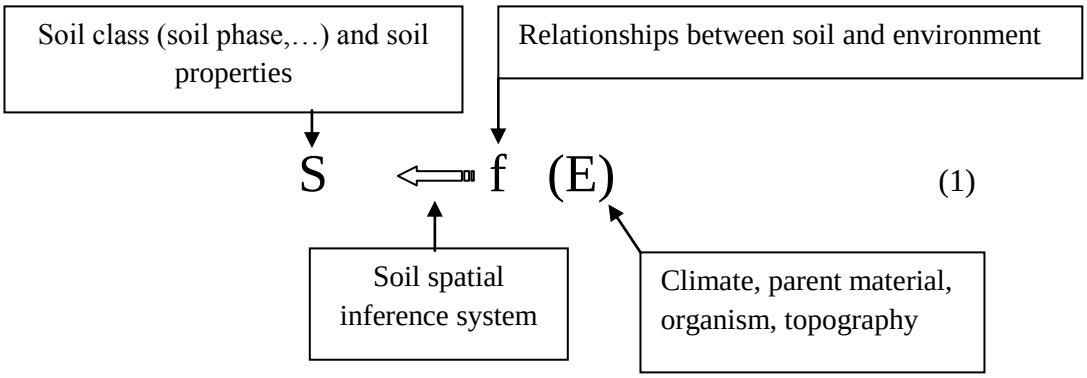


Figure1. Soil is a function of its environmental factors (Zhu et al., 1997)

In Iran, most soil surveys have been carried out using traditional methods and some areas have not yet been mapped at any scale. Recently, industry, agriculture, and mining sectors have increasingly focused on the application of geographic information systems and, as a result, digital soil data are now being collected more systematically. Foreseen intensive applications in agriculture and hydrological management demand high quality soil maps. Although predictive soil mapping studies are still at an introductory stage in Iran, they provide a start point especially in arid regions, where traditional soil survey methods are difficult to undertake.

The primary focus of this dissertation is to develop and test DSM methods using soil and environmental data at the study site located in southeastern Iran. This study is limited to those landforms commonly associated with the desert environment of southeastern Iran. Soils of

many parts of Iran have never been mapped; particularly there is no soil map available for the area study. Since, the knowledge on the soil resources is vital for land management decisions in arid regions, the specific objective of this dissertation is to develop models of spatial soil information (soil map unit, soil taxa, and soil properties) that can be used to produce more accurate soil maps for the survey area and preliminary maps of non-mapped regions. Little attempt has been made to evaluate the capability of DSM approach in Iran (Hengl, et al. 2007), and no investigation has been carried out in desert landscape of central and southeastern Iran.

To carry out digital soil mapping, Iranian soil scientists are faced with areas without any data and soil map. On the other hand, based on the principal of digital soil mapping techniques, there should be soil observations (soil profile data). Under such circumstances and also, due to the high cost of field work, digital soil mapping should be performed with limited data.

We attempted to produce a map of the USDA soil great groups, based on a limited set of field data as a desirable starting point at a scale of 1:50,000 to provide a base for mapping at the soil series scale which is common in Iran. In the present study, we evaluate the suitability and performance of logistic regression methods and boosted regression tree as potential techniques for soil mapping using a limited point dataset in an arid region of Iran. Furthermore, we evaluate the trade-off between the quality of maps produced at different taxonomic levels using neural networks and the information value of these maps. Since a vast area of the world is covered by similar desert soils, the results of this research could be used for soil mapping in such area for which very few data sources are available.

Chapter 2

Spatial prediction of USDA-soil great groups in arid Zarand region, Iran: comparing logistic regression approaches to predict diagnostic horizons and soil types

Accepted by European Journal of Soil Science

2.1. Introduction

Conventional soil mapping methods are efficient in medium to low intensity surveys because they use relationships among soil properties and more readily observable environmental features as a basis for mapping. However, the implicit predictive models are qualitative, complex and rarely described in a clear manner. Therefore, developing an explicit analogue of conventional survey practice suited to medium to low intensity surveys is of great importance (McKenzie and Ryan, 1999). A key feature is the use of quantitative environmental variables from digital terrain analysis and remote sensing to predict the spatial distribution of soil properties and classes. The use of these technologies for quantitative soil survey has been illustrated in various studies (McBratney et al., 2003; Lagacherie et al., 2007). Prediction of soil properties based on information from environmental variables has always been the basis for all the soil mapping methods. Unfortunately, traditional methods do not yield quantifiable soil-landscape information that describes robustly actual soil variation (Scull et al., 2003).

Soils cannot be separated from the landscapes in which they form. As landscapes evolve, soils develop through the interaction of pedogenic and geomorphic processes. The differences in soil type with landscape position are usually attributed to differences in runoff, erosion and deposition processes which affect soil genesis (Canton et al., 2003). Several studies in arid and semi-arid areas indicate that soils have a wide range of spatial variability resulting from differences in parent material, age of land surface, topography, water distribution, amount and intensity of rainfall and plant heterogeneity (Shmida and Burgess, 1988; Canton et al., 2003). Therefore, information obtained from these differences could be used to study and identify various soils.

The predictive soil distribution models can be generated on the base of the relationships between soil and relief classes and geomorphological units (Grinand et al., 2008). Logistic regression methods have been successfully used in soil science for determining the probability of occurrence of drainage classes (Campling et al., 2002) or to relate the soil types with terrain attributes (Debella-Gilo et al., 2009) for example. Hengl et al. (2007) found that the success of multinomial logistic regression in predicting WRB soil groups heavily depends on the correlation with predictors for all classes.

Although some researchers have successfully used logistic regression, this technique has rarely been used to map taxonomic classes, particularly for areas with limited data and where no reference soil map is available. Logistic regression models are well-known, conceptually simple and easy to interpret: therefore, to facilitate the preparation of digital soil maps, their use seems appropriate.

2.2. Objectives

We attempted to produce a map of the USDA soil great groups, based on a limited set of field data as a desirable starting point at a scale of 1:50,000 to provide a base for mapping at

the soil series scale which is common in Iran. This raised the following questions: (i) how to obtain a great-group map of sufficient quality using a limited point data set? (ii) what environmental co-variables have large predictive power? and (iii) is direct prediction of soil type better than the prediction of diagnostic horizons followed by classification into great groups?

This study was therefore conducted to answer these questions by implementing and comparing selected digital soil mapping techniques both to predict soil classes directly and, alternatively, to predict the occurrence of relevant diagnostic horizons followed by classification of the indicator maps with the USDA soil classification system.

2.3. Methods

2.3.1. Study area

The study area is located in the Zarand region, southeast Iran, about 70 km from the city of Kerman, between 56-57° E longitude and 30-31° N latitude and covers an area of about 90 000 ha. This area is surrounded by mountains (limestone, dolomite, shale) from northwest to southeast. Major landforms in the study area include alluvial fans, coalescing alluvial fans (Bajadas), salt plain (playa), gypsiferous hills and sand dunes (Figure 1). The soil moisture regime of the study area is Aridic. The mean annual precipitation, temperature and potential evapotranspiration are 61 mm, 22° C and 2500 mm, respectively.

2.3.2. Sampling design and profile description

A stratified sampling scheme was adopted for the study area using digital maps of geology, geomorphology and topography for stratification. The sampling design aimed to provide a good spatial coverage of the area and also cover the spatial variability of environmental variables to be used for prediction.

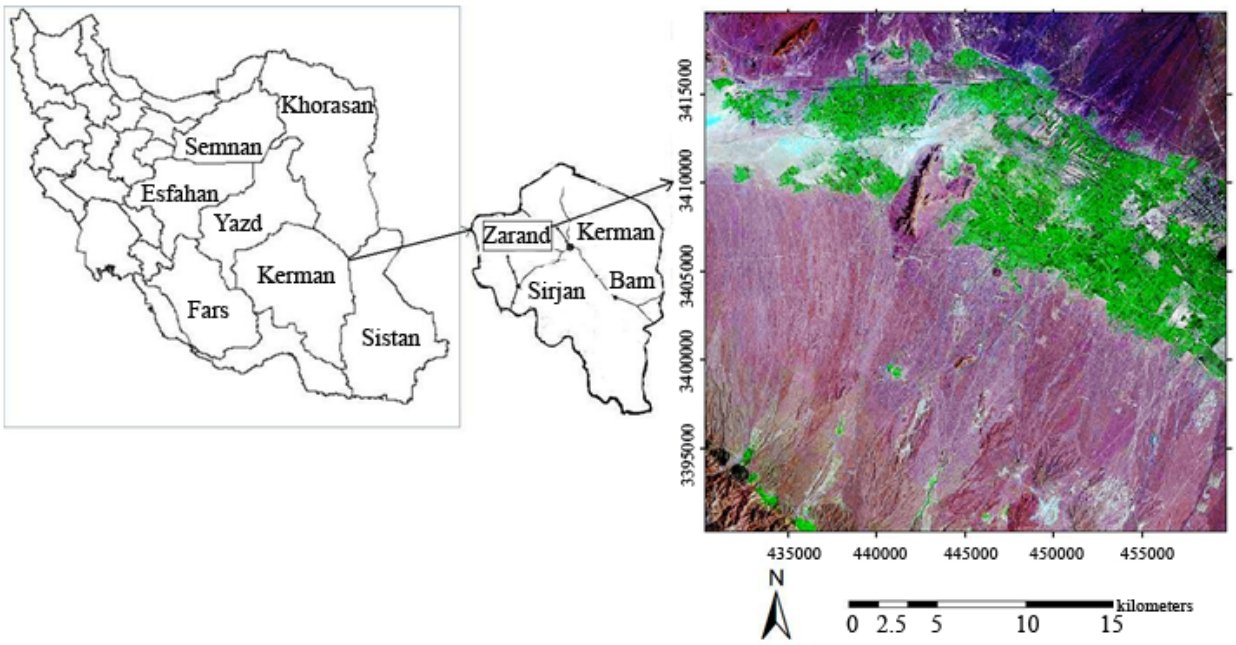


Figure 1. The study area located near Kerman city, Iran (Landsat ETM+ image (RGB: 243))

The sampling strata were defined to represent the differences in landforms (geomorphology), topography (DEM) and lithology. Within each stratum, sampling locations were randomly chosen so that the sample size was proportional to the stratum area. This resulted in 126 profiles, which were then described, sampled, analyzed and classified by using the USDA soil classification system (Soil Survey Staff, 2006). The sampling locations are shown on the geomorphology map in Figure 2. The most abundant great groups in each geomorphic unit are shown in Table 1 and the profile description was used to compile a list of diagnostic surface and subsurface horizons likely to occur in the area.

2.3.3. Ancillary spatial variables

Three groups of ancillary variables were employed in mapping as follows:

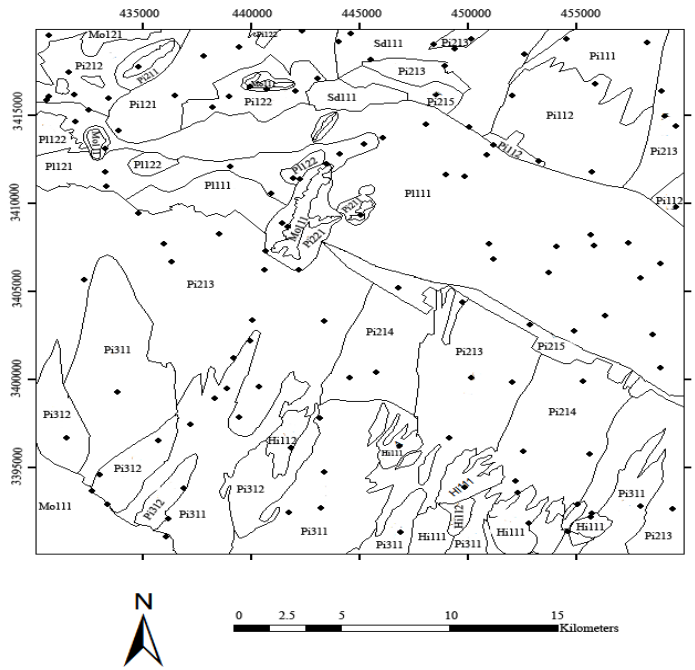


Figure 2. Geomorphology map and sampling points for mapping and validation. Codes refer to Table 1

(i) Terrain attributes

A digital elevation model (DEM) was compiled from the Aster Global Digital Elevation Model website (METI and NASA, 2009). Seven terrain attributes were obtained from the DEM including slope (SLOPE), mean curvature (MEANC), wetness index (WI), plan curvature (PlCur), profile curvature (PrCur), topographic wetness index (TWI) and Multi-resolution valley bottom flatness index (MrVBF) (Gallant and Dowling, 2003). All the terrain characteristics were derived by using the Saga GIS (Olaya, 2004).

(ii) Remote sensing indices

The Landsat 7 ETM data were used to extract remote sensing indices such as normalized difference vegetation index (NDVI) (Rouse et al., 1973), ratio vegetation index (RVI) (Pearson and Miller, 1972) and perpendicular vegetation index (PVI) (Richardson and Wiegand, 1977).

Table 1. Geomorphology map hierarchy and the major soil great group per geomorphic surface (profile description) for the study area in Zarand region, Iran

No	Landscape	Landform	Lithology	Geomorphic surface	Code	Major soil great group observed
1	Mountain	Rock outcrop	Dolomite-limestone	Rock surface	Mo111	Torriorthents
2	Mountain	Eroded outcrop	Sandstone, shale	Eroded surface	Mo121	Torriorthents
3	Hill	Eroded outcrop	Conglomerate-sandstone-gypsum	Dendrite drainage system with high topography	Hi111	Haplogypsids
4	Sand hills	Dune	Wind sediments	Parabolic stream	Sd111	Torripsamments
5	Playa	Clay flat	Silt, Clay, Salt	Cultivated clay flat	Pl111	Haplosalids
6			Fine and coarse alluvial sediments	Clay flat, highly salty and wetness	Pl121	Haplosalids
7				Salty and wetness, dense stream	Pl122	Haplosalids
8	Piedmont	Alluvial fan	Alluviums of limestone, shale, sandstone, Igneous rocks	Active fan, upper section	Pi111	Haplocalcids
9				Active fan, lower section, low slope	Pi112	Haplocalcids
10			Alluviums of siltstone, shale, sandstone, quartz	Active fan, upper section	Pi121	Calcigypsids
11				Active fan, lower slope	Pi122	Haplocalcids
12	Piedmont	Bajada	Alluviums of siltstone, shale, sandstone, limestone, igneous rocks	Upper section, high slope	Pi211	Haplocalcids
13				Upper section, dense drainage system, low slope	Pi212	Haplogypsids
14				Lower section, low slope,	Pi213	Haplocalcids
15				Lower section, low slope, new parallel streams, new deposits	Pi214	Haplocambids
16				Cultivated bajada	Pi215	Haplosalids
17	Piedmont	Dissected bajada	Alluviums of siltstone, shale, sandstone, limestone, igneous rocks	Flat and lower topography with dense streams	Pi311	Haplocambids
18				Higher topography and deep streams, upper section	Pi312	Haplocambids

(iii) Geomorphology map

Air photo interpretation (API) was used to differentiate geomorphical entities on the basis of their formation processes, general structure and morphometry. The entities were defined through a nested geomorphic hierarchy as defined by Toomanian et al. (2006). This approach uses a four-level geomorphic hierarchy to breakdown the complexity of different landscapes. Therefore, the geomorphology map has four levels including landscape, landform, lithology, and geomorphic surface. This hierarchy was delineated on aerial photos (1:40 000).

Stereoscopically interpreted aerial photos of the study area were imported into a GIS environment and after ortho-photo geo-referencing, geomorphic surfaces were mapped and inserted in GIS via on-screen digitization. There was a total of 18 geomorphic surfaces in the study area (Figure 2 and Table 1).

All the maps of the terrain attributes, remote sensing images and geomorphology were projected to the same geographic reference system (such as WGS 84 UTM 40N). The values of the terrain attributes, remote sensing indices and levels of geomorphology map were then converted into a table for all the point locations with observed soil horizons and USDA soil great groups. This table was imported to R software (R Development Core Team, 2011) for predictive mapping.

2.3.4. Predictive mapping with logistic regression

We considered two approaches for mapping the target variable: indirect prediction and direct prediction of the great groups.

In indirect prediction, the occurrence of relevant diagnostic horizons was first mapped, and subsequently, the indicator maps were combined on a pixel basis by using the presence or absence of diagnostic horizons. The primary target variable was thus binary, and the auxiliary variables could be either quantitative or categorical. In such a situation, binary logistic regression is an appropriate prediction method.

(i) Binary logistic regression as an indirect approach

For predicting soil classes by binary logistic regression, the soil diagnostic horizons were first predicted and then these horizons were combined for the prediction of soil great groups.

The theory and applications of the logistic model in soil science has been reviewed by Lane (2002). For the logistic model, the diagnostic horizon as our response variable is treated as a binomial having values of 0 or 1 for the absence or presence of the horizon, respectively:

$$ln\left(\frac{p}{1-p}\right) = \beta_0 + \sum_{i=1}^{n} \beta_i * x_i \qquad (2)$$

where p is the probability of occurrence of a diagnostic horizon, β_0 is the intercept, β_i is the regression coefficient, and x_i is an independent variable.

In cases of the absence of diagnostic horizons, soils might be classified as Entisols. Therefore, we added indicator variables with the names of Entisols sub-orders as predictants (for example, Psamments).

To select among the fitted models, the Akaike Information Criterion (AIC) (Akaike, 1973) was used (Guisan et al., 2007). This adjusts the residual deviance for the number of predictors, thus favouring parsimonious models. Thus, the model with the smallest AIC and residual deviance was selected; ANOVA was conducted to evaluate predictor importance as it requires the minimal processing time (Behrens et al., 2010).

To allow combination of the indicator maps for classification into great groups, a decision tree must be defined which links the occurrence of diagnostic horizons to a soil great group. This tree was formulated with the classification key of the Soil Taxonomy (Soil Survey Staff, 2006). The occurrence of diagnostic horizons was decided on the basis of exceedance of threshold values for predicted probabilities of occurrence. The threshold values for each indicator map were selected based on receiver operator characteristic (ROC) curve (Figure 3).

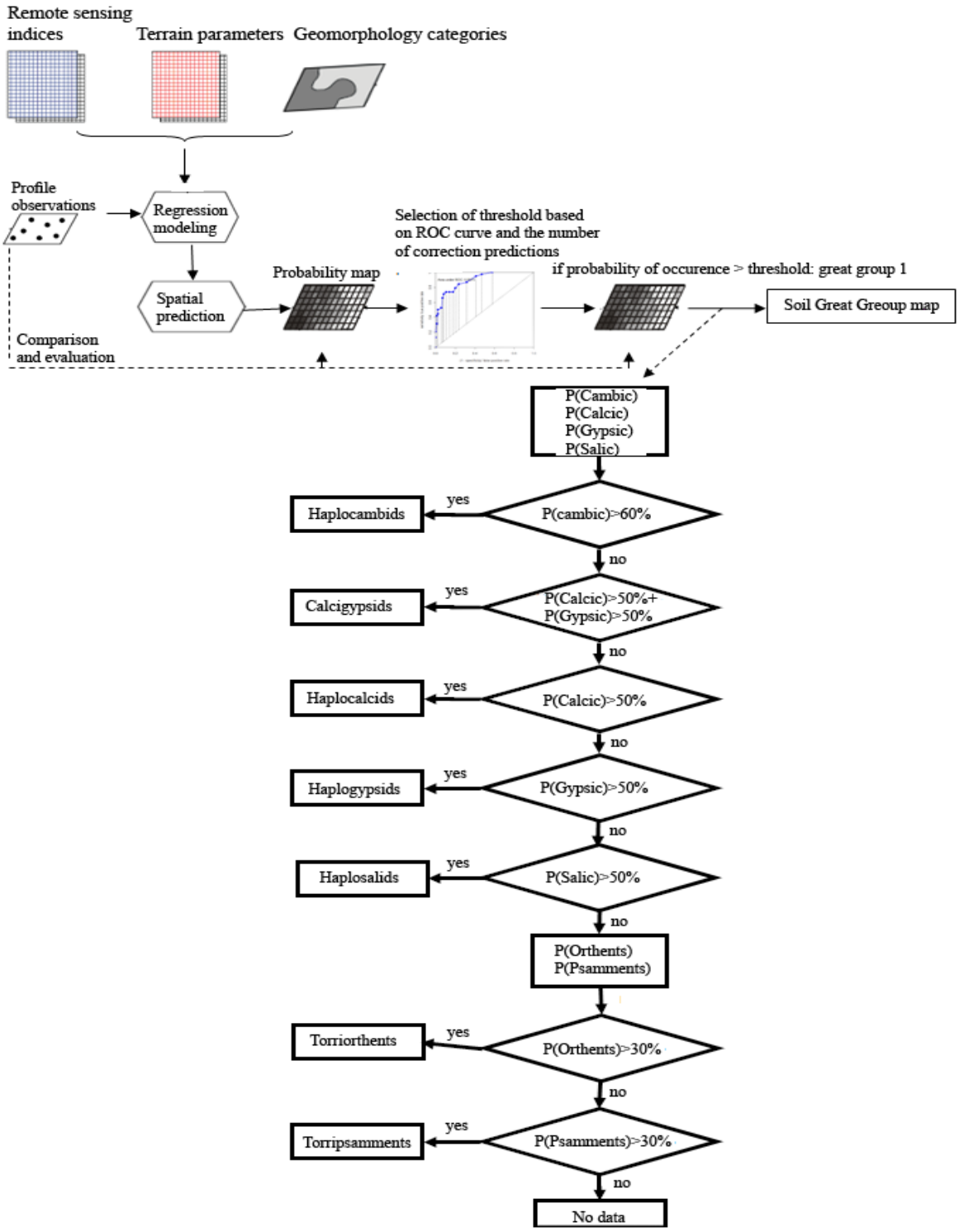


Figure 3. Flowchart of activities for the indirect prediction of soil great groups from mapped occurrences of diagnostic horizons of soils from the study area, using profile observations and auxiliary information in a binary logistic regression

The ROC allows the determination of the best combination of sensitivity and specificity in the identification of occurrences (Pontius and Schneider, 2001). We used the ROC to graphically compare the distributions of probability for recorded presence and absence at the sample locations by scanning the probability range [0; 1] for a series of cut-off values. For each cut-off value, the true and false positives were counted as follows:

(i) the true positives (sensitivity) of the probability distribution of recorded presence:

$$Sensitivity = 1 - \Pr(y \le c | D = 1) \qquad (3)$$

and (ii) the false positives (1-specificity) of the probability of recorded absence:

$$(1 - Specificity) = 1 - \Pr(y \le c | D = 0) \qquad (4)$$

where y is the test value (in this case, the mapped probability of the occurrence of a diagnostic horizon), c is the cut-off value and D is the field value taking the value 1 for presence and 0 for absence (Finke et al., 2008).

These counts were plotted in the ROC curve. Subsequently, the area under the curve (AUC) (Rossiter and Loza, 2010) was estimated by numerical integration as a measure of the overall quality of the map.

An AUC of 1 indicates a perfect map, and AUC of 0.5 indicates that the map does not have any discriminative power to detect presence or absence of a diagnostic horizon. The cut-off values were chosen so that AUC was maximal and the number of correct predictions at field positions was maximal (Manel et al., 2001).

(ii) Multi-nomial logistic regression (MLR) as a direct approach

Multi-nomial logistic regression was used to model the relationships the soil great groups (categorical dependent variables) and the terrain attributes, remote sensing indices and levels of the geomorphology map (quantitative predictors) with the 'nnet' package of R. This

estimated iteratively coefficients for all predictors. A multi-nomial logistic regression model with reference category is expressed as follows:

$$log\left(\frac{\pi_{ij}}{\pi_{iJ}}\right) = \alpha_j + \beta_j x_i, j = 1,2, \ldots\ldots, J-1 \qquad (5)$$

where α_j is a constant, β_j is a vector of regression coefficients, for j=1, 2, …, J-1 and x_i is a vector of explanatory variables. This model is analogous to a logistic regression model, except that the probability distribution of the response is multi-nomial instead of binomial and there are J-1 equations instead of one so that

$$P(y_i = j) = \pi_{ij} = \frac{exp(\beta_j x_i)}{1+\sum_{j=1}^{j} exp(\beta_j x_i)} \qquad (6)$$

Then, the probability of reference category is given by

$$P(y_i = 0) = \pi_{iJ} = \frac{1}{1+\sum_{j=1}^{j} exp(\beta_j x_i)} \qquad (7)$$

The dependent variable has more than two categories. All the soil great groups occurring in the data set are possible categories. Before running the model, the reference class must be selected; in this case we used Calcigypsids (first in the default alphabetic order). The significance of the regression coefficient of each predictor variable for each dependent variable was evaluated using the Wald statistic. This statistic tests whether changes in a given predictor variable lead to significant change in the odds ratio of the dependent variable (Hosmer and Hjort, 2002). Thus, we can infer the significance of each predictor. The greater the absolute value, the greater significance it has.

2.3.5. Validation and statistical inference

(i) Model validation

The most extreme form of cross-validation, known as the leave-one-out approach, was used. Each regression model was fitted by using n-1 observations and the soil group/horizon

was predicted at the observation site which was not used. The prediction with the unused observation was validated and this sequence was repeated n times for the other observation sites. Then the validation indices were estimated using the n validation results. The result of validation was an indicator variable taking value 1 if the predicted soil great group was equal to the observed soil great group and was 0 otherwise.

(ii) Statistical inference

The observation sites are a stratified simple random sample (Kempen et al., 2009). Strata were defined by ancillary variables such as geomorphology, geology and topography maps. The resulted geomorphic surfaces were representative of the differences in geomorphology, geology and topography of the landforms. Subsequently, these resulted in 18 strata and 126 locations allocated to the strata in proportion to their area, with a minimum of two per stratum. According to sampling design and defined strata a weighted purity should be calculated as the actual purity. The actual purity was estimated as suggested by Kempen et al. (2009) in the following form:

$$\hat{f} = \sum_{h=1}^{l} w_h \hat{f}_h \,, \qquad (8)$$

where $w_h$is the weight (relative area) of stratum h, $\hat{f}_h$ is the estimated areal fraction of stratum h correctly classified, and l is the number of strata. The stratum fractions were estimated by the fraction of correctly predicted locations in each stratum,

$$\hat{f}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} y_i \,, \qquad (9)$$

where n_h is the number of random sampling locations in the stratum h, and y_i is the indicator variable at sampling location i, taking the value 1 if the predicted soil great group was equal to the observed soil great group and was 0 otherwise.

The simplest statistical criterion in design based sampling is the unweighted purity used to estimate the global purity (P_g) which is the estimated fraction of soil taxon x or stratum h correctly predicted. Therefore

$$(P_g)_x = \frac{\sum_{i=1}^{n_x} y_i}{n_x}, \quad (10)$$

where $(P_g)_x$ is global purity of the soil taxon x and n_x is sample size of the soil taxon x.

The bulk purity (P_b) is the estimated fraction of soil taxa correctly predicted in total sampling units

$$(P_b) = \frac{\sum_{i=1}^{n} y_i}{n}, \quad (11)$$

where n is number of total sampling units in the area.

The difference in purity between two methods for the whole area can be assessed with indicators calculated from data. The indicators were determined by comparing the field classification with binary and multinomial classification and then the difference between indicators was calculated by $d_i = y_i^b - y_i^m$ where y_i^b is an indicator for the correct prediction by the binary method and y_i^m is a similar indicator for the multinomial method. This variable d can have values of -1, 0 and 1 and was used to estimate $\hat{d}$, which is the mean difference in actual purity of the binary and multinomial methods. Using the calculated standard error of the estimated mean of difference d, we tested whether differences were significant or not. The variance of the mean value for d ($\hat{V}\left(\hat{\bar{d}}_{St}\right)$) was calculated according to De Gruijter et al. (2006):

$$\hat{V}\left(\hat{\bar{d}}_{St}\right) = \sum_{h=1}^{H} a_h^2 \, \hat{V}(\hat{\bar{d}}_h), \quad (12)$$

where $a_h$is the relative area of stratum h, $\hat{\bar{d}}_h$ is the stratum mean of d and $\hat{V}(\hat{\bar{d}}_h)$ is the estimated variance of $\hat{\bar{d}}_h$ that can be calculated as follows

$$\hat{V}\left(\hat{\bar{d}}_h\right) = \frac{1}{n_h(n_h-1)}\sum_{i=1}^{n_h}(d_{hi} - \hat{\bar{d}}_h)^2 \,, \qquad (13)$$

where n_h is the sample size in stratum h and d_{hi} is indicator variable in stratum h. $\hat{V}(\hat{\bar{d}}_h)$ can be used to construct confidence intervals and to test if d significantly differs from 0, in which case the binary and multi-nomial classification methods perform significantly different.

2.4. Results and discussion

2.4.1. The soil-geomorphology-terrain relations modelled by binary logistic regression (an indirect approach)

The geomorphology map at the fourth scale (geomorphic surface) was a powerful predictor unlike the other scales (landscape, landform and lithology) that were not important in model fitting (Table 2). This can be explained by the fact that the geomorphic surfaces have formed recently, or during a geological period with soil formation with conditions close to those of current processes in the arid regions.

Table 2. The variables used to predict soil diagnostic horizons (indirect approach) and soil great groups (direct approach)

	Soil horizon or soil class	Variables in modelling	Most significance variables based on likelihood test
Indirect approach	Salic	GS+ NDVI+ MrVBF+ PVI	GS***, NDVI**,PVI*,MrVBF**
	Gypsic	GS+ NDI+ PlCur+ NDVI+ WI+ PVI	GS***,WI**, PlCur**,PVI*,NDI*
	Calcic	GS+ PVI	GS***,PVI*
	Cambic	El+ Sl+ PVI+ TWI+ MrVBF	El*,Sl*,PVI**,MrVBF**,TWI*
	Psamments	GS	GS***
	Orthents	El+ GS	El***,GS*
Direct approach	Soil great groups	GS+ MrVBF+ El+ WI	GS***,MrVBF***, WI**, El*

GS: Geomorphic Surface, PC: Plan Curvature, El: Elevation, Sl: Slope. Significance code: '***P< 0.001, **P< 0.01, '*P< 0.05

For most soil horizons, the logistic model was significant at P<0.05. Table 2 gives the most significant correlations between the spatial distribution of the diagnostic horizons and the geomorphic surface, elevation, PVI and MrVBF. The terrain attributes were the most effective

characteristics in predicting the diagnostic horizons (Table 3), indicating that the relief is the most important factor explaining the formed soils. All the diagnostic horizons, except cambic, were significantly correlated with the geomorphic surface.

Table 3. Purity of maps of the probability of occurrence for 4 diagnostic horizons (H) and 2 great groups, made by the binary logistic (indirect) approach

Data set	**Salic H.**	**Gypsic H.**	**Calcic H.**	**Cambic H.**	**Orthents**	**Psamments**
Proportion of profiles correctly predicted	27/38	50/59	25/37	35/47	4/6	7/7
Actual purity	0.52	0.73	0.41	0.61	0.52	1.00

.The indicator maps of the occurrence of diagnostic horizons are presented in Figure 4. The probabilities that a given diagnostic horizon or soil class occurs at a given pixel are represented by values between 0 and 1, where 0 (black areas) indicates a probability of 0% and 1 (white areas) a probability of 100% for a diagnostic horizon or soil class. Also, because Psamments only occur in sand dunes, the prediction of this soil class was limited to this geomorphic surface (Table 2). Thus, the geomorphic surfaces predict Psamments directly.

The prediction maps of gypsic and salic diagnostic horizons have high probabilities at the boundary of playa and bajada regions (centre to east areas of Figure 4). It appears that this area receives soluble salts washed out from upper areas. As expected, the large probability of the presence of salic horizon occurred in the playa landform (geomorphic surfaces Pl111, Pl121 and Pl122) (Figure 4). Among these geomorphic surfaces, the greatest probability of salic horizon was observed in the western side of playa (Figure 4), showing the greatest degree of salinity. Also, the poor vegetation cover is a clear evidence for the occurrence of salic horizon (Figure 1).

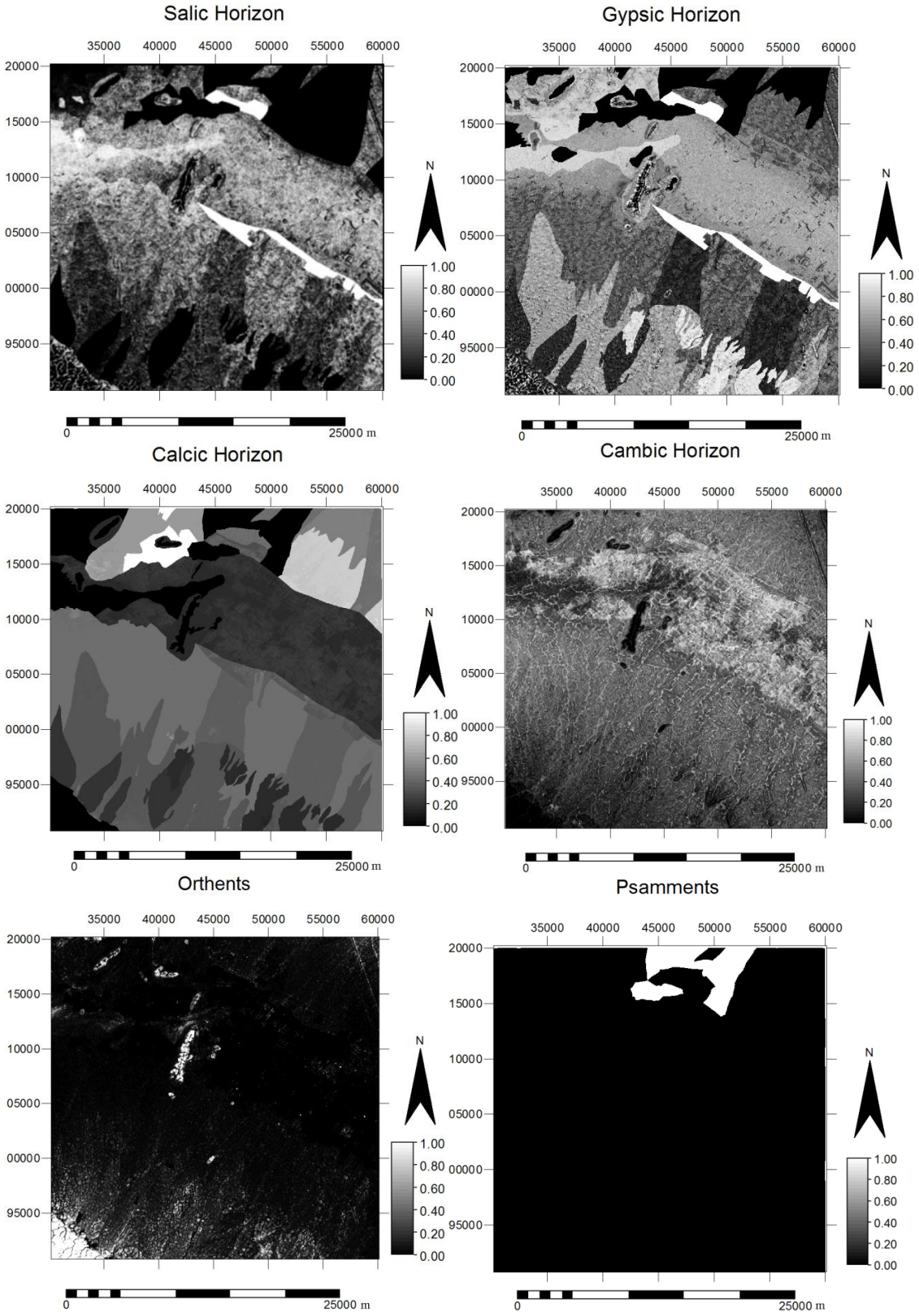


Figure 4. The probability distribution of diagnostic horizons of soils used in the prediction of soil great groups in the study area

The greatest probability of the gypsic diagnostic horizon occurs in gypsiferous hills that consist of 50-70% gypsum (data not shown) (Figure 4). Also, these hills are the likely source of gypsum for low-lying areas. This is only based on field observation and profile descriptions.

Calcic horizons were predicted in alluvial fan and bajada landforms (Figure 4), which is in accordance with observed profiles. The presence of calcic horizon in these landforms depends on soil stability in the sampling location as calcic horizons were not observed in lateral slopes.

Therefore, the geomorphic processes and surfaces should be differentiated in more detail and the geomorphologic criteria should be better defined to include the major processes. The resulting geomorphology map as the base of sampling and the representation of variation will be more accurate for determining the sampling design. Prediction maps of Orthents and Psamments, as expected, reflected the occurrence of eroded rock surfaces and presence of unconsolidated deposits, respectively. Rocky surfaces (mountain) and unconsolidated deposits (sand dunes) are clearly identified in Figure 1.

Finally, the diagnostic horizons were combined (Figure 4) to classify soils into great groups (Figure 5). Hengl et al. (2007) reported the distribution of Gypsisols and Solonetz soils with a high probability in southeast and central Iran at the regional scale. The results of this study are consistent with their findings but at a more detailed scale.

2.4.2. Determining threshold value by using ROC curve

The ROC curve is a graph of the sensitivity (proportion of true positives) of the model prediction plotted against the complement of its specificity (the proportion of false positives), at a series of threshold values for a positive outcome (Rossiter and Loza, 2010).

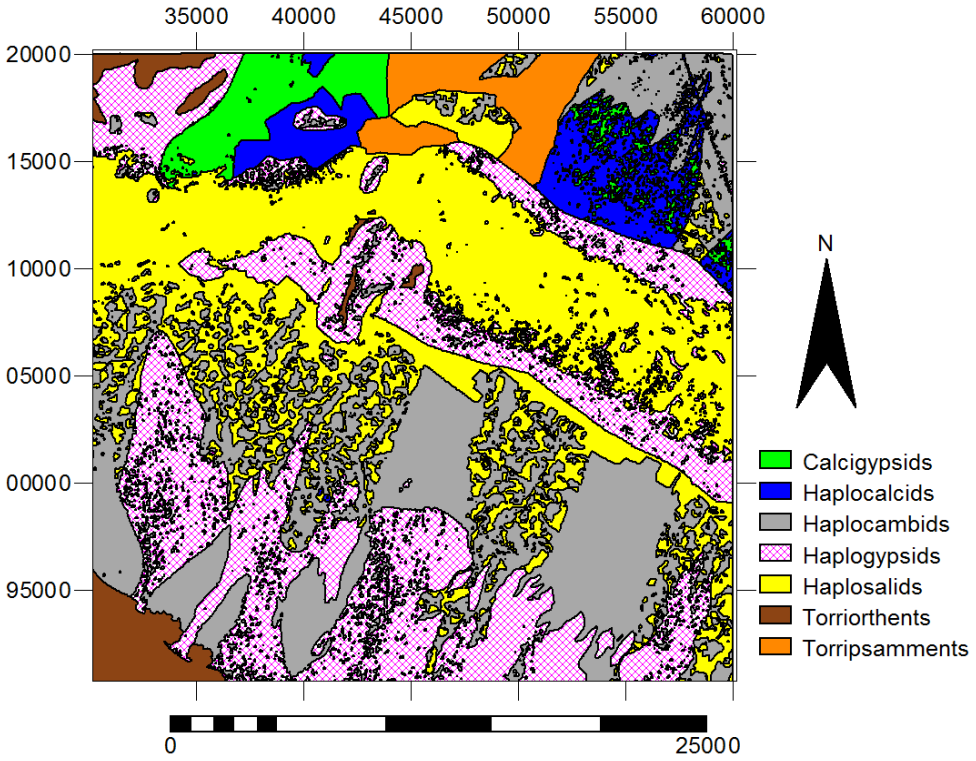


Figure 5. The spatial distribution of soil great groups in the study area derived from binary logistic regression

The selection of best model was based on the smallest AIC, the largest AUC and the number of correctly predicted classes. We therefore had to select the threshold value so that these options were fulfilled. As an example, we explain how to select the threshold value for the classification of gypsic horizon. Figure (6) shows the AUC for selected logistic model of the gypsic horizon. The graph clearly shows that this criterion for model with the variables given in Table 3 is much better than when the geomorphic surface is used as the single predictor. The AUC had the largest value among other fitted models for the gypsic horizon (results not shown), while AIC had the smallest for this model (Figure 6). Therefore, we selected the model with largest AUC and smallest AIC. However, this does not provide the best threshold value or the threshold with the smallest error.

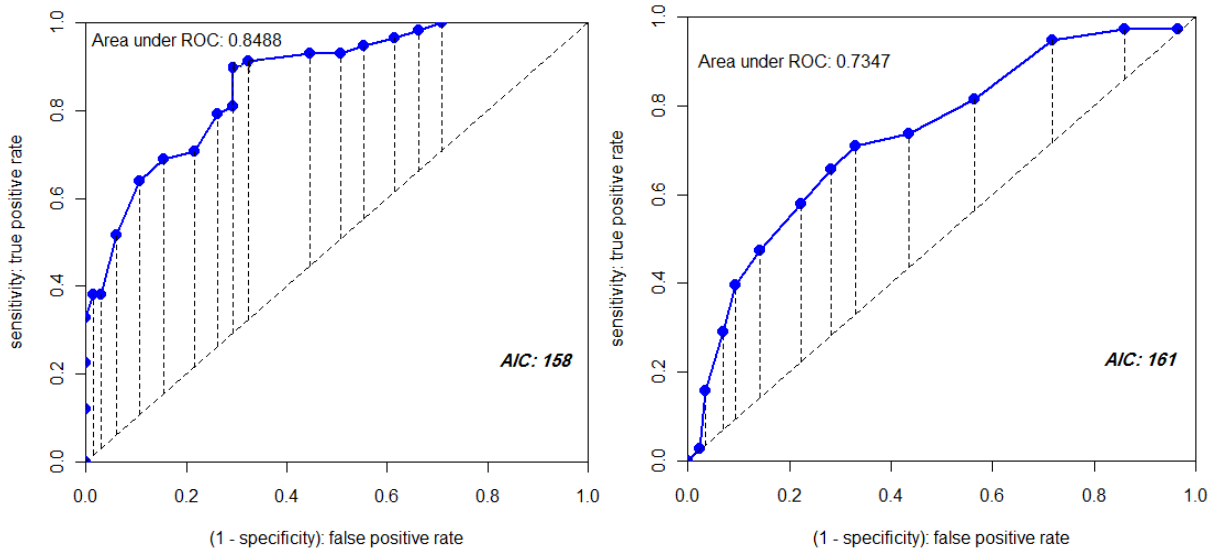


Figure 6. Receiver Operator Characteristic (ROC) curve and Area Under Curve (AUC) for prediction of gypsic horizon by binary logistic regression (left-hand curve for logistic model with variables shown in Table 3 and right-hand curve for logistic model using only geomorphic surfaces

At any threshold, we can compute the sensitivity and specificity, by comparing the predicted and actual classes. The relationships of the threshold with sensitivity and specificity are shown in Figure (7). This indicates that the best result is obtained at the threshold 0.5, because in this example we have 58 observed gypsic horizon (presence=1) and 65 observed non-gypsic horizon (absence=0) (Figure 8). Figure 8 shows at a threshold of p = 0.5, that 46 (of 58 present) and 48 (of 65 absent) were correctly predicted. With any increase or decrease of the threshold value, the predicted correct number of classes decreases (Figure 7). We therefore selected a threshold value of 0.5 for classification of gypsic horizon. The same methodology was followed for other diagnostic horizons.

2.4.3. The soil-geomorphology-terrain relationships modelled by multi-nomial logistic regression (a direct approach)

Multi-nomial logistic regression directly predicts the soil great groups from the predictors. The parsimonious model for prediction was selected in a similar way to binary

logistic regression on the basis of the smallest AIC and residual deviance. Therefore, important predictors were identified including geomorphic surface, MrVBF, elevation and wetness index (Table 2).

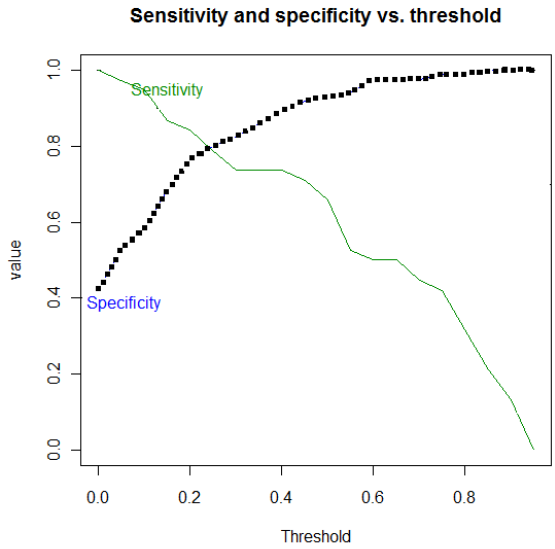
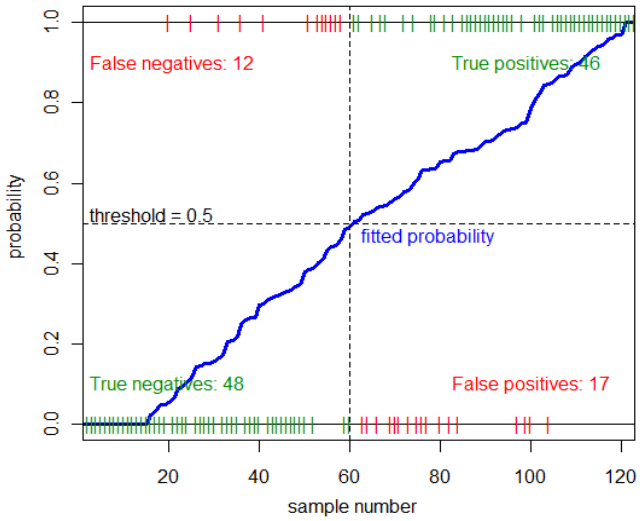


Figure 7. Sensitivity and specificity plotted against thresholds values for the predictions by the logistic model for the gypsic horizon



Sensitivity: 0.7931 ; Specificity: 0.7385

Figure 8. The logistic model prediction for the gypsic horizon arranged by probability with the samples corresponding to each probability either at the top (if sample is actually present) or the bottom (if sample is actually absent)

None of the remote sensing indices were involved in the prediction. The geomorphology map had the potential to increase mapping efficiency as categorical predictor in both direct and indirect approaches. After the geomorphic surface, MrVBF and WI were the most important predictors (Table 2). Multi-resolution Valley Bottom Flatness index (MrVBF) identifies flat valley bottoms and WI indicates the degree of wetness (Wang and Laffan, 2009). These indices indicate potential zones of transport for many materials, particularly sediment and other materials in excess water flow (Whiteway et al., 2004). Therefore, MrVBF and WI act indirectly in the identification of some of great groups particularly Haplosalids, Haplocalcids and Calcigypsids. The effects of geomorphology processes on soil distribution pattern and development have been widely recognized (Birkeland et al., 2006; Golosov et al., 2008). The dominant role of geomorphologic processes in determining soil classes has been shown by Scull et al. (2005). Using multi-nomial logistic regression, Debella-Gilo and Etzelmuller (2009) showed that the terrain attributes exerted a significant effect on the distribution of soil groups. This terrain attributes and geomorphic processes help to predict soil types in both regression approaches. In turn, soil types result from different formation factors in the study area. The soil great group map derived from multi-nomial regression is shown in Figure 9.

2.4.4. Comparison of predictive models

The prediction purity was reasonably good for all the diagnostic horizons (Table 4), as the binary logistic model identified fairly strong relationships between the predictors and diagnostic horizons. In other words, predictors were useful to predict diagnostic horizons as purities for each diagnostic horizon are above 0.5, with the exception of the calcic horizon which had an actual purity of 0.41. The poor purity of calcic horizons is probability due to the

lack of a clear relation between current geomorphology and the formation of a calcic horizon, which is a slow pedogenic process. While a relatively poor prediction of Orthents may result from having only few sampling points (six) in relation to the area, the field identification of Orthents is easier in comparison with that for calcic horizons because this is a hidden characteristic.

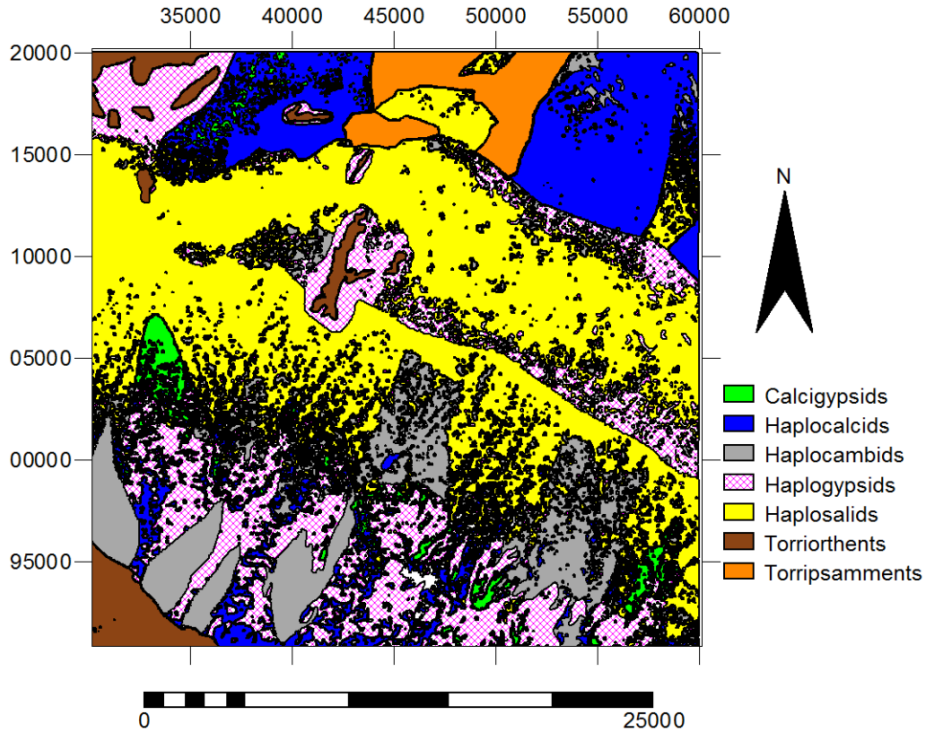


Figure 9. The spatial distribution of soil great groups in the study area derived from multi-nomial logistic regression.

The map purity for soil taxa and strata is summarized in Table 4. The bulk purity of multi-nomial logistic was 66%, which was 7% larger than the binary logistic which is probably due to the classification resulting from the decision tree. The purity of predictions was not the same for all the soil classes in both approaches (Table 4). Generally, the actual purity was less than the global purity for the soil taxa (Table 4). This can be related to the effect of the size of sampling units relative to the total study area. The purities for Haplogypsids and Haplosalids with larger sample size are better than those for other soil

groups. Poorer performance (global and actual purity) was observed for Calcigypsids and Haplocambids in both binary and multinomial methods (Table 4). It is likely that predictions with a high degree of uncertainty are the result of an incomplete conceptual model. There were no diagnostic properties for Haplocambids that can help us classify it clearly. Also, because intensive erosion and sedimentation occur in arid regions, soil differentiation is very difficult and even more powerful models would not provide great accuracy.

Table 4. Predictive quality of binary logistic regression (indirect approach) and multinomial logistic regression (direct approach) for soil taxa and geomorphic strata in the study area

		Binary Logistic Reg.			Multinomial logistic Reg.		
	n	Bulk purity	Global purity	Actual purity	Bulk purity	Global purity	Actual purity
	126	0.59			0.66		
Soil taxa							
Calcigypsids	6		0.28	0.02		0.40	0.03
Haplocalcids	21		0.60	0.29		0.43	0.32
Haplocambids	21		0.60	0.24		0.43	0.22
Haplogypsids	28		0.78	0.51		0.56	0.43
Haplosalids	36		0.70	0.60		0.80	0.75
Torriorthents	6		0.83	0.08		1.00	0.10
Torripsamments	7		1.00	0.06		1.00	0.06
Geomorphic strata							
Mo111	3		0.67	0.02		1.00	0.02
Mo121	3		1.00	0.02		1.00	0.02
Hi111	7		0.86	0.06		0.57	0.04
sd111	7		1.00	0.06		1.00	0.06
Pl111	21		0.80	0.40		0.70	0.30
Pl121	6		0.83	0.10		0.83	0.10
Pl122	2		0.00	0.00		1.00	0.03
Pi111	2		0.00	0.00		0.00	0.00
Pi112	5		0.20	0.01		0.60	0.04
Pi121	5		0.20	0.02		0.60	0.06
Pi122	4		1.00	0.03		1.00	0.03
Pi211	6		0.67	0.08		0.83	0.10
Pi212	5		0.60	0.12		0.6	0.14
Pi213	23		0.56	0.37		0.56	0.37
Pi214	7		0.71	0.08		0.71	0.08
Pi215	3		0.67	0.02		0.67	0.02
Pi311	12		0.50	0.20		0.40	0.20
Pi312	5		0.60	0.04		0.80	0.08

Weak prediction for Calcigypsids can be attributed to several factors. First, the small sample size of this soil class was because it was difficult to identify because of the undifferentiated surfaces and processes. Therefore, the delineations do not occur in the stratification and thus are not imported in the sampling design and subsequently, the small number of locations involves very uncertain purity estimates (Kempen et al., 2009). The soil great group Calcigypsids was observed in geomorphic surfaces Pi111, Pi112 and Pi121 with the small number of sampling units.

Subsequently, the purities of these geomorphic surfaces are also poor (Table 4) confirming the poor prediction of Calcigypsids. Second, designing the decision tree in the binary logistic method makes it difficult to identify the threshold values used for the classification of soil classes in such a vast area (Figure 3). This introduces uncertainty in the identification of the prediction model and subsequent classification. In addition, bias in sampling (lack of complete characterization of predictor space with respect to the response variable) contributes to uncertainty (Beaudette and O'Geen, 2009).

Sample size, size of study area and map scale, all affect the prediction performance. Because the distribution of the samples in this study was stratified randomly over strata determined from ancillary data (Table 2), not all the soil types present were equally represented. This may have affected the results and explains why the prediction of soils with very limited presence in the research area had poorer accuracy than those with greater representation in the sample data because they had a larger spatial representation (Table 4).

Comparison of the probability maps of gypsic horizon and Orthents (Figure 4) indicates that, in some parts, the same regions with large probability ($P>0.5$) exist. It is possible that some Torriorthents derived from binary logistic (indirect approach) under the decision tree (Figure 3) were predicted as Haplogypsids (Figure 5). This prediction may have been

influenced by the spatial distribution of light reflectance caused by high altitude gypsiferous hills and mountains. Poor reliability may also result from the small number of samples and, consequently, the lack of good relationships between predictors and the soil great groups, particularly in the indirect approach using the decision tree.

Both methods provided good prediction for Haplosalids, as shown by large values for global and actual purity (Table 4). The accurate prediction of Haplosalids is explained by good spatial correlation with indices such as the wetness index, NDVI and, especially, the playa land-form in the centre of the study area. Debella-Gilo and Etzelmuller (2009) showed that the high probability areas for each soil great group coincided with the theoretically known landscapes. Remote sensing (light reflectance) parameters were influential in predicting Haplosalids, Haplogypsids and Torriorthents, because of overlapping areas in the probability map of diagnostic horizons (Figure 4). Remote sensing separated Haplosalids from Haplogypsids (Figure 5), because of different altitude and as a result, different light reflectance. Haplosalids are mostly found in playa, which has a salty surface, and occur at the lowest altitudes while Haplogypsids commonly occur in gypsiferous hills at a much higher altitudes. The greatest purities were observed in the geomorphic surfaces with the most sampling units (Table 4). One of these units was the Pl111 geomorphic surface which had the largest number of sampling units (Table 4) and the maximum number of Haplosalids that had the greatest the purity amongst the soil taxa (Table 4).

Another criterion for accuracy assessment is sensitivity or producer's accuracy of map. Sensitivity values of the soil taxa of binary and multi-nomial regression are presented in Table 5. Soil classes with larger number of samples, followed the same trend in sensitivity and actual purity. Thus, Haplosalids had both high purity and sensitivity. The variability and heterogeneity of ancillary properties have also influenced these parameters. For instance, the

variability of properties such as slope, curvature, surface reflectance and altitude in the flat stratum of Pl111 was less than that in other strata such as Pi121, Pi122, Mo111, and Mo121.

Table 5. Estimated sensitivity (true positives) of the binary and multinomial logistic regression for soil taxa

	sensitivity	
Soil taxa	Binary logistic	Multinomial logistic
Calcigypsids	0.31	0.45
Haplocalcids	0.50	0.34
Haplocambids	0.61	0.44
Haplogypsids	0.77	0.42
Haplosalids	0.67	0.73
Torriorthents	0.74	1
Torripsamments	1	1

The validation results obtained from both models showed that they did not have similar predictive capability at a confidence level of 0.05% (Table 6). The multi-nomial direct approach results in a significantly better weighted purity than the binary indirect approach. The variances presented in Table 6 indicate that the difference is significant. The purity resulting from both methods differs for each soil taxon (Table 4). The difference between largest and smallest estimated purities, respectively attributed to Calcigypsids and Haplosalids is large in both models. The discrepancy between purities of two models could be explained by model training and also the design of the decision tree, especially for Calcigypsids where the purity resulting from indirect model (binary logistic) is less than that of the direct model (multi-nomial logistic) (Table 4).

Table 6. The mean and variance difference in actual purity obtained from the models

			Difference between methods				
Stratum (h)	a_h	n	no difference[a] d=0	difference d<>0	$\hat{\bar{d}}_h$	$\hat{V}(\hat{\bar{d}}_h)$	$a_h^2\hat{V}(\hat{\bar{d}}_h)$
Mo111	0.034	3	2	1	0.67	0.092	0.000108
Mo121	0.035	3	3	0	1.00	0.093	0.00012
Hi111	0.031	7	5	2	0.71	0.011	1.03E-05
sd111	0.008	7	7	0	1.00	0.058	4.08E-06
Pl111	0.202	21	12	9	0.57	0.004	0.000195
Pl121	0.025	6	5	1	0.83	0.033	2.11E-05
Pl122	0.008	2	0	2	0	1.816	0.000126
Pi111	0.020	2	0	2	0	1.316	0.000532
Pi112	0.042	5	2	3	0.40	0.106	0.000188
Pi121	0.032	5	1	4	0.21	0.081	8.37E-05
Pi122	0.014	4	4	0	1.00	0.092	1.88E-05
Pi211	0.017	6	3	3	0.51	0.019	6.29E-06
Pi212	0.015	5	3	2	0.62	0.025	6.33E-06
Pi213	0.233	23	8	15	0.34	0.013	0.000699
Pi214	0.071	7	4	3	0.57	0.143	0.000724
Pi215	0.021	3	2	1	0.67	0.833	0.000382
Pi311	0.145	12	5	7	0.42	0.030	0.000638
Pi312	0.041	5	3	2	0.60	0.150	0.000254
	sum	126	69	57	0.50		0.004117

[a]:number of samples that have same classifications with both methods, and this classification is the same as the field classification

The best prediction result was obtained when characteristics derived from terrain, remote sensing and geomorphologic processes were used together and when differentiation of geomorphologic processes and overall heterogeneity identification and stratification of the study area was made. In areas where the distribution of predictors was more homogenous, the models can better understand and connect predictors and response. In general, the direct method, which is a black box approach, produced a slightly better map in terms of MP (Table 4) than the indirect method: the difference was largest for the Calcigypsids. As the calcic and gypsic diagnostic horizons are predicted fairly accurately (Table 4), the error may originate partly from the decision tree (Figure 3) which translates the occurrences of diagnostic horizons

into soil great groups. Application of decision trees for prediction of soil types by mapped diagnostic horizons therefore, looks to be promising alternative. An advantage of the indirect methods is that it gives insight into the causes of errors in prediction at the level of diagnostic horizons, which helps in the search for better covariates.

Chapter 3

Spatial prediction of soil great groups by boosted regression trees using a limited dataset in an arid region, southeastern Iran

Submitted to Geoderma

3.1. Introduction

Numerical information of soils based on new processing tools and digital data is continuously increasing. In the context of a growing demand of high-resolution spatial soil information for environmental protection and management, fast and accurate prediction methods are needed. Recent publications indicate that digital soil mapping has been tested in a wide range of soils and mapping scales throughout the world (McBratney et al., 2003; Grunwald, 2006; Dobos et al., 2001; Hengl et al., 2007). In digital soil mapping, soil observations are related to readily available ancillary spatial data. The relationship is quantified by different prediction methods using geographic information science, statistics and pedological approaches. Therefore, digital soil mapping relies on advances in computing and information processing occurred over the last 30 years. Recent soil landscape predictive algorithms such as neural networks, fuzzy logic or tree model tools develop mainly from machine learning fields (Fayyad et al., 1996; Grinand et al., 2008).

The Classification and Regression Tree (CART) algorithm was applied for predictive soil mapping using data and maps from a reference area by Lagacherie (1992). Recent statistical advances were implemented on decision tree models, namely stochastic gradient boosting

(Freidman, 1999). Boosted Regression Tree (BRT) is one of the several new techniques which aim to improve the performance of a single model by fitting many models and combining them for prediction. Boosting, or more precisely, stochastic gradient boosting, increases the predictive performance by reducing the over-learning, or overfitting, that commonly occurs with simple regression trees. Fitted BRT functions may be linear, curvilinear or non-linear, where the choice of error distribution includes normal, binomial and Poisson (De'ath, 2007; Elith et al., 2008). However, unlike the GLM (generalized linear model), in fitting a BRT model, there is no concern regarding outliers, the number or order of predictors, missing predictor values and variable selection. Given these advantages of the BRT method, there has been recent interest in tree-based models for soil mapping applications (Brown et al., 2006; Grinand et al., 2008). Recent studies have recognized advantages of using boosted trees as compared with simple trees which include the improvement of accuracy (Moran and Bui, 2002; Lawrence et al., 2004), little tuning needed and high robustness (Friedman and Meulman, 2003). Because it is more flexible, a boosted model tends to fit more realistic than a linear model and; therefore, inferences made based on the model may have more credibility.

Bauer and Kohavi (1999) made an extensive comparison of boosting to several other competitors on 14 dataset and found boosting as the best algorithm. Friedman et al. (2000) compared several boosting variants to the CART method and found that all the boosting variants outperform the CART algorithm on eight datasets. This technique showed significant improvements in the classification accuracy compared to unboosted classification and regression trees. Grinand et al. (2008) evaluated the ability of boosted tree model to provide accurate soil landscape prediction at an unsampled area. They found that the predictive capacity of models was quite low when extrapolated to an independent validation area.

3.2. Objectives

In this chapter, we evaluate the suitability and performance of boosted regression tree for soil mapping using a limited point dataset in an arid region of Iran.

3.3. Methods

3.3.1. Study area and soil sampling

The study area and soil sampling were described in the chapter 2.

3.3.2. Ancillary spatial variables

The ancillary variables were explained in the chapter 2.

3.3.3. Statistical models

In order to predict the soil great groups (target variable) by logistic-BRT, the occurrence of relevant diagnostic horizons was first mapped. Various maps were subsequently combined for a pixel-wise classification by combining the presence or absence of diagnostic horizons. We refer to this method as an 'indirect approach'. In other method, the 'direct approach', the dependent variable (the great group) is a categorical variable for which multiclass-BRT could be applied.

(i) Boosted regression trees

Boosted regression trees are a combination of two powerful statistical techniques: boosting and regression trees. Boosting is a machine learning technique similar to model averaging, where the results of several competing models are merged. Boosting uses a forward, stage-wise procedure, where tree models are fitted iteratively to a subset of the training data. Subsets of the training data used at each iteration of model fitting are randomly selected without replacement, where the proportion of the training data used is determined by

the modeler, the "bag fraction" parameter. This procedure, known as stochastic gradient boosting, introduces an element of stochasticity that improves model accuracy and reduces overfitting (Elith et al., 2008).

Initially, 50 trees were fitted in the normal manner, using recursive binary partitioning of the data. Residuals from the initial fit were then fitted with another set of 50 trees. These residuals were then fitted with another set of trees, and so forth, whereby the process focused on extreme observations. Trees were fitted iteratively until a specific loss function was minimized, verified through n-fold cross-validation. In the case of regression trees, the loss function minimized was the model deviance. Final fitted values were based on the entire dataset and computed as the sum of all the trees multiplied by the learning rate (Elith et al., 2008; De'ath, 2007).

In fitting a BRT, two parameters were specified, the learning rate and the tree complexity. The learning rate determines the contribution of each successive tree to the final model, as it proceeds through the iterations. The tree complexity shows whether the model would represent the main effects only (tree complexity =1), or whether interactions should have been included (tree complexity= 2, 3 …). Ultimately, the combination of the learning rate and tree complexity determines the total number of trees in the final model.

Fitted BRT models were obtained using the BRT script provided by Elith et al. (2008), which references the "gbm" library (Ridgeway, 2007) in R software (R Development Core Team, 2005). To fit a BRT model, default parameters of the BRT script were used, where learning rate= 0.01 and tree complexity= 1 and cross-validation= 10-fold. However, as a rule of thumb, because the fit lowers the deviance, the bag fraction was changed from the default value of 0.75 to 0.5. Parameters for the two- or three-way interaction model were the same as those above, except that the tree complexity was 2 or 3. Predictors that are weakly correlated

with the response variable may become strong predictors when taken together in the two- or three-way interaction model. Therefore, to predict some of the diagnostic horizons, main-effect BRT models were fitted, while two or three-way BRT models were used for the others.

Before fitting BRT model for diagnostic horizons (indirect approach), we tested which model settings (learning rate and tree complexity) had a better performance. The setting with the lowest deviance was selected.

To allow the combination of indicator maps for classification into great groups, a decision tree must be defined which links the occurrence of diagnostic horizons to a soil great group. This tree was explained in chapter 2, Figure 3.

In multiclass-BRT as a direct approach, a general strategy is the one-versus-all technique, where each individual class (coded as 1), is modeled against all the remaining classes (each coded as zero), and k different ensembles are constructed and then the maximum probability is given as the class label (Friedman et al., 2000)

3.3.4. Validation and statistical inference

(i) Model validation

The most extreme form of cross-validation, known as leave-one-out approach, was used. Each regression model was fitted by using n-1 observations and the soil great group/diagnostic horizon was predicted at the observation site which had not been used. The prediction with the unused observation was validated and this was repeated n times for the other observation sites. The validation indices were then estimated using the n validation results. The result of validation was an indicator variable taking value 1 if the predicted soil great group was equal to the observed soil great group and was 0 otherwise.

(i) Statistical inference

The observation sites are a stratified simple random sample (De Gruijter et al., 2006; Kempen et al., 2009). Strata were defined by ancillary variables such as geomorphological, geological and topographical maps. Therefore, the resulted geomorphic surfaces were representative of the differences in geomorphology, geology and topography of the landforms. Subsequently, these resulted in 18 strata and 126 locations allocated to the strata in proportion to their area, with a minimum of two per stratum. Based on the sampling design and the defined strata, a weighted purity should be calculated. The purity was estimated following the method proposed by Brus et al. (2011) as follows:

$$\hat{p} = \sum_{h=1}^{l} w_h \hat{p}_h \qquad (14)$$

where w_h is the weight (relative area) of stratum h, $\hat{p}_h$ is the estimated areal fraction of stratum h correctly classified, and l is the number of strata. The stratum fractions were estimated by the fraction of correctly predicted locations in each stratum.

$$\hat{p}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} y_i \qquad (15)$$

where n_h is the number of random sampling locations in the stratum h, and y_i is the indicator variable at sampling location i, taking the value 1 if the predicted soil great group was equal to the observed soil great group and was 0 otherwise.

The difference in purity between two methods for the whole area can be assessed using indicators calculated from data. The indicators were determined by comparing the field classification with logistic-BRT and multiclass-BRT classification and then difference between indicators was calculated by $d_i = y_i^l - y_i^m$ where y_i^l is an indicator for the correct prediction by the logistic-BRT method and y_i^m is a similar indicator for the multiclass-BRT

method. This variable d can have values -1, 0, and 1 and was used to estimate $\hat{d}$, which is the mean difference in purity of the logistic-BRT and multiclass-BRT methods. Using the calculated standard error of the estimated mean of difference d, we tested whether differences were significant or not. The variance of the mean value for $d(\hat{v}(\hat{\bar{d}}_{st}))$ was calculated according to De Gruijter et al. (2006);

$$\hat{v}(\hat{\bar{d}}_{st}) = \sum_{h=1}^{H} w_h^2 \hat{v}(\hat{\bar{d}}_h) \qquad (16)$$

where w_h is the relative area of stratum h, $\hat{\bar{d}}_h$ is the stratum mean of d and $\hat{v}(\hat{\bar{d}}_h)$ is the estimated variance of $\hat{\bar{d}}_h$ that can be calculated as follows

$$\hat{v}(\hat{\bar{d}}_h) = \frac{1}{n_h(n_h-1)} \sum_{i=1}^{n_h} (d_{hi} - \hat{\bar{d}}_h)^2 \qquad (17)$$

where n_h is the sample size in stratum h and d_{hi} is indicator variable in stratum h. $\hat{v}(\hat{\bar{d}}_h)$ can be used to construct confidence intervals and to test if d significantly differs from 0 in which case the logistic-BRT and multiclass-BRT classification methods perform significantly different.

The performance of prediction for each soil great group was calculated as the purity for the soil class in the digital map. The purity for soil class s of the digital soil map was estimated by:

$$\hat{p}_s = \frac{\sum_{h=1}^{l} A_h \bar{y}_h^{(s)}}{\sum_{h=1}^{l} A_h \bar{x}_h^{(s)}} \qquad (18)$$

Where A_h is the area of stratum h, $\bar{y}_h^{(s)}$ is the sample mean of indicator $y_{i,h}^{(s)}$ taking value 1 if the mapped and observed soil group at sampling location i equal soil group s and 0 else, and $\bar{x}_h^{(s)}$ is the sample mean of indicator $x_{i,h}^{(s)}$ taking value 1 if the mapped soil group equals soil group s and 0 otherwise.

Another statistic for accuracy assessment is the sensitivity of the map. The sensitivity of map unit s of the digital soil map was estimated by:

$$\hat{S}^{(s)} = \frac{\sum_{h=1}^{l} A_h \bar{y}_h^{(s)}}{\sum_{h=1}^{l} A_h \bar{z}_h^{(s)}} \qquad (19)$$

Where A_h is the area of stratum h and $\bar{z}_h^{(s)}$ is the sample average of the indicator $z_{i,h}^{(s)}$ taking value 1 if the observed soil group equals soil group s and 0 else.

The Kappa index is a robust index which takes into account the probability that a class is classified by chance (Girard and Girard, 1999). It is a simple derived statistic that measures the proportion of all possible cases of presence or absence that are predicted correctly by a model after accounting for chance predictions. A higher kappa index indicates a high model performance (D'heygere et al., 2006). Kappa has been used extensively in map accuracy work (Congalton, 1991; Freeman and Moisen, 2008).

3.4. Results and Discussion

3.4.1. Model-building

(i) Logistic-BRT model as an indirect approach

Generally, to predict each phenomenon, factors affecting its formation and evolution should be considered. Modeling a phenomenon solely using software and model would lead to weak or unexpected results. This is true for predictions of soil and its properties. Therefore,

pedogenic processes and environmental conditions affecting soil and its properties should be considered in the predictions related to soil as already mentioned by Jenny (1941).

We describe the model-building process and fitting of logistic-BRT for diagnostic salic horizon with a viewpoint of pedology, because the final model should be reliable both statistically and pedologically. Use of knowledge of soil-landscape system should be fully integrated throughout the process of model-building. For the other diagnostic horizons we followed a similar approach.

Salic horizon is the most frequently occurring diagnostic horizon in the study area. It has been formed due to (1) the accumulative, low-lying areas of playa which receive substantial amount of soluble salts from the surrounding areas (2) the presence of hardpans in soils and irrigation of agricultural lands which lead to the increase of groundwater level, and (3) the heavy texture of soils in playa. Therefore, we expect such factors as the form of land, geological materials and processes to affect the formation of salic horizons in the study area; and therefore, these factors were employed in modeling. For BRTs, an index of relative influence was calculated in summing the contribution of each variable, which is equivalent to summing the branch length for each variable in the regression tree (Figure 10).

For BRT model fitted for the salic horizon, the five most influential variables were: 1) geomorphic surface (GS) (39.4%), 2) MrVBF (17.5%), 3) wetness index (8.06%), 4) clay index (7.79%), and 5) slope (5.5%). Among predictors, geomorphic surface (GS) was identified an important predictor for all the diagnostic horizons (Table 7). This emphasizes the role of geomorphology processes in soil development as reported in many soil-geomorphology studies (Cantón et al., 2003; Toomanian et al., 2006).

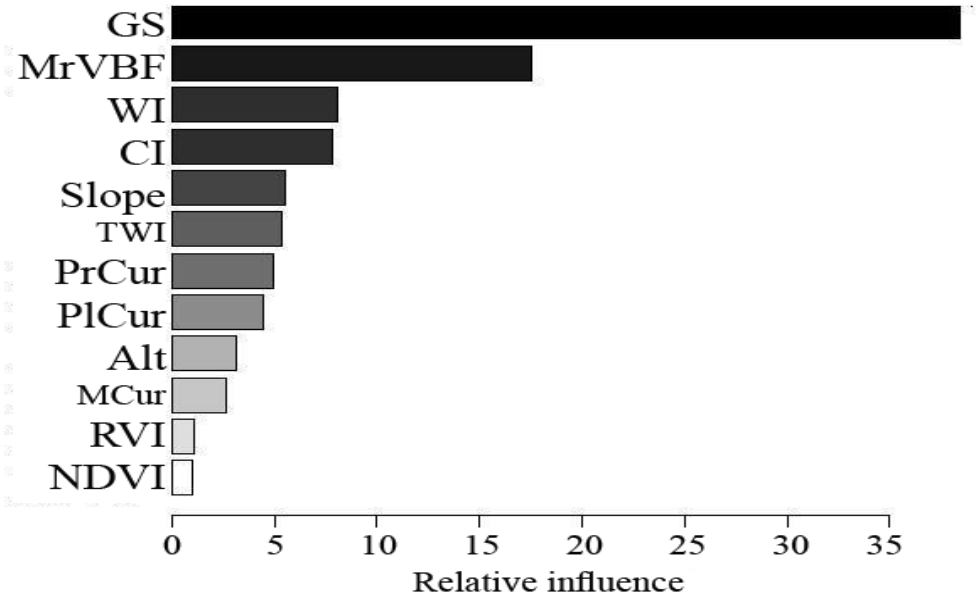


Figure 10. Relative influence of model terms calculated by the contribution of each term in reducing the overall model deviance for the salic horizon

Table 7. The selected variables and Area Under Curve (AUC) for fitted BRT model of each diagnostic horizon and soils without diagnostic horizons (indirect approach)

Diagnostic horizon/ Great group	BRT model	Variables	AUC
Salic	3-way	GS+ MrVBF+ WI+ CI+ Sl+ TWI+ PlCur	0.95
Gypsic	2-way	GS+ WI+ CI+ NDVI+ TWI+ PrCur	0.87
Calcic	2-way	GS+ WI+ CI+ PrCur	0.91
Cambic	3-way	GS+ TWI+ Sl+ El+ PlCur	0.81
Psamments	1-way	GS+ MCur	1
Orthents	2-way	GS+ Sl+ MCur+ PlCur+ El	0.96

For abbreviations, refer to Table 1.

For the BRT model of the salic horizon, in addition to geomorphic surface, MrVBF and WI were the most influential predictors (Figure 10). Multi-resolution Valley Bottom Flatness index (MrVBF) is intended to identify flat valley bottoms and WI indicates the degree of wetness (Wang and Laffan, 2009). These parameters present zones of transport for a wide range of materials, particularly fluxes of sediment and other entrained materials (Whiteway et

al., 2004). Therefore, MrVBF and WI indirectly act as driving forces for the formation of salic horizon, as high values of MrVBF and WI correspond to high probabilities of the occurrence of salic horizon (Figure 11). For many types of statistical model, partial dependency plots (Friedman, 2001) can be used to visualize dependencies between the response and one or more predictors.

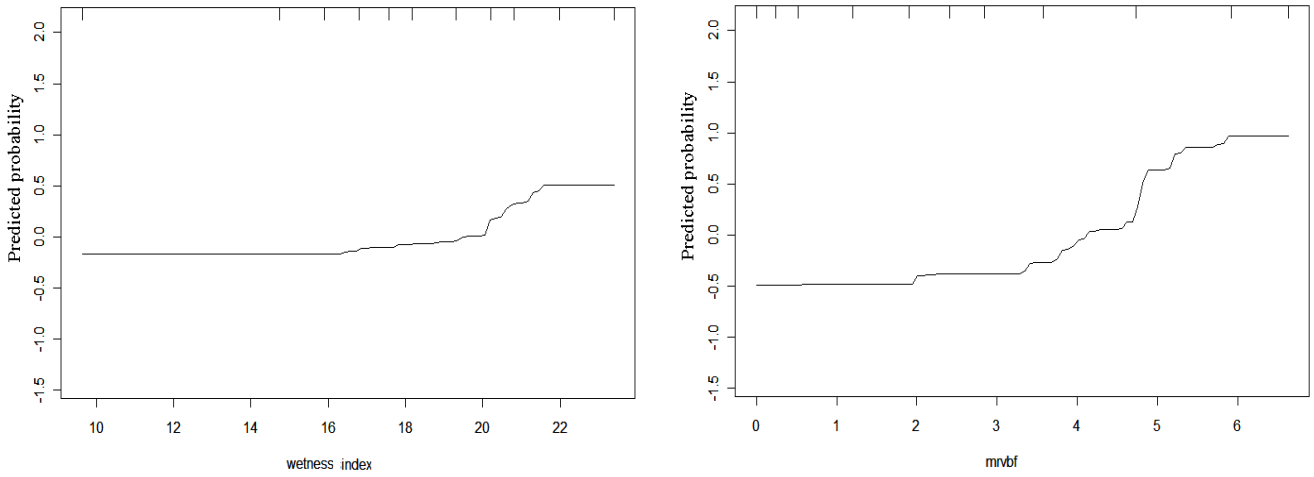


Figure 11. Predicted probabilities (BRT) for the occurrence of the salic horizon as a function of MrVBF and WI

Figure 10 shows that the clay index and slope are effective predictors and have important role in the formation of salic horizons. The formation of playa, alluvial fan, and other landforms has greatly controlled the parent materials and matter fluxes. Given the arid climate, soils are expected to show a close relation with geomorphology, both via the landscape units (GS) and topographic factors derived from the DEM. In this strategy, the model identifies the driving factors and processes controlling pedogenesis and soil spatial distribution. It was also proved that the model retrieved relevant and accurate soil-landscape relationships.

In fitting BRT model, the most effective predictors were selected based on the decrease of deviance, the correlation between response and independent variable and the area under roc

curve. Some variables did not contribute to the fitted model for salic horizon, as deviance did not change significantly with their deletion (Figure 12).

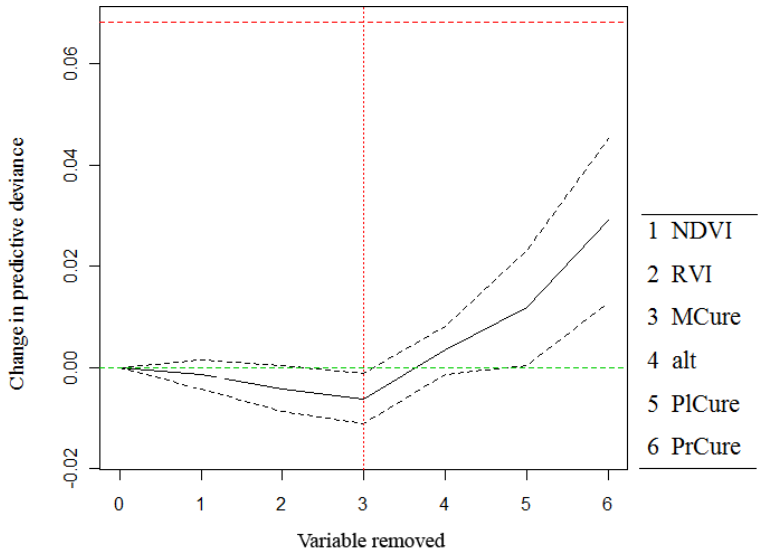


Figure 12. Change in predictive deviance with removal of parameters for the salic horizon

Furthermore, the area under ROC curve (AUC) increased with the relevant predictors (Figure 13). The AUC (0.89) for BRT model of salic horizon with the relevant variables is greater than when less relevant variables such as curvature (0.74) are used. AUC values and selected variables for fitted BRT model of diagnostic horizons are presented in Table 7.

Finally, after selecting predictors for the BRT model of salic horizon, the model was fitted with a learning rate=0.001, interaction depth=3, bag fraction=0.5 and the optimal number of trees = 2300 (Figure 14). Similarly, BRT model was fitted for the other diagnostic horizons.

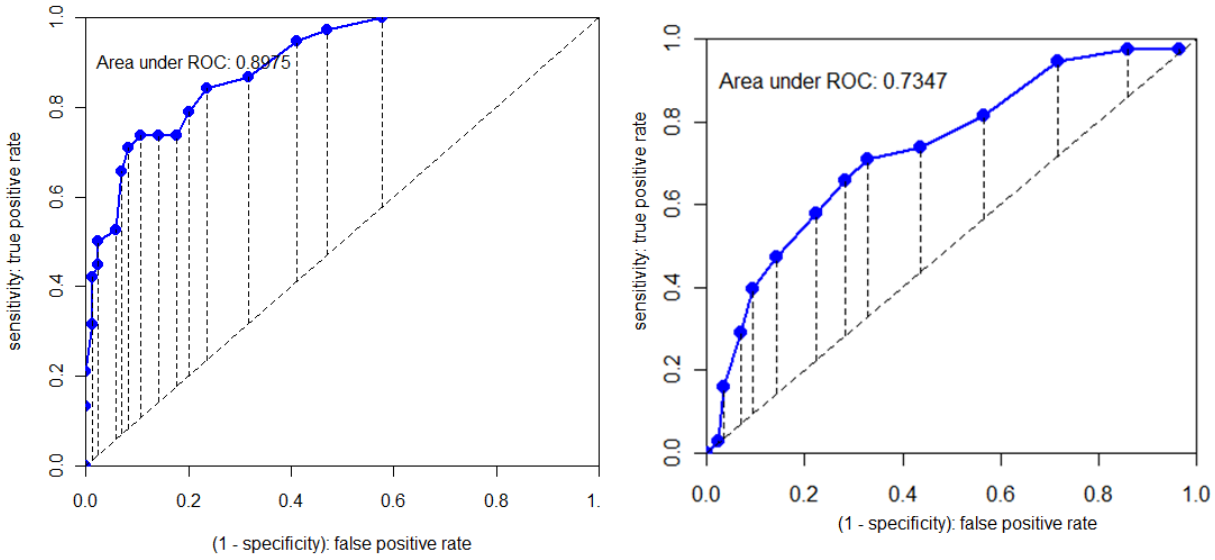


Figure 13. Area under ROC (AUC) for prediction of salic horizon by logistic-BRT (right-hand curve for model with variables shown in Table 3 and left-hand curve for model with variables shown in Table 3 plus mean curvature

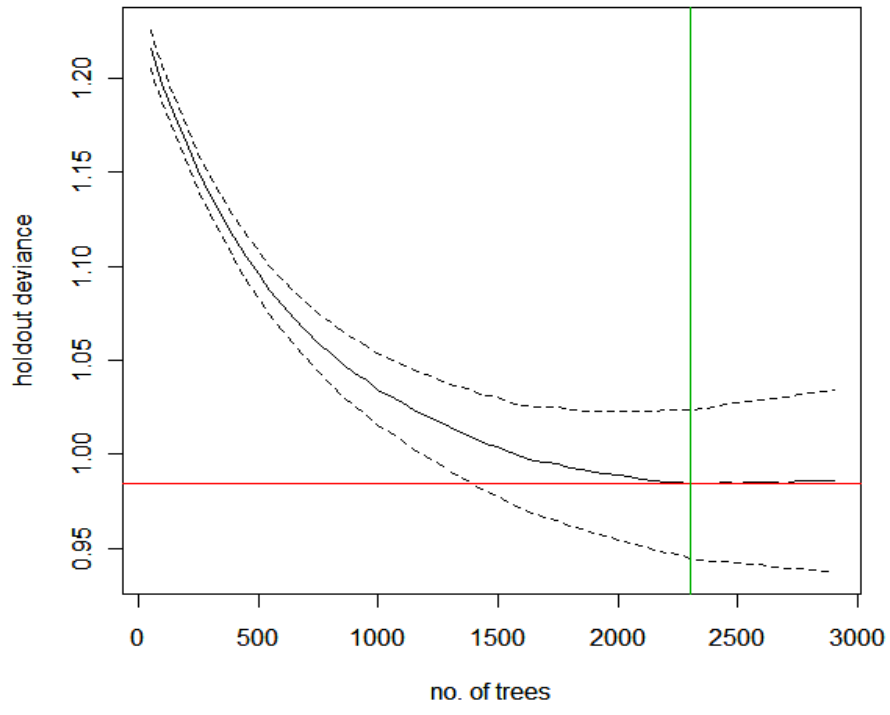


Figure 14. Optimization plot for the Boosted Regression Tree (BRT) model for the salic horizon. The solid black curve is the mean changes in predictive deviance and the dotted curves indicate 1 standard error zones. The red horizontal line shows the minimum of the mean, and the green vertical line the number of trees at which it occurs.

Partial purity for diagnostic horizons is shown in Table 8. It is the proportion of horizons correctly classified. Since soil great groups were predicted by combining diagnostic horizons in the indirect approach, purity obtained from the prediction of diagnostic horizons was called partial purity. High partial purity resulted from the prediction showed good spatial distribution for diagnostic horizons. For the salic horizon, a partial purity of 0.8 was obtained which shows high correlation between spatial distribution of salic horizon and the covariates.

Table 8. Prediction quality of boosted regression trees for diagnostic horizons and soils without diagnostic horizons (indirect method)

	Salic H.	Gypsic H.	Calcic H.	Cambic H.	Orthents	Psamments
Proportion of profiles correctly predicted	30/38	55/58	27/37	37/47	9/9	7/7
Partial purity	0.79	0.95	0.73	0.79	1	1

The results obtained for the partial purity of Orthents is not justified because the low occurrence of Orthents in the dataset may cause chance effects. Considering the low presence of Orthents in the dataset, it is likely that the model has not been properly trained and therefore, poorer predictions were obtained.

The indicator maps of the occurrence of diagnostic horizons are presented in Figure 15. The probability that a given diagnostic horizon occurs at a given pixel is represented by values between 0 and 1, where 0 is absolutely no chance and 1 (white areas) indicates the presence of a diagnostic horizon or soil group. The method predicted high probability of salic horizon in playa landform, gypsic horizon in gypsiferous hills, calcic horizon in alluvial fans, Psamments in sand dunes and Orthents in mountains (Figure 15). Finally, the diagnostic horizons were combined to classify into the soil great group (Figure 16). We expect that the prediction of diagnostic horizons with high purity would lead to good predictions of soil great groups.

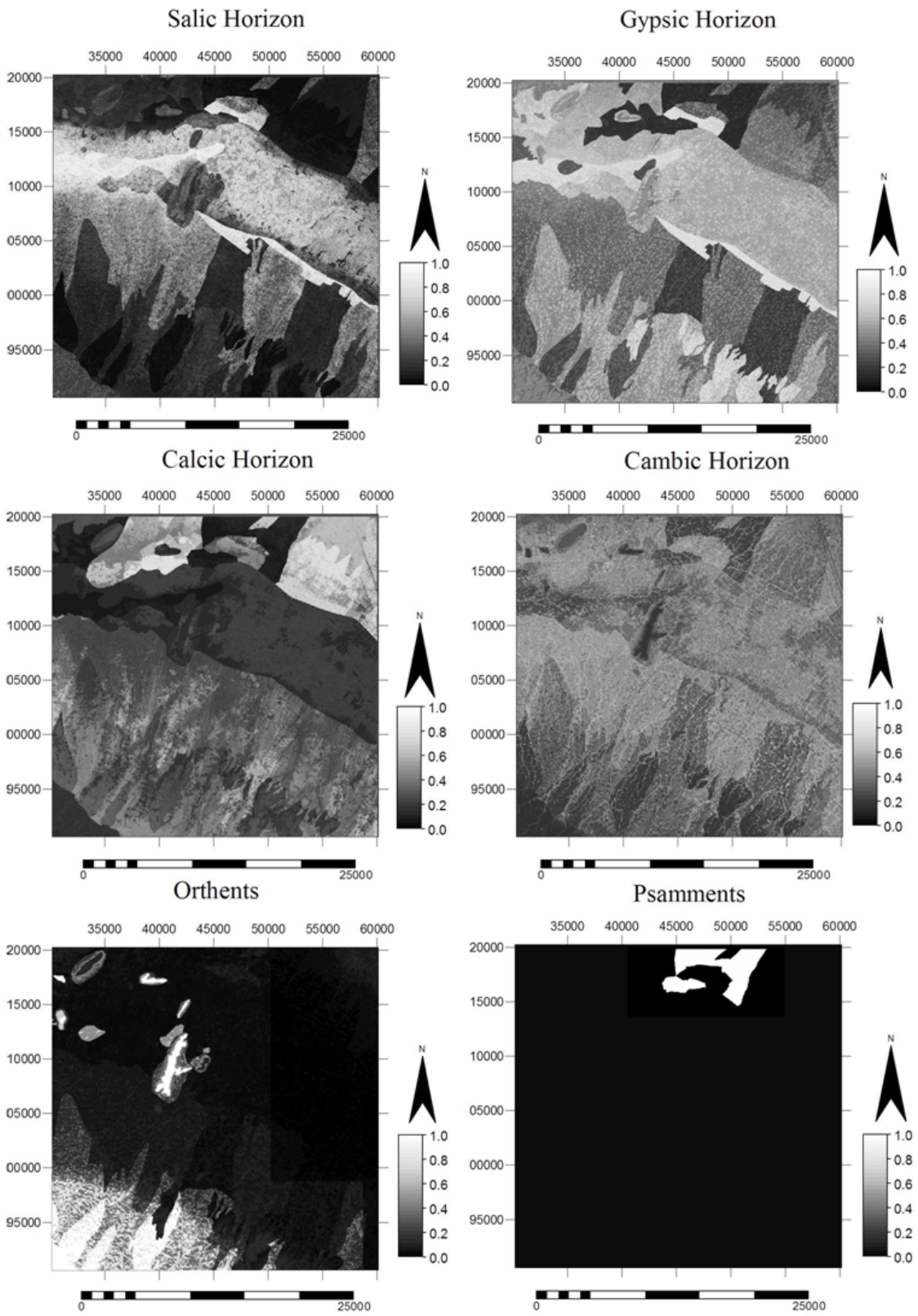


Figure 15. The mapped probability of occurrence of diagnostic horizons derived from boosted regression trees (indirect approach)

(ii) Multiclass-BRT model as a direct approach

The multiclass-BRT model directly predicts each soil great group from the predictors. The best model for prediction was selected similar to logistic-BRT on the basis of the highest AUC and lowest deviance. The number of predictors was different among the seven BRT models of soil great groups. The number of predictors showed a clear relationship with the number of observations of soil great groups (Table 9).

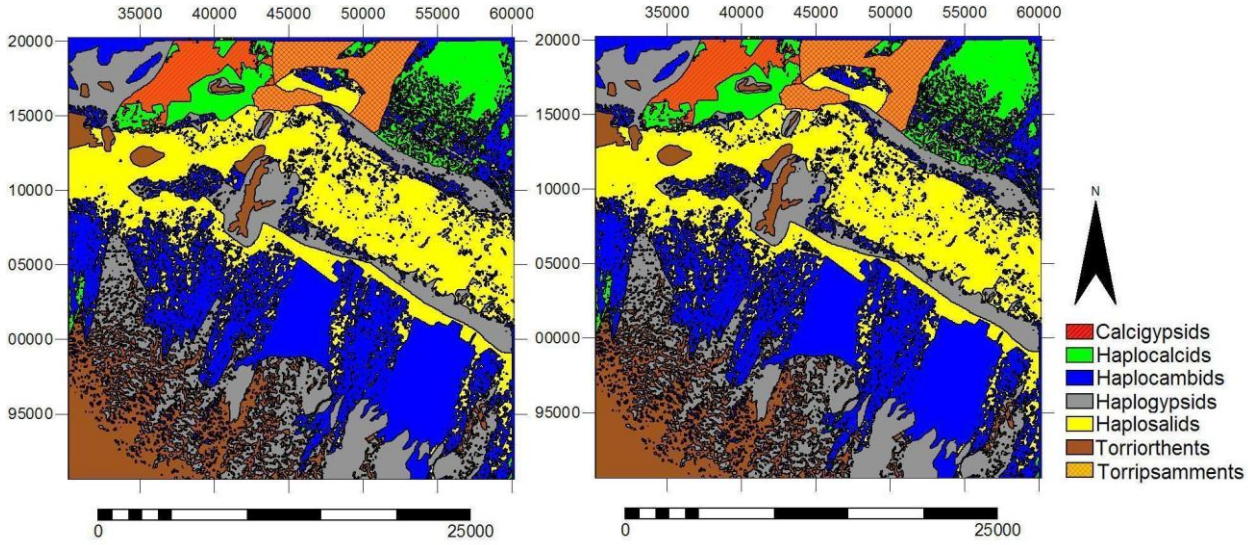


Figure 16. Spatial distribution of the soil great groups derived from logistic-BRT (right), and multiclass-BRT (left)

Only two predictors were imported for fitting the BRT model for Torriorthents and Torripsamments with 6 and 7 observations, respectively whereas there were eight predictors in model for Haplosalids with 35 observations. The same result was also found for diagnostic horizons in the indirect approach. More predictors contributed to the fitting of logistic-BRT model of gypsic horizon (67 observations) in comparison with that of the calcic horizon (44 observations).

Table 8. Selected variables and Area Under Curve (AUC) for fitted BRT model of each soil great group (direct approach)

Great group	BRT model	Variables	AUC
Haplosalids	3-way	GS+ MrVBF+ WI+ NDVI+ PVI+ TWI+ PlCur	0.85
Haplogypsids	2-way	GS+ WI+ El+ PlCur + TWI+ PrCur	0.88
Haplocalcids	2-way	GS+ WI+ PVI+ PrCur	0.78
Haplocambids	3-way	GS+ TWI+ Sl+ El+ PlCur	0.79
Torripsamments	1-way	GS+ MCur	1
Torriorthents	2-way	GS+ Sl+ El	1

For abbreviations, refer to Table 1.

Also, the presence of predictors depends on the spatial distribution of observations and that, in turn, depends on the presence of soil great groups or diagnostic horizons in the space. For example, in the study area, the soil great group Torriorthent is limited to high slope and elevated areas and Torripsamment to sand dunes, whereas Haplosalid is distributed in the landforms with different geomorphic processes and terrain parameters such as playa, alluvial fan and bajada. Therefore, under such conditions, the model has to cover different spatial distributions with different predictors and consequently, soil patterns hardly distinguishable. Grinand et al. (2008) found a strong relationship between pixels having high uncertainty with mixed soil distribution.

The geomorphology map was identified as the most important predictor for all the soil great groups. This re-emphasizes the effective role of geomorphological processes on the soil development. Each geomorphic surface encompasses unique characteristics and distribution of special predictors which results in better representation of the soil-landscape relationships. Therefore, on the basis of the results obtained from modeling, the undeniable role of geomorphology processes is confirmed in this study area. Due to the high frequency of Haplosalids occurrence in playa and also the presence of vegetation in this landform, remote

sensing indices such as PVI and NDVI can be imported for fitting Haplosalids which can be observed in the fitted BRT model (Table 8). For the other soil great groups, terrain attributes were mainly selected. Dobos et al. (2001) selected DEM and it derivatives for soil classification, highlighting their importance for soil-landscape characterization.

The soil great group maps derived from the direct approach are presented in Figure 16. As expected, the method well predicted the great group Torripsamment in sand dunes, Haplogypsid in gypsiferous hills, and Haplocalcid in alluvial fans.

3.3.2. Spatial prediction and prediction accuracy

The overall purity of the soil map derived from indirect and direct approaches is 49% and 58%, respectively (Table 9). In general, the purity is 10% lower in the indirect approach, possibly because of the decision tree. Some the diagnostic horizons were incorrectly translated to the soil great groups under the decision tree. Therefore, the increase in map purity of the direct approach compared to that of the indirect approach is largely attributed to the nature and performance of the approach.

At the level of geomorphic surface, purity distribution is better represented (Table 9). Among strata, stratum Mo111, Mo112, and Sd111 showed high purity. This is likely due to the presence of a single soil type in these strata. The relationship between predictors and soil great groups can be easily detected in more homogenous areas in comparison with less homogenous locations. Different geomorphological and pedological processes have led to heterogeneity in the area and created different soil classes. More homogeneous strata with low number of sample have high purity (e.g. Mo111, Mo112, Sd111), while strata with larger sample size have low purity (e.g. Pl111). In stratum Pl111, there are also different soil great groups such as Haplosalid, Haplogypsid and Haplocambid, while there may be the same

ancillary parameters. This is a problem in arid zones, at least at this scale, where there are large areas with low variable topographic and reflectance properties which could be used as ancillary data. Therefore, the relationship between different soil great groups and predictors could be confusing for the model.

Table 9. The estimated purity in each stratum and soil map purity derived from indirect and direct approaches

		Logistic-BRT (indirect)	Multiclass-BRT (direct)
	n	purity	purity
Map purity		0.49	0.58
stratum			
Mo111	3	1	1
Mo121	3	1	1
Hi111	7	0.57	0.57
Sd111	7	1	1
Pl111	21	0.43	0.62
Pl121	6	0.83	0.83
Pl122	2	1	1
Pi111	2	0	0
Pi112	5	0.51	0.67
Pi121	5	0.4	0.4
Pi122	4	1	1
Pi211	6	0.67	0.83
Pi212	5	0.8	0.4
Pi213	23	0.39	0.57
Pi214	7	0.57	0.56
Pi215	3	1	1
Pi311	12	0.2	0.4
Pi312	5	0.6	0.6

Table 9 also shows that sampling intensity did not influence the classification accuracy. For example, Pl111 and Pi213 with greatest sampling points did not show high purity. Moran and Bui (2002) and Grinand et al. (2008) found the same results.

The results of the statistical assessment of the final BRT models (direct and indirect) for soil great groups are presented in Table 10. The overall purity of direct approach shows that

the model predicts the soil great groups better than the indirect approach, possibly due to the decision tree which translates the occurrence of diagnostic horizons into soil great groups in the indirect approach. In the direct approach, soil great groups are directly predicted.

The lowest purity was observed for Calcigypsids in both approaches (Table 10). It seems that the sample size influenced the prediction accuracy about Calcigypsids. Sample size is of major importance in the accuracy of assessment process (Foody, 2002). The small number of sample locations results in the weak association of predictors with the soil great groups. In the indirect approach, to predict Calcigypsids under the decision tree, the high probability of calcic and gypsic horizons should be combined. However, this probability rarely occurs due to the low sample size and the generalization process. When the model uses lower sample size, generalization error increases, particularly in the boundaries of soil great groups.

Table 10. The kappa index, estimated purity and sensitivity of soil great groups predicted from direct and indirect approaches

		Logistic-BRT (indirect)			Multiclass-BRT (direct)		
Soil stratum	%n	Sensitivity	Kappa	purity	Sensitivity	Kappa	purity
Torriorthents	4	1	1	1	1	1	1
Torripsamments	4	1	1	1	1	1	1
Haplosalids	30	0.59	0.54	0.62	0.70	0.69	0.72
Haplogypsids	22	0.55	0.43	0.41	0.55	0.59	0.50
Haplocalcids	18	0.24	0.35	0.50	0.37	0.49	0.67
Haplocambids	17	0.47	0.22	0.34	0.69	0.4	0.44
Calcigypsids	5	0.11	0.24	0.30	0.16	0.12	0.30

For Torripsamment, because of the direct relationship between this soil great group and explanatory data, even with a low sample size, the models showed a high purity (Table 10). The purity of BRT for Torriorthents was high. This is due to accordance of spatial distribution

of very contrasting predictors and Torriorthents. The spatial distribution of Torriorthents is largely limited to mountains and areas that encompass very contrasting ancillary properties.

The purity of the soil great group Haplosalid shows that the model correctly predicts at 62% (logistic-BRT) and 72% (multiclass-BRT) of locations which is more than other soil great groups. High purity of Haplosalids can be related to high purity of the salic horizons. Playa landform is an area with high presence probability of salic horizon and; therefore, the high probability of Haplosalids. Both gypsic and salic horizons were found at the edges of playa (Figure 8). This is a transition zone where the relationship between soil great groups and predictors are weak. Subsequently, the presence probability of salic horizon decreases and prediction purity of Haplosalids will be lower in such areas. At these locations, the model easily confuses among soil great groups. Kempen et al. (2009) also concluded that prediction uncertainty is larger in topographic transition zones than areas with stronger relationships between soil groups and predictors and the models confuse at these locations.

Generally, the BRT models predicted the spatial prediction of soil great groups fairly well. The variation in purity of soil great groups can be mainly related to the spatial distribution of samples over the strata that can present actual properties of the stratum.

Therefore, it seems that BRT model is not sensitive to the sample size, while the relationship target variable and explanatory variables influence its performance. Therefore, it is very important to identify explanatory variables, having a causal relation with the target variable. With use of suitable explanatory variables, models can accurately identify these relationships and therefore, chance proportion in prediction automatically decreases. Therefore, boosted regression trees can produce reliable results in soil modeling of large datasets.

Assessment of soil maps accuracy is imperative for any soil mapping study including traditional and digital. A statistic parameter for accuracy assessment that provides valuable information is the sensitivity or producer's accuracy of the map. This statistic is often used in image classification studies (Foody, 2002) but is hardly reported for digital soil maps. Sensitivity values of the soil great groups are presented in Table 10. An example of the sensitivity statistic is given for Haplosalid. This map unit has a high purity (72%), which tells the user that soil great group Haplosalid is found at 72% of the area. However, the sensitivity of Haplosalid is 70%, meaning that only 70% of the true area of soil great group Haplosalid is mapped as Haplosalid. The sensitivity of Haplosalid is higher in the direct approach (70%) compared to the indirect approach (59%). It seems that results of the direct approach are more consistent with reality and user can have more confidence for the produced map. Therefore, sensitivity parameter can appears to better present the reliability of produced map for users. Sensitivity and purity of soil map were lowest for Calcigypsid in this study.

Kappa index ranges from 0.12 to 0.72 in direct approach and 0.11 to 0.59 in indirect approach (Table 10). Values of purity and kappa index suggest that the predictive ability of the direct model is greater and more satisfactory for most of soil groups. High discrepancies between accuracy and kappa index suggest larger influence of chance factor (Grinand et al., 2008). Therefore, chance effect is greater in classifying Haplocalcids and Haplocambids in indirect model and Haplocalcids and Calcigypsids in direct model.

In the study of Luoto and Hjort (2005), kappa index was one of the criteria for evaluating the predictive performance in geomorphological modelling. They reported value of kappa index 0.49-0.56 and implied the model's reliable predictions. Therefore, the predictions derived from the direct model have more reliable than the indirect model (Table 10). Overall, a better performance was detected for the direct models. Giassen et al. (2006) implied the

unsatisfactory results with values 45% and 31% overall occuracy and kappa indices, respectively.

In geomorphic-stratum, the higher putrity of Pl122 than Pi111 might be attributed to the chance, because there were only two sampling sites in this stratum.

The results obtained indicate that there is not any significant difference between the two approaches. However, high values of the estimated purity and sensitivity were not identified for the same soil great groups and the same strata in both approaches. Some errors are likely associated with the decision tree in the indirect approach. We did not only apply the regression models for the evaluation capability of the models in prediction, but it would be worthwhile in easy and quick making a view from the regional soils in arid areas that there is not any data and map and therefore, soil survey is difficul such as the study area.

Chapter 4

Selection of taxonomic level for soil mapping using diversity and map purity indices, a case study from an Iranian arid region

Ready for submitted to Geomorphology

4.1. Introduction

From traditional to digital soil mapping, visualized product has almost always been a map displaying the spatial distribution of soil classes. In traditional soil mapping, spatial distribution of soils is described based on field observations and the use of landscape features, which are related to soil patterns. The digital soil mapping is a quantitative approach to produce digital maps of soil type and properties. It is based on the use of mathematical and statistical models that combine information from soil observations with information contained in correlated environmental variables and remote sensing images (Schull et al., 2005). The assumptions underlying the rules of development are as follow: (1) Soil distribution reflects the long-term interactions between terrain variables, geology, and vegetation in landscape. (2) A digital elevation model (DEM) and derived terrain attributes can represent factors of soil formation. (3) The existing soil maps have captured soil–landscape interactions in the area mapped.

Prediction of soil classes can be made at all the soil taxonomy levels including order, suborder, great group, subgroup, soil family, and soil series. Digital soil maps, similar to conventional soil maps, are not perfect and contain errors (Brus et al., 2011). Prediction of soil

classes is thus associated with uncertainty and impurity results in maps at each category level. Lower categoric levels are defined by a greater number of diagnostic criteria than higher categoric levels. An apparently pure mapping unit in a high category classification may still contain a high impurity when evaluated at a lower category of classification. Olaniyan and Ogunkunle (2007) investigated the purity of mapping units at the subgroup level. They showed the high purity values resulted from the very broad definitions of the mapping units. Kempen et al. (2009) reported that prediction uncertainty is smaller in areas with stronger relationships and correlations. An accurate prediction of lower category levels needs more detailed environmental variables or very contrasting variables that act at finer scale such as endogenic processes. Therefore, the purity of predicted soil map is expected to change at different levels of soil taxonomy. High impurity in the soil map seems to be related to high soil diversity in the region. Ibañez et al. (1998) reported that soil mapping in the Mediterranean region was particularly prone to uncertainty, because of high soil diversity. Kempen et al. (2009) used Shannon entropy to quantify the uncertainty of an updated soil map. They reported that very heterogeneous areas showed high entropy and low purity and their prediction was associated with high uncertainty.

The application of the diversity concept to soil taxonomic units is a different approach to soil quantitative characterization (Ibanez et al., 1995; Phillips, 2001, 2002; Saldana and Ibanez, 2004; Martin et al., 2005; Phillips and Marion, 2005). Pedodiversity, as well as biodiversity, may be considered as a framework to analyze spatial patterns. It was recognized as a novel pedometric tool by McBratney et al. (2000). Soil diversity is determined by diversity indices such as richness, Shannon entropy and evenness (Ibanez et al., 1994; Phillips, 2001; Gue et al., 2003). Gue et al. (2003) studied pedodiversity by calculating Shannon's entropy for various taxonomic levels including soil order, suborder, great group, subgroup,

family, and series. They showed that richness and Shannon's entropy increased with increasing taxonomic detail. Toomanian et al. (2006) studied soils on different landscapes in central Iran and reported a similar behavior of diversity in soil taxonomy and soil geomorphologic categories.

We have to change understanding the scale from performance of various processes at different category levels. Specific soil processes are determined by soil forming factors and are expressed in diagnostic horizons, properties, and materials, which are then used to classify the soils. Generally, the soil forming factors define the state of soil system and the soil forming processes characterize specific pathway of soil development. This implies that a soil type or a particular soil property is the outcome of a nonlinear dynamic system (Ibanez et al., 1990, Toomanian et al., 2006, Caniego et al., 2007) under unstable and chaotic conditions. Therefore, small local chaotic and short-lived perturbations may lead to more diversity in a finer scale or at a lower category level.

The variation of soil properties mainly depends on genesis processes (Liu et al., 2006). The genesis processes are used to determine soil classification. Thus, soil types indicate the variations in soil genesis processes. Soil genesis and properties are a function of the climatic, biotic, geomorphological, lithological, aquatic and anthropogenic conditions being affected at any given time in a given space. This is manifested in the spatial heterogeneity and diversity of soils (Degórski, 2003). Soil maps represent a visual synthesis of soil heterogeneity and diversity and should be a good object for diversity and pattern analysis. The quality of soil map is a function of the taxonomic system used, the level in the taxonomic system and the scale of study.

The representation capability of soil diversity by soil map depends on the scale of the soil map. Large-scale soil maps may include more detailed soil classes and represent higher

soil diversity and heterogeneity compared to small-scale soil maps. In other words, the results obtained on small-scale soil maps can be considered only as rough estimates. In such conditions, soil map may not meet user's expectations. Also, production of large-scale soil maps needs more time and budget compared to small-scale soil maps. Soil spatial patterns have an important effect on landscape management, allocation of resources for different land uses and intensive applications in agricultural and hydrological management. This is very important in arid regions, where soil mapping is a hard work to perform.

4.2. Objectives

The objective of this study was to make reliable soil maps by using digital soil mapping which represent as much (taxonomic) diversity as possible. To achieve this goal, diversity indices and map quality indicators are used in combination with final optimal taxonomic level for a certain area with a certain sampling effort.

4.3. Methods

4.3.1. Description of the study area and soil sampling

The study area and soil sampling were described in the chapter 2.

4.3.2. Data configuration

(iii) Geomorphic hierarchy

The geomorphology map and geomorphic hierarchy were explained in the chapter 2.

(iv) Ancillary spatial variables

The ancillary variables were explained in the chapter 2.

4.3.3. Mapping methodology

Artificial neural networks (ANNs) are nonlinear mapping structures based on the function of the human brain. They have been shown to be universal and highly flexible function approximates for any data (Luoto and Hjort, 2005). Neural networks have received considerable attention as a mean to build accurate models for prediction when the functional form of the underlying equations is unknown (Venables and Ripley, 2002).

Lek and Guegan (1999) proposed that ANN models are more powerful than multiple regression models when modeling nonlinear relationships. The full classification procedure in ANN is a complex nonparametric process that is sometimes seen as a black box, even by computer scientists (Venables and Ripley, 2002).

The application of an ANN consists of two stages. During the first stage, the network is trained, which means that it learns the conditions on which a certain feature (e.g., a soil unit) occurs. This stage comprises of the training data set, which is soil sample data and the input unit (cell or neuron), which is the soil predictors (i.e., a terrain attribute and/or a geomorphic or geological unit). The output unit represents the target variable as the desired output, i.e., the mapped soil unit (Figure 17). The connections between neurons are described by the weights w_i ($w_{i1}\ldots$, w_{in}). The adjustment of these weights is based on the learning process. As each attribute combination (in terms of pixels of a grid map) is put into the network in succession, the weights are adjusted iteratively if the predicted output does not match the output of a training data set. During the second stage, the learned knowledge in terms of the calibrated weights can be applied to prediction areas, for which the same input parameters (i.e., terrain attributes, remote sensing indices, geomorphological and geological units) are available, but no soil map has been surveyed. The network then predicts the soil units based on the learned weights (Moonjun, 2007).

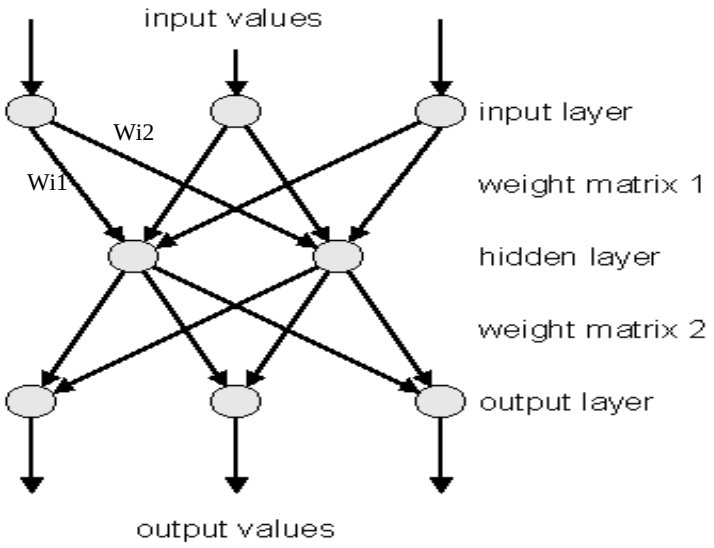


Figure 17. Exemplified topology of a feed-forward multilayer artificial network

The general workflow applied in this study to predict soil classes based on ANN is shown in Figure 18. ANNs are nonlinear models; therefore they can be applied in nonlinear and complex systems such as soils. Behrens et al., (2005) used feed-forward ANN to spatially predict soil units. They showed the suitability of ANN to identify characteristic structures in the distribution of soil units and high prediction power of ANN. Fidêncio et al. (2001) applied two types of neural networks (radial basis function networks and self-organizing maps) to classify soil samples from different geographical regions in Sao Paulo, Brazil by means of their near-infrared (diffuse reflectance) spectra. Zhu (2000) applied and developed an ANN approach to predict the probability of soil classes from soil environment factors.

ANN can be run in R using the package "neuralnet" (Fritsch and Guenther, 2010). "Neuralnet" package contains a very flexible function to train feed-forward neural networks. "Neuralnet" was built to train neural networks in the context of regression analysis and focuses on multiple layer perceptrons, which are well applicable when modeling functional relationships. An arbitrary number of covariates and response variables as well as of hidden layers can theoretically be included. The feed-forward back propagation algorithm used in this

study has multi-layer as input layer, hidden layer and output layer. A detailed description on the subject can be found in Gunther and Fritsch (2010).

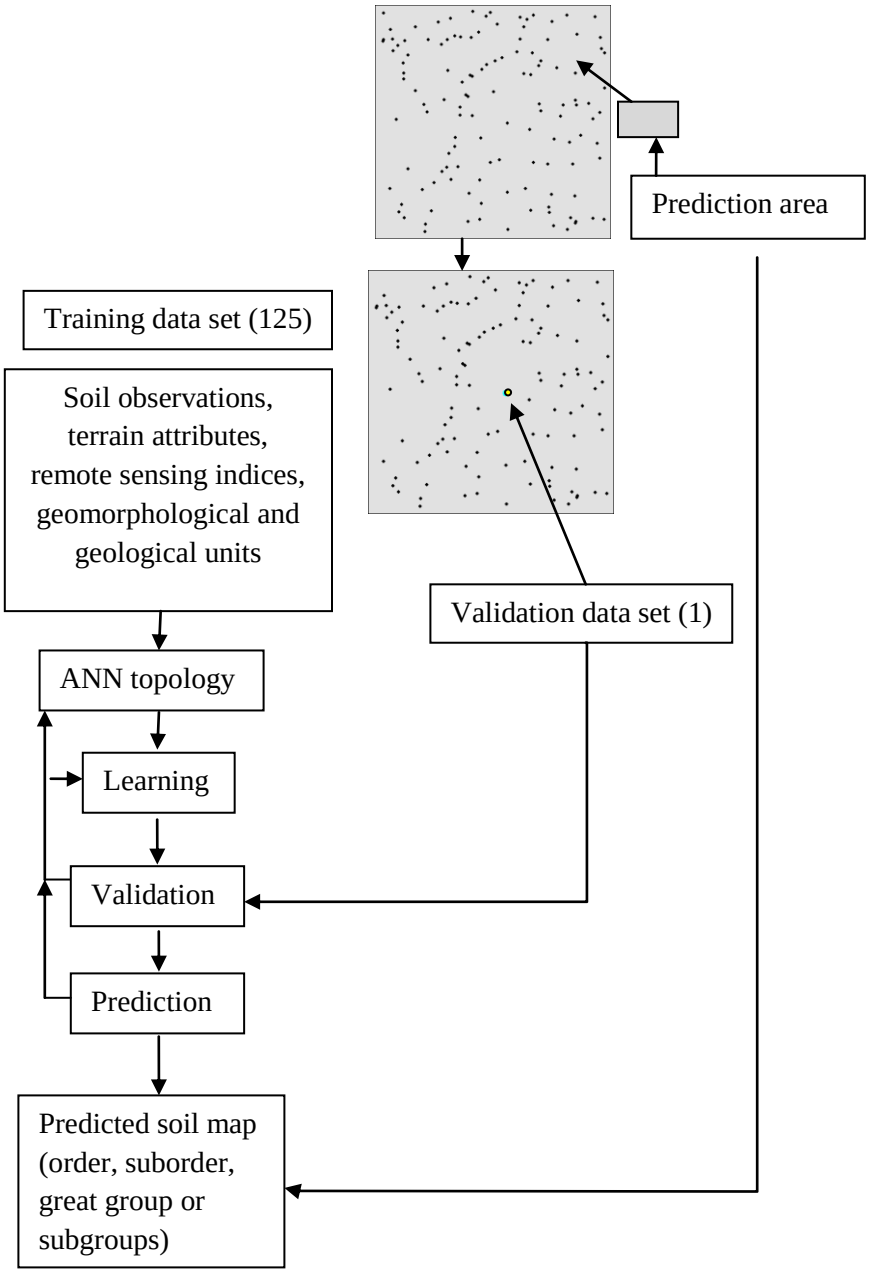


Figure 18. Workflow, applied to learn and predict soil units using artificial neural network

In this study, the response variable was a categoric variable. Therefore, a general strategy of one-versus-all technique was used, where each individual class (coded as 1) is modeled against all the remaining classes (each coded as zero), and k different ensembles are

constructed and then the class with maximum probability is given as the class label (Friedman et al., 2000). The prediction was done for taxonomy categories of order, suborder, great group, and subgroup.

4.3.4. Model validation

When no budget or time is available for additional probability sampling, the best option for cross validation is leave-one-out cross validation (Brus et al., 2011), which is the most common form of n-fold cross-validation (Efron and Tibshirani, 1993). In such a validation, the data set is split n times into a set of n – 1 locations for calibration and one for validation (Figure 18). For each sampling location, the model is refitted leaving that location out of the calibration dataset. The target variable is then predicted for that location. This is done for all the sampling locations and error function or other parameters are computed.

4.3.5. Map purity index

In stratified simple random sampling, the area is sub-divided into sub-areas referred to as strata, and from each stratum a simple sample is selected. In this study, the area was divided into sub-areas based on geomorphic surfaces and there are; therefore, 18 strata. The map purity was estimated as the weighted average of the overall accuracies per stratum $\hat{p}_h$, as also recommended by Brus et al., (2011) for stratified simple random sampling.

$$\hat{p} = \sum_{h=1}^{H} w_h \hat{p}_h = \sum_{h=1}^{H} w_h \frac{\sum_{u=1}^{U} n_{huu}}{n_h} \qquad (20)$$

where w_h is the relative area of stratum h, n_{huu} is the number of sampling locations in stratum h correctly mapped as u, and n_h is the total number of sampling locations in stratum h.

4.3.6. Soil diversity indices

In this study, the taxonomic diversity at the order, suborder, great group and subgroup is discussed. The diversity indices were calculated for both local and global approaches.

In local approach, pedodiversity indices including Shannon entropy, richness and evenness for each geomorphic category were calculated by summation of indices of all patterns incorporated in each category. To calculate the diversity indices in each geomorphic surface, the number of profiles belonging to a given geomorphic surface (n_i) and the total number of profiles in the study area (N) were taken into account.

(i) Richness index

The number of different objects or entities such as soil great groups in a certain ecosystem or predefined territory (e.g. geomorphic categories) was considered as the richness of species.

(ii) Proportional indices

The diversity indices are measured by relative abundance of soil categories to total sampled points in geomorphic units (Ibanez et al., 1995; Phillips, 2001). The proportional abundance of objects is the most frequently used method to estimate the diversity. Evenness refers to the relative abundance of each object in a defined area. Logically, when the evenness of objects is equally probable, the diversity is highest when the richness of comparing units is the same (Ibanez et al., 1995). The most frequently used proportional abundance index is the Shannon index (H′) (Longuet-Higgins, 1971), which is mathematically defined as follows:

$$H' = -\sum_{i=1}^{n} p_i * lnp_i \qquad (21)$$

where H' is the entropy or diversity of the population, and p_i is the proportion of individuals found in *i*th unit. In calculations, the n_i/N was used instead of p_i, where n_i is the number of individuals of the objects belonging to *i*th unit, and N is the total number of individuals

collected. H_{max} (the richness when all objects in reference area are equally probable) is used to measure the evenness (*E*). If the following condition is fulfilled:

$$H' = H_{max} = \ln S \quad (22)$$

Then, the evenness is:

$$E = H'/H_{max} = H'/\ln S \quad (23)$$

Where S is the richness, the number of individuals in each category or map units.

The global approach is based on the moving window technique for the entire digital soil map. For each pixel, the surrounding area (window) is analyzed in terms of spatial structure and the diversity indices are calculated.

4.4. Results and discussion

4.4.1. Digital soil mapping using ANN

Using the feed forward back-propagation algorithm, a number of three layer ANNs as input layer, hidden layer and output layer were trained for the soil map prediction in each category level. In the input layer the number of input nodes was fixed as the number of predictors, which are soil forming factors. The number of hidden layers was adjusted for each soil class. Faussett (1994) found that a topology with one hidden layer is theoretically sufficient to extract the relevant knowledge from a learning data set. Also, Behrens et al. (2005) used one hidden layer in network topology in their study. The other network parameters including the optimum iteration learning rates, the number of hidden-layer nodes and transfer functions were adjusted after stage of learned to train the network and selected the parameters which gave the best fit while the training error < 0.01. The selection of input parameters was determined in respect to the network performance. In "neuralnet", this is done by the criteria such as training error, AIC (Akaike Information Criterion) and BIC

(Bayesian Information Criterion). According to smallest amounts of these parameters, the most influential predictors were selected (Table 11).

Table 11. The selected variables for fitted ANN model of soil class based on soil taxonomy

Soil category level	Covariates
Order	GS*** + El***
Suborder	GS** + MrVBF** + WI
Great group	GS** + PVI +MrVBF* + WI** + El + TWI
Subgroup	GS* + PVI + MrVBF* + WI* + PlCur + Sl + TWI + NDVI

For abbreviations, refer to Table 3, Significance code: '***P< 0.001, **P< 0.01, '*P< 0.05

In most predictions, geomorphic surfaces were the most effective predictors and also, a combined use of terrain attributes and geomorphic surfaces as predictors resulted in the best results. As expected, the number of predictors increased from order to subgroup level (Table 11). Pattern recognition is easier when the model trains for two soil types (at the order level) compared to when the model trains for more different soil types toward soil subgroup (12 subsets at the subgroup level). Therefore, there are not the same soil types for pattern recognition at lower category levels (i.e., Typic Calcigypsid). In such conditions, the relationship of the soil classes and soil covariates could not be learned satisfactorily by the ANN.

It seems that the sample size can affect the prediction performance. In other words, classification criteria between soil classes are very contrasting in higher category levels and also they occur in the larger understanding scale. Therefore, these criteria could be learned precisely based on the covariates such as geomorphology map, relief, and remote sensing indices. Toward lower category levels, differentiation between classification criteria becomes more difficult and, also, classification criteria occur in finer scale. Therefore, some classification criteria for some soil classes might not be learned precisely based on the covariates such as relief, geology, geomorphology, and remote sensing indices used for

prediction. Descending the taxonomic system introduces more properties that might be related to local conditions and natural selection (Toomanian et al., 2007) at lower category levels and can lead to the complexity of system. Therefore, it might not be recognized by the applied covariates.

Validation is based on a single observation from the original sample. Therefore, the learning ability and the prediction ability of the ANN were tested as map purity.

(i) Soil map purity

Table 5 presents the estimated purity of the soil map and also of each stratum based on soil taxonomic hierarchy. Generally, as the category changes from order to subgroup, the purity decreases. This can be related to several reasons which are discussed below:

1) Weak or even no relationship exists between soil classes and environmental factors. Different numbers and types of criteria are necessary for soil classification at the subgroup category. The number and type of soil classification criteria at the order category is less than those at suborder, great group, subgroup, and so on. Considering the fact that a soil order consists of many suborders, a suborder consists of many great groups, a great group consists of many subgroups and a subgroup consists of many soil families, this implies that there are many criteria in the classification of soil units. Descending the taxonomic system introduces more properties and criteria. Therefore, some properties might not be included in the applied covariates and disconnection occurs between soil classes and covariates at lower category. Soil categorical maps were correlated with environmental attributes for each soil taxonomic hierarchical level. Significant relationship between soil categories and environmental attributes particularly geomorphic surfaces, mrvbf, incoming solar radiation and digital elevation model was observed (Table 11). The results showed that the correlation coefficient

for digital elevation model, mrvbf and geomorphic surfaces decreases from order to subgroup taxonomy category.

Generally, soil orders change in large spatial scale according to soil forming factors and processes. Digital soil mapping techniques learn and apply the relationships between soil classes and soil forming factors (or soil covariates). Therefore, the prediction of soil orders is easier than other levels. In the study area, two orders were mapped including Entisol and Aridisol. Entisol was found in Mo111, Mo121, and Sd111 geomorphic surfaces in mountains and sand dunes landscapes (Table 1), while Aridisol occurs in other geomorphic surfaces. The differences between the two orders are mainly related to their topography and age which have been indirectly differentiated in prediction by geomorphology map and elevation. Therefore, the effect of differences in topography and age could lead to the soil map with a high purity at the order level. Generally, map units with more than one soil type are considered less refined than those identified by a single type of soil (Ibanez et al., 1995). Therefore, the map purity is similar for Mo111, Mo121 and Sd111 geomorphic surfaces at order to subgroup levels (Table 12).

The number and diversity of factors and processes affecting soil formation increase toward the lower categories and soil class variability is not well correlated with variations in soil forming factors and processes at lower category levels. Hugget (1998), Phillips et al. (1999), Phillips (2001, 2005), Saldana and Ibanez (2004), and Phillips and Marion (2005) suggest that minor variations in initial sedimentation conditions, small perturbations, weathering, additions, losses, transfers, and transformations could grow unstably over time and bring about an unstable and chaotic condition for soil genesis. In such conditions, pedologic evolution increases independent of variation in other environmental factors. Digital

soil mapping relies on the relationships between soil observations and environmental factors. Therefore, weak predictions are made when weak relationships exist.

Table 12. The estimated purity of soils in each stratum based on soil taxonomy

		Purity of predicted soil map			
Stratum	n	Order	Suborder	Great group	Subgroup
Soil map		0.99	0.71	0.58	0.34
Mo111	3	1	1	1	1
Mo121	3	1	1	1	1
Hi111	7	1	0.86	0.53	0.23
Sd111	7	1	1	1	1
Pl111	21	1	0.76	0.55	0.24
Pl121	6	1	0.83	0.83	0.83
Pl122	2	1	1	1	1
Pi111	2	1	0	0	0
Pi112	5	1	0.6	0.4	0.2
Pi121	5	1	0.4	0.2	0.2
Pi122	4	1	1	1	0.5
Pi211	6	1	0.67	0.5	0.17
Pi212	5	1	0.6	0.4	0.2
Pi311	23	0.95	0.57	0.43	0.13
Pi312	7	1	0.71	0.57	0.14
Pi313	3	1	1	1	1
Pi411	12	1	0.6	0.42	0.17
Pi412	5	1	0.8	0.6	0.2

2) Contrasting soil units: The number of different soil units in each geomorphic surface increases from order to lower categories. Table 13 shows that there are 4 suborders, 5 great groups, and 9 subgroups for Aridisol in the study area. Different soils were mapped at order to subgroup levels, but the number of contrasting soil types decreases and inherent similar soil types increase at order to subgroup levels. For example, Entisols and Aridisols are very highly different soil types, while Gypsic Haplosalid and Calcic Haplosalid are not as much different.

Soils have different minor properties toward lower categories that were created due to detailed processes. Consequently, the ability of model to distinguish between soil units and

thus, soil map purity decreases due to low contrasting soil units from order to subgroup category (Table 12). Olaniyan and Ogunkunle (2007) investigated the purity of the soil map of Nigeria produced by the Federal Department of Agricultural Land Resources. They reported that soil mapping units with high purity included very contrasting soil types.

Table 13. The percent of soil classes in each category level

Order	%N	**Suborder**	%n	**Great groups**	%n	**Subgroups**
Entisol	10	Orthent	5	Torriorthent	5	Typic Torriorthent
		Psamment	5	Torripsamment	5	Typic Torripsamment
Aridisol	90	Salid	30	Haplosalid	30	Typic Haplosalid
						Gypsic Haplosalid
						Calcic Haplosalid
		Calcid	20	Haplocalcid	20	Typic Haplocalcid
						Sodic Haplocalcid
		Cambid	20	Haplocambid	20	Typic Haplocambid
						Sodic Haplocambid
		Gypsid	30	Haplogypsid	24	Typic Haplogypsid
						Sodic Haplogypsid
				Calcigypsid	6	Typic Calcigypsid

On the other hand, the ability of model to predict contrasting soil units depends on the sample size. The results show that the model has weak performance in the stratum with the lower number of soil types. For example, the map purity is very low in the stratum Pi111 in comparison with the stratum Pl122 with the same samples size (Table 12). There are different soil classes including Haplocambid and Haplocalcid in the stratum Pi111, unlike the stratum Pl122 with similar number of soil classes including Haplosalid.

3) Soil diversity: Figure 19 shows the predicted soil map and the Shannon's entropy and richness maps based on soil taxonomic hierarchy. As the predicted number of soil classes increases, the richness and entropy increases from suborder to subgroup category. Therefore, the diversity indices are closely related to the number of soil classes.

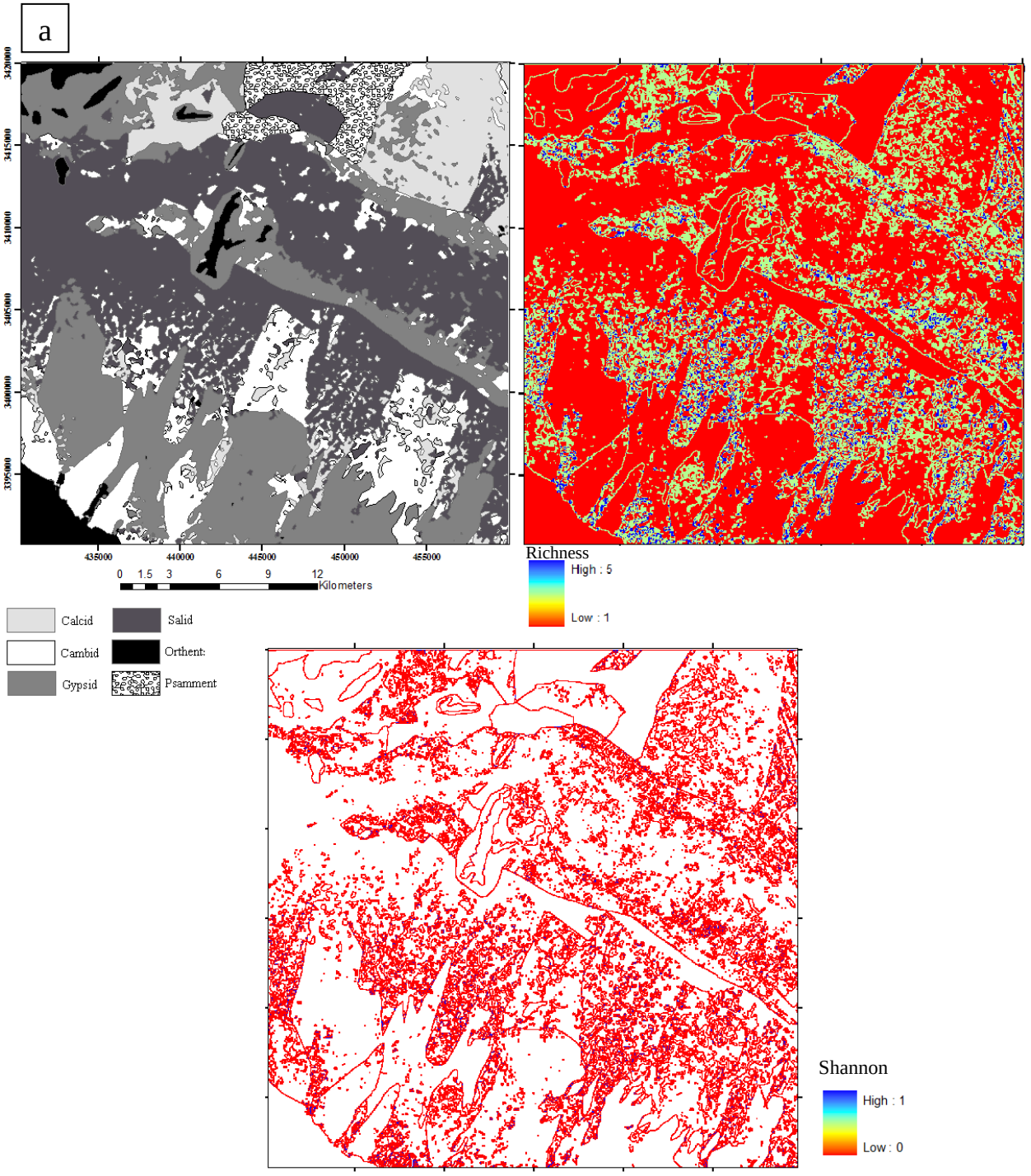
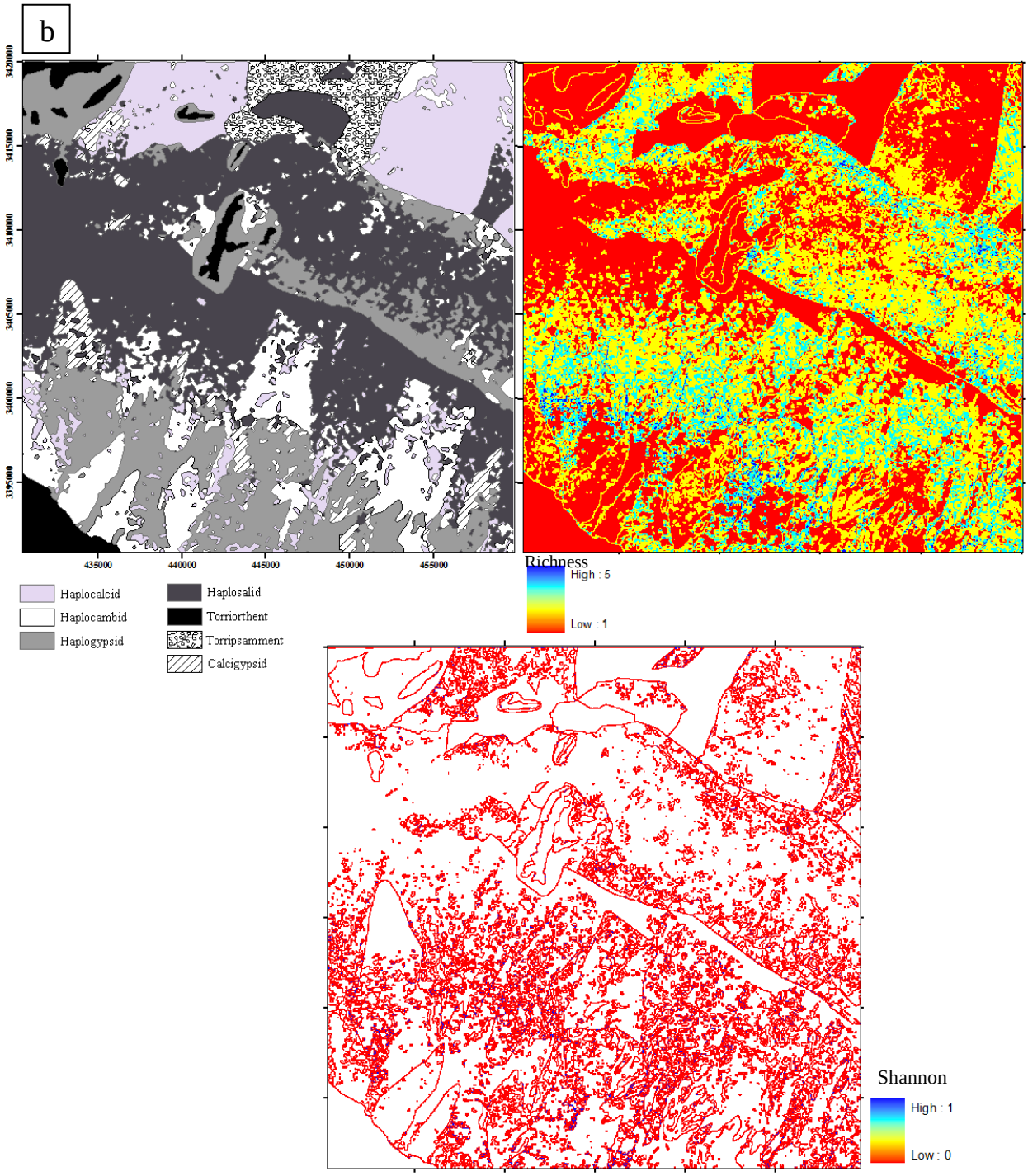
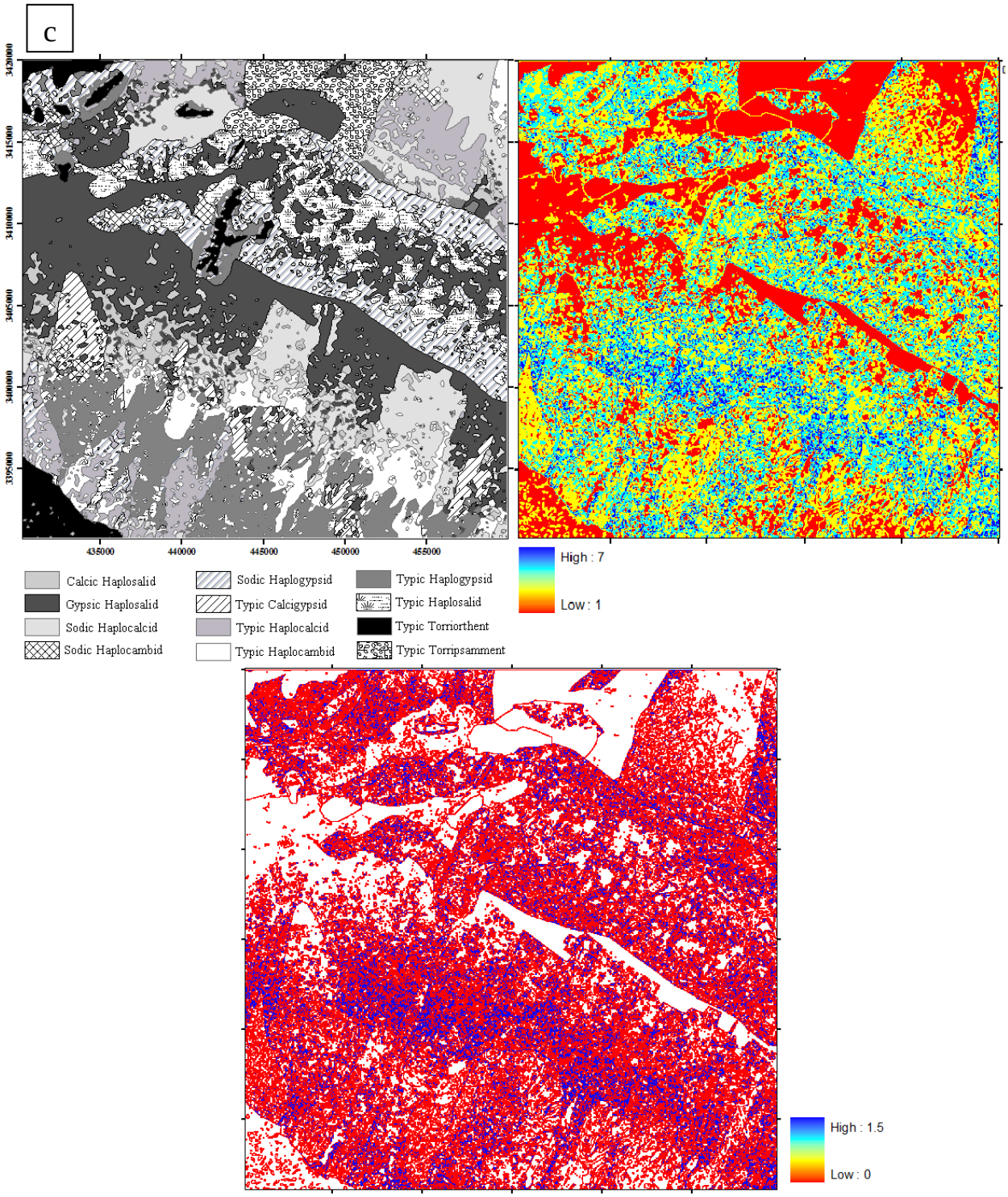


Figure 19. The soil map and the map of richness and Shannon index based on soil taxonomic hierarchy (a: suborder)



Continue of Figure 19. The soil map and the map of richness and Shannon index based on soil taxonomic hierarchy (b: great group)



Continue of Figure 19. The soil map and the map of richness and Shannon index based on soil taxonomic hierarchy (c: subgroup)

Minasny et al. (2010) showed that the areas with more soil mapping units exhibit the largest pedodiversity for the world soil map. They stated that the measure of pedodiversity

depends on the coverage or density of the soil map. Therefore, the greater the number of soil units or the higher density of the soil maps, the higher the diversity at the subgroup category. However, Peterson et al. (2010) argued that direct comparison of pedodiversity measures between studies is not possible as it depends on the type of classification used, the scale and the soil survey intensity.

The prediction uncertainty was quantified by Shannon's entropy (Kempen et al., 2009). An increase in entropy index means an increase in the prediction uncertainty and a decrease in the map purity based on soil taxonomy hierarchy. The richness and entropy map were prepared based on the predicted soil map at each category level. A clear relationship was found between pedodiversity indices and the environmental variables. The highest correlation was observed between the wetness index and pedodiversity indices (Figure 20). No variation in pedodiversity indices was found at the locations with the lowest wetness index. Highly different soil units occur in bajada and playa (Figure 20). The highest entropy was obtained for playa, dissected bajada, and dissected old bajada and the lowest entropy for mountains and sand dunes.

An increase in the diversity and a decrease in the rate of map purity based on soil taxonomic hierarchy confirm that heterogeneity is progressing in the area. This is also proven when we change the spatial scale and focus on the diversity trend in finer scale through geomorphic hierarchy, in other words, local diversity.

Pedodiversity indices based on taxonomic hierarchy are presented in Table 14. Since the diversity indices at the soil order category was the same for all of the geomorphic surfaces, they were all presented in a single column in Table 14.

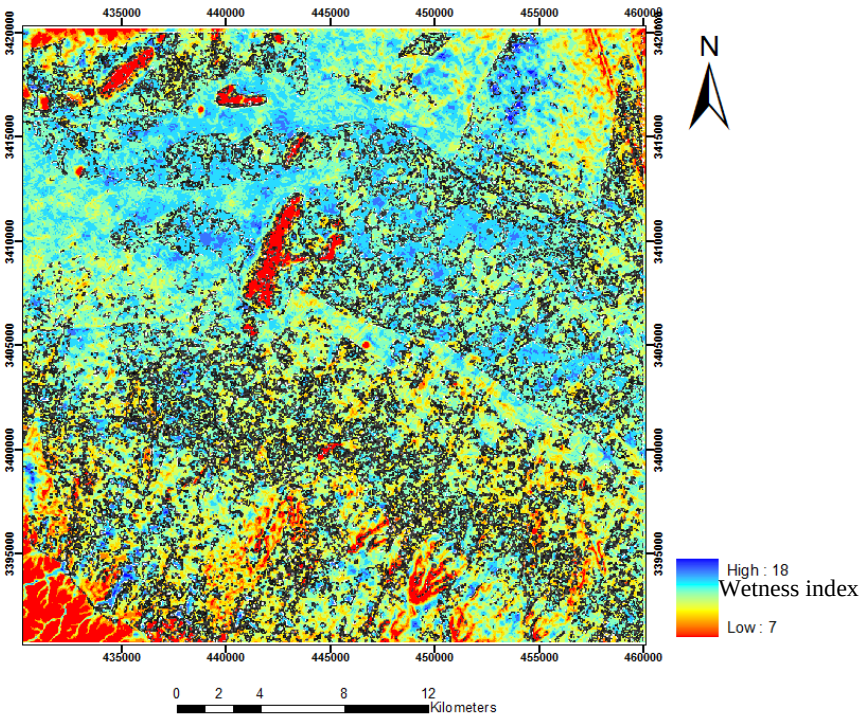


Figure 20. the correlation map wetness index and Shannon index on the map of wetness index. The black lines are high correlation of wetness and Shannon indices.

Table 14. Pedodiversity of geomorphic surfaces based on taxonomy hierarchy

	Order	Suborder			Great groups			Subgroups		
Stratum	H′, E	S	H′	E	S	H′	E	S	H′	E
Mo111	0	1	0	0	1	0	0	1	0	0
Mo121	0	1	0	0	1	0	0	1	0	0
Hi111	0	1	0	0	2	0.41	0.59	2	0.41	0.59
Sd111	0	1	0	0	1	0	0	1	0	0
Pl111	0	3	0.9	0.82	4	1.3	0.92	6	1.72	0.96
Pl121	0	2	0.45	0.65	2	0.45	0.65	3	0.63	0.92
Pl122	0	1	0	0	1	0	0	1	0	0
Pi111	0	2	0.69	0.99	2	0.69	0.99	2	0.69	0.99
Pi112	0	2	0.69	0.99	3	0.95	0.86	4	1.04	0.96
Pi121	0	3	1.05	0.95	3	1.05	0.95	4	1.33	0.96
Pi122	0	1	0	0	1	0	0	2	0.56	0.81
Pi211	0	3	0.87	0.79	3	0.87	0.79	4	1.32	0.95
Pi212	0	3	0.95	0.86	3	0.95	0.86	4	1.33	0.96
Pi311	0	4	1.22	0.88	4	1.22	0.88	7	1.89	0.97
Pi312	0	3	0.96	0.87	3	0.96	0.87	4	1.24	0.89
Pi313	0	1	0	0	1	0	0	1	0	0
Pi411	0	4	1.24	0.89	5	1.51	0.97	7	1.85	0.99
Pi412	0	3	0.95	0.86	3	0.95	0.86	3	0.95	0.86

The diversity indices increase soil order to soil subgroup. This is probably due to simultaneous increase in the richness and evenness through this hierarchical method. Toomanian et al. (2006) studied the soil diversity in geomorphic surfaces of Zayandeh-rud valley and reported similar results. The highest diversity indices were observed for the subgroup category.

In Mo111, Mo121, Sd111, Pl122, and Pi313 units, the general trend of increase in the diversity indices from order to subgroup is not similar to what observed in other units. The results of low diversity indices for mountain landscapes are in line with the results of Behrens et al. (2009). Behrens et al. (2009) indicated that very low pedodiversities in mountain landscapes is not expected since the diversity in such landscapes is generally high due to varying parent materials and high soil-forming relief energy and, hence, various denudation and accumulation processes. They attributed the low diversity to problems of mapping small, elongated structures at small scales. In addition to these problems, the dry climate of the study area might be a reason for the low diversity in these units. The diversity indices increase at the soil family category due to the effect of different parent materials on carbonates and soil mineralogy (data not shown).

The sand dune landscape has formed during Quaternary and is the youngest landscape in the study area. Therefore, age and climate are the major factors responsible for the low diversity in sand dunes. The diversity indices increase at the soil family category in this geomorphic surface which can be mainly attributed to irregular sedimentation of fine and coarse materials.

An important factor in pedodiversity analysis is sampling density. The presentation of soil diversity is more accurate at high sampling density, particularly in well developed landforms. Landforms playa (Pl111) and bajada (Pi311) with largest sample size show the

highest diversity parameters (Table 14). The sampling density influences the density of soil map or the presence of different soil classes and, therefore, could affect the presentation of soil diversity.

The trend in diversity indices is opposite to that of soil map purity based on taxonomic hierarchy in geomorphic surfaces (Figures 21 and 22). Decreasing slope of the soil map purity is exactly similar to increasing slope of the Shannon index. These illustrate that the indices are inter-related via a parameter. The key parameter is the number of different soil types in each stratum (richness). Minasny et al. (2010) showed that Shannon entropy is closely related to the number of soil classes. When the number of different soil classes or richness increases, greater number of fractions is summed in the Shannon index. Therefore, geomorphic surfaces with more richness have more entropy and diversity (Table 14).

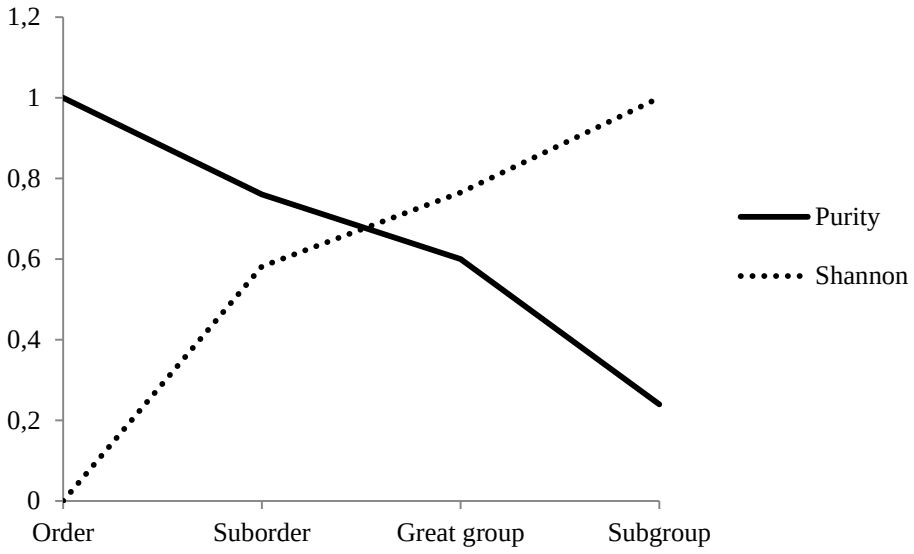


Figure 21. Relationship between pedodiversity indices and map purity based on taxonomic hierarchy for Pl111 geomorphic surface

Among the geomorphic surfaces, playa (Pl111) and bajada (Pi311 and Pi411) are highly diverse. Therefore, the number of different soil types (richness) increases based on soil taxonomic hierarchy and as a result Shannon index increases (Table 14). For example in

stratum Pi411 with 12 sample points, there are 4 different soil classes at great group level and 7 ones at subgroup category. Therefore, 4 fractions are summed at great group category and 7 fractions are summed at subgroup category, while the denominator is constant.

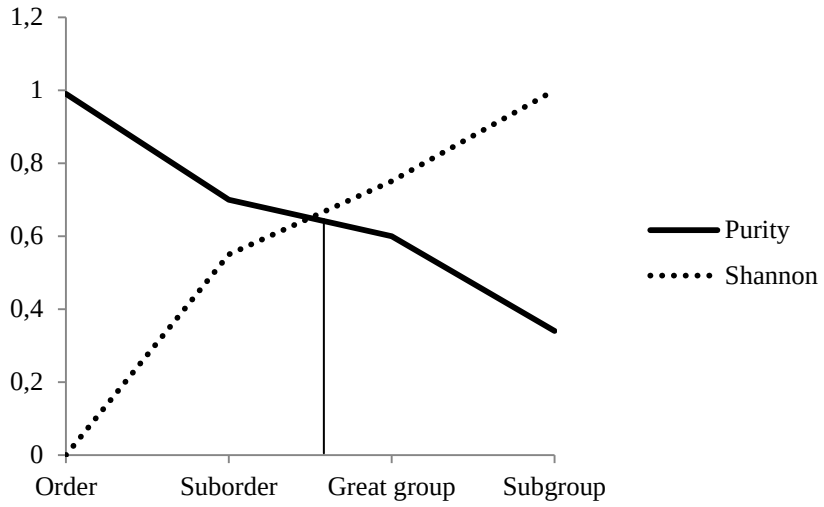


Figure 22. Relationship between pedodiversity indices and purity of the predicted soil map based on taxonomic hierarchy

The entropy indices are representative of deterministic soil complexity (Phillips, 1996). Therefore, any increase in entropy in the study area from order to subgroup category represents higher complexity of the soil system. Moreover, an increase in entropy and the number of different soil types influences the prediction ability of the model. When the system complexity increases, the number of different soil classes actually increases and, therefore the model should be trained for more different soil classes. It means that there are fewer observations per class for training of the model. Consequently, uncertainty increases for prediction of each soil class and soil map purity decreases for the soil category and geomorphic surfaces with more different soil classes or highly diverse classes (Table 14). Therefore, soil diversity directly influences the soil map purity as it is a reflection of the density of soil maps (Minasny et al., 2010).

In terms of management practices, we need a soil map with high purity that adequately represents soil diversity. The pedodiversity measures are related to the density of soil map or presence of various soil units. The soil maps with more different soil units have a higher diversity. Therefore, where the soil mapping is faced with restrictions such as hard working conditions in arid regions, we have to project the most efficient way for the soil mapping in terms of applicability for users.

(ii) The combined index

Soil mapping should be carried out at the soil category with high purity and also, it should represent the real soil diversity. The obtained results showed that excessive increase or decrease of one of them leads to loss of useful information. Relationships between pedodiversity indices and purity based on taxonomic hierarchy are presented in Figs. 21 and 22. These relationships are illustrated for a geomorphic surface (Pl111) and for the predicted soil map of each taxonomic category in Figs. 21 and 22, respectively. In both figures, the highest purity and lowest entropy are observed at the order level. Soil mapping at the order category is not appropriate for management purposes in the study area. On the other hand, soil mapping at the level of subgroup illustrates a high diversity and low purity (Figs. 21 and 22). In such circumstances, uncertainty and impurity of soil map are very high which are not acceptable for users.

Figs. 21 and 22 indicate that the best category for soil mapping is the intersection point of the pedodiversity and purity graphs. This point lies between the suborder and great group categories and includes 60 percent of diversity and purity. Which category is better: suborder or great group?

We introduced a logical index via both diversity and purity. It seems that multiple Shannon entropy and purity are a more logical index which is presented in Figs. 23 and 24.

This index shows that the best category for soil mapping in the study area is the great group. However, there is not a big difference between suborder and great group categories, because there is not much difference between diversity indices at these two levels. The same results were also obtained by Toomanian et al. (2006). The increasing trend of the entropy through hierarchical downscale method shows that diversity and heterogeneity is progressing in the study area. Therefore, it seems that soil mapping will be more efficient at the great group category. The soil map at this level includes soil classes such as Calcigypsid which is not mapped at the suborder level.

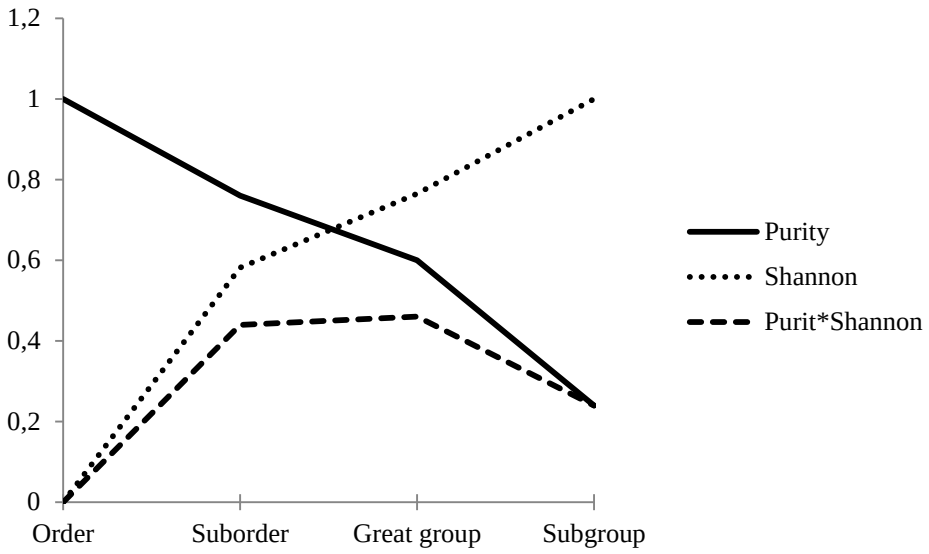


Figure 23. variation of logical index based on soil taxonomic hierarchy in Pl111 geomorphic surface

We propose a framework for purposive soil mapping for dry areas with similar geomorphology 1) Mountains and sand dunes do not show diversity due to dry climate and the low age. They have been omitted in some studies in arid regions such as Toomanian et al. (2006). Therefore, it seems that the suborder level is good enough for mapping such land forms. Besides, the mapping of these land forms is essential because of natural resources and erosion hazards.

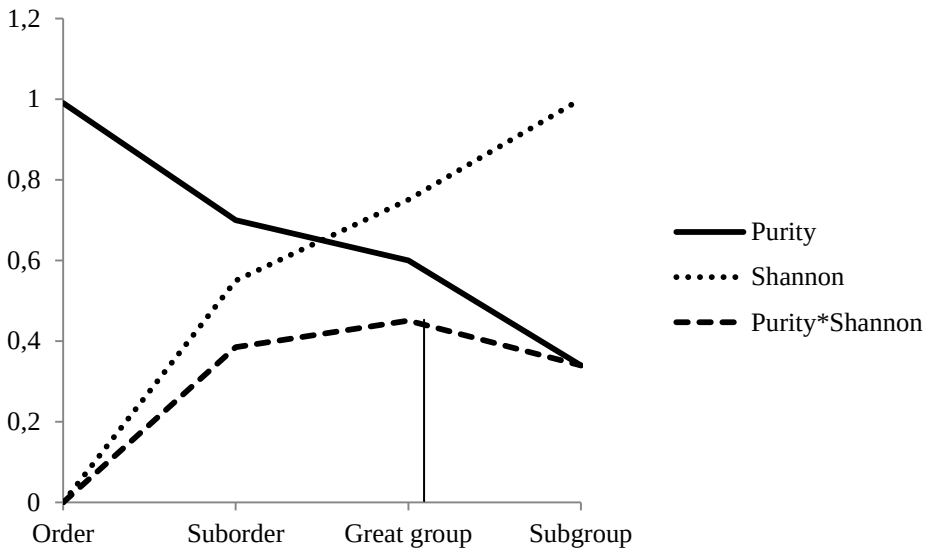


Figure 24. variation of logical index of the predicted soil map based on soil taxonomic hierarchy in the study area

2) Gypsiferous hills rich in gypsum are classified as Gypsid suborder with no diversity at the suborder category (Table 14). On the other hand, the gypsiferous hills belong to Neogene period based on the geology of the study area (Nazemzadeh and Azizani, 1991). Therefore, they are older and more diverse than the sand dunes and the entropy increases from suborder to great group category (Table 14). The study of Toomanian et al. (2006) in the arid region showed that with an increase in the age of geomorphic surfaces, the entropy and the richness indices increase. Again, dry climate and low age are the limiting factors for higher diversity from great group to subgroup category. Therefore, soil mapping at the great group level is appropriate in the gypsiferous hills and also more purity is achieved. This confirms the results obtained by the use of the logical index.

3) Playa, alluvial fan and bajada: In most geomorphic surfaces of these landforms, increasing entropy and decreasing purity were observed based on soil taxonomic hierarchy. The defined logical index showed that soil mapping at great group category is efficient. Because decreasing purity is high at subgroup category and uncertainty increases in the

predicted map. Consequently, the predicted map in subgroup category will not be very useful, while, the results present significant entropy and purity in great group category at this scale. This analysis is conditioned to a chosen sampling density that in the study area based on the sampling density applied; great group level was the best category for soil mapping.

Chapter 5
Conclusions

The following conclusions can be drawn from this study:

- The use of the geomorphology map greatly improves the prediction accuracy of digital soil map. The best predictions in this investigation could be achieved when soil forming factors were simultaneously used in the modelling approach. The spatial distribution of soils in the study area followed the distribution pattern of most geomorphic and terrain attributes.
- Soils that are highly influenced by topographic and geomorphic characteristics in the study area such as Haplosalids, Haplogypsids and Torripsamments, were predicted more accurately than those only slightly influenced by topographic and geomorphic characteristics such as Haplocambids and Calcigypsids.
- As a reliable and flexible approach, logistic regression could successfully be used to prepare continuous digital soil maps. The application of decision trees for prediction of soil types could be a promising alternative.
- In addition to their application in land-use change studies, ROC curves could be successfully used for grouping soil classes.
- The size and the spatial distribution of samples in different soil classes greatly influence the quality of digital soil maps.

- The integration of GIS data into R software provides the opportunity to predict soil classes in adjacent areas. The GIS based softwares such as SAGA and powerful statistical softwares such as R can easily support soil survey and mapping tasks. Altogether, an extended digital terrain analysis approach and clear description of geomorphological, geological and pedological processes could be a promising key technology in future soil mapping.
- The soil map purity is affected by soil diversity as their trend changing was opposite from order to subgroup category level.
- The diversity indices maps are based on soil map, so diversity measures depend on the density of the soil map. The density of soil map is density of different soil units. Therefore, higher soil diversity was achieved with greater density of different soil units.
- Based on the use of combined index, the best category for soil mapping in the study area is the great group. Soil mapping at the level of subgroup illustrates high diversity and low purity. In such circumstances, uncertainty and impurity of soil map are very high which are not appropriate for users.
- Sampling density is an important factor for determination of spatial variations of entropy.

References

Akaike, H. 1973. Information theory and an extension of the maximum likelihood principle. In: Second International Symposium on Information Theory. pp. 267–281. Akademiai Kiado, Budapest.

Bauer, E., Kohavi, R., 1999. An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. Machine Learning 36, 105–142.

Beaudette, D.E., O'Geen, A.T. 2009. Quantifying the aspect effect: An application of solar radiation modeling for soil survey. Soil Science Society of American Journal, 73, 1345-1352.

Behrens, T., Scholten, T., 2006. Digital soil mapping in Germany–a review. Journal of Plant Nutrition & Soil Science 169, 434–443.

Behrens, T., Zhu, A.X., Schmidt, K., Scholten, T. 2010. Multi-scale digital terrain analysis and feature selection for digital soil mapping. Geoderma, 155, 175-185.

Behrens, T., Forster, H., Scholten, Th., Steinrucken, U., Spies, Goldschmitt Michael., E.D., 2005. Digital soil mapping using artificial neural networks. Journal of Plant Nutrition & Soil Science 168, 21-33.

Behrens, Th., Schneider, O., Losel, G., Scholten, Th., Hennings, V., Felix-Henningsen, P., Hartwich, R., 2009. Analysis on pedodiversity and spaltial subset representativty-the German soil map 1:1000,000. Journal of Plant Nutrition & Soil Science 172, 91-100.

Birkeland, P.W., Shroba, R.R., Burns, S.F., Price, A.B., Tonkin, P.J. 2006. Integrating soils and geomorphology in mountains-an example from the Front Range of Colorado. Geomorphology, 55, 329-344.

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

Brus, D.J., Kempen, B., Heuvlink, G.B.M., 2011. Sampling for validation of digital soil maps. European Journal of Soil Science 62, 394-407.

Burrough, P. A., 1993. Soil variability: a late 20th century view. Soils Fertilizers 56, 529–562.

Campling, P., Gobin, A. & Feyen, J. 2002. Logistic modeling to spatially predict the probability of soil drainage classes. Soil Science Society of America Journal 66, 1390-1401.

Caniego, F.J., Ibanez, J.J., San Jose Martinez, F., 2007. Renyi dimensions and pedodiversity indices of the earth pedotaxa distribution. Nonlinear Processes Geophysics 14, 547-555.

Cantón, Y., Solé-Benet, A., Lázaro, R., 2003. Soil-geomorphology relations in gypsiferous materials of the Tabernas Desert (Almería, SE Spain). Geoderma 115, 193-222.

Congalton, R.G., 1991. A review of assessing the accuracy of classifications of remotely sensed data. Remote Sensing Environment 37 (1), 35-46.

Costantini Edoardo, A.C., Barbetti, R., Righini, G., 2002. Managing the uncertainty in soil mapping and land evaluation in areas of high pedodiversity. Methods and strategies applied in the province of Siena (Central Italy). 7th International Meeting on Soils with Mediterranean Type of Climate. Options Méditerranéennes, Série A n.50. Valenzano (Bari) Italia.

De Gruijter, J.J., Brus, D.J., Bierkens, M.F.P., Knotters, M., 2006. Sampling for Natural Resource Monitoring. Springer, Berlin.

De'ath, G., 2007. Boosted trees for ecological modeling and prediction. Ecology 88 (1), 243-251.

Debella-Gilo, M., Etzelmuller, B. 2009. Spatial prediction of soil classes using digital terrain analysis and multinomial logistic regression modeling integrated in GIS: Examples from Vestfold County, Norway. Catena 77, 8-18.

Degórski, M., 2003. Pedodiversity as a part of geodiversity in creation of landscape structure. Multifunctional Landscapes. Monitoring, Diversity and Management, vol. II. WIT PRESS, Southampton, Boston, pp. 105–121.

Deventer, van A.P., Ward, A.D., Gowda, P.H., Lyon, J.G., 1997. Using thematic mapper data to identify contrasting soil plains and tillage practices. Photogrammetric Engineering Remote Sensing 63 (1), 87-93.

D'heygere, T., Goethals, P., De Pauw, N., 2006. Genetic algorithms for optimisation of predictive ecosystems models based on decision trees and neural networks. Ecological Modelling 195, 20-29.

Dobos, E., Montanarella, L., Negre, T., Micheli, E., 2001. A regional scale soil mapping approach using integrated AVHRR and DEM data. International Journal of Applied Earth Observation Geoinformation 3 (1), 30-42.

Efron, B. and Tibshirani, R. J. 1993. An Introduction to the Bootstrap. Chapman & Hall/CRC, London.

Elith, J., Leathwick, J.R., Hastie, T., 2008. A working guide to boosted regression trees. Journal of Animal Ecology 77, 802-813.

Faussett, L. V. (ed.) 1994. Fundamentals of neural networks: architectures, algorithms, and applications. Prentice Hall, Englewood Cliff, N. J.

Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., 1996. From data mining to knowledge discovery in databases. American Association for Artificial Intelligence Press, Menlo Park. pp. 1-34.

Fidêncio, P.H., Ruisanchez, I., Poppi, R.J., 2001. Application of artificial neural net works to the classification of soils from Sao Paulo state using near-infrared spectroscopy. Analyst 126, 2194–2200.

Finke, P.A., Meylemans, E., Van de Wauw, J., 2008. Mapping the possible occurrence of archaeological sites by Bayesian inference. Journal Archaeological Science 35, 2786-2796.

Foody, G.M., 2002. Status of land cover classification accuracy assessment. Remote Sensing Environment 80, 185–201.

Freeman, E.A., Moisen, G.G., 2008. A comparison of the performance of threshold criteria for binary classification in terms of predicted prevalence and kappa. Ecological Modelling 217, 48-58.

Freidman, J.H., 1999. Stochastic gradient boosting. Technical Report. Department of Statistics, Stanford University.

Friedman, J.H, Hastie, T., Tibshirani, R., 2000. Additive Logistic Regression: A Statistical View of boosting. Analysis Statistic 28 (2), 337-407.

Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. Analysis Statistic 32 (2), 407-499.

Friedman, J.H., Meulman, J.J., 2003. Multiple additive regression trees with application in epidemilogy. Statistics in Medicine 22, 1365–1381.

Fritsch, S., Guenther F., 2010. Package 'neuralnet'. R-News. 2010:02:23.

Gallant, J.C., Dowling, T.I., 2003. A multiresolution index of valley bottom flatness for mapping depositional areas. Water Resource Research 39 (1), 1347.

Giassen, E., Clarke, R.T., Junior, A.V.I., Merten, G.H., Tornquist, C.G. 2006. Digital soil mapping using multiple logistic regression on terrain parameters in Southern Brazil. Science Agriculture 63, 262-268.

Girard, M.C., Girard, C.M., 1999. Traitements des données de télédétection. Dunod, Paris. 529p.

Golosov, V., Sidorchuk, A. & Walling, D.E. 2008. Nikolay I. Makkaveev and development of fluvial geomorphology in Russia and the former Soviet Union. Catena, 73, 146-150.

Grinand, C., Arrouays, D., Laroche, B., Martin, M.P., 2008. Extrapolating regional soil landscapes from an existing soil map: Sampling intensity, validation procedures, and integration of spatial context. Geoderma 143, 180-190.

Grunwald, S., 2006. Environmental Soil-Landscape Modelling, Geographic Information technologies and Pedometrics. Taylor and Francis Group.

Gue, Y., Gong, P., Amundson, R., 2003. Pedodiversity in the United States of America. Geoderma 117, 99-115.

Guisan, A., Zimmermann, N.E., Elith, J., Graham, C.H., Philips, S., Peterson, A.T. 2007. What matters for predicting the occurrences of trees: Techniques, data, or species characteristics? Ecological Monographs, 77, 615-630.

Gunther, F., Fritsch, S., 2010. Neuralnet: Training of Neural Networks. The R Journal, Vol.2/1, June 2010. 30-38.

Hengl, T., Toomanian, N., Reuter, H., Malakouti, M.J., 2007. Methods to interpolate soil categorical variables from profile observations: Lessons from Iran. Geoderma 140, 417-427.

Hosmer, D.W., Hjort, N.L. 2002. Goodness-of-fit processes for logistic regression: simulation results. Statistics in Medicine, 21, 2723–2738.

Hudson, B.D., 1992. The soil survey as paradigm-based science. Soil Science Society of American Journal 56, 836–841.

Hugget, R.J., 1998. Soil chronosequences, soil development, and soil evolution: a critical review. Catena 32, 155–172.

Ibanez, J.J., De Alba, S., lobo, A., Zucarello, V., 1998a. Pedodiversity and global soil patterns at coarse scales (with discussion). Geoderma 83, 171–214.

Ibanez, J.J., De-Alba, S., Bermudez, F.F., Garcia-Alvarez, A., 1995. Pedodiversity: concepts and measures. Catena 24, 215-232.

Ibanez, J.J., Jimenez-Ballesta, R., Garcıa Alvarez, A., 1990. Soil landscapes and drainage basins in Mediterranean mountain areas.Catena 17, 573–583.

Ibanez, J.J., Perez-Gonzalez, A., Jimenez-Ballesta, R., Saldana, A., Gallardo, J., 1994. Evolution off luvial dissection landscapes in Mediterranean environments. Quantitative estimates and geomorphological, pedological and phytocenotic repercussions. Z. Geomorphology 38, 105–119.

Jafari, A., Finke, P.A., Van De Wauw, J., Ayoubi, S., Khademi, H., 2012. Spatial prediction of USDA-soil great groups in arid Zarand region, Iran: comparing logistic regression approaches to predict diagnostic horizons and soil types. Eur. J. Soil Sci. (In press)

Jenny, H., 1941. Factors of Soil Formation. McGraw-Hill Book Company Inc., New York.

Kempen, B., Brus, D.J., Heuvlink, G.B.M., Stoorvogel, J.J., 2009. Updating the 1:50000 Dutch soil map using legacy soil data: A multinomial logistic regression approach. Geoderma 151, 311-326.

Lagacherie, P., 1992. Formalisation des lois de distribution des sols pour automatiser la cartographie pédologique à partir d'un secteur pris comme référence. Mémoire de Thèse, Université de Montpellier. Institut National de la Recherche Agronomique, France. 175p.

Lagacherie, P., 2005. Using a fuzzy pattern matching algorithm for allocating soil individuals to pre-existing soil classes. Geoderma 128, 274-288

Lagacherie, P., McBratney, A.B., Voltz, M. 2007. Digital Soil Mapping: An Introductory Perspective. Developments in Soil Science, Vol. 31. Elsevier, Amsterdam.

Lane, P.W. 2002. Generalized linear models in soil science. European Journal of Soil Science, 53, 241-251.

Lawrence, R., Bunn, A., Powell, S., Zambon, M., 2004. Classification of remotely sensed imagery using stochastic gradient boosting as a refinement of classification tree analysis. Remote Sensing Environment 90, 331–336.

Lek, S., Guegan, J., 1999. Artificial neural networks as a tool in ecological modelling, an introduction. Ecological Modelling 120, 65–73.

Liu, T.L., Juang, K.W., Lee, D.Y., 2006. Interpolating soil properties using kriging combined with categorical information of soil maps. Soil Science Society of American Journal 70, 1200-1209.

Longuet-Higgins, M.S., 1971. On the Shannon–Weaver index of diversity, in relation to the distribution of species in bird censuses. Theoretical Population Biology 2, 271–289.

Luoto, M., Hjort, J., 2005. Evaluation of current statistical approaches for predictive geomorphological mapping. Geomorphology 67, 299–315.

Manel, S., Williams, H.C. & Ormerod, S.J. 2001. Evaluating presence-absence models in ecology: the need to account for prevalence. Journal of Applied Ecology, 38, 921-931.

Martin, M.A., Pachepsky, Y.A., Perfect, E., 2005. Scaling, fractals and diversity in soils and ecohydrology. Ecological Modeling 182, 217–220.

McBratney, A.B., Mendonca-Santos, M.L., Minasny, B., 2003. On digital soil mapping. Geoderma 117, 3–52.

McBratney, A.B., Odeh, I.O.A., Bishop, T.F.A., Dunbar, M.S., Shatar, T. M., 2000. An overview of pedometric techniques for use in soil survey, Geoderma 97, 293–327.

McGarigal, K., Marks, B.J., 1995. FRAGSTATS: spatial pattern analysis program for quantifying landscape structure. USDA For. Serv. Gen. Tech. Rep. PNW-351.

McKenzie, N.J. & Ryan, P.J. 1999. Spatial prediction of soil properties using environmental correlation. Geoderma, 89, 67-94.

Minasny, B., McBratney, A.B., Hartemink, A.E., 2010. Global pedodiversity, taxonomic distance, and the World Reference Base. Geoderma 155, 132-139.

Ministry of Economy, Trade and Industry of Japan (METI) and the National Aeronautics and Space Administration (NASA). 2009. Aster Global Digital Elevation Model (Aster GDEM). NASA Official. http://www.gdem.aster.ersdac.or.jp.

Mitchell, T., 1997. Machine Learning. McGraw Hill, New York.

Moonjun, R., 2007. Application of artificial neural network and decision tree in a GIS-based predictive soil mapping for landslide vulnerability study. A case study of Hoi Num Rin Sub-watershed, Thailand. Master Thesis.

Moran, J.M., Bui, E.N., 2002. Spatial data mining for enhanced soil map modelling. International Journal of Geographic Information Science 16 (6), 533–549.

Nazemzadeh, M., Azizani., H., 1991. The report of Geological deposits during fourth era for the Zarand plain. Geological management of southeast region, Kerman.

Olaniyan, J.O., Ogunkunle, A.O., 2007. An evaluation of the soil map of Nigeria: II. Purity of mapping unit. Journal of World Association of Soil and Water Conservation, J2: 97-108.

Olaya, V.F., 2004. A gentle introduction to Saga GIS. The SAGA User Group e.V, Göttingen, Germany, p. 208.

Paton, T.R., Humphreys, G.S., Mitchell, P.B., 1995. Soils. In: A New Global View. UCL Press, London.

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. In: Proceedings

of the 8th International Symposium on Remote Sensing of Environment, pp. 1357-1381. Environmental Research Institute of Michigan, Ann Arbor, Michigan, USA.

Peterson, A., Grongroft, A., Miehlich, G., 2010. Methods to quantify the pedodiversity of 1 km2 areas-results from southern Africa drylands. Geoderma 55, 140-146.

Phillips, J. D., 2001. Divergent evolution and the spatial structure of soil landscape variability. Catena 43, 101–113.

Phillips, J. D., 2005. Weathering instability and landscape evolution. Geomorphology 67, 255–272.

Phillips, J.D., 1998. On the relation between complex systems and the factorial model of soil formation (with discussion). Geoderma, 86: 1-42.

Phillips, J.D., 1999. Methodology, scale, and the field of dreams. Department of Geology, Texas A$M University. Annals AAG 754-759.

Phillips, J.D., 2002. Global and local factors in earth surface systems. Ecological Modeling 149, 257–272.

Phillips, J.D., Gares, P.A., Slattery, M.C., 1999. Agricultural soil redistribution and landscape complexity. Landscape Ecology 14, 197–211.

Phillips, J.D., Marion, D., 2005. Biomechanical effects, lithological variations and local pedodiversity in some forest soils of Arkansas. Geoderma 124, 73-89.

Pontius, R.G., Schneider, L.C., 2001. Land-cover change model validation by an ROC method for the Ipswich watershed, Massachusetts, USA. Agric. Ecosystem Environment 85, 239-248.

R Development Core Team. 2011. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Richardson, A.J., Wiegand, C.L., 1977. Distinguishing vegetation from soil background information. Photogrammetric Engineering Remote Sensing 43 (12), 1541-1552.

Ridgeway, G., 2007. Gbm: Generalized Boosted Regression Models, R Package version 1.6-3. URL http://www.ipensieri.com/gregr/gbm.shtml.

Ripley, B., 1996. Pattern Recognition and Neural Networks. Cambridge University Press, New York. 403pp.

Rossiter, D.G. & Loza, A.V. 2010. Technical note: Analyzing land cover change with logistic regression in R (Version 2.2, First version April 2004). ITC, Enschede, The Netherland.

Rouse, J.W., Hass, R.H., Schell, J.A., Deering, D.W., 1973. Monitoring vegetation systems in the Great Plains with ERTS. In: S.C., Freden, E.P., Mercanti, M.A. Becker (Eds.), NASA SP-351: Proc. Third Earth resources Tech. Satellite-Symp. Vol. 1: Technical Presentations Sec. A. Washington, DC: NASA Science and Technology Information Office, pp. 309-317.

Saldadna, A., Ibaanez, J.J., 2004. Pedodiversity analysis of three fluvial terraces of the Henares River (central Spain). Geomorphology 62, 123–138.

Scull, P., Franklin, J., Chadwick, O.A. 2005. The application of classification tree analysis to soil type prediction in a desert landscape. Ecological Modelling, 181, 1-15.

Scull, P., Franklin, J., Chadwick, O.A., McArthur, D. 2003. Predictive soil mapping: a review. Progress in Physical Geography, 27, 171-197.

Shmida, A., Burgess, T.L. 1988. Plant growth-form strategies and vegetation types in arid environments. In: Plant Form and Vegetation Structure (ed. M.J.A. Werger), pp. 211 –241. SPB Academic Publishing, The Hague, Netherlands.

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

Stern, H., 1996. Neural networks in applied statistics. Techno metrics 38, 205–220.

Toomanian, N., Jalalian, A., Khademi, H., Karimian Eghbal, M., Papritz, A. 2006. Pedodiversity and pedogenesis in Zayandeh-rud Valley, Central Iran. Geomorphology, 81, 376-393.

U.S. Geology Survey (USGS), 2004. Geology.com/news/2010/free-lansat-images-from-USGS-2.shtml. URL http://glovis.usgs.gov.

Venables, W., Ripley, B., 2002. Modern Applied Statistics with S. Springer-Verlag, Berlin. 495 pp.

Wang, D., Laffan, S.W. 2009. Characterisation of valleys from DEMs. In: Proceedings of 18th World IMACS Congress and MODSIM09 International Congress on Modelling and Simulation, July 13-17, Cairns, Australia. http://mssanz.org.au/modsim09/F4/wang_d.pdf.

Whiteway, T.G., Laffan, S.W., Wasson, R.J., 2004. Using sediment budgets to investigate the pathogen flux through catchments. Environ. Manag. 34, 516-527.

Zhu, A. X., 2000. Mapping soil landscape as spatial continua: The neural network approach. Water Resources Research 36, 663-667.

Zhu, A. X., Hudson, B., Burt, J., Lubich, K. & Simonson, D. 2001. Soil mapping using GIS, expert knowledge, and fuzzy logic. Soil Science Society of America Journal 65, 1463–1472.