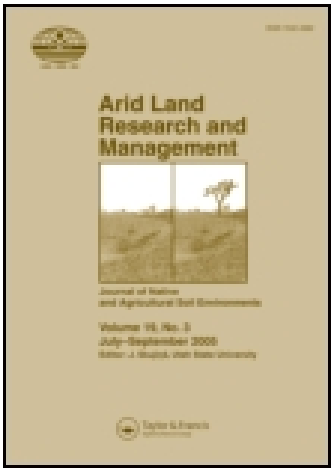
This article was downloaded by: [University of Tasmania]
On: 13 November 2014, At: 00:56
Publisher: Taylor & Francis
Informa Ltd Registered in England and Wales Registered Number: 1072954 Registered office: Mortimer House, 37-41 Mortimer Street, London W1T 3JH, UK



Arid Land Research and Management
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/uasr20

Digital Mapping of Soil Classes Using Decision Tree and Auxiliary Data in the Ardakan Region, Iran

R. Taghizadeh-Mehrjardi [a], F. Sarmadian [b], B. Minasny [c], J. Triantafilis [d] & M. Omid [b]

[a] Faculty of Agriculture and Natural Resources, University of Ardakan, Ardakan, Iran

[b] Faculty of Agricultural Engineering and Technology, University College of Agriculture and Natural Resources, University of Tehran, Karaj, Iran

[c] Faculty of Agriculture, and Environment, The University of Sydney, Australia

[d] School of Biological, Earth and Environmental Sciences, The University of New South Wales, Australia

Published online: 28 Jan 2014.

To cite this article: R. Taghizadeh-Mehrjardi, F. Sarmadian, B. Minasny, J. Triantafilis & M. Omid (2014) Digital Mapping of Soil Classes Using Decision Tree and Auxiliary Data in the Ardakan Region, Iran, Arid Land Research and Management, 28:2, 147-168, DOI: 10.1080/15324982.2013.828801

To link to this article: http://dx.doi.org/10.1080/15324982.2013.828801

PLEASE SCROLL DOWN FOR ARTICLE

Taylor & Francis makes every effort to ensure the accuracy of all the information (the "Content") contained in the publications on our platform. However, Taylor & Francis, our agents, and our licensors make no representations or warranties whatsoever as to the accuracy, completeness, or suitability for any purpose of the Content. Any opinions and views expressed in this publication are the opinions and views of the authors, and are not the views of or endorsed by Taylor & Francis. The accuracy of the Content should not be relied upon and should be independently verified with primary sources of information. Taylor and Francis shall not be liable for any losses, actions, claims, proceedings, demands, costs, expenses, damages, and other liabilities whatsoever or howsoever caused arising directly or indirectly in connection with, in relation to or arising out of the use of the Content.

This article may be used for research, teaching, and private study purposes. Any substantial or systematic reproduction, redistribution, reselling, loan, sub-licensing,

systematic supply, or distribution in any form to anyone is expressly forbidden. Terms & Conditions of access and use can be found at http://www.tandfonline.com/page/terms-and-conditions

Downloaded by [University of Tasmania] at 00:56 13 November 2014

Arid Land Research and Management, 28:147–168, 2014
Copyright © Taylor & Francis Group, LLC
ISSN: 1532-4982 print/1532-4990 online
DOI: 10.1080/15324982.2013.828801

Taylor & Francis
Taylor & Francis Group

Digital Mapping of Soil Classes Using Decision Tree and Auxiliary Data in the Ardakan Region, Iran

R. Taghizadeh-Mehrjardi[1], F. Sarmadian[2], B. Minasny[3], J. Triantafilis[4], and M. Omid[2]

[1]Faculty of Agriculture and Natural Resources, University of Ardakan, Ardakan, Iran
[2]Faculty of Agricultural Engineering and Technology, University College of Agriculture and Natural Resources, University of Tehran, Karaj, Iran
[3]Faculty of Agriculture, and Environment, The University of Sydney, Australia
[4]School of Biological, Earth and Environmental Sciences, The University of New South Wales, Australia

Downloaded by [University of Tasmania] at 00:56 13 November 2014

Digital soil mapping (DSM) involves acquisition of field soil observations and matching them with environmental variables that can explain the distribution of soils. The harmonization of these data sets, through computer-based methods, are increasingly being found to be as reliable as traditional soil mapping practices, but without the prohibitive costs. Therefore, the present research developed decision tree models for spatial prediction of soil classes in a 720 km^2 area located in an arid region of central Iran, where traditional soil survey methods are difficult to undertake. Using the conditioned Latin hypercube sampling method, the locations of 187 soil profiles were selected, which were then described, sampled, analyzed, and allocated to six Great Groups according to the USDA Soil Taxonomy system. Auxiliary data representing the soil forming factors were derived from a digital elevation model (DEM), Landsat 7 ETM$^+$ images, and a map of geomorphology. The accuracy of the decision tree models was evaluated using overall, user, and producer accuracy based on an independent validation data set. Our results showed some auxiliary variables had more influence on the prediction of soil classes which included: topographic wetness index, geomorphological map, multiresolution index of valley bottom flatness, elevation, and principal components of Landsat 7 ETM$^+$ images. Furthermore, the results have confirmed the DSM model successfully predicted Great Groups with overall accuracy up to 67.5%. Our results suggest that the developed methodology could be used to predict soil classes in the arid region of Iran.

Keywords decision tree analysis, digital soil mapping, haplosalids, salinity

Received 5 March 2013; accepted 22 July 2013.

The authors would like to acknowledge the University of Tehran and the National Soil Salinity Center in Iran. The authors thank Ms. Adrianne Ryan for providing useful feedback on the early draft of this manuscript.

Address correspondence to F. Sarmadian, Faculty of Agricultural Engineering and Technology, University College, of Agriculture and Natural Resources, University of Tehran, Karaj, Iran. E-mail: Fsarmad@ut.ac.ir

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/uasr.

Downloaded by [University of Tasmania] at 00:56 13 November 2014

Introduction

Traditionally, soil mapping in Iran involves soil sampling, classification, and the extrapolation of this information using aerial-photograph interpretation and expert knowledge. However, this approach is labor-intensive, subjective, time-consuming, and expensive. This is problematic in countries such as Iran where large agricultural areas are yet to be mapped with fine scales. In order to develop detailed knowledge regarding the spatial distribution of the soil resource, digital soil mapping (DSM) techniques are increasingly being employed to add value to traditional soil maps (McBratney et al., 2003). The basic premise that underlies DSM is various soil forming factors need to be considered during the development of a soil map. In order to generate maps of these factors cost-effectively, auxiliary data variables that are available throughout the area are used as surrogate information (McBratney et al., 2003). In arid areas, soil forming factors such as parent material, relief, and age of land surface can easily be obtained from geological maps, remote sensing data, and digital elevation models (DEM) (e.g., Canton et al., 2003).

A further requirement in DSM is that the relationship between soil and auxiliary variables is implemented by applying empirical models (Kempen et al., 2009). Various modeling techniques have been used for the digital mapping of soil classes. These methods include logistic regressions (Hengl et al., 2007; Jafari et al., 2012), artificial neural networks (McBratney et al., 2003), machine learning systems (Lacoste et al., 2011), and decision tree analysis, which is perhaps the most commonly employed. One of the earliest studies using this approach (Bui et al., 1999) extracted predictive soil decision rules from a geological map (i.e., parent material), climate maps, and terrain attributes (i.e., relief). This approach was further improved upon by Moran and Bui (2002) who added remotely sensed satellite imaging data. Luoto and Hjort (2005) used several methods, including regression, decision tree analysis, and neural networks for mapping geomorphic surfaces in Finland. They concluded that decision tree analysis performed much more strongly than the other methods. Similarly, Moonjun et al. (2010) tested whether artificial neural network and decision tree model could be used to predict soil classes in Thailand. They point out decision tree model had higher performance. Scull et al. (2005) reported similar success with decision tree analysis when using quite different auxiliary data to predict soil units in a desert ecosystem in the USA. Caten et al. (2012) also reported high performance of decision tree model for spatial prediction of soil classes in Brazil. Furthermore, Grinand et al. (2008) applied the decision tree model to predict the soil units at an unvisited area. Despite the rapid progression of DSM across varied landscapes and land uses, few studies have attempted to map soil classes in arid regions of Iran. At present, the only available soil map in Iran is a recently prepared national soil map, at a scale of 1:1,000,000. Hengl et al. (2007) applied different approaches to predict World Reference Base (WRB) soil groups in Iran. They used the soil data base of Iran to test different soil-class interpolators such as supervised classification using maximum likelihoods, multinominal logistic regression, regression kriging on membership, and classification of taxonomic distances. They concluded the best prediction was achieved using regression kriging of memberships. However, although these maps are suitable for national planning, they lack fine detail. Jafari et al. (2012) used indirect method to predict the presence of diagnostic horizons using decision tree and binary logistic models, and a direct method that used multinomial logistic regression approaches to predict soil Great Groups in Zarand,

Iran. In the present research we proposed a simpler method to predict directly soil Great Groups using decision tree analysis. Jafari et al. used simple stratified random sampling method but we used conditioned Latin hypercube sampling that targets the full distribution of covariates. They also used a 30 m DEM from ASTER to compute terrain attributes but our DEM is derived from more accurate RADAR. In addition, we conducted wavelet analysis to remove noise and artifacts from the DEM. Overall, in the present study, our objective is to predict a soil map of taxonomic level up to Great Groups using digital soil mapping technique (i.e., decision tree analysis) in an arid area where traditional soil survey methods are difficult to undertake.

Downloaded by [University of Tasmania] at 00:56 13 November 2014

Material and Methods

Study Area

The study area was the Ardakan region in the province of Yazd located in central Iran. It covers an area of 72,000 ha (Figure 1a). The region experiences an arid climate with a mean annual precipitation of 75 mm and minimum and maximum temperature of 7.2°C and 43°C, respectively. The soil moisture and temperature regimes are aridic and thermic, respectively. The major geological units are composed of red gypsiferous marls and brown to grey limestone. The major landforms of the region are—from east to west—mountain, alluvial fans, salt plain, coalescing alluvial fans (Bajadas), and gypsiferous hills, respectively (Figure 1b).

Acquisition of Auxiliary Data

The digital soil mapping approach used in this paper is the *scorpan* model (McBratney et al., 2003):

$$S = f(s, c, o, r, p, a, n) + \varepsilon,$$

where S, the soil class to predict, is a function of *s*oil (s), *c*limate (c), *o*rganisms (o), *r*elief (r), *p*arent materials (p), *a*ge (a), and spatial positio*n* (n); and where ε is the error. In a DSM approach, a set of one or more continuous (i.e., digital elevation model and remote sensing) or categorical (i.e., geomorphology map) variables could represent each of the soil forming factors.

Terrain Attributes

Primary and secondary DEM variables or terrain attributes are commonly used in predictive soil models. In the present research, multiresolution ridgetop flatness index (MrRTF; Gallant and Dowling, 2003), multiresolution index of valley bottom flatness (MrVBF; Gallant and Dowling, 2003), valley depth (Abdel-Kader, 2011), elevation (National Cartographic Center, 2010), altitude above channel network (Olaya, 2004), modified catchment areas (Olaya, 2004), mid-slope position (Bohner and Antonic, 2009), topographic wetness index (Moore et al. 1991), and catchment slope (Scull et al., 2005) were derived from a DEM (Figure 2a) (grid size of 10×10 m; National Cartographic Center, 2010). Figure 2a illustrated that most part of the study area (i.e., especially in the middle part) is dominantly flat. Given that

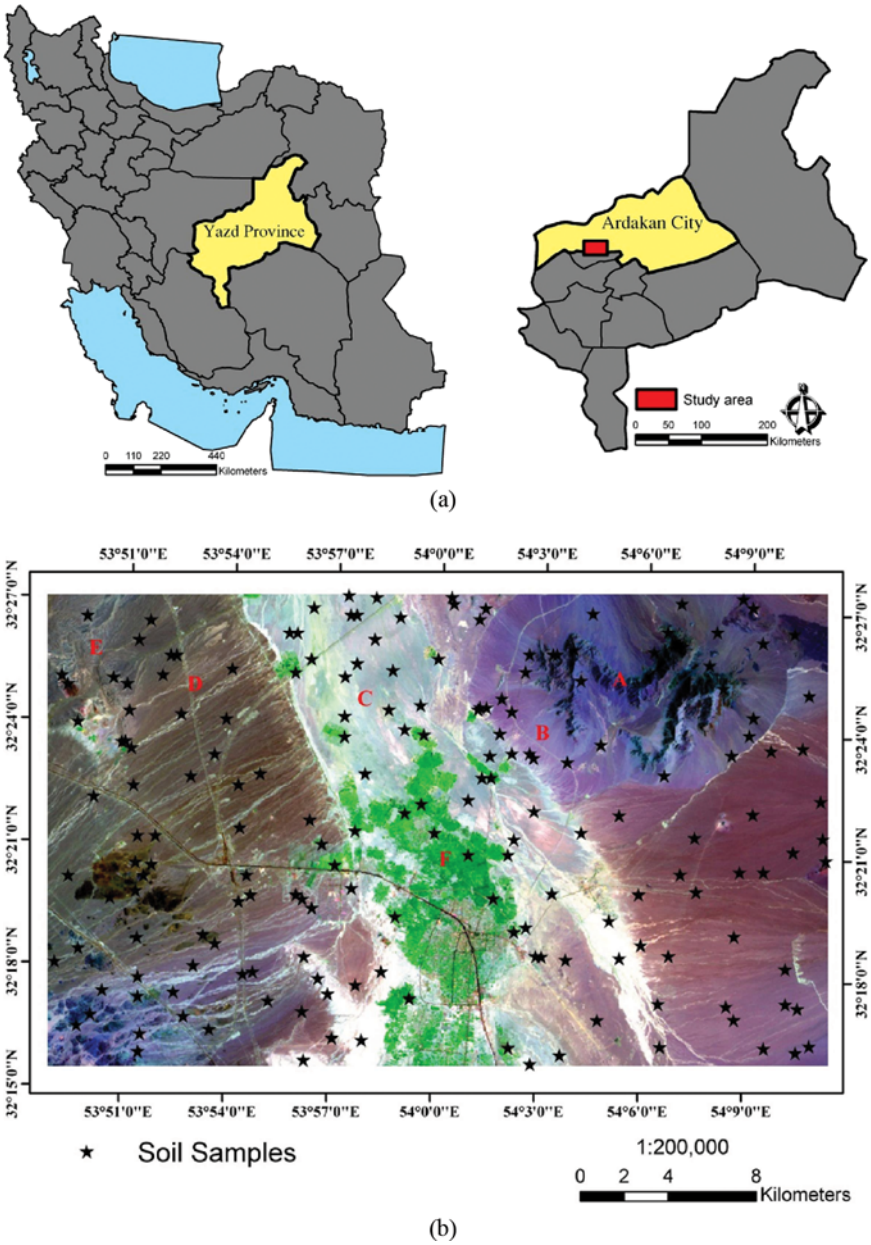


Figure 1. (a) Location of the Ardakan region study area in central Iran and (b) spatial distribution of soil units draped over Landsat ETM^{+} image (False color composite of bands 1, 2, and 3) [A: Mountain landscape with rock surfaces, B: Alluvial fan, C: Playa, with fine and to some extent coarse alluvial sediment, D: coalescing alluvial fans (Bajadas), E: gypsiferous hills, F: Pistachio orchard].

soils occurring in a flat area are not strongly influenced by local topographic characteristics, we do not regard local terrain attributes such as slope or slope curvatures as important auxiliary data (Scull et al., 2005).

Downloaded by [University of Tasmania] at 00:56 13 November 2014

Downloaded by [University of Tasmania] at 00:56 13 November 2014

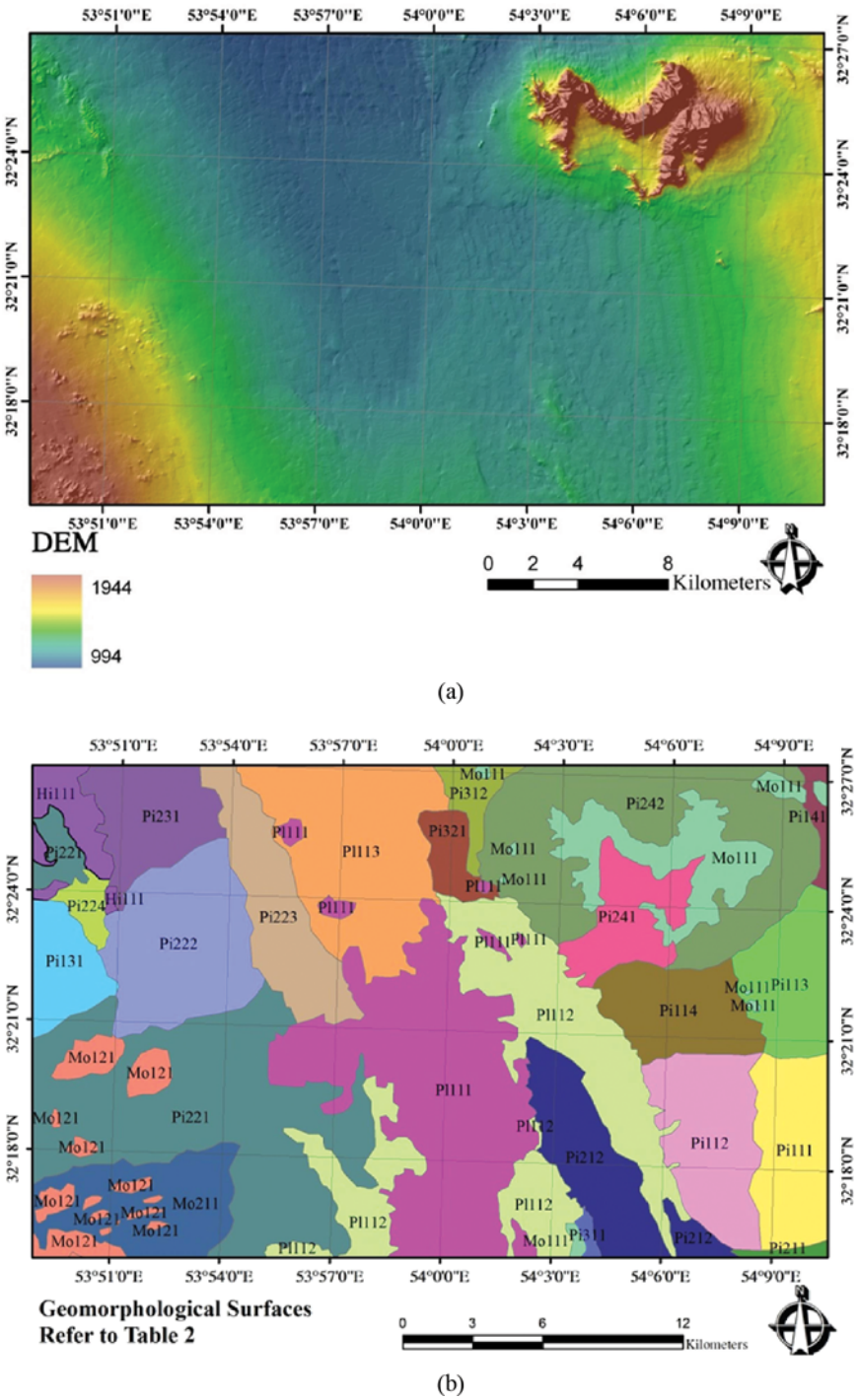


Figure 2. Digital elevation models (a) and delineated geomorphic surfaces (b) (codes refer to Table 2).

Downloaded by [University of Tasmania] at 00:56 13 November 2014

Table 1. Geomorphological map hierarchy and the major soil Great Group for each geomorphological surface

Landscape	Landform	Lithology	Geomorphological surface	Code	Major Great Group soil observed
Mountain	Dissected ridge	Dolomite-limestone	Rock outcrop	Mo111	Torriorthents
	Rock outcrop	Gray to green andesitic, and limestone	Rock outcrop	Mo121	Torriorthents
	Rock pediment	Eroded calcareous and dark shale	Eroded surface	Mo211	Torriorthents
Hill land	Eroded rock outcrop	Sandstone-gypsum	Dendrite drainage system with high topography	Hi111	Haplogypsids
Playa	Ardakan basin	Fine and coarse alluvial sediments	Soft clay flat, salty and cultivated	Pl111	Haplocalcids
			Clay flat, dense stream, salty	Pl112	Haplosalids
			Clay flat, highly salty and wetness	Pl113	Haplosalids
Piedmont	Alluvial fan	Alluvium of limestone	Active fan, upper section	Pi111	Haplosalids
			Active fan, lower section, salty	Pi112	Haplosalids
			Active fan, upper section with more drainage system, desert pavement	Pi113	Haplosalids
			Active fan, lower section with more drainage system, desert pavement	Pi114	Haplosalids

Downloaded by [University of Tasmania] at 00:56 13 November 2014

	Alluvium of grey limestone with red sandstone at the base	Active fan, lower section	Pi131	Calcigypsids
	Alluvium sandstone, shale and limestone	Active fan, lower section	Pi141	Haplogypsids
Bajada	Alluvium of grey limestone with red sandstone at the base	Upper section, high slope, dense drainage system	Pi211	Haplosalids
		Lower section, high slope	Pi212	Haplosalids
		Upper section, coarse, salty	Pi221	Calcigypsids
		Upper section, parallel streams	Pi222	Haplosalids
		Lower section, new deposits	Pi223	Haplocalcids
		Cultivated bajada, salty	Pi224	Haplosalids
	Alluvium of gypsum hill lands	Upper section, coarse, salty	Pi231	Haplosalids
	Alluvium of dolomite-limestone	Coarse, calcareous	Pi241	Haplocalcids
	Alluvium of dolomite-limestone	Coarse, with dense drainage network	Pi242	Haplocalcids
Old bajada	Alluvium of siltstone, shale, sandstone, gypsum	Flat and lower topography	Pi311	Haplosalids
		Higher topography and deep streams	Pi312	Haplosalids
	Alluvium of gypsum	Higher topography and deep streams, coarse	Pi321	Haplogypsids

Remote Sensing

We initially computed some band ratios and normalized difference vegetation index (Boettinger et al., 2008) was found to be the best representation of the vegetation variable. Other band ratios (i.e., clay index (Boettinger et al., 2008), carbonate index (Boettinger et al., 2008), gypsum index (Nield et al., 2007), salinity index (Metternicht and Zinck, 2003), and brightness index (Metternicht and Zinck, 2003) were also computed to represent parent material and soil factors at the study area. In addition, principal component analysis (PCA) was computed on the ETM^+ bands, based on its correlation matrix (Nield et al., 2007). PCA is an effective approach to discriminate saline soils in arid regions (Metternicht and Zinck, 2003).

Downloaded by [University of Tasmania] at 00:56 13 November 2014

Geomorphology Map

A geomorphology map for the study area was also prepared, based on a nested geomorphic hierarchy approach defined by Toomanian et al. (2006). In this approach, aerial photographs (1:50,000) were used to delineate geomorphological entities in four levels, which included: landscape (1:250,000), landform (1:100,000), lithology (1:100,000), and geomorphological surface (1:50,000). After ortho-photo geo-referencing of aerial photographs, delineated boundaries of geomorphological surfaces were inserted in a GIS environment. The study area had 25 geomorphological units (Figure 2b and Table 1).

For representing continuous spatial variations and data modeling, all data layers were registered to a common grid of 30 m spacing.

Wavelet Analysis

In this study, the DEM was originally prepared from RADAR images (National Cartographic Center, 2010), and consequently contained a substantial amount of noise. Therefore, a two-dimensional discrete wavelet transformation was carried out using MATLAB software (MathWorks, 2010) in order to spatially decompose the terrain attribute layers (Lark and Webster, 2004; Lark and McBratney, 2002) and remove the noise and artifacts. Finally, the data layers were decomposed into four levels: L1, L2, L3, and L4. These levels corresponded to pixel sizes of 20, 40, 80, and 160 m, respectively.

Data Collection and Soil Sample Analysis

Latin hypercube sampling (LHS) is a procedure that ensures a full coverage of the range of each covariate by maximally stratifying the marginal distribution. LHS involves sampling n values from the prescribed distribution of each of k (in this case = 10) covariates $X_1, X_2, \ldots X_k$. The cumulative distribution for each covariate is divided into n equi-probable intervals. A value is selected randomly from each interval. The n values obtained for each variable are matched randomly with those of the other variables (Minasny and McBratney, 2006). The locations of 187 soil samples were selected based on the Latin hypercube sampling method. The model used ten auxiliary variables showing the most variation based on the coefficient of variation. Auxiliary variables used for sample selection include; Length slope factor, stream power, slope length, slope, aspect, geomorphology units, and ETM^+ images

Downloaded by [University of Tasmania] at 00:56 13 November 2014

Table 2. Morphological and physico-chemical properties of representative profiles up to Great Group level

Soil horizon	Depth (cm)	Sand, %	Sill, %	Clay, %	Texture	Gypsum, %	CCE[a], %	pH	ECe (dSm^{-1})	SAR
Haplocalcids (53°58′33″ and 32°16′30″)										
AP	0–15	38.0	26.00	36.00	C.L	–	16.90	7.5	12.4	6.0
Bk_1	15–55	38.0	26.00	36.00	C.L	–	23.70	7.9	16.4	32.1
Bk_2	55–150	55.0	13.00	32.00	Sa.C.L	–	40.60	8.0	16.0	35.5
Haplocambids (53°58′30″ and 32°18′00″)										
A	0–25	38.0	28.0	34.0	C.L	–	21.50	7.8	15.5	20.0
BW_1	25–50	34.0	25.0	41.0	C	–	20.30	7.8	8.8	17.0
BW_2	50–90	40.0	29.0	31.0	C.L	–	21.00	7.8	11.0	22.1
C	90–150	40.0	29.0	31.0	C.L	–	22.00	7.7	27.8	32.7
Calcigypsids (53°55′00″ and 32°20′00″)										
A	0–15	65.0	6.0	29.0	Sa.C.L	–	24.50	7.7	2.2	2.9
Bk	15–50	65.0	6.0	29.0	Sa.C.L	2.0	29.90	7.8	18.4	35.3
By	50–150	68.0	13.0	19.0	Sa.L	9.1	21.10	7.9	13.8	22.3
Haplogypsids (53°47′14″ and 32°29′11″)										
A	0–7	56.0	19.0	25.0	Sa.C.L	1.0	24.00	7.6	5.6	7.8
By_1	7–20	46.0	13.0	41.0	Sa.C	15.0	20.80	7.7	5.6	8.0
By_2	20–60	63.0	8.0	29.0	Sa.C.L	15.3	15.90	7.7	3.9	3.1
C	60–200	72.0	4.0	24.0	Sa.C.L	15.7	17.90	7.8	19.3	34.1
Haplosalids (53°58′30″ and 32°28′12″)										
A	0–25	11.0	32.0	57.0	C	–	20.60	7.7	30.2	31.5
Bz_1	25–80	20.0	28.0	52.0	C	–	13.50	7.7	67.7	69.0
Bz_2	80–200	13.0	35.0	52.0	C	–	14.00	7.7	67.7	69.0
Torriorthents (54°03′30″ and 32°25′16″)										
A	0–20	82.0	8.0	10.0	L.Sa	–	29.50	7.92	1.7	2.5
C	20–50	88.0	2.0	10.0	L.Sa	–	34.00	7.74	0.1	2.6

[a]Calcium carbonate equivalent.

(bands 3, 4, 5, and 7). Figure 1b shows the location of the 187 soil profiles. At each site, pedons were dug down to 1.5 m and then the horizon characteristics were described and allocated to classes according to the US Soil Taxonomy up to Great Group level (Soil Survey Staff, 2010). The soil profiles were allocated into two orders (i.e., Aridisols and Entisols), five sub-orders (i.e., Calcids, Cambids, Gypsids, Salids, and Orthents), and six Great Groups (Haplocalcids, Haplocambids, Calcigypsids, Haplogypsids, Haplosalids, and Torriorthents). The samples, taken from all genetic horizons, were air-dried at room temperature and ground to pass through a 2-mm sieve prior to analysis. The particle size distribution was determined using the hydrometer method (Gee and Bauder, 1986). Electrical conductivity (EC) and soil reaction (pH) were measured using a conductivity meter (PW-9527 Philips, Poland) and pH meter (EYELA-2000, Rikakikai, Japan), respectively, after preparation of saturation pastes of the soil samples. Organic carbon was determined using the Nelson and Sommers (1982) method. The saturation percentage (SP) was measured using the gravimetric method and the calcium carbonate equivalent using the volumetric method. The level of soluble ions (i.e., carbonates, bicarbonates, sodium, potassium, chlorine, sulfate, calcium, and magnesium) was determined using common experimental methods (Sparks et al., 1996). Table 2 shows the basic soil morphological and physico-chemical properties of representative Great Group soil profiles.

Downloaded by [University of Tasmania] at 00:56 13 November 2014

Decision Tree Analysis

A decision tree correlates several independent variables (i.e., auxiliary variables) with direct or indirect relationships to a target variable (i.e., soil classes) with a tree structure, generated by partitioning the data recursively into a number of groups. Here, the See5 decision tree analysis software (Quinlan, 2001) was used to predict soil classes from the auxiliary data.

Evaluation of Models

In order to test the accuracy of our predictions, the data was divided randomly into two sets. The larger set was used for training (i.e., 150 points; 80%) and the smaller set was set aside for validation (i.e., 37 points; 20%) (Schmidt et al., 2008). Accuracy of the decision tree model for prediction of soil classes (i.e., Great Groups level) was evaluated in the error data matrix using descriptive statistical methods such as user accuracy, producer accuracy, and overall accuracy (Jensen, 1996). The simplest descriptive statistical method is overall accuracy which is computed by dividing the total correct (i.e., the sum of the major diagonal) by the total number of pixels in the error matrix. Producer accuracy, a measure of omission or exclusion errors, shows how successful the model is in prediction. It is calculated by dividing the total number of correctly predicted pixels of an individual category by the total number of pixels given to that category from the reference data. User accuracy, a measure of commission or inclusion errors, shows how well these map predictions are represented in reality. It is calculated by dividing the total number of correctly predicted pixels of a category by the total number of pixels that were actually classified in that category (Stehman and Czaplewski, 1998; Elnaggar, 2007; Brus et al., 2011).

Results

Descriptions of Soil Classes

Rather than presenting descriptive data for all sampling points, short descriptions are provided for six pedons (i.e., Great Group level) which were selected as representative (Table 2).

- The mountainous area was dominated by Torriorthents. These profiles are largely non-saline, with electrical conductivity of saturation paste (ECe) lower than $2\,dS\,m^{-1}$.
- Haplocalcids are mainly formed on limestone parent materials. The diagnostic features of these profiles are a calcic subsurface horizon, overlain by an ochric surface horizon. The accumulation of carbonates as nodules and pendants can be found at depths of 15–70 cm. The texture in the surface horizon is clay loam; and changes to sandy clay loam with depth. Most of the soil observations related to these taxa are located in a pistachio orchard in the central part of the study region (Pl111).
- Haplocambids having a cambic horizon covered just 3% of the soil observations and to some extent was located in combination with Typic Haplocalcids.
- Calcigypsids were characterized by a calcic and gypsic subsurface horizon. Accumulation of gypsum occurred as pendants mostly at depths of 50–100 cm, while pendants and nodules of carbonates were present in the upper part of profile (15–50 cm). The surface had finer texture (sandy clay loam) compared with deeper horizons (sandy loam).
- Haplogypsids exhibited a gypsic horizon characterized by accumulation of gypsum as pendants, with sizes ranging from 1 cm to over 10 cm. Most soil observations related to these taxa were found in the gypsiferous hills.
- Haplosalids have a fine texture throughout the profile and the salinity level particularly in the subsurface horizons was very high, sometimes exceeding $60\,dS \cdot m^{-1}$.

Descriptions of Auxiliary Data

Principal Component Analysis of the Landsat ETM$^+$ Images

To reduce the total number of ETM$^+$ data, a principal component analysis (PCA) based on a correlation matrix was computed. Principal component analysis of the Landsat ETM$^+$ images revealed that the first three components (PC1, PC2, and PC3) represent 99% of the variation within the images. The first component defined 90% of image variation, and this one is correlated to band 2, which covers the green range of the spectrum. Figure 3 shows the spatial distribution of the first two principal components (PC1 and PC2) of the Landsat ETM$^+$ images. According to Figure 3a, the largest values for PC1 coincide with the playa landform in middle of the study area, while the lowest values are associated with mountainous landform. In PC2 (Figure 3b), the largest values coincide with the irrigated areas and some of the vegetated areas in the central part of the study area.

Spatial Distribution of the Auxiliary Data

From all the auxiliary variables used at this study, only two of them, which are considered as the most important predictors and calculated from DEM, are explained here.

Downloaded by [University of Tasmania] at 00:56 13 November 2014

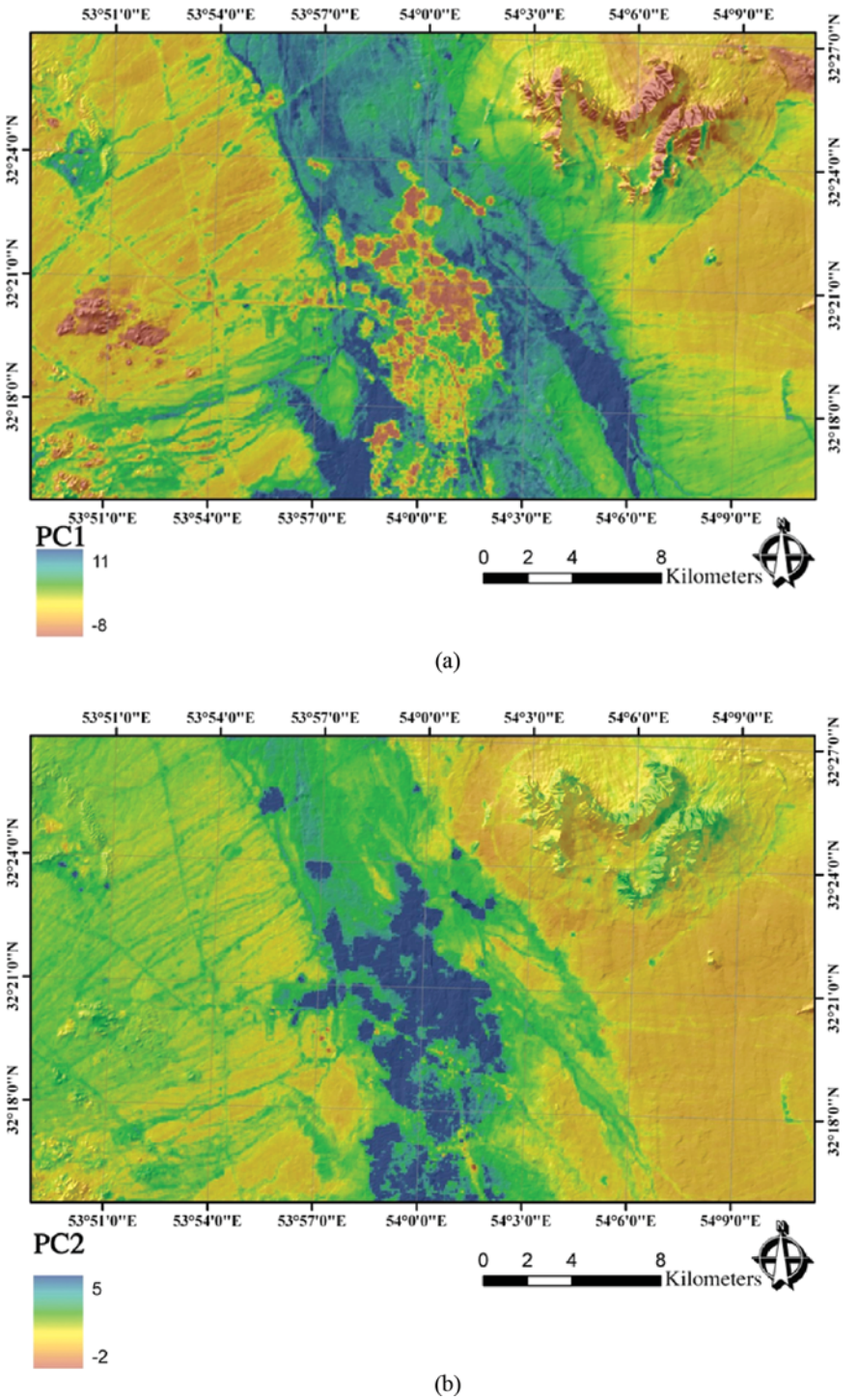


Figure 3. Some auxiliary data of the Ardakan plain derived from ETM^{+} images which included: PC1 (a) and PC2 (b).

Downloaded by [University of Tasmania] at 00:56 13 November 2014

Downloaded by [University of Tasmania] at 00:56 13 November 2014

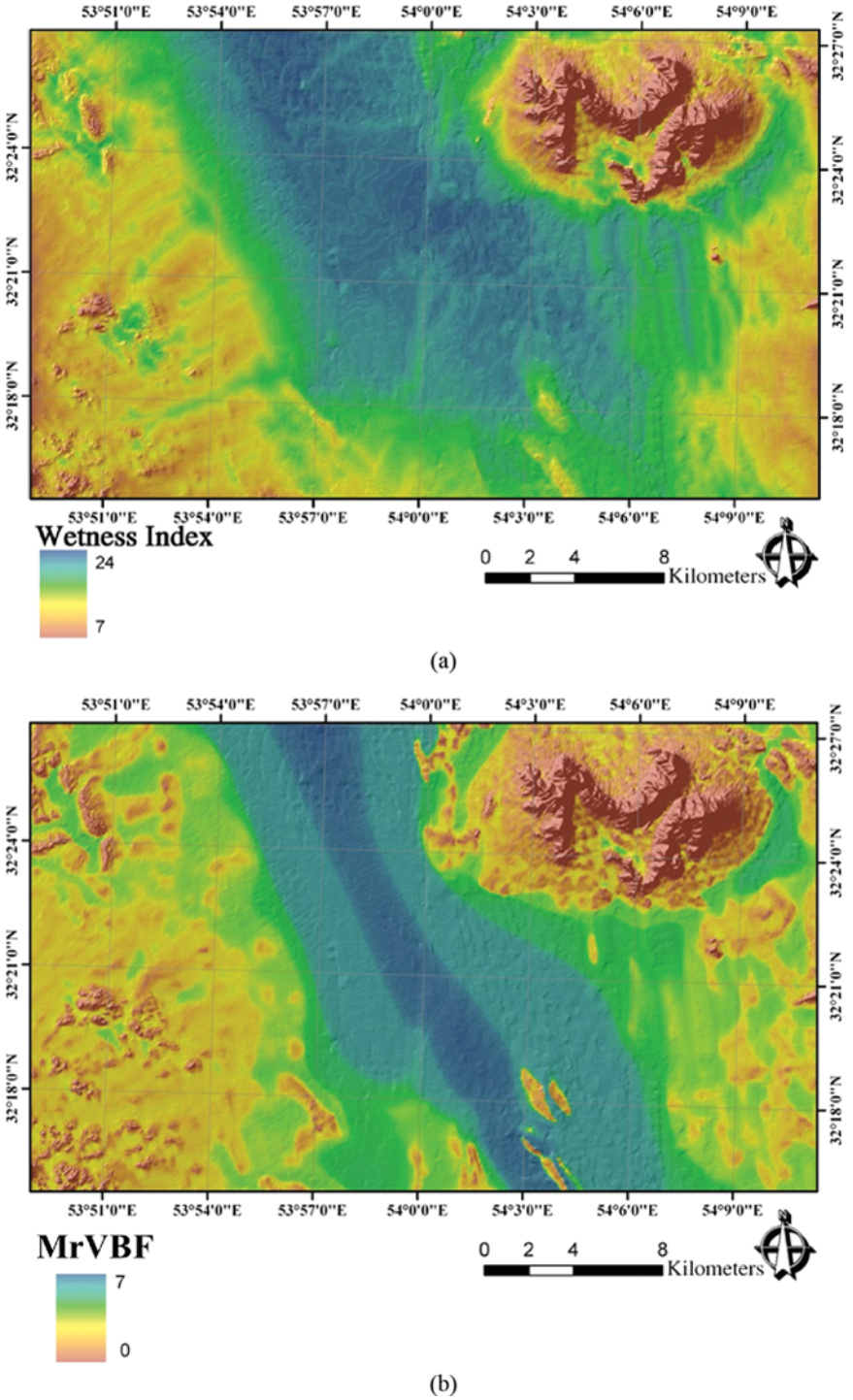


Figure 4. Some auxiliary data of the Ardakan plain derived from DEM which included: wetness index (a) and MrVBF (b).

TWI: Figure 4a shows the spatial distribution of the topographic wetness index. Small values (≅7) are generally associated with the mountainous area. Intermediate-large values (between 7 and 14) are associated with parts of the bajada landforms. High values (>14) corresponded with the playa landform, which showed a high potential for accumulation of salt materials. The location of the playa region at the outlet of Ardakan basin suggests that it has received considerable additions of soluble materials, washed out from the entire watershed. Studies by Moore et al.

Downloaded by [University of Tasmania] at 00:56 13 November 2014

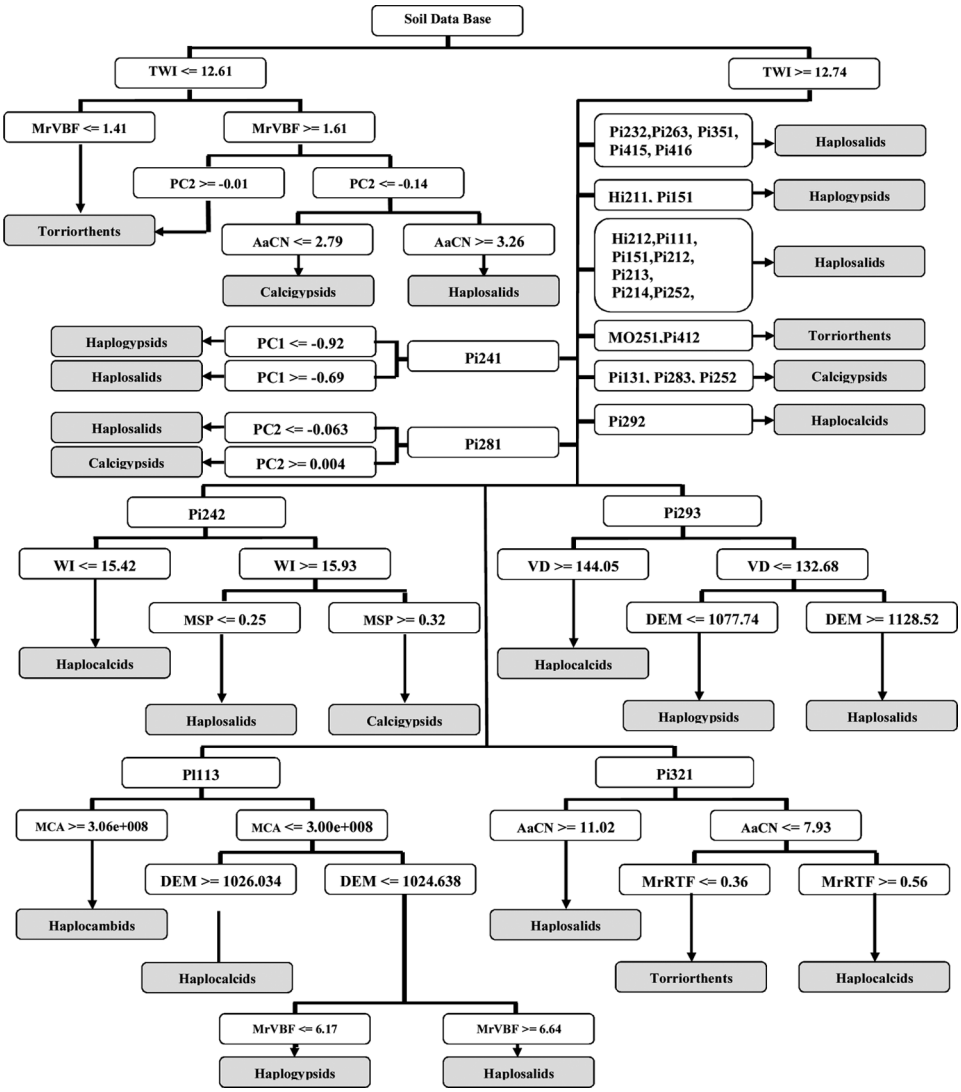


Figure 5. The rules defined by decision tree to predict soil Great Group. (WI: wetness index, MrVBF: Multi-resolution index of Valley Bottom Flatness, PC: Principal component, DEM: digital elevation model, AaCN: Altitude above channel network, MCA: Modified catchments area, VD: Valley depth, MSP: Mid-slop position, MrRTF: Multi-resolution ridgetop flatness index; Refer to Table 1 for definition of geomorphology codes).

(1991) and Jafari et al. (2012) showed that there is a high correlation between soil salinity and the wetness index. MrVBF: The multi-resolution valley bottom flatness index (MrVBF) indicates flat valley bottoms, the depositional areas within landscapes. Spatial distribution of MrVBF (Figure 4b) shows a similar trend to that exhibited for the wetness index. For example, the lowest value of MrVBF was strongly associated with the more elevated parts of the study area. The highest values corresponded with the central study area, which could be a potential zone of transport for many materials during excess water flow.

Selection of Auxiliary Data for Decision Tree Analysis

The significance of each type of auxiliary data represented as an attribute percentage, essentially the percentage of training cases for which the value of that covariate is used in predicting a class. Analysis of decision tree models showed that some auxiliary variables, including wetness index (100%), geomorphology surfaces (84%), MrVBF (27%), PC2 (13%), DEM (11%), and altitude above channel network (11%), had the strongest influence (i.e., relative influence of model) on the prediction of soil Great Groups. Topographic wetness index was the most powerful predictor, and was utilized by the model for every prediction. Figure 5 also showed the decision rules for soil groups; it can be easily inferred from this figure that topographic wetness index was the most important predictor. The second most important predictor was the geomorphological surface map, which was used by the model in 84% of Great Groups predictions. This emphasizes the role of geomorphologic processes in soil development as reported in many soil-geomorphology studies (Jafari et al., 2012; Toomanian et al., 2006). The MrVBF was also incorporated in the model, though at a lower rate of 27%. However, indices derived from the Landsat images such as NDVI, clay index, gypsum index, salinity index, and brightness index had very little influence in mapping of soil classes.

Spatial Distribution of Soil Classes

For prediction of soil classes at the Great Group level, we initially calculated terrain parameters from original DEM (i.e., 10 m pixel size) and its four levels of decomposed (i.e., 20, 40, 80, and 160 m pixel sizes). Then, we compared the original DEM and its four levels of decomposed terrain attribute layers. Our results showed

Table 3. Overall accuracy (%) of training and validation data for original DEM and each of the decomposed levels

DEM	Training (%)	*RI_T (%)	Validation (%)	RI_V (%)
Original DEM (10 m)	73.3%	0	54.1%	0
L_1 (20 m)	84.7%	16.0%	59.5%	9.0%
L_2 (40 m)	81.3%	11.0%	62.2%	15.0%
L_3 (80 m)	71.4%	−2.0%	56.8%	5.0%
L_4 (160 m)	86.0%	17.0%	67.6%	25.0%

RI_T: Relative Improvement of Training data set; RI_V: Relative Improvement of Validation data set.

$RI = \frac{(Overall\ accuracy\ of\ original\ DEM) - (Overall\ accuracy\ of\ decomposed\ Layers)}{(Overall\ accuracy\ of\ original\ DEM)}$.

Downloaded by [University of Tasmania] at 00:56 13 November 2014

Downloaded by [University of Tasmania] at 00:56 13 November 2014

Table 4. Confusion matrix of soil Great Group [training and validation data sets (%)]

	Great group soils	Haplocalcids	Haplocambids	Calcigypsids	Haplogypsids	Haplosalids	Torriorthents	UA%
Training data set	Haplocalcids	**13**				4	2	68.4%
	Haplocambids		**2**			2		50.0%
	Calcigypsids	1		**10**	1	3	1	62.5%
	Haplogypsids			1	**9**	3	1	64.2%
	Haplosalids				1	**72**		98.6%
	Torriorthents	1					**23**	95.8%
	PA%	86.6%	100.0%	90.9%	81.8%	85.7%	85.1%	OA% = 86.0
Validation data set	Haplocalcids	**1**				1		50.0%
	Haplocambids	1						0.0%
	Calcigypsids	2		**1**	2	1		16.6%
	Haplogypsids				**2**	1		66.6%
	Haplosalids			1	1	**19**	1	86.3%
	Torriorthents					1	**2**	66.6%
	PA%	25.0%	0.0%	50.0%	40.0%	82.6%	66.6%	OA% = 67.67

The rows and columns show the Field identified and Predicted Great Group, respectively.
UA: User accuracy; PA: Producer accuracy; OA: Overall accuracy.

that for prediction of the target variable (Great Group soils), the decomposed data layers (L_1, L_2, and L_4) had larger accuracy than data layers that were derived from the original DEM. The data in Table 3 indicates that decomposed data layer L_4 produced the best model and could enhance prediction accuracy by about 25% and 17% for validation and training data sets compared to the original DEM.

Overall, this model successfully predicted six Great Groups with reasonable accuracy, up to 67.5% and 92% based on the validation and training data sets, respectively (Table 4). The confusion matrix for the training data set showed that the highest accuracy belonged to Haplosalids, with 98% user accuracy, followed by Torriorthents with 95% user accuracy. In contrast, the worst predictions were for Haplocambids, with just 50% user accuracy. This low level of accuracy occurs because the Haplocambid Great Group constituted the smallest number of observations, and only covered less than 2% of the study area. Similar to training data set, the confusion matrix for the validation data set (Table 4) showed that the highest accuracy belonged to Haplosalids, with 86% user accuracy, followed by Torriorthents and Haplogypsids with 66% user accuracy; Meanwhile, Haplogypsids have 44% producer accuracy, which is less than producer accuracy of Torriorthents (i.e., 66%). Similar to training data set, the worst predictions were obtained for Haplocambids, with 0% user and producer accuracy (Table 4). It means that the decision rules could not classify this soil Group Group properly due to that the Haplocambid Great Group constituted the smallest number of observations. The rules defined by this decision tree analysis (Figure 5) were applied to predict Great Groups across the study area using the kriging method (Figure 6). It can be seen that

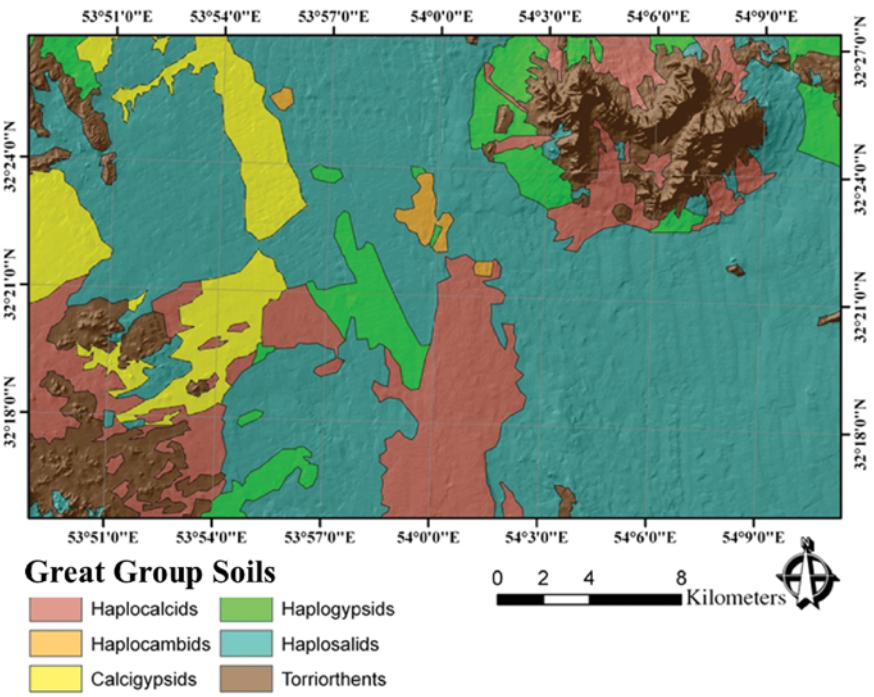


Figure 6. Digital soil map of Great Groups using decision tree analysis.

Downloaded by [University of Tasmania] at 00:56 13 November 2014

Entisols occur primarily in the highlands, where they cover 12% of the area, whereas Aridisols are mostly located in the lower parts of the region in different landforms, including bajada, alluvial fans, and playa landforms. Haplosalids with salic horizons can be found across most of the study area and are particularly prevalent in the center, which partially coincides with the playa landform. Calcigypsids and Haplogypsids having just gypsic horizons were strongly associated with geological units. Most of the Haplocalcids with calcic horizons occur in the center of the study area (Pl111); wheras, other Haplocalcids occur to some extent in combination with other Great Groups across the study area.

Discussion

Auxiliary Data Used in Predictive Models

Terrain attributes, such as wetness index and MrVBF, were found to be the most effective characteristics for explaining the distribution of soil classes. Geomorphological surface was also important, as an indicator of parent material. Remote sensing data (PC2) helped to characterize parent materials and the distribution of vegetation, as well as various soil properties across the study area.

Topographic wetness index, which could depict stationary water content in soils, indicated the potential areas where salic horizons may be present. MrVBF was also an effective index in the flat areas, especially for identifying flat valley bottoms and, consequently, indicated potential zones of transport for sediment and other materials. Jafari et al. (2012) also confirmed the potential of these covariates for discrimination of saline soils. According to Figure 4, the soils in the middle part of the region have the highest potential to receive materials washed out from upper lands and, as a result, this part of the area was occupied with saline soils (Haplosalids) (Figure 6). Our results also indicated that soils found in upper land were coarse textured and showed low levels of salinity, whereas soils located in lower part of the area were more likely to be saline and finely textured (Table 2). Therefore, these indices (MrVBF and wetness index) helped to discriminate some Great Group soils (i.e., Torriorthents and Haplosalids).

Results also suggested that the geomorphological map provided the second most important auxiliary data which is in line with results of Jafari et al. (2012). Their findings showed that geomorphological surfaces are the most important factor for spatial modeling of soil classes in Zaran, Iran. This may result from the fact that the geomorphological surfaces have formed recently, and hence have good relationship with soil processes in the arid regions. Our results also indicated that most of the Aridisols are located in the lower part of the region, across different landforms which included bajada, alluvial fans, and playa. These landforms have received more soluble salts washed out from upper areas. In addition, saline and gypsiferous parent materials in these landforms helped the formation of salic and gypsic horizons. Therefore, Gypsic Haplosalids with gypsic and salic horizons dominated the study area, which is consistent with the findings of Hengl et al. (2007) and indicates a high probability of the occurrence of Gypsisols and Solonetz soils in central Iran. Similarly, Jafari et al. (2012) indicated high probability of saline soils in arid regions of Iran. Furthermore, these landforms had poor vegetation coverage (Figure 1b), which could further indicate the presence of salic horizons. Calcigypsids and Haplogypsids have strong association with geomorphical units because they were

Downloaded by [University of Tasmania] at 00:56 13 November 2014

frequently found where the lithology is gypsiferous marls; most of the Haplogypsids were found in gypsiferous hills. Haplocalcids can be found in the middle part of the area; but also can be found in combination with other Great Groups that have calcareous parent materials.

Accuracy of Predictive Models

Applying these auxiliary data, decision tree analysis predicted the spatial location of soil classes in Great Group level with overall accuracy of 67.5%, which is similar to the results obtained by other researchers (Moran and Bui, 2002; Bui et al., 1999; Henderson et al., 2005; Luoto and Hjort, 2005; Jafari et al., 2012). Scull et al. (2005) reported 65–70% accuracy for prediction of Great Groups in the Mojave Desert eco-region of California in the United States of America. An advantage of using tools such as decision tree analysis is that it can be applied to areas with restrictive conditions, such as central Iran, and the results may help to design a bigger project in the future. Decision trees are easy to interpret and can handle both continuous and categorical data.

Table 4 further showed that the highest accuracy for prediction of Great Groups belonged to Haplosalids with 98% and 86% user accuracy based on training and validation data sets. That this class can be found in 51% of the soil observations might be one reason why Haplosalids had the highest accuracy. Supporting this suggestion is the poor prediction of Haplocambids, which were only represented by a few observations (2% of total). Jafari et al. (2012) mentioned that the size of sampling units relative to total study area is an important factor determining the purities of a map; hence the smaller sample size contributes to uncertainty. Furthermore, results indicated that Torriorthents were also predicted with high levels of accuracy (95% user accuracy) as compared with soil classes in flat areas (with the exception of Haplosalids, as previously discussed). This may occur because soils that evolved in the mountain landforms have a good relationship with terrain attributes. This result is consistent with findings of Scull et al. (2005) who concluded that soil units in mountainous areas had better relationships with terrain parameters compared with soil units located in flat regions. Similarly, Jafari et al. (2012) mentioned that Haplosalids, Haplogypsids and Torripsamments that are highly influenced by topographic and geomorphic characteristics in the study area were predicted more accurately than those only slightly influenced by topographic and geomorphic characteristics. However, our results showed better accuracy (i.e., overall accuracy based on validation data set: 0.67%) than obtained by Jafari et al. (2012) (i.e., purity of soil map in Great Group level: 0.58). This can be attributed to several factors. First, we used wavelet analysis to remove the artifact from the DEM. We showed in Table 3 that with wavelet filter, DEM at a broader scale gives a better prediction for flat areas. In fact, the filtering and pre-processing of DEM is important, and wavelet analysis is a useful tool to achieve this. Second, the higher overall accuracy at present research than obtained by Jafari et al. (2012) might be related to the effect of the size of sampling units relative to the total study area (i.e., $0.26/km^2$ vs. $0.148/km^2$). In fact, the small number of locations involves uncertain purity estimates (Kempen et al., 2009). In addition, the soil sampling method (i.e., Latin hyper cube method versus a simple stratified random method) is an important factor. In the Jafari et al. (2012) work, the distribution of the samples was stratified randomly over strata determined from ancillary data, and not all the soil types present were equally

Downloaded by [University of Tasmania] at 00:56 13 November 2014

represented. Meanwhile, in the present research, we applied LHS to cover the distribution of all covariates and, hence, it provided better accuracy.

Conclusions

This study described the prediction of the spatial distribution of soil classes in Ardakan-Yazd plain of central Iran, using a decision tree technique. During the application of this method, we used a variety of auxiliary variables derived from different sources.

Expansion of agricultural production in the future is likely to encroach on dry lands, where irrigation will be necessary for success. The annual rainfall of just 75 mm in the Yazd-Ardakan plain is considered to be the major challenge for agricultural production. The low rainfall is far more prohibitive than other natural factors such as desertification or human factors such as increasing emigration. Therefore, one of the major projects that the Iranian government has determined to address these issues is to develop qualitative and quantitative land suitability maps for drip irrigation. One of the main requirements of such a project is detailed knowledge regarding soil distribution, which can be achieved through DSM. Our results confirmed that decision tree analysis was a reliable approach that could be successfully used to prepare continuous soil maps, up to the Great Group level. Therefore, we recommend the use of this approach to map the soils in other parts of Iran. Future work is needed to define the level of confidence in the maps predicted using this technique.

Downloaded by [University of Tasmania] at 00:56 13 November 2014

References

Abdel-Kader, F. H. 2011. Digital soil mapping at pilot sites in northwest coast of Egypt: A multinomial logistic regression approach. *The Egyptian Journal of Remote Sensing and Space Science* 14: 29–40.

Boettinger, J. L., R. D. Ramsey, J. M. Bodily, N. J. Cole, S. Kienast-Brown, S. J. Nield, A. M. Saunders, and A. K. Stum. 2008. Landsat spectral data for digital soil mapping, pp. 193–203, in A. E. Hartemink, A. B. McBratney and M. L. Mendonca-Santos, eds., *Digital soil mapping with limited data*. Springer Science, Rio de Janiero, Brazil.

Bohner, J., and O. Antonic. 2009. Land-surface parameters specific to topo-climatology, pp. 192–226, in T. Hengl and H. I. Reuter, eds., *Geomorphometry: Concepts, software, applications*. Elsevier, Amesterdam.

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

Bui, E. N., A. Loughhead, and R. Corner. 1999. Extracting soil-landscape rules from previous soil surveys. *Australian Journal of Soil Research* 37: 495–508.

Canton, Y., A. Sole-Benet, and R. Lazaro. 2003. Soil-geomorphology relations in gypsiferous materials of the Tabernas Desert (Almeria, SE Spain). *Geoderma* 115: 193–222.

Caten, A. T., R. S. D. Dalmolon., L. F. C. Ruiz, and M. L. Mendonca-Santos. 2012. Digital soil mapping: Strategy for data pre-processing, p. 466, *Digital soil assessment and beyond-Minasny, Malone and McBratney*. Taylor and Francis Group, Sydney, Australia.

Elnaggar, A. A. 2007. Development of predictive mapping techniques for soil survey and salinity mapping. Doctoral dissertation, Oregon State University, Corvallis, Oregon, p. 148. Retrieved from http://hdl.handle.net/1957/5754

Gallant, J. C., and T. I. Dowling. 2003. A multiresolution index of valley bottom flatness for mapping depositional areas. *Water Resources Research* 39: 1347–1360.

Gee, G. W., and J. W. Bauder. 1986. Particle-size analysis, pp. 404–408, in A. Klute, ed., *Methods of soil analysis part 1. Physical and mineralogical methods. Monograph 9*. 2nd ed. *Agronomy Society of America and Soil Science Society of America, Madison, Wisconsin.*

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

Henderson, B. L., E. N. Bui, C. J. Moran, and D. A. P. Simon. 2005. Australia-wide predictions of soil properties using decision trees. *Geoderma* 124: 383–398.

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

Jafari, A., P. A. Finke, J. V. de Wauw, S. Ayoubi, and H. Khademi. 2012. Spatial prediction of USDA-great soil groups in the arid Zarand region, Iran: Comparing logistic regression approaches to predict diagnostic horizons and soil types. *European Journal of Soil Science* 63: 284–298.

Jensen, J. R. 1996. Introductory digital image processing: *A remote sensing perspective*, 2nd ed. Prentice Hall, Inc., Upper Saddle River, New Jersey.

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

Lacoste, M., B. Lemercier, and C. Walter. 2011. Regional mapping of soil parent material by machine learning based onbased on point data. *Geomorphology* 133: 90–99.

Lark, R. M., and A. B. McBratney. 2002. Wavelet analysis. Chapter 1–Soil sampling and statistical procedures, pp. 184–195, in J. H. Dane and G. C. Topp, eds., *Methods of soil analysis. Part 4, Physical methods, SSSA Book Series 5*. Soil Science Society of America, Madison, Wisconsin.

Lark, R. M., and R. Webster. 2004. Analysing soil variation in two dimensions with the discrete wavelet transform. *European Journal of Soil Science* 55: 777–797.

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

Mathworks. 2010. Matlab version 7.0. *The Mathworks Inc.*, Natick, MA.

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

Metternicht, G., and J. A. Zinck. 2003. Remote sensing of soil salinity: Potentials and constraints. *Remote Sensing of Environment* 85: 1–20.

Minasny, B., and A. B. McBratney. 2006. A conditioned Latin hypercube method for sampling in the presence of ancillary information. *Computer & Geosciences* 32: 1378–1388.

Moonjun, R., A. Farshad, D. P. Shrestha, and C. Vaiphasa. 2010. Artificial neural network and decision tree in predictive soil mapping of Hoi Num Rin Sub-Watershed, Thailand, in J. L. Boettinger, D. W. Howell, A. C. Moore, A. E. Hartemink, and S. Kienast-Brown, eds. *Digital soil mapping. Progress in Soil Science 2*, Springer, Utah, USA.

Moore, I. D., R. B. Grayson, and A. R. Ladson. 1991. Digital terrain modeling: Review of hydrological, geomorphological and biological applications. *Hydrology Processing* 5: 3–30.

Moran, C. J., and E. N. Bui. 2002. Spatial data mining for enhanced soil map modeling. *International journal of Geographical Information Science* 16: 533–549.

National Cartographic Center. 2009. Research Institute of NCC, Tehran, Iran. www.ncc.org.ir

Nelson, D. W., and L. E. Sommers. 1982. Total carbon, organic carbon and organic matter, pp. 539–579, in L. A. Page, ed., *Methods of soil analysis, part 2, Chemical and microbiological properties*, 2nd ed. *Monograph 9. Agronomy Society of America and Soil Science Society of America, Madison, Wisconsin.*

Downloaded by [University of Tasmania] at 00:56 13 November 2014

Nield, S. J., J. L. Boettnger, and R. D. Ramsey. 2007. Digitally mapping Gypsic and nitric soil areas using Landsat ETM data. *Soil Science Society America Journal* 71: 245–252.

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

Quinlan, J. R. 2001. *See 5: An Informal Tutorial*. http://www.rulequest.com

Schmidt, K., T. Behrens, and T. Scholten. 2008. Instance selection and classification tree analysis for large datasets in digital soil mapping. *Geoderma* 146: 138–146.

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

Soil Survey Staff. 2010. *Keys to Soil Taxonomy*, 10th ed. *United States Department of Agriculture, Washington.*

Sparks, D. L., A. L. Page, P. A. Helmke, R. H. Leoppert, P. N. Soltanpour, M. A. Tabatabai, G. T. Johnston, and M. E. Summer. 1996. *Methods of soil analysis*. Soil Science Society of America, Madison, Wisconsin.

Stehman, S. V., and R. L. Czaplewski. 1998. Design and analysis for thematic map accuracy assessment: Fundamental principles. *Remote Sensing of Environment* 64: 331–344.

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

Downloaded by [University of Tasmania] at 00:56 13 November 2014