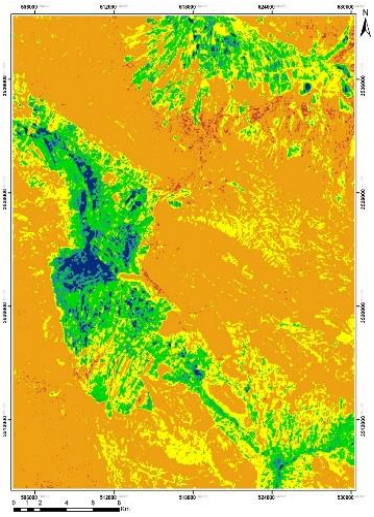
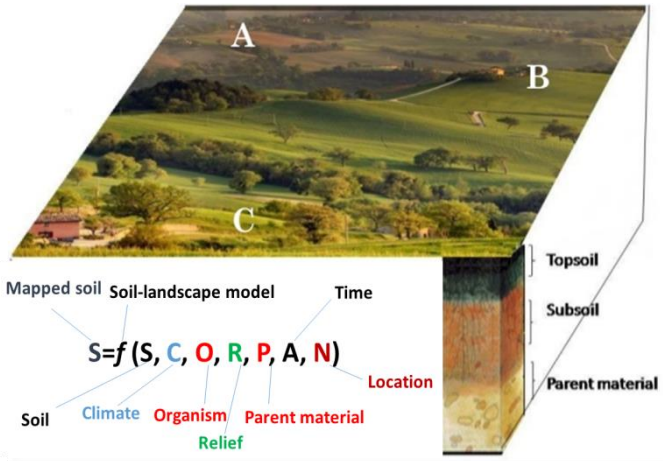
Digital soil mapping, downscaling and updating conventional soil maps using GIS, RS, Statistics and auxiliary data





Mojtaba Zeraatpisheh

Thesis submitted in fulfilment of the requirements for the degree of Doctor (PhD) of Science: Geology

GHENT UNIVERSITY

FACULTY OF SCIENCES

Academic Year 2016 - 2017

Digital soil mapping, downscaling and updating conventional soil maps using GIS, RS, Statistics and auxiliary data

Digitale bodemkartering, neerschalen en actualiseren van conventionele bodemkaarten gebruikmakend van GIS, remote sensing, statistiek en hulpkaarten

Mojtaba Zeraatpisheh

Promoters:

Prof. Dr. Shamsollah Ayoubi (Department of Soil Science, Isfahan University of Technology)

Prof. Dr. Peter Finke (Department of Soil Management, Ghent University)

Examination committee:

Prof. Dr. Hossein Shariatmadari (Department of Soil Science, Isfahan University of Technology)

Prof. Dr. Ir. Olivier Thas (Department of Mathematical Modelling, Statistics and Bioinformatics, Ghent University)

Acknowledgments

First and foremost, I would like to thank Allah for providing the opportunity of study and crossing my road with supervisors, friends and colleagues.

I would like to send my sincere thanks to my promoters Prof. Dr. Peter Finke and Prof. Dr. Shamsollah Ayoubi for the valuable advice and for persisting and inspiring me with his positive attitude. I want to dedicate my special thanks to my advisors, Prof. Dr. Hossein Khademi and Dr. Azam Jafari for their advice, patience, and support throughout my research.

I would like to express my deepest appreciation and thanks to Prof. Dr. Ir. Olivier Thas and Prof. Dr. Hossein Shariatmadari, as the members of the Examination board of the PhD defense for critical feedback, spending plenty of their valuable times, for making constrictive comments and providing useful suggestions.

I would like to acknowledge all my colleagues in Ghent University and Isfahan University of Technology: Meisam, Shahrokh, Samaneh, Vajieh, Emmanuel, Saba, Thanh Thuy Doan, Mesfin, Okky, Nguyen Min Phoung. Outside University, a special thanks goes to my friends and their families, Dr. Maximilian Witt, Prof. Dr. Dieter Witt, Sandro Lobina, Michael Mayer, Alexander Odgen, Jonas Mertens, Dr. Yvonne Witt, Luc, Hilde and Steffi Mertens for nice parties, for nice traveling and warm hospitality.

To my mom and dad, a super special thanks to them for their supports and words of encouragement. I heartily appreciate my wife Maedeh for her love, tolerance and understanding towards me.

This work was partly funded by the Ministry of Science, Research and Technology (MSRT) of Iran, I am grateful for their support. Lastly, I would like to extend my sincere gratitude to all those in the Department of Soil Science at Isfahan University of Technology, the Department of Geology and Soil Science and the Department of Soil Management at Ghent University in providing me the opportunity to pursue the Joint PhD program.

August 2017

Mojtaba Zeraatpisheh

DEDICATION

To Maedeh, My Sweet Home;

To my family and to each of those who have stood with me when I did not think that I could stand.

Summary

Spatial distribution of soil types and soil properties in the landscape are important in many environmental researches. Conventional soil surveys are not designed to provide the high-resolution soil information required in environmental modelling and site-specific farm management. The objectives of this study were to investigate the relationship between soil development, soil evolution in the landscape, updating legacy soil maps and pedodiversity in an arid and semi-arid region. The application of Digital Soil Mapping (DSM) techniques was investigated with a particular focus to predict soil taxonomic classes and spatial distribution of soil types by soil observations and covariate sets representative of s,c,o,r,p,a,n factors.

In the first study, focus is on establishing relationships between pedodiversity and landform evolution in a 86,000 ha region in Borujen, Chaharmahal-Va-Bakhtiari Province, Central Iran. From an overview study, we could conclude that landform evolution was mainly affected by topography and its components.

A second study compares various DSM-methods and a conventional soil mapping approach for soil class maps in terms of accuracy, information value and cost in central Iran. Also, the effects of different sample sizes were investigated. Our results demonstrated that in most predicted maps, in DSM approaches, the best results were obtained using the combination of terrain attributes and the geomorphology map. Furthermore, results showed that the conventional soil mapping approach was not as effective as DSM approach.

In the third study, different models of the DSM approach were compared to predict the spatial distribution of some important soil properties such as clay content, soil organic carbon and calcium carbonate content. Among all studied models, the terrain attribute “elevation” is the most important variable to predict soil properties. Random forest had promising performance to predict soil organic carbon. But results revealed that all models could not predict the spatial distributions of clay content properly.

The minimum area of land that can be legibly delineated in a traditional (printed) map is highly dependent upon mapping scale. For example, this area at a mapping scale of 1:24,000 is about 2.3 ha but at a mapping scale of 1:1,000,000 it is about 1000 ha. A mapping scale of 1:1,000,000 is just too coarse to show a fine-scale pattern or soil type with any degree of legibility, but finer-scale soil maps are more expensive and time-consuming to produce. Thus, spatial variation is often unavoidably obscured. The fourth study of this dissertation focuses on downscaling and updating soil map methods. Thus, the objectives were to apply supervised and unsupervised disaggregation approaches to disaggregate soil polygons of conventional soil map at a scale of 1: 1,000,000 in the selected area. Therefore, soil subgroups and great groups were selected because it is a basic taxonomic level in regional and national soil maps in Iran.

In general, we conclude that DSM approach and also disaggregation approach are capable to predict soil types and properties, produce and update legacy soil maps. However, still a number of challenges need to be evaluated e.g. influence of expert knowledge on CSM approach, resolution of ancillary data, georeferenced legacy soil samples data to validate disaggregated soil maps.

Samenvatting

De ruimtelijke distributie van bodemklassen en bodemeigenschappen over het landschap is van belang in veel omgevingsstudies. Conventionele bodemkarteringen zijn niet toegespitst op het verkrijgen van bodeminformatie in hoge resolutie, zoals nodig voor omgevingsmodellering, and voor locatie-specifieke landbouw. De doelstellingen van deze studie waren het doen van onderzoek naar bodemontwikkeling in een landschappelijke context, naar actualisatie van bestaande bodemkaarten en bodemdiversiteit in een aride en semi-aride regio. De toepassingsmogelijkheden van "Digital Soil Mapping" (DSM)

technieken werden onderzocht met speciale aandacht voor de voorspelling van taxonomische bodemklassen en de ruimtelijke verdeling van bodemklassen, gebruikmakend van bodemobservaties en kaarten van hulpvariabelen welke de s,c,o,r,p,a,n factoren weerspiegelen.

In een eerste studie lag de focus op het vaststellen van relaties tussen de bodemdiversiteit en de evolutie van de landvormen in een 86.000 ha regio in Borujen, Chaharmahal-Va-Bakhtiari Provincie, Centraal Iran. Uit een overzichtsstudie konden we concluderen dat de evolutie van de landvormen hoofdzakelijk werd bepaald door de topografie en daarvan afgeleide componenten.

Een tweede studie vergelijkt verschillende DSM-methoden en een conventionele bodemkartering voor het maken van kaarten van de bodemklasse in termen van accuraatheid, informatiewaarde en productiekosten in Centraal Iran. Hierbij werden ook de effecten van de steekproefomvang mee beschouwd. Onze resultaten demonstreerden dat, in de meeste kaarten geproduceerd met DSM, de beste resultaten werden bereikt gebruikmakend van een combinatie van terreinattributen en de geomorfologische kaart. Tevens lieten de resultaten zien dat de conventionele bodemkartering niet zo effectief was als de DSM-benadering.

In de derde studie werden verschillende DSM-modellen vergeleken om de ruimtelijke distributie van een aantal belangrijke bodemkenmerken te voorspellen, zoals het gehalte aan klei, organische koolstof en calciumcarbonaat. Over alle bestudeerde modellen was het terreinattribuut "terreinhoogte" de belangrijkste variabele bij het voorspellen van bodemeigenschappen. De Random Forest methode had veelbelovende prestaties bij de voorspelling van organische koolstof. Echter, resultaten lieten zien dat geen van de modellen de ruimtelijke verdeling van het kleigehalte goed kon voorspellen.

De minimale oppervlakte die leesbaar kan worden afgegrensd in een traditionele (gedrukte) kaart hangt sterk af van de karteringsschaal. Bijvoorbeeld, bij een karteringsschaal

van 1:24 000 is deze oppervlakte ongeveer 2.3 ha, maar bij een karteringsschaal van 1;1 000 000 is deze circa 1 000 ha. Een schaal van 1:1 000 000 is te grof om de fijnschalige patronen van het bodemtype leesbaar weer te geven, echter gedetailleerdere bodemkaarten kosten meer geld en tijd om te produceren. Dientengevolge is het onvermijdelijk dat de ruimtelijke variatie niet wordt weergegeven. De vierde studie van deze dissertatie richt zich op neerschalings- en actualisatiemethoden. De doelstellingen waren hier om gecontroleerde ("supervised") en ongecontroleerde disaggregatiemethoden toe te passen om bodemkaartvlakken neer te schalen, uitgaande van een conventionele bodemkaart 1:1 000 000 van het geselecteerde gebied. Hiertoe werd gewerkt met de taxonomische niveaus "subgroup" en "great group" welke in regionale en nationale bodemkaarten in Iran worden gebruikt.

De algemene conclusie luidt dat de DSM-benadering en ook neerschalingsmethoden in staat zijn om bodemklassen en bodemkenmerken ruimtelijk te voorspellen en om bestaande kaarten te actualiseren. Echter, aandacht blijft vereist voor bijvoorbeeld de invloed van expertkennis bij een traditionele bodemkartering, effecten van de ruimtelijke resolutie van de hulpinformatie en de aanwezigheid van bodemgegevens met georeferentie om neergeschaalde bodemkaarten te valideren.

Table of Contents

Chapter 1

General Introduction, Research Necessity, and Objectives

1.1 Introduction

Soil is the natural part of the surface of the earth, characterized by layers or horizons parallel to the surface resulting from the modification of parent materials by physical, chemical and biological processes operating under varying conditions during varying periods of time (Thornbury, 1969; White, 2013).

Soil is limited in quantity and degradable in quality; therefore, this resource is a non-renewable and irreplaceable capital good in all the productive activities of humans and plays a key role in the natural environment (Bouma, 2006). It is important to know the spatial distribution of soil types and soil properties in the landscape. Standard soil surveys are not designed to provide the high-resolution soil information required in environmental modelling and site-specific farm management (Petersen, 1991). Conventional approaches to soil mapping produce maps that delineate neither the inherent variability of the soil nor the variation of the attributes mapped. Moore et al., (1993) have proven these approaches were expensive and often unreliable to mapping the spatial variability soil resources. Therefore, there is a vital need for the improvement of accurate and inexpensive approaches of mapping the spatial variability of the soil types and soil properties. In the last 30 years, improvement of technology and accuracy of Geographic information science (GIS) prepared great potential to develop the efficiency and quality of methods used to gather spatial soil information (McBratney et al., 2000; Scull et al., 2005).

In contrast to conventional approach of soil mapping which is expensive, time-consuming and unreliable (Moore et al., 1993), digital soil mapping (DSM) can reduce the production

costs by establishing relationships between expensive soil observations and inexpensive ancillary data (Kempen et al., 2012; McBratney et al., 2003).

DSM has matured to become a legitimate branch within the soil science domain. DSM has evolved from a science-driven research phase of the early 1990s to presently, a fully operational and functional process. This evolution is evidenced by the increasing extents of DSM projects from small research areas towards regional, national and even continental extents (Malone, 2013).

DSM is defined by Lagacherie et al., (2006) as the creation and population of spatial soil information systems by numerical models inferring the spatial and temporal variations of soil types and soil properties from soil observation, remotely sensed data, terrain analysis data, vegetation, climate and expert knowledge.

Jenny (1941) proposed that soil development is a function of climate, organisms, topography, parent material and time; this hypothesis can be the basic assumption of DSM. Hence, soil properties of a location can theoretically be estimated if information about those variables is available for the location. McBratney et al., (2003) introduced further variables space (spatial position) and soil information derived from other investigations, as soil can be predicted from its own properties in the so-called "*SCORPAN* model". The *SCORPAN* approach is expressed as by the equation:

$$S_c = f(S, C, O, R, P, A, N) + \varepsilon$$

where S_c is soil types or soil properties, S is the soil information such as obtained from a prior soil map or expert knowledge, C refers to climate, O is organisms such as human activity, R is relief, P is parent material, A is age, N refers to neighborhood or spatial position and ε is spatial dependent residuals.

These ancillary data used in DSM usually can be obtained relatively cheaply over large areas, e.g. a digital elevation models (DEM) and its derivatives, satellite images and

geospatial information (e.g. geology map and geomorphology map). The ancillary data are assumed to have some relationship with the soil variables (Goovaerts, 1997). Figure 1-1 shows the process of DSM, where geo-referenced soil observations are joined with environmental variables from the input data.

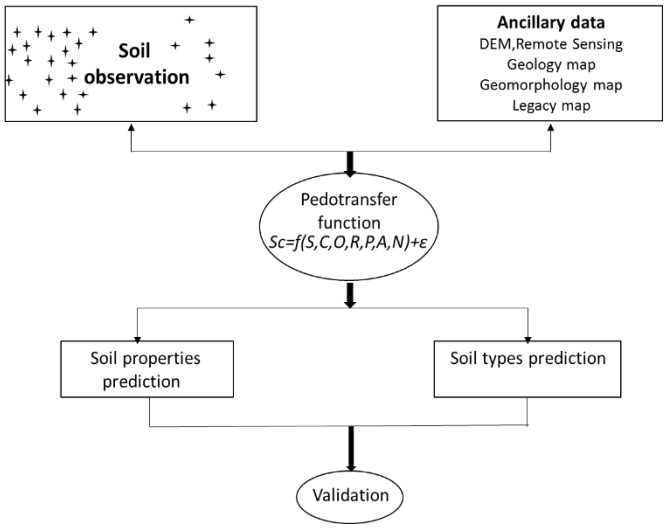


Figure 1-1- Workflow of Digital soil mapping.

In Iran, like the most developing countries in the world, precision farming, and agriculture are growing activities. Precision agriculture and farm-based planning require high-resolution soil maps and reliable application methods. DSM provided this pathway to predict and produce soil maps and soil information by the efficient cost and accuracy.

Although soil survey studies started in 1950 in Iran, some areas have not yet been mapped at any scale and most soil surveys have been carried out using traditional methods and also only 20 to 25 percent of the country has already been mapped. Iran has a total of 22 million ha of agricultural land, while agriculture has played a key role in the economy of the country, only 3.3 million ha has been mapped at a scale of 1:20,000, 16 million ha at a scale of 1:50,000, 1.3 million ha at a scale of 1:100,000 and 1 million ha at other scales. It is obvious that more than 50 percent of the agricultural land has been mapped at the scale coarser than the scale of 1:50,000 (Banaei et al., 2005).

To map the entire country, traditional methods are costly and time-consuming, consequently, alternative approaches such as DSM methods can accelerate the preparation of soil maps of the country.

In recent years, the national map of soil types at a scale of 1:1,000,000 was published by the Soil and Water Research Institute of Iran (Banaei et al., 2005; Mohammad, 2000). This map was prepared based on physiography units and did not illustrate soil variation properly.

Nevertheless, soil mapping has made certain progress. Conventional soil maps produced in the past decades are the major data sources for information on the spatial variation of soil, but they are limited in terms of both the levels of spatial detail and the accuracy of soil attributes as well as have high requirements of costs and time.

Regarding the limitations and shortcomings mentioned, this thesis deals with concepts and techniques developed in three separate bodies of knowledge: pedodiversity and soil landscape, digital soil mapping of soil class and soil properties and finally disaggregation of the legacy soil map.

1.2 Research objectives

The objectives of this dissertation were to investigate the relationship between soil development, soil evolution in the landscape and pedodiversity in an arid and semi-arid region (chapter 2).

The application of DSM techniques was investigated with a particular focus to predict soil taxonomic classes and spatial distribution of soil types by soil observations and covariate sets representative of s,c,o,r,p,a,n factors. Cost, accuracy, and efficiency in the study area, and the effect of sample size were also investigated (chapter 3).

Chapter 4 compares different models of the DSM approach to predict the spatial distribution of some important soil properties such as clay content, soil organic carbon and calcium carbonate content.

The minimum area of land that can be legibly delineated in a traditional (printed) map is highly dependent upon mapping scale. For example, this area at a mapping scale of 1:24,000 is about 2.3 ha but at a mapping scale of 1:1,000,000 it is about 1000 ha (Soil Survey Division Staff, 1993). A mapping scale of 1:1,000,000 is just too coarse to show a really fine-scale pattern or soil type with any degree of legibility, but finer-scale soil maps are more expensive and time-consuming to produce. As a result, spatial variation is often unavoidably obscured. Thus, the objectives of Chapter 5 were to apply supervised and unsupervised disaggregation to disaggregate soil polygons of conventional soil map at a scale of 1:1,000,000 in the selected area. In this chapter soil subgroups and great groups were selected because it is a basic taxonomic level in regional and national soil maps in Iran.

Chapter 2

Relationships between pedodiversity and landform evolution in a semi-arid region, central Iran

Abstract

This study was conducted to explore the role of landform forming processes on soil development, and to investigate soil homogeneity within landform units in terms of soil properties and soil classes in a semi-arid region of central Iran. The landform delineation was carried out by air photo interpretation (API). A total number of 125 profiles were described and classified up to the subgroup level according to U.S Soil Taxonomy. The investigated soil properties within the study area in 0-30 cm layer showed high variability. The coefficients of variation for all soil properties were high, which indicated high variability in soil forming factors across the study area. Almost all coefficients of variation (CV) for soil properties in different landforms decreased as compared to overall CV. The CV indicates that landform delineation contains information on soil variability and that similar pedogenesis processes which occurred within land units led to similar soils. Diversity indices increased from the soil order to the soil subgroup, with an abrupt increase from great group to subgroup level. Pedodiversity indices-to-area relationships showed that there are additional soil classes if the number of observations would increase. It can be deduced that probably the number of observations were insufficient in the studied area. It is also proved that soil diversity increases with the landforms' area. The results suggested that pedodiversity could be considered as an effective method to break down soil and landform complexity.

2.1 Introduction

Effective soil management needs an understanding of soil distribution patterns within the landscape (McBratney et al., 2000). Pedogenetic processes are the main causes of variations

in soil properties (Liu et al., 2006). These processes commonly are used for soil classification; therefore, soil types represent (at least part of) the variations in pedogenesis processes.

The importance of understanding landscape dynamics, heterogeneity, and environmental changes is increasingly acknowledged (Wilson and Forman, 1995). Quantitative techniques are needed to analyze landscape and diversity, with emphasis on landscape structure and spatial patterns (Turner and Gardner, 1991). Several quantitative approaches of soil diversity have been proposed (McBratney, 1995; Saldaña and Ibáñez, 2004).

Currently, pedodiversity analysis is considered as a branch of pedometrics or mathematics applied to pedology. Pedodiversity studies were firstly started by analyzing soil series-area relationships (Beckett, 1978). In the end of 20th century, McBratney (1992) and Ibañez et al., (1995) explained pedodiversity and expressed a need for it regarding soil conservation. It can be simply defined as the variation of soil properties or soil classes within a study area. Pedodiversity, as well as biodiversity, is a method to evaluate soil variability, spatial patterns and soil species (McBratney, 1992) usually using taxa from soil classification (Minasny and McBratney, 2007). Moreover, it is also a way to quantify and understand the structure of the soil variation. Ibáñez and Effland (2011) considered pedodiversity analysis as an interesting mathematical tool in soil geography.

The ecological diversity indices introduced as measures of soil diversity, comprise richness, abundance, evenness and proportional abundance (e.g. Shannon entropy) (Ibañez et al., 1995; Phillips, 2001). Recently, also taxonomic distance was recommended as a pedodiversity measure (Minasny and McBratney, 2007; Petersen et al., 2009). Diversity can be analyzed in any context where it is possible to establish a classification or taxonomy (Ibáñez et al., 2013). Ibañez et al. (1998) calculated the diversity indices at the worldwide level for continents based on the FAO Soil Map. McBratney et al. (2000) showed that soils in

some continents like Australia are less diverse than South and Central America. In order to evaluate pedodiversity in the USA, Guo et al. (2003) calculated Shannon's entropy for different taxonomic levels from order to suborder, great group, subgroup, family, and series (U.S. Soil Taxonomy). They found an increase in taxonomic richness and Shannon's entropy with increasing taxonomic level. Toomanian et al. (2006) and Jafari et al. (2013) studied soils on different landscapes in central Iran and reported an increase of diversity indices at lower Soil Taxonomy classification levels and soil geomorphologic categories.

Many researchers used landform analysis to determine and describe soil variability and also to delineate homogeneous parts of the landscape (Hengl and Rossiter, 2003; Jafari et al., 2013; Moore et al., 1993; Toomanian et al., 2006). Landforms should be as homogeneous as possible for interpretation of soil as a part of the dynamic ecosystem.

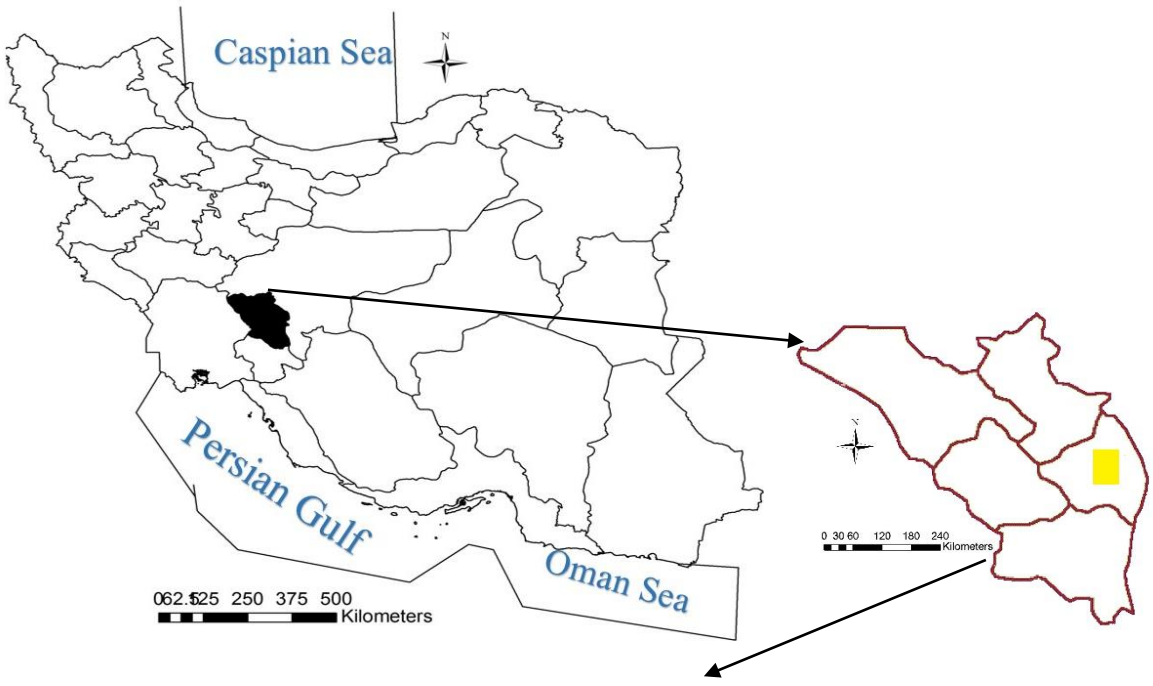
2.1.1 Objectives

Unfortunately, the information on spatial distribution of soils in Iran is frequently insufficiently available. Generally, the available information is restricted to the soil type, soil texture, descriptions of soil diagnostic horizons, and subsoil texture, and is based on the conventional soil survey. Therefore, there is little information about soil diversity, landscape evolution, soil pedogenesis process and their relationships in arid to semi-arid regions of Iran. Thus, the objectives of this study were i) to understand the relationship of soil-landform evolution and pedogenesis processes across a toposequence and ii) to analyze pedodiversity indices in different landform types in order to explore heterogeneity/homogeneity of soils and landforms at various taxonomic levels in a semi-arid region of central Iran.

2.2 Materials and Methods

2.2.1 Description of the study area

The study area is located between 51° 19′ 9″ to 51° 20′ 45″ E longitude and 31° 41′ 00″ to 32° 00′ 00″ N latitude, and area covering approximately 86,000 ha in the Borujen region, Chaharmahal-Va-Bakhtiari Province, Central Iran (Figure 2-1). The mean annual precipitation is 255 mm, mean annual temperature is 10.7°C, and mean elevation of the selected area is 2277 m a.s.l (Esfandiarpoor et al., 2009). The main land uses in this area include irrigated wheat cropping, dryland farming, and pasture. According to the US Soil Taxonomy (Soil Survey Staff, 2014), the study area has a Xeric soil moisture regime and a Mesic soil temperature regime. Major landscape units in the study area consist of mountains, hills, piedmonts, and lowlands.



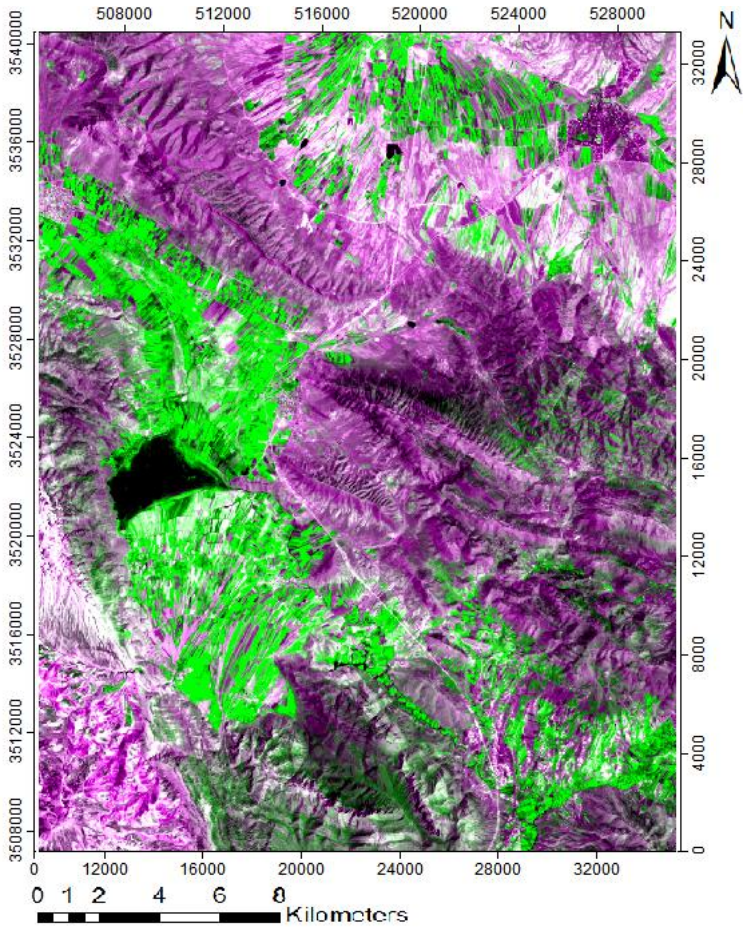


Figure 2-1- The location of the study area (Landsat ETM+ image; RGB: 243). The black area in the upper left of the figure identifies the Chaharmahal-Va-Bakhtiari Province among all of the provinces in Iran. The upper right part of the figure shows districts in the Province of Chaharmahal-Va-Bakhtiari and the location of the study area.

2.2.2 Landform delineation

Air photo interpretation (API) was done to delineate landform patterns using the aerial photo (1:50,000), based upon the knowledge on landform formation processes, general structure, and morphometry. Hengl and Rossiter (2003) proposed that in the flat or low relief regions, where high-resolution images are highly cost demanding, it is superior to use the traditional air photo interpretation (API) method to differentiate the landscape patterns. Therefore, it could be the cheapest way to delineate the accurate landform boundaries by using the reflectance contrasts on aerial photos (Toomanian et al., 2006). The study area consists of four dominant landscape units, which comprise of hill land, piedmont, mountain, and lowland. Mountain is the main landscape which is composed of mainly rock outcrop

landform and occupied 34.90 % of the study area. The other different landforms were identified including eroded hill land (Hi1, 4.55%), developed hill land (Hi2, 7.28%), alluvial fan (Pi1, 30.36%), wetland (Pl1, 0.35%), lagoon (Pl2, 0.95%), piedmont plain (Pi5, 6.64%), river plain (Pi4, 0.66%), pediment (Pi3, 9.20%) and dissected alluvial fan (Pi2, 6.08%). The spatial distribution of described landforms and their legends are presented in Figure 2-2 and Table 2-1, respectively.

Table 2- 1- Landscape and landform units of the study area.

Landscape	Landform	Area (km^2)	Percentage of total area	Landform codes	Description
Hill	Eroded hill lands	39.21	4.54	Hi1	Single and low topography hills
	Developed hill lands	62.91	7.28	Hi2	Continuous hills with high topography
Mountain	Rock outcrop	292.94	33.91	Mo1	Eroded rock surface, shallow soils
Piedmont	Alluvial fan	262.30	30.36	Pi1	Active fan, cultivated plain in lower slope
	Dissected alluvial fan	52.57	6.09	Pi2	Dissected and undulating red alluvial fan
	Pediment	79.64	9.22	Pi3	Shallow soil, colluvial material
	River plain	5.74	0.66	Pi4	Low drainage, young terraces
	Piedmont plain	57.41	6.65	Pi5	Young terraces, moderate depth soils
Lowland	Wetland	3.02	0.35	Pl1	Low drainage, wetness
	Lake (Lagoon)	8.17	0.95	Pl2	Covered by hydrophilic plants

2.2.3 Soil surveying method, Sampling, and profile description

The soil sampling scheme was carried out by applying the conditioned Latin hypercube sampling (Minasny and McBratney, 2006) algorithm using Matlab software (MathWorks, 2009) with all covariates to be mentioned in section 3.2.3 (Table 3-1, chapter 3). Location coordinates of 100 soil profiles were acquired by Latin hypercube sampling and 25 legacy profiles were added to our dataset. Figure 2-2 shows the distribution of the soil profiles

described in the study area. All locations were excavated to a depth of 100–150 cm, described, sampled, analysed, and classified up to the subgroup level of the US Soil Taxonomy (Soil Survey Staff, 2014).

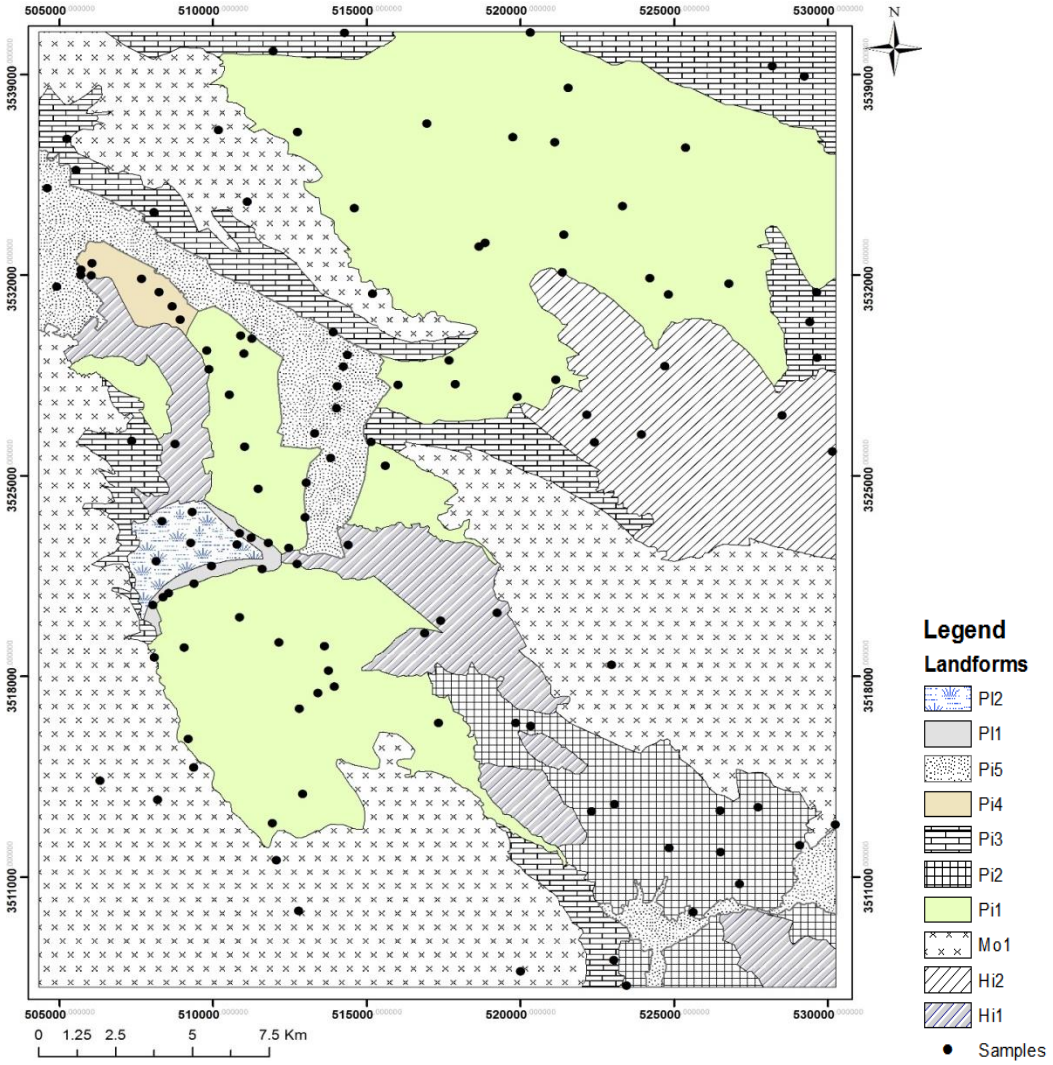


Figure 2-2. Delineated landform map and sampling points in the study area. Geomorphic surfaces in the legend are explained in Table 2-1.

2.2.4 Diversity indices

For the analysis of soil diversity (e.g. pedodiversity), all 125 observed profiles in the study area were analyzed at different (hierarchical) landform levels. The studied soils were classified into four categories including order, suborder, great group and subgroup according to U. S. Soil Taxonomy (Soil Survey Staff, 2014).

Pedodiversity indices comprising Shannon entropy, richness and evenness were calculated for each landform category. To calculate the diversity indices in each landform, the number of profiles belonging to a given landform (n_i) and the total number of profiles in the study area (N) were considered. Richness (S) is a measure of the number of different objects or soil types in the site. Therefore, the number of soil classes (at the different taxonomic levels) in each landform was considered as the richness of species. Abundance is defined as the distribution of the number of soil individuals. The proportional abundance of objects is the most frequently used method to estimate the diversity (Toomanian et al., 2006). The Shannon diversity index (H′) is the most frequently used proportional abundance index (Ibáñez et al., 2013; Longuet-Higgins, 1971; Toomanian et al., 2006). The Shannon diversity index is defined mathematically as follows:

$$\mathbf{H'} = -\sum_{\mathbf{i=0}}^{\mathbf{n}} \boldsymbol{p_i} * \boldsymbol{lnp_i} \qquad \text{(Eq. 2.1)}$$

where H' is the negative entropy, negentropy, or diversity of a population and p_i is the proportion of individuals found in the i^{th} object. In calculations, instead of the p_i, the proportion of n_i/N was used.

The maximum possible Shannon diversity index (H_{max}) occurs in situations where all objects are equiprobable, therefore H_{max} is used to measure the evenness (E). The evenness index can be used as a measure of the heterogeneity of the distribution of taxa within soilscapes (Pielou, 1966) if the following condition is fulfilled:

$$H' = H_{max} = \ln S \qquad \text{(Eq. 2.2)}$$

Then, evenness takes the following form:

$$E = H'/H_{max} = H'/\ln S \qquad \text{(Eq. 2.3)}$$

S here is the richness, the relative number of individuals in each category or landform. The Evenness index varies from 0 to 1. The maximum value of H is defined as $H_{max}=lnS$, a value

close to H_{max} indicates an even proportional contribution of each class (Costantini and L'Abate, 2016; Martín and Rey, 2000).

The O'Neill dominant index (*D*) was used to examine the deviation of the calculated Shannon diversity index from the maximum diversity H_{max} (O'Neill, 1988) as follows:

$$D = \ln(S) + \sum_{i}^{S} p_i * ln(p_i) \quad \text{(Eq. 2.4)}$$

Several indices have been derived from the number of taxa recorded and the total number of individuals of all species (numerical species richness). Two of them are presented here, the first one is the Margalef's index (Magurran, 1988) and is calculated by the following equation:

$$\boldsymbol{Dmg = \frac{(S-1)}{lnN}} \quad \text{(Eq. 2.5)}$$

The other is The Menhinick's index (Whittaker, 1977) is calculated as follows:

$$Dmn = \frac{S}{\sqrt{N}} \quad \text{(Eq. 2.6)}$$

where *S* is the richness (i.e., soil types) recorded and *N* is the total number of individuals (i.e., the number of samples summed in every landform) overall *S* species. Both indices varied from 0 to l (Saldaña and Ibáñez, 2004).

Two kinds of functions could display the relationship between the number of species and the area: (1) a linear relationship between log S (richness) and log area (power function) and (2) a linear relationship between S and log area (logarithmic function) (Saldaña and Ibáñez, 2004). Toomanian et al. (2006) recommended the richness–area curve for the soil family richness versus the area of geomorphic surfaces. In the current study, the relationships between the area of landforms and the soil richness (richness-area relationship) and also the Shannon diversity index (Entropy–area relationship) at the subgroup level were examined.

2.2.5 Statistical analysis

For statistical analysis of some soil surface (0-30 cm depth) properties, descriptive statistical analyses including standard deviation (Std.Dev.), variance, the coefficient of variation (CV), mean, minimum and maximum were calculated using the SPSS software (version 17.0). One way analysis of variance (ANOVA) was used to analyze the significance of landform types on soil properties (Duncan's test at the 5% level of significance).

2.3 Results and Discussion

Table 2- 2 presents a brief statistical description of some soil properties of surface soil layer (0-30 cm) for the study area. According to this table, the investigated soil properties for the entire studied area in 0-30 cm layer had high variability. Soil organic carbon (OC) ranged from 0.18% to 8.88%, with a CV of 96.99%. Soil texture analysis also showed high variability for particle size distribution, where clay ranged from 15.96% to 64.00%, with a CV of 27.48%; silt ranged from 6.00% to 61.36%, with a CV of 33.37% and sand ranged from 4.00% to 53.88%, with a CV of 39.74%. Soil calcium carbonate equivalent (CCE) showed relatively low variability (CV= 29.90%) with a range of 11% to 72.50%. This low variability of CCE might explain that all soils of the studied area are affected by limestone parent material, rather than other soil forming factors. Wilding (1985) categorized CV values into 3 classes with high (CV > 35%), moderate (15% < CV < 35%) and low variability (CV < 15%). According to this classification, the coefficients of variation for selected soil properties were high, which indicates a broad range of values across the study area.

Table 2- 2. The coefficient of variation (CV), minimum and maximum for some soil properties in the study area for 0-30 cm layer.

Statistical criteria	OC (%)	CCE (%)	Sand (%)	Silt (%)	Clay (%)
CV (%)	96.99	29.90	39.74	33.37	27.48
Minimum	0.18	11.30	4.00	6.00	15.96
Maximum	8.88	72.50	53.88	61.36	64.00
Mean	1.46	35.74	21.82	39.58	38.49

Descriptive statistics of some soil properties determined in this study within distinguished landforms are presented in Table 2-3. Almost all CVs for soil properties within landforms decreased as compared to overall CV throughout the study area, which indicates that landform delineation leads to homogeneous units and could explain a part of soil variability (Table 2-3). Post et al.,(2008) stated that the CVs or the variances may be used for comparison of the variability between the different sources. Other researchers, Borujeni et al., (2010) Moore et al., (1993) and Toomanian et al., (2006) also reported that landform delineation can partly explain soil variability and therefore can be used to produce more homogeneous soil units.

Table 2-3. Descriptive statistics of some surface soil properties (0-30 cm) in different landforms.

Landforms	Soil properties (%)	Std.Dev.	Variance	CV (%)	Min	Max	Mean
Hi1	OC	0.29	0.08	34.56	0.53	1.27	0.85
	CCE	11.25	126.61	38.42	11.30	42.30	29.29
	Sand	3.58	12.80	15.83	19.00	28.00	22.60
	Silt	5.86	34.30	14.79	34.00	49.00	39.60
	Clay	5.93	53.20	15.70	31.00	44.00	37.80
Hi2	OC	0.33	0.11	32.20	0.72	1.58	1.03
	CCE	5.57	31.01	28.05	12.25	27.75	19.85
	Sand	1.72	2.97	10.88	14.00	18.00	15.83
	Silt	4.23	17.87	10.48	32.00	44.00	40.33
	Clay	3.49	12.17	7.96	40.00	50.00	43.83
Mo1	OC	0.18	0.03	30.82	0.36	0.95	0.59
	CCE	8.81	77.56	22.83	25.30	56.00	35.58
	Sand	8.08	65.33	31.49	12.00	39.00	25.67
	Silt	9.63	92.70	24.58	26.00	58.00	39.17
	Clay	6.37	40.52	18.10	60.00	52.00	35.17
Pi1	OC	1.37	1.87	104.44	0.18	5.40	1.31
	CCE	8.67	75.16	22.89	19.75	55.25	37.88
	Sand	9.02	81.42	44.07	4.00	52.00	20.47
	Silt	6.96	48.45	17.64	24.00	60.36	39.45

	Clay	7.54	56.91	18.93	24.00	57.00	39.86
	OC	0.40	0.16	39.53	0.45	1.81	1.00
	CCE	13.78	189.82	40.07	15.00	49.80	34.38
Pi2	Sand	7.61	57.97	40.43	11.00	38.00	18.83
	Silt	8.28	68.57	20.32	30.00	57.00	40.75
	Clay	7.94	62.99	19.64	26.00	50.00	40.42
	OC	0.19	0.04	27.39	0.41	1.05	0.69
	CCE	10.81	116.81	32.76	18.80	60.80	32.99
Pi3	Sand	8.01	64.09	31.57	13.00	38.00	25.36
	Silt	5.26	27.65	13.31	30.00	49.00	39.50
	Clay	3.94	15.52	11.21	30.00	43.00	35.14
	OC	0.85	0.72	31.48	1.23	3.90	2.70
	CCE	1.70	2.88	5.24	29.70	35.00	32.40
Pi4	Sand	3.55	12.57	16.49	14.00	26.00	21.50
	Silt	6.30	39.64	12.05	44.00	61.00	52.25
	Clay	8.74	76.41	33.78	18.00	42.00	25.88
	OC	0.68	0.46	59.61	0.41	2.23	1.13
	CCE	17.00	289.11	41.38	23.50	72.50	41.09
Pi5	Sand	14.28	203.84	58.99	10.20	53.88	24.20
	Silt	8.96	80.23	20.22	30.00	60.00	44.30
	Clay	10.49	109.96	33.29	15.96	44.00	31.50
	OC	2.63	6.90	46.89	1.58	8.88	5.60
	CCE	9.56	91.47	25.10	28.50	53.80	38.11
Pl1	Sand	5.99	35.86	43.61	7.84	22.00	13.73
	Silt	13.69	187.46	28.79	30.00	61.36	47.56
	Clay	9.87	97.37	25.49	28.52	50.00	38.71
	OC	0.11	0.02	3.44	3.33	3.60	3.36
	CCE	1.44	2.07	387	35.00	39.00	37.16
Pl2	Sand	1.67	2.80	5.40	29.00	33.00	31.00
	Silt	1.60	2.57	22.35	6.00	10.00	7.16
	Clay	1.72	2.97	2.79	60.00	64.00	61.83

It could be argued that the landform unit with low CV (e.g. Pl2 unit) and landform units with high CV (e.g. Pi1 and Pi5 units) are the most homogeneous and heterogeneous units,

respectively. Based on the field observations, some part of soil spatial variability could be explained by land use management due to different agriculture practices and local-scale soil variability. Additional statistical analysis showed that there are significant differences for the mean of soil properties among different landforms except for sand content (Table 2-4). However, in general, there was no significant difference for sand content.

Table 2-4. Comparison of the mean values of some soil properties (0-30 cm) in different landforms.

Landforms	OC (%)	CCE (%)	Sand (%)	Silt (%)	Clay (%)
Hi1	0.85cd	29.29ab	22.60a	39.60ab	37.80bcd
Hi2	1.031cd	19.85b	15.83a	40.33ab	43.83b
Mo1	0.592d	38.58a	25.67a	39.16b	35.16bcd
Pi1	1.311d	37.88a	20.47a	39.45b	39.85b
Pi2	1.007d	33.26ab	18.83a	40.75b	40.41bc
Pi3	0.699d	32.99ab	25.35a	39.50b	35.14bcd
Pi4	2.705bc	32.41ab	21.50a	52.25a	25.87d
Pi5	1.134cd	41.09a	24.20a	44.3ab	31.49cd
Pl1	5.605a	38.11ab	13.73a	47.55ab	38.71bcd
Pl2	3.330b	37.50ab	30.00a	6.00c	64.00a

Note: Values in the same column followed by the same letter(s) are not significant at the 5% level of significance according to ANOVA. Landforms defined in Table 1.

2.3.1 Soil and landform evolution

The results of soil description and classification indicated the occurrence of three soil orders including Inceptisols, Entisols and Mollisols within ten different landform units in the study area. The numbers of observed profiles for different Soil Taxonomy categories comprising order, suborder, great group, and subgroup are presented in Table 2-5. Five suborders, seven great groups and twelve subgroups in the selected area were identified. A sequence of evaluated soils throughout the main distinguished landforms is shown in Figure 2-3. The majority of Mollisols occurred in Pl1 and Pl2 landforms (lowland landscape), where the groundwater table is close to the soil surface. Shallower water tables in an arid landscape may correspond to increased biomass production which might lead to organic carbon

accumulation via humification (see high OC content for Pl1 and Pl2 units in Table 2-4). These are two main processes to form a mollic epipedon (Bockheim, 2015). The organic carbon (OC) accumulation has a strong relationship with clay contents and climate condition (Zeraatpishe and Khormali, 2012). Inceptisols were mainly present in piedmont landscapes (Pi) and all Entisols were located in the mountain, piedmont and hill land landscapes (Mo, Pi and Hi, respectively).

Table 2-5. Number of soil classes observed in the study area in different soil taxonomy levels.

Order	No. profile	Suborder	No. profile	Great group	No. profile	Subgroup	No. profile
Entisols	36	Orthents	36	Xerorthents	36	Typic Xerorthents	13
						Lithic Xerorthents	23
Inceptisols	70	Aquepts	3	Endoaquepts	3	Typic Endoaquepts	3
		Xerepts	67	Calcixerepts	45	Petrocalcic Calcixerepts	5
						Typic Calcixerepts	33
						Aquic Calcixerepts	7
				Haploxerepts	22	Aquic Haploxerepts	7
						Typic Haploxerepts	15
Mollisols	19	Aquolls	9	Endoaquolls	9	Typic Endoaquolls	9
		Xerolls	10	Haploxerolls	8	Aquic Haploxerolls	4
						Typic Haploxerolls	4
				Calcixerolls	2	Typic Calcixerolls	2

Based on the field work, observations and laboratory analyses, mollic and ochric epipedons were identified. Regarding the abundance of calcium carbonate in the parent material (see Table 2-4), some soil surface horizons have high pH and the required cation exchange capacity (CEC) and Base Saturation and soil organic matter content to qualify for mollic horizon according to U.S Soil Taxonomy (Soil Survey Staff, 2014).

Within the study area, different subsurface diagnostic horizons were identified: calcic horizon (Bk), cambic horizon (Bw), petrocalcic horizon (Bkm) and also subsoil redoximorphic features. Regarding the geology map of the study area, soils formed primarily from weathered geological formations such as limestone, conglomerates with marl, Quaternary deposits, shale, alluvium and colluvial deposits. Moreover, all rock outcrops observed in the field were limestone (Borujen Geology Map, 1990).

The eroded hill land (Hi1) and rock outcrop (Mo1) landforms are distinctly recognizable in the study area with shallow soils and in some part with limestone rock outcrops. Most of the soils in those landforms are Lithic Xerorthents (Figure 2-3). Developed hill land (Hi2) is situated in the eastern part of investigated area, which is continuous hills with high topography and deeper soil that is categorized as Typic Calcixerepts, with developed calcic (Bk) and cambic (Bw) diagnostic horizons. Among all landforms in the piedmont, the alluvial fan (Pi1) is dominant. It is located in an agricultural region with low topography and slope. Apart from the development of calcic and cambic horizons in these landforms, agriculture practices, and stable topography led to accumulation and cementation of calcium carbonate partly and formation of Petrocalcic horizon (Bkm). Typic Calcixerepts, Aquic Haploxerepts, Typic Haploxerepts and Aquic Calcixerepts are four main suborders among eight suborders observed in piedmont landform (Figure 2-3).

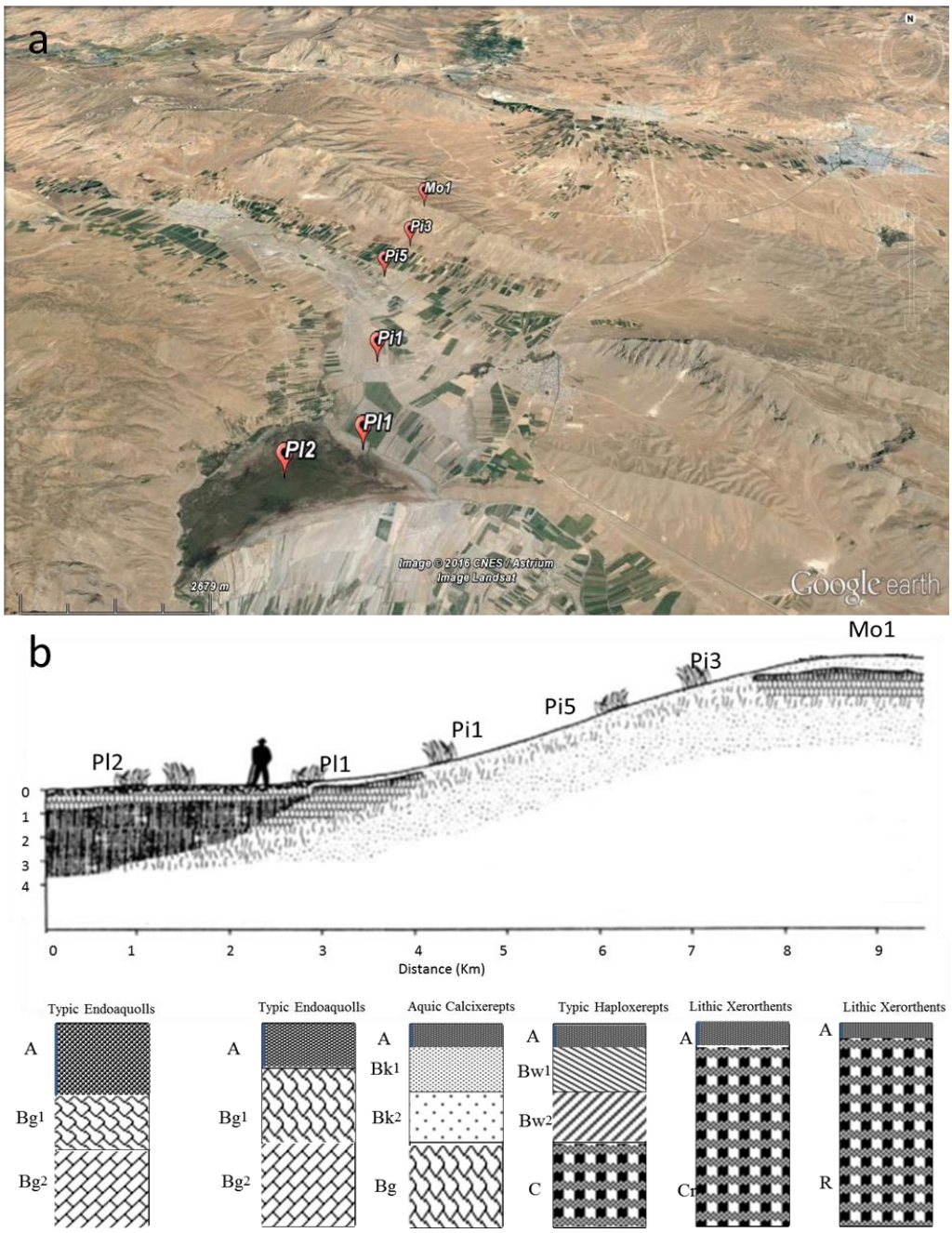


Figure 2-3. a) 3D image of the study area with the main landforms in a sequence of soil evolution, and b) a schematic transect of soil on the landforms with observed profiles.

The gently undulating but dissected alluvial fan (Pi2) has formed during the Quaternary and is one of the youngest landforms in the study area. Because of the enrichment of calcium carbonate in parent material, most of the soils have calcic horizons. The differences between the landforms of pediment (Pi3) and piedmont plain (Pi5) are mainly related to their topography, where Pi3 commonly occurred at higher slope and elevation positions (Figures 2-2 and 2-3). The river plain landform (Pi4) is located in the north-west of study area affected

by underground water. It contains gley properties (Bg) and calcic horizons (Aquic Calcixerepts and Typic Endoaquepts). The lagoon landform (Pl2, Figures 2-2 and 2-3) has been created by overland flow and underground drainage. Soils located in wetlands and lagoon landforms (Pl1 and Pl2) mainly included Typic Endoaquolls and Aquic Haploxerolls. These receive water under closed basin conditions, and have fine textured materials and exhibit waterlogging conditions. The landform of the lagoon (Pl2) is located in a geological depression (graben) which was created after the uplifting of Zagros Mountains (Alpine Orogeny).

The Zagros range spans the whole length of the western and southwestern Iranian plateau (Kelishadi et al., 2014; Mehnatkesh et al., 2013) with a total length of 1,500 km. The Tertiary to Holocene events included strong uplift processes, magmatism and volcanism, erosional processes, and the associated deposition of extensive alluvial fans from the uplifted mountains that occurred in the Zagros zone (Alsharhan et al., 2001). As a result of uplift and folding, the Tethys Sea, which once covered the entirety of Iran, was closed. With frequent uplifting, a large number of shallow water bodies were created (Hojati and Khademi, 2011). These tectonic and geological processes could be the main reason for the existence of lagoon landforms in the region (Figure 2-3 a and b).

Accumulation and cementation by secondary carbonates are the main discriminating properties within stable landforms like alluvial fan (Pi1), dissected alluvial fan (Pi2) and piedmont plain (Pi5). Hill land and mountain landscapes are pedogenetically undeveloped and poorly differentiated due to the dominance of physical weathering process. Soils within the river plain landform (Pi4) and the lowland landscape mostly affected by water logging, have redoximorphic features and organic carbon accumulation. Generally, in the arid and semi-arid regions, soil classes and properties are determined by parent materials and landscape position (Jafari et al., 2013; Sağlam and Dengiz, 2015; Toomanian et al., 2006). In

general, landform delineations could successfully identify the homogeneous soils that are controlled by similar pedogenesis process.

2.3.2 Pedodiversity indices and landform evolution

The pedodiversity indices were used to investigate soil homogeneity and heterogeneity in Soil Taxonomy hierarchic categories within landform units. The pedodiversity indices at different taxonomic levels of the entire study area are presented in Table 2-6. As can be seen, diversity indices increase from the soil order to the soil subgroup, but their change and increase at soil subgroup level are sudden. It could be because of the simultaneous increase of the richness and evenness through soil suborder level (Jafari et al., 2013; Toomanian et al., 2006). Also, it is concluded that the higher the detail in Soil Taxonomy level, the higher the pedodiversity values.

The Margalef's and the Menhinick's indices at all soil taxonomic levels showed the same tendency with an increase from the soil order to the soil subgroup (Table 2-6). Many researchers have shown that how the richness and Shannon indices increased through taxonomic hierarchical organizations in small and large scales (Guo et al., 2003; Ibañez et al., 1998; Jafari et al., 2013; Saldaña and Ibáñez, 2004; Toomanian et al., 2006).

The diversity indices of landforms at all Soil Taxonomy levels are presented in Table 2-7. Pedodiversity indices increased in most of the landforms from the soil order to the soil subgroup level. The general trend of increase in the diversity indices in the Hi1 and Mo1landform units are not similar to that observed in other units from the order to the subgroup. The low diversity in the hill land and mountain landscapes was expected due to a limitation in the soil pedogenesis process. These results are in accordance with findings of other researchers such as Behrens et al. (2009) and Jafari et al. (2013).

Table 2- 6. Diversity indices based on U.S. Soil Taxonomic hierarchy.

Level of study	N	S	H'	Hmax	E	Dmg	Dmn	D
Order	125	3	0.97	1.09	0.88	0.41	0.27	0.13
Suborder	125	5	1.17	1.61	0.73	0.83	0.45	0.43
Great group	125	7	1.55	1.94	0.80	1.24	0.63	0.39
Subgroup	125	12	2.17	2.48	0.87	2.28	1.07	0.32

N: number of observations; S: richness index; H': Shannon index; Hmax: maximum Shannon index; E: evenness; Dmg: Margalef's index; Dmn: Menhinick's index; D: O'Neill dominant index

At the lower Soil Taxonomy categories, the diversity values might increase due to variability in parent materials and soil characteristics as reported by Saldaña and Ibáñez (2004) and Jafari et al. (2013) for soil family category. Diversity indices of some landforms were constant or low increasing from soil order to subgroup (Table 2-7), presumably because of low dominance of few pedogenesis processes which occurred within the landform and subsequently led to the formation of more homogeneous soils within the landforms.

Table 2- 7. Pedodiversity of landform units based on U.S. Soil Taxonomic hierarchy.

		Order							Suborder						
Landforms	N	S	H'	Hmax	E	Dmg	Dmn	D	S	H'	Hmax	E	Dmg	Dmn	D
Hi1	5	2	0.67	0.69	0.97	0.62	0.89	0.02	2	0.67	0.69	0.97	0.62	0.89	0.02
Hi2	6	2	0.69	0.69	1.00	0.56	0.82	0.00	2	0.69	0.69	1.00	0.56	0.82	0.00
Mo1	12	2	0.29	0.69	0.41	0.40	0.58	0.41	2	0.29	0.69	0.41	0.40	0.58	0.41
Pi1	47	3	0.93	1.10	0.85	0.52	0.44	0.17	3	0.93	1.10	0.85	0.52	0.44	0.17
Pi2	12	2	0.45	0.69	0.65	0.40	0.58	0.24	2	0.45	0.69	0.65	0.40	0.58	0.24
Pi3	14	2	0.63	0.69	0.91	0.38	0.53	0.06	2	0.69	0.69	1.00	0.38	0.53	0.00
Pi4	8	1	0.00	0.00	-	0.00	0.35	0.00	2	0.66	0.69	0.95	0.48	0.71	0.03
Pi5	10	2	0.33	0.69	0.47	0.43	0.63	0.37	2	0.33	0.69	0.47	0.43	0.63	0.37
Pl1	5	2	0.50	0.69	0.72	0.62	0.89	0.19	3	0.95	1.10	0.86	1.24	1.34	0.15
Pl2	6	1	0.00	0.00	-	0.00	0.41	0.00	1	0.00	0.00	-	0.00	0.41	0.00
Landforms				Great group								Subgroup			
Hi1	5	2	0.67	0.69	0.97	0.62	0.89	0.02	3	1.05	1.10	0.96	1.24	1.34	0.04
Hi2	6	3	0.71	1.10	0.65	1.12	1.22	0.39	4	1.33	1.39	0.96	1.67	1.63	0.06
Mo1	12	2	0.29	0.69	0.41	0.40	0.58	0.41	2	0.29	0.69	0.41	0.40	0.58	0.41
Pi1	47	4	1.19	1.39	0.86	0.78	0.58	0.20	8	1.84	2.08	0.88	1.82	1.17	0.24
Pi2	12	3	0.87	1.10	0.79	0.80	0.87	0.23	5	1.42	1.61	0.88	1.61	1.44	0.19
Pi3	14	3	0.99	1.10	0.90	0.76	0.80	0.11	4	1.20	1.39	0.86	1.14	1.07	0.19
Pi4	8	3	0.97	1.10	0.89	0.96	1.06	0.12	3	0.97	1.10	0.89	0.96	1.06	0.12
Pi5	10	3	0.90	1.10	0.82	0.87	0.95	0.20	4	1.17	1.39	0.84	1.30	1.26	0.22
Pl1	5	3	0.95	1.10	0.86	1.24	1.34	0.15	3	0.95	1.10	0.86	1.24	1.34	0.15
Pl2	6	1	0.00	0.00	-	0.00	0.41	0.00	1	0.00	0.00	-	0.00	0.41	0.00

N: number of observations; S: richness index; H': Shannon index; Hmax: maximum Shannon index; E: evenness; Dmg: Margalef's index; Dmn: Menhinick's index; D: O'Neill dominant index. Landforms defined in Table 2-1.

Among the studied landforms, soils of the alluvial fan (Pi1) were highly diverse. Soil pedogenesis processes strongly defined the values of the pedodiversity. In alluvial fan landform (Pi1) as compared to other landforms, some soil formation factors such as calcification, decalcification and leaching had higher impacts on the divergence of soils. Furthermore, the high diversity could be explained by the number of observations and the size of the area of study. Firstly, the number of observed soil profiles (*N*) has a significance

effect on diversity and, additionally, Minasny et al. (2010) showed that H′ is closely related to the number of soil classes. When the number of different soil classes or richness increases, a greater number of fractions are summed in H′. Secondly, the alluvial fan landform is distributed in all subareas in the study area (Figure 2-2), therefore, when an area is large enough, divergent pedogenesis processes (e.g. calcification and decalcification) lead to diverse soils. An increase in soil heterogeneity verifies the existence of divergent soil evolution (Saldaña and Ibáñez, 2004).

Soils within the lagoon landform (Pl2) were homogeneous at all Soil Taxonomy levels (Table 2-6). In the local-scale and homogeneous landform, convergent soil pedogenesis processes are dominant and lead to similar soils during the time. Number of soil classes also increased from the soil order to the suborder level, which means soil richness increased through hierarchy Soil Taxonomy (Table 2-6 and 2-7). Rannik et al. (2016) stated that the large number of soil species (*S*) also expressed high heterogeneity.

The O'Neill dominant index (D) showed the deviation of the Shannon diversity indices (*H'*) from the maximum diversity (*Hmax*) therefore when the difference of deductions is null, maximum diversity would happen (Table 2-7).

When the number of species (*S*) are equal in different landforms, the evenness index (*E*) could show how diversity is (Table 2-7). Ibañez et al. (1995) showed that when the evenness of objects is equally probable, the diversity is highest when the richness of comparing units is the same.

Low pedodiversity indices were reported in the several studies for mountainous landscapes (e.g. (Behrens et al., 2009; Jafari et al., 2013)). In our study, mountainous landscape (rock outcrop landform, Mo1) covered approximately 34% of the study area, therefore, because of low diversity values, richness–area and Shannon index-area relationships plotted with excluding of mountain diversity values. Richness–area relationships were plotted to

recognize the close-fitting model (power or logarithmic) in the study area. Figure 2-4 illustrates the logarithmic and power functions of richness-area relationship. The results indicated that both the logarithmic function (Figure 2-4a) and the power function (Figure 2-5b) had a good fit in the study area. Both figures show an increasing trend with increasing area. It is concluded that more additional soil types can be expected in the study area, as similar studies elsewhere indicate: Saldaña and Ibáñez (2004) stated that a larger area should be sampled to capture the spatial variability of the area. The positive relations between richness and area are in agreement with the results of Ibáñez et al. (2005) in the Aegean Islands in Greece, the study of Saldaña and Ibáñez (2004) on fluvial terraces in central Spain and Toomanian et al. (2006) on Zayandeh-rud Valley, central Iran.

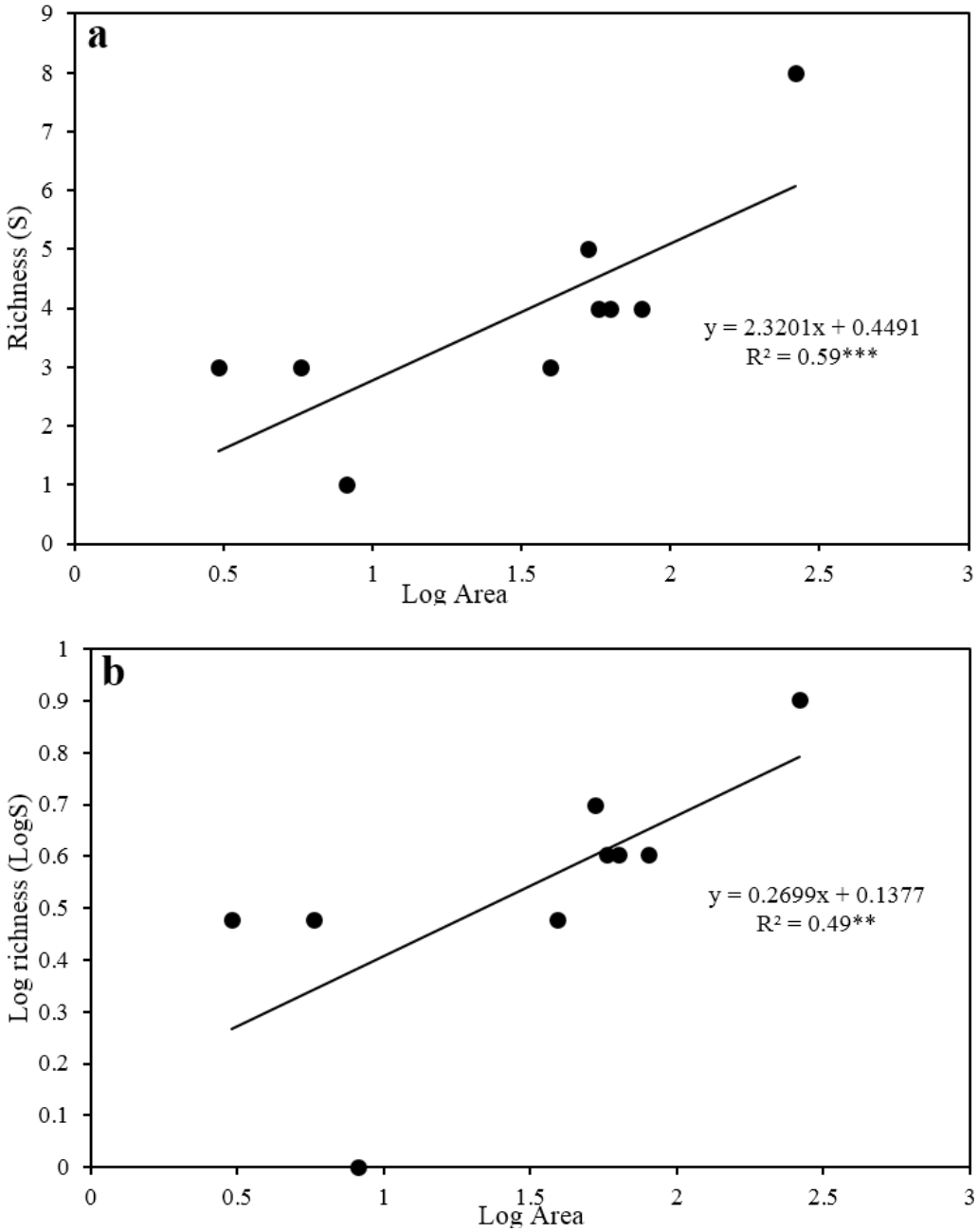


Figure 2-4. a) Richness–area relationships; logarithmic functions and b) a power functions. ***statistically significant (α=0.01), **statistically significant (α=0.05).

The complexity of environment could be monitored and measured by Shannon index-area relationship (Guo et al., 2003; Ibañez et al., 1998; Ibañez et al., 1995; Saldaña and Ibáñez, 2004; Toomanian et al., 2006). The relationship between Shannon index and the area of landforms showed a positive relationship (Figure 2-5a) which was in agreement with the results of previous investigations by Guo et al. (2003), Ibañez et al. (1998) and Toomanian et al. (2006). Shannon diversity index showed a linear positive relationship with the number of

profiles in landforms (Figure 2-5b). It is ascertained that higher pedodiversity and variability in the survey area could be achieved if the number of observations or the sample density increases.

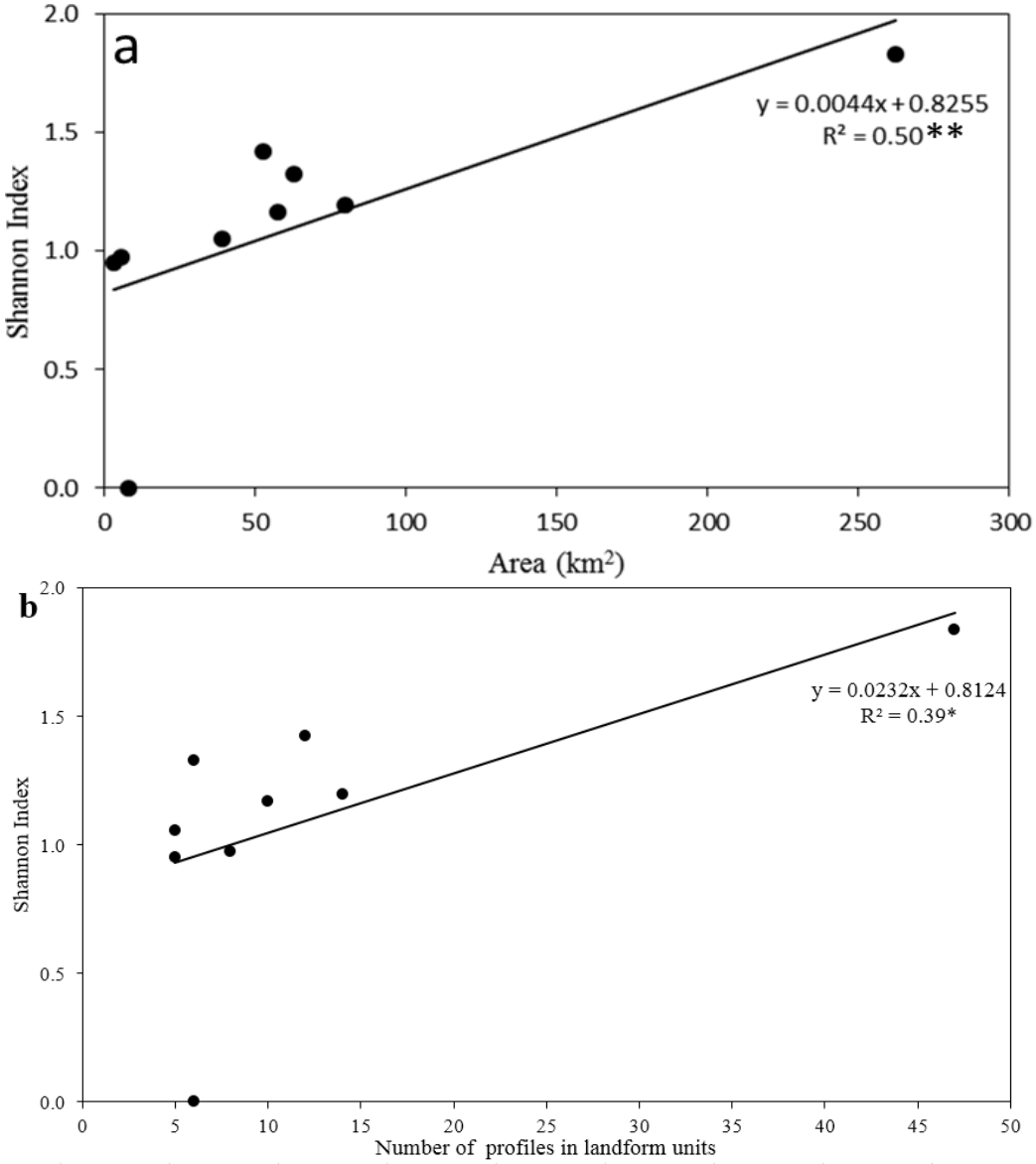


Figure 2-5. a) Shannon index-area relationship and b) number of sampling in landform units. **statistically significant (α=0.05), *statistically significant (α=0.10)

2.4 Conclusions

Soil-landform evolution and pedodiversity-landform analysis as an easy and well-known approach to breaking down the complexity of soils throughout landscape were evaluated in this study for a semiarid region of Iran. It was confirmed that landform delineation generates

more homogeneous units of soil properties and soil types. Furthermore, within landform units, soil evolution is better interpretable and comparable regarding the pedogenesis process and soil development. Results revealed that pedogenesis processes lead to form uniform soil types and properties at the local-scale. On the other hand, within large-scale or extensive landform units, divergent processes tended to create distinct soil types and properties, and consequently, the differences of soils between landform units were considerable.

It has already been shown that pedodiversity indices could be used to separate heterogeneous/homogeneous soils at different levels. Diversity indices decreased from the soil orders to the soil subgroup. The logarithmic and power functions fitted well and satisfactory for richness–area relationships. Pedodiversity indices-area relationships showed that there are additional soil classes if the numbers of observations or the sampling density increase. It seems probable that the number and the intensity of observations were insufficient in the study area. Further investigations are suggested to be done for a more profound understanding of relationships between soil and landscape evolution in the semi-arid regions of Iran.

Chapter 3

Comparing the efficiency of digital and conventional soil mapping to predict soil types in a semi-arid region in Iran

Based on:

M. Zeraatpisheh, S. Ayoubi, A. Jafari, P. Finke. 2017. Comparing the efficiency of digital and conventional soil mapping to predict soil types in a semi-arid region in Iran. Geomorphology, 285: 186–204, DOI: 10.1016/j.geomorph.2017.02.015.

Abstract

The efficiency of different digital and conventional soil mapping approaches to produce categorical maps of soil types is determined by cost, sample size, accuracy and the selected taxonomic level. The efficiency of digital and conventional soil mapping approaches were examined in the semi-arid region of Borujen, central Iran. This chapter aimed to (i) compare two digital soil mapping approaches including Multinomial logistic regression and random forest, with the conventional soil mapping approach at four soil taxonomic levels (order, suborder, great group and subgroup levels), (ii) validate the predicted soil maps by the same validation data set to determine the best method for producing the soil maps, and (iii) select the best soil taxonomic level by different approaches at three sample sizes (100, 80, and 60 point observations), in two scenarios with and without a geomorphology map as a spatial covariate. In most predicted maps, using both digital soil mapping approaches, the best results were obtained using the combination of terrain attributes and the geomorphology map, although differences between the scenarios with and without the geomorphology map were not significant. Employing the geomorphology map increased map purity and the Kappa index, and led to a decrease in the 'noisiness' of soil maps. Multinomial logistic regression had better performance at higher taxonomic levels (order and suborder levels); however, random forest showed better performance at lower taxonomic levels (great group and

subgroup levels). Multinomial logistic regression was less sensitive than random forest to a decrease in the number of training observations. The conventional soil mapping method produced a map with larger minimum polygon size because of traditional cartographic criteria used to make the geological map 1:100,000 (on which the conventional soil mapping map was largely based). Likewise, conventional soil mapping map had also a larger average polygon size that resulted in a lower level of detail. Multinomial logistic regression at the order level (map purity of 0.80), random forest at the suborder (map purity of 0.72) and great group level (map purity of 0.60), and conventional soil mapping at the subgroup level (map purity of 0.48) produced the most accurate maps in the study area. The Multinomial logistic regression method was identified as the most effective approach based on a combined index of map purity, map information content, and map production cost. The combined index also showed that smaller sample size led to a preference for the order level, while a larger sample size led to a preference for the great group level.

3.1 Introduction

Soil information of good quality and high spatial resolution is essential for adequate support of land use management practices, precision agriculture, and ecosystem research. In spite of more than 50 yr of soil survey history in the world, in Iran there are just few maps at scales appropriate for land use planning and agricultural practices. As an example, the conventional soil map of Iran (1:1,000,000) recently was prepared by the Soil and Water Research Institute of Iran (Banaei et al., 2005; Mohammad, 2000) based on landform delineations of the main physiographic regions, is not sufficiently informative (Hengl et al., 2007). Detailed maps supporting many applications, however, exist in some countries with soil maps at spatial resolutions of 100 m (The Netherlands; De Vries et al. 2003; Kempen et al. 2015), 10 m (one-third of Germany; Lösel, 2003), 100-400 m (Germany; McBratney et al. 2003) and 200-500 m (France; King et al. 1999). Therefore, it is necessary for Iranian soil

scientists and decision makers to produce soil maps at finer scales that provide more detailed information.

Conventional methods of soil mapping are currently considered to be ineffective to produce detailed soil maps at a reasonable cost and time (Kempen et al., 2012). Digital soil mapping (DSM) is a powerful technique which is increasingly applied by soil scientists and environmentalists to map soil types and/or properties using ancillary data (Lagacherie et al., 2006; McBratney et al., 2003). These ancillary data, termed environmental covariates, can be obtained from digital elevation models (DEM), satellite imagery (remote sensing data), maps of geology and geomorphology, and legacy soil maps (categorical maps) (Krasilnikov et al., 2011).

The basis of DSM is the application of pedometric techniques that predict the spatial distribution of soil types and soil properties (Wulf et al., 2015). Here, we focus on making maps of soil types because these have been mapped at incomplete coverage until now and the desire exists to finalize soil mapping in the most economically feasible way. Recently, several novel models have been developed to produce soil type maps from profile observations by utilizing auxiliary data (Brungard et al., 2015; Heung et al., 2014; Nelson and Odeh, 2009). Many such methods have been investigated for digital soil mapping of soil types, including Random Forest (RF) (Brungard et al., 2015; Reza Pahlavan Rad et al., 2014), Multinomial Logistic Regression (MLR) (Abdel-Kader, 2011; Brungard et al., 2015; Jafari et al., 2012; Kempen et al., 2012), Artificial Neural Networks(Brungard et al., 2015; Jafari et al., 2013), Support Vector Machine (Kovačević et al., 2010), Nero-Fuzzy approach (Viloria et al., 2016), and Genetic Algorithms (Nelson and Odeh, 2009).

DSM models are divided into simple, intermediate, and complex models (Brungard et al., 2015) based on their interpretability and the number of parameters required. In the present study, two DSM models including RF (a complex model), MLR (a simple model), and the

conventional soil mapping method were compared for predicting soil types. RF and MLR compared favourably to other methods in earlier studies in Iran (Jafari et al., 2013; Reza Pahlavan Rad et al., 2014).

RF can be regarded as an ensemble of classification and regression trees (CART) which are aggregated to provide the final prediction (Breiman, 2001; Breiman et al., 1984; Cutler et al., 2007). RF has several advantages over other statistical modelling approaches (Breiman, 2001; Liaw and Wiener, 2002). Its input and output variables can be both continuous and categorical (Grimm et al., 2008). Moreover, RF has the advantage of incorporating 'randomness' into its predictions through reiterative bootstrap sampling and randomized variable selection when generating each decision tree (Heung et al., 2014). The RF algorithm is considered a powerful modelling technique for predicting soil types because (i) it is quite robust to noise in predictors, (ii) it shows no over-fitting, (iii) it produces predictions with low bias and low variance, and (iv) since it is also fairly fast, it does not require the pre-selection of variables (Díaz-Uriarte and De Andres, 2006; Prasad et al., 2006; Wiesmeier et al., 2011). RF also identifies the most important covariates (Archer and Kimes, 2008; Hua et al., 2005).

Abdel-Kader (2011) reported that the MLR model is the most frequently used statistical model for spatial prediction of soil types and spatial modelling in land use and ecology studies (Abdel-Kader, 2011; Jafari et al., 2012; Kempen et al., 2012; May et al., 2008; Rhemtulla et al., 2007; Suring et al., 2008. However, in recent years only some studies have used MLR for digital soil mapping (Abdel-Kader, 2011).

DSM and CSM approaches are similar in that they both make use of relationships between soil properties and more readily observable land surface properties (shape, position, and reflectance). Conventional soil maps are limited by the scale of the base map, their inability to represent continuous soil classes and spatial variation (Roecker et al., 2010).

Production of maps using CSM techniques is also labour-intensive and expensive. DSM-based maps suffer less from these limitations, thus DSM is generally assumed to be more efficient than CSM (Kempen et al., 2012).

3.1.1 Objectives

Although several papers have been published on the benefits of DSM compared to CSM in recent years, few examples exist that compare DSM techniques with CSM approaches for predicting soil types in the same area, especially in arid and semi-arid regions. The objective of this chapter, therefore, was to compare two different DSM techniques (MLR and RF) with a conventional soil survey for producing soil maps at different taxonomic levels in a semiarid region of Iran. This comparison assessed not just map accuracy, but also information content and production cost, with the purpose of selecting the most efficient method as a function of the taxonomic level of the maps. Because results may depend on sampling density, we evaluated the effect of three different sample sizes on our conclusions.

3.2 Materials and Methods

3.2.1 Description of the study area

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

3.2.2 Soil sampling scheme and profile description

The soil sampling scheme and profile description were explained in chapter 2.

3.2.3 Environmental covariates

Environmental covariates were represented by categorical maps of geomorphology and geology (scale of 1:100,000), by quantitative maps representing topographic attributes, and by remote sensing data. Topography and parent material are the main soil forming factors in arid and semi-arid regions (Florinsky et al., 2002; Mehnatkesh et al., 2013; Tajik et al., 2012). Therefore, to obtain the topographic attributes, we downloaded a DEM with the cell size of

30 × 30 m derived from the Aster GDEM database (Ministry of Economy, Trade and Industry of Japan, National Aeronautics and SpaceAdministration, 2009). The terrain attributes obtained from the DEM included elevation, the topographic wetness index, the SAGA (System for Automated Geoscientific Analysis) wetness index, a multi-resolution of ridge top flatness index, a multi-resolution valley bottom flatness index (Gallant and Dowling, 2003), curvature, profile curvature, plan curvature, aspect, and slope (Table 3.1).

Remote sensing auxiliary variables included the normalized difference vegetation index (NDVI; Boettinger et al., 2008), the ratio vegetation index (Pearson and Miller, 1972), the perpendicular vegetation index (Richardson and Wiegand, 1977), the clay index (Boettinger et al., 2008), and the soil adjusted vegetation index (SAVI; Huete, 1988). These indices were derived from the Landsat Enhanced Thematic Mapper acquired in 2008 (U.S. Geology Survey, 2004). All extracted environmental covariates were used in the Latin hypercube sampling scheme and soil type prediction (Table 3.1). The SAGA GIS was used to derive environmental covariates (Olaya, 2004).

Table 3-1. Environmental covariates used as predictors in the study area.

Environmental covariates	Nature of the soil variable derived	Name of covariate	Definition	Type	Reference/source
Topographic attributes	DEM	El	Elevation (m)	Quantitative	SAGA GIS
		TWI	Topographic Wetness Index		SAGA GIS
		WI	Wetness Index		SAGA GIS
		MrRTF	Multi-resolution of ridge top flatness index		(Gallant and Dowling, 2003)
		MrVBF	Multi-resolution Valley Bottom Flatness Index		(Gallant and Dowling, 2003)
		Cu	Curvature		SAGA GIS
		PrCu	Profile Curvature		SAGA GIS
		PlCu	Plan Curvature		SAGA GIS
		As	Aspect		SAGA GIS
		Sl	Slope angle (%)		SAGA GIS
Remote sensing attributes	Landsat ETM	NDVI	Normalized Difference Vegetation Index	Quantitative	(Boettinger et al., 2008)
		RVI	Ratio Vegetation Index		(Pearson and Miller, 1972)
		PVI	Perpendicular Vegetation Index		(Richardson and Wiegand, 1977)
		CI	Clay Index		(Boettinger et al., 2008)
		SAVI	Soil Adjusted Vegetation Index		(Huete, 1988)
Geomorphology map	Landform	GEM	Hierarchical four level classification (31 geomorphic surfaces)	Categorical	Arc GIS
Geology map	Geologic unit	GEO	Lithological units (12 units)	Categorical	Arc GIS

3.2.4 Geomorphology map

The geomorphological units on the map represent the soil forming factors of parent material as well as relief, and are thus expected to be highly relevant in the complex process of soil survey. In this study, the geomorphic units were prepared based on a nested hierarchical method proposed by (Toomanian et al., 2006). Air photo interpretation was applied to represent the complexity of landscapes using four hierarchical geomorphic levels: landscapes, landforms, lithology, and geomorphic surfaces that have been formed by a unique set of processes in a period of time. The pedological expert knowledge on the landscape and relationship between soils and soil forming factors was employed by stereoscopic delineation from aerial photo pairs. After ortho-photo geo-referencing of stereoscopically interpreted aerial photos, 31 geomorphic surface classes (Table 3-2 and Figure 3-1) were delineated on-screen and imported into an ArcGIS environment.

Table 3-2 .Four hierarchical levels of the geomorphology map in the study area and the associated units.

Landscape	Landform	Lithology (Codes)	Geomorphic surface	Code
Hill	Eroded	Thick bedded conglomerate with marl (Plcb)	Single and low topography hills	Hi111
	Developed hill lands	Dark grey massive limestone (Kt2)	Continuous hills with high topography	Hi211
		Marl, limestone (Klm1)	Continuous hills with high topography	Hi221
Mountain	Rock outcrop	Dark grey massive limestone (Kt2)	Eroded rock surface	Mo111
		Massive dark grey limestone, marl, shale (Kld)	Eroded rock surface	Mo121
		Mixed deposit of limestone, marl and shale (Jklmk)	Eroded rock surface	Mo131
		Marl and rare sandstone (Kshgu)	Eroded rock surface	Mo141
Piedmont	Alluvial Fan	Older terraces and alluvial fans (Qt1)	Active fan, upper section, high slope	Pi111
			Active fan, upper section, low slope	Pi112
		Young terraces and alluvial fans (Qt2)	Active fan, lower section	Pi121
			Active fan, lower section, Cultivated	Pi122
		Silty clay flat (Qt3)	Cultivated plain	Pi131
			Cultivated plain, high slope	Pi132
			Cultivated clay flat	Pi133
		River channel and recent alluvium (Qat)	Dryland farming, high slope	Pi141
	Dissected Alluvial Fan	Older terraces and alluvial fans (Qt1)	Dissected Red Alluvial Fan	Pi211
		Young terraces and alluvial fans (Qt2)	Dissected Red Alluvial Fan	Pi221
		Silty clay flat (Qt3)	Dissected Red Alluvial Fan	Pi231
	Pediment	Dark grey massive limestone (Kt2)	High slope pediment, shallow soil	Pi311
		Young terraces and alluvial fans (Qt2)	High slope pediment, shallow soil	Pi321
		Grey marl and conglomerate, Lake deposit (Qmc)	High slope pediment, shallow soil	Pi331
		River channel and recent alluvium (Qat)	High slope pediment, shallow soil	Pi341
		Older terraces and alluvial fans (Qt1)	High slope pediment, shallow soil	Pi351
			High slope pediment, deeper soil	Pi352
		River channel and recent alluvium (Qat)	High slope pediment, shallow soil	Pi361
		Massive dark grey limestone, marl, shale (Kld)	High slope pediment, shallow soil	Pi371
	River Plain	Silty clay flat (Qt3)	Cultivated clay flat, low drainage	Pi411
	Piedmont Plain	Young terraces and alluvial fans (Qt2)	Cultivated, shallow soil, high slope	Pi511
		Silty clay flat (Qt3)	Cultivated, deep soil, low slope	Pi521
Low land	Wet land	Silty clay flat (Qt3)	Low drainage, wetness	Pl111
	Lake (Lagoon)	Silty clay (Sc)	Seasonal lagoon	Pl211

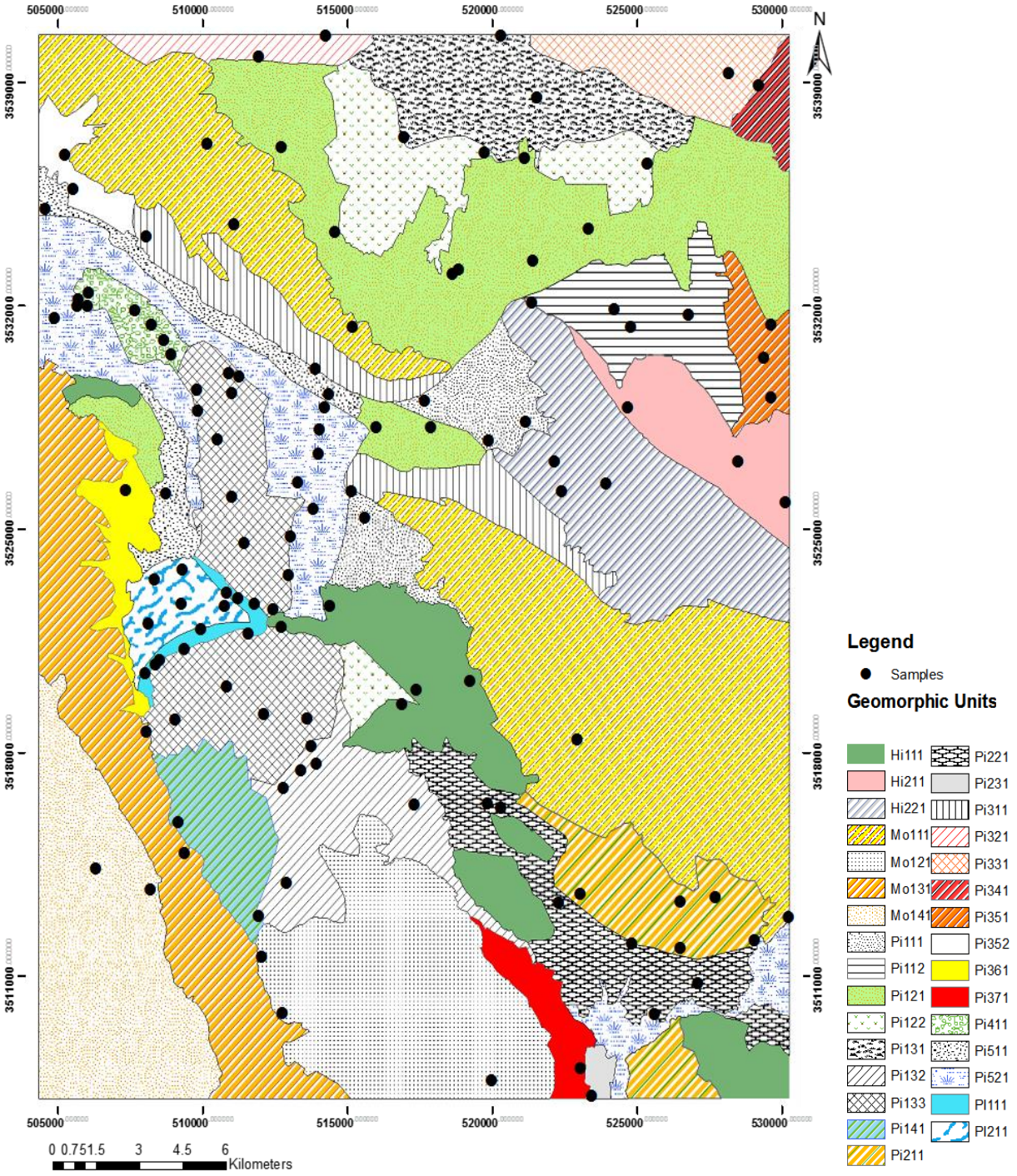


Figure 3-1 Geomorphology map and sampling points in the study area. Geomorphic surfaces in the legend are explained in Table 3-2.

3.2.5 Mapping strategy and scenarios

We used two conditions of prediction, two DSM approaches (MLR and RF), a CSM approach, and three training (point) datasets, resulting in twelve DSM prediction scenarios and one CSM scenario:

Two conditions of prediction: Modelling with and without the geomorphology map as a covariate.

Three point datasets: First, 125 samples were randomized and divided into a training dataset of 80% (100 observations) and a validation dataset of 20% (25 observations). Modelling was done using 100, 80, and 60 samples taken from the training data set. All DSM prediction scenarios were applied at four soil taxonomic levels (order, suborder, great group, and subgroup), and produced maps with pixels of 30 × 30 m.

3.2.6 Digital Soil Mapping Approaches (DSM)

3.2.6.1 Random Forest (RF)

Random Forest (RF) is currently among the most successful methods to predict soil types from numerical and categorical data (Breiman, 2001). The RF method involves two steps. First, each tree (of numerous decision trees, e.g., 1000) of the random forest is constructed independently using bootstrap sampling from the original data. For each tree, a different bootstrap sample is taken from the sample, which is approximately 66% of the available observations, n. In this condition, the bootstrap sampling makes RF less sensitive to over-fitting in comparison with decision trees. Second, those observations (33%) not in the bootstrap sample are referred to as out-of-bag and the proportion of misclassification of these samples (out-of-bag error) is a measure of the precision of the method (Peters et al., 2007). The nodes of each tree are split based on the best environmental covariate chosen from a randomly selected subset of the

input environmental covariates. A second parameter is the number of trees in the forest. The selection of optimal input environmental covariate values for modelling at each taxonomic level was performed by iterating over environmental covariate values from 1 to the total number of covariates (Peters et al., 2007). For each environmental covariate value, the number of trees in the forest increased from 100 to 1000 by increments of 100.

The RF algorithm provides two measures of variable importance: the mean decrease in accuracy and the mean decrease in Gini. The node impurity is measured by the Gini index, the mean decrease in Gini refers to the improvement of the splitting criterion which measures the reduction in class impurity from partitioning the data set (Myles et al., 2004). Mean decrease in accuracy is a permutation based measure of variable importance derived from evaluating the contribution of a variable to the prediction accuracy. The mean decrease in Gini also measures variable importance by permuting the values of each environmental covariate in the out-of-bag sample. Environmental covariates associated with a comparatively large increase in out-of-bag error are more important. A variable that produces high homogeneity in the descendent nodes results in a high mean decrease in Gini (Breiman, 2001).

Modelling at different Soil Taxonomy levels was performed for all scenarios presented in section 3.2.5 using the "randomForest" package (Liaw and Wiener, 2002) in R 3.0.1 (R Development Core Team, 2013). In case there is no an independent validation dataset, out-of-bag error rates may be used for validation (Heung et al., 2014), but in this study 20% of the observed dataset was used for independent validation.

3.2.6.2 Multinomial logistic regression (MLR)

The logistic model belongs to the family of generalized linear models and is used when the response variable is a categorical variable. Multinomial logistic regression (MLR) describes the

relationship between a combination of environmental predictive variables and a binary response variable by means of a link function (Hosmer Jr and Lemeshow, 2004).

MLR was used to model the relationships between the different Soil Taxonomy levels (orders, suborders, great groups and subgroups) as categorical dependent variables and the remote sensing indices, terrain attributes (quantitative predictors), geomorphic units, and geology maps (qualitative predictors) with the same scenarios for RF as explained in section 3.2.5. A MLR model (Hosmer Jr and Lemeshow, 2004) with reference category *J* is expressed as follows:

$$\mathbf{log}\left(\frac{\boldsymbol{\pi_{ij}}}{\boldsymbol{\pi_{iJ}}}\right) = \boldsymbol{\alpha_j} + \boldsymbol{\beta_j x_i} = \mathbf{1}, \mathbf{2}, \ldots \ldots \mathbf{J} - \mathbf{1} \qquad \text{(Eq. 3.1)}$$

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 *y* is multinomial instead of binomial and there are J-1 equations instead of one so that:

$$P(y_i = j) = \pi_{ij} \frac{\exp(\beta_j x_j)}{1+\sum_{j=1}^{j} \exp(\beta_j x_j)} \qquad \text{(Eq. 3.2)}$$

Then, the probability of reference category is given by:

$$P(y_i = 0) = \pi_{ij} = \frac{1}{1+\sum_{j=1}^{j} \exp(\beta_j x_j)} \qquad \text{(Eq. 3.3)}$$

At each classification level of the Soil Taxonomy, the dependent variables have more than two categories. At the beginning of modelling, in the MLR, the reference class must be selected. In this condition, we selected Entisols (order), Aquepts (suborder), Calcixerepts (great group) and Aquic Calcixerepts (subgroup) which are the first in the default alphabetical order, by using "nnet" package in R 3.0.1 (R Development Core Team, 2013). The best MLR model selection at each level was determined by minimizing the Akaike information criterion (Akaike, 1974). In

this chapter, both the backward and forward stepwise Akaike information criterion approaches were used for model selection. The selected model had the fewest independent variables and the lowest Akaike information criterion and then was employed for prediction of soil types. The stepwise Akaike information criterion was done with the "MASS" package in R 3.0.1 (R Development Core Team, 2013).

3.2.7 Conventional soil mapping (CSM) approach

According to Qi and Zhu (2003) and Qi et al. (2008), soil polygons can be delineated using the perceived distribution of landscape units through aerial photo interpretation. The resulting map polygons for a soil type then correspond to the spatial patterns of the landscape units. Following this assumption, a conventional soil map was produced based on a physiographic survey. First, physiographic units were delineated by air photo interpretation and consultation of the geology map. Afterwards during the field survey, boundaries of physiographic units were verified and adjusted. The delineated polygons were then used to produce the conventional soil map at four Soil Taxonomic levels (order, suborder, great group, and subgroup). The CSM was based on 80% of the dataset (100 samples), so that the same validation data set was available as for the DSM-derived soil maps (next section). The CSM was performed by one co-author who was not involved in the construction of the DSM maps to ensure that the DMS and CSM maps were produced independently. We did not repeat CSM for different sampling densities as with DSM because any CSM mapper would, perhaps unknowingly, use experience from the previous mapping during repetitions and it would also involve large additional efforts.

3.2.8 Cost evaluation

One of the aims of this study was to compare the cost efficiency of DSM and CSM for soil survey and mapping. Therefore, we calculated the costs of field work, laboratory analysis, and

the cost of preparing the geomorphology maps and the conventional soil map based on the prices declared and approved by National Cartographic Centre (2015) and Soil and Water Research Institute (2015) of Iran. Because of the local currency and price levels, we avoided presenting the real costs but calculated relative cost (RC) for each soil taxonomic level and scenario (section 3.2.5) as:

$$RC = \frac{(MaxC - RealC)}{(MaxC - MinC)} \quad \text{(Eq. 3.4)}$$

where MaxC is the maximum cost, RealC is the real cost, and MinC is the minimum cost. The cost models for DSM and CSM are different because different decisions were made in both approaches to determine what analyses were necessary.

3.2.9 DSM cost analysis

DSM scenarios comprised DSM with a geomorphology map and DSM without a geomorphology map as covariate. Based on the expected soil types in the research area (Borujeni et al., 2010; Jalalian and Mohmmadi, 1989) for each soil taxonomic level, different laboratory analyses were required to establish diagnostic epipedons, subsurface horizons and other properties for the soil classification (Soil Survey Staff, 2014). We did all the required soil analyses of organic carbon, calcium carbonate, gypsum, soil texture, and soil reaction (pH) for all 125 samples at the order level. After distinguishing the order in the study area, some additional quantitative and laboratory information were needed to classify at the suborder level. At this stage, based on possible suborders, 54 samples needed to be analysed for the exchangeable sodium percentage and the sodium adsorption ratio. For classification at the great group and subgroup levels there was no need for additional quantitative laboratory information,

as qualitative information collected in the field sufficed. The maximum cost (MaxC) was calculated by:

$$MaxC = Gp + ACsg + MoC \quad \text{(Eq. 3.5)}$$

where Gp is cost of producing the geomorphology map, ACsg is the maximum absolute cost for the most expensive level (which is great group or subgroup), and MoC is modelling cost.

3.2.10 CSM cost analysis

This approach was based upon the experts' field knowledge assuming that experts were able to decide on the number of samples and sort of analyses to be conducted. CSM at the order level included soil analyses for organic carbon, calcium carbonate equivalent, and soil texture for all 125 soil samples, along with gypsum content and soil reaction (pH) for 85 of the soil samples. At the suborder level, the situation was the same as suborder in DSM approach (see section 3.2.9). At the great group level, 24 additional soil samples were analysed for gypsum characterization compared to the order level to distinguish Calciaquolls and Calcixerolls within the suborders of Aquolls and Xerolls. To classify at the subgroup level, qualitative information collected in the field sufficed, and no additional quantitative laboratory analyses were required.

3.2.11 Validation strategy and performance indicators

The performance in terms of map quality of each soil map predicted by the three different approaches (MLR, RF, and CSM) was evaluated using the same independent validation dataset, being a random subset of 20% of the 125 sampled field profiles (Table 3). Even though this subset is not strictly a probability sample because it is drawn from a latin hypercube sample combined with a legacy sample, we consider it suitable to evaluate for comparison of the performance of the different approaches. For instance, (Brus et al., 2011) suggested that RF validation with an independent dataset would be more reliable than RF validation with out-of-bag error. For assessing the quality of the predicted soil maps, map purity was used based on the

confusion matrix (Brus et al., 2011). Map purity (MP) is defined as the proportion of the samples or soil types that were correctly predicted over the total number of validation locations:

$$MP = \sum_{u=1}^{U} \frac{A_{uu}}{A} \quad \text{(Eq. 3.6)}$$

where U is the number of classes, A_{uu} is the number of correctly classified observations of map unit u, and A is the total number of observations in the study area (validation dataset). Map purity measures a range between 0 to 1 where a good map has a value of map purity close to 1 (Behrens et al., 2010).

Additionally, Cohen's Kappa coefficient of predicted maps was determined for the two DSM approaches and the CSM approach. Kappa (k) indicates the agreement percentage between two maps, corrected for chance agreement (Fleiss et al., 1969):

$$\boldsymbol{k} = \frac{(\boldsymbol{P_o - P_e})}{(\boldsymbol{1 - P_e})} \quad \text{(Eq. 3. 7)}$$

where P_e is the expected proportion, the hypothetical probability of chance agreement, and P_o is the observed proportion, that is the relative observed agreement between the maps.

The detail depicted on a map informs whether it is useful for subsequent spatial studies. A map with high spatial detail may be useful in spatial analysis if its accuracy is high as well. We derived a proxy for map intricacy using the Shannon entropy index (S) using a moving window. The degree of spatial detail was assessed by mapping the entropy using a 3×3 cell moving window on the soil type map produced by DSM or CSM. A high entropy in a window coincides with a higher map intricacy. The average of the mapped entropies was taken as a proxy for the information detail of the entire map.

The Relative Purity (RP) is calculated as:

$$\boldsymbol{RP} = \frac{(\boldsymbol{P - MinP})}{(\boldsymbol{MaxP - MinP})} \quad \text{(Eq. 3.8)}$$

where *P* is purity, MinP is minimum purity, and *MaxP* is maximum purity. The relative diversity (RD) is:

$$\boldsymbol{RD} = \frac{(\boldsymbol{S}-\boldsymbol{MinS})}{(\boldsymbol{MaxS}-\boldsymbol{MinS})} \quad \text{(Eq. 3.9)}$$

where *S* is the average Shannon entropy index mapped with a 3x3 cell moving window, *MinS* is the minimum Shannon entropy index, and *MaxS* is the maximum Shannon entropy index. In addition to the above performance indicators, we assessed a combined index, which is the product of relative cost, relative diversity for map detail, and relative purity for map quality. By multiplying these three factors, we assume that they equally and independently contribute to the performance of a map. The combined index was calculated for all scenarios, predicted maps, and the conventional map. The schematic illustration of the modelling strategies in this research is presented in Figure 3-2.

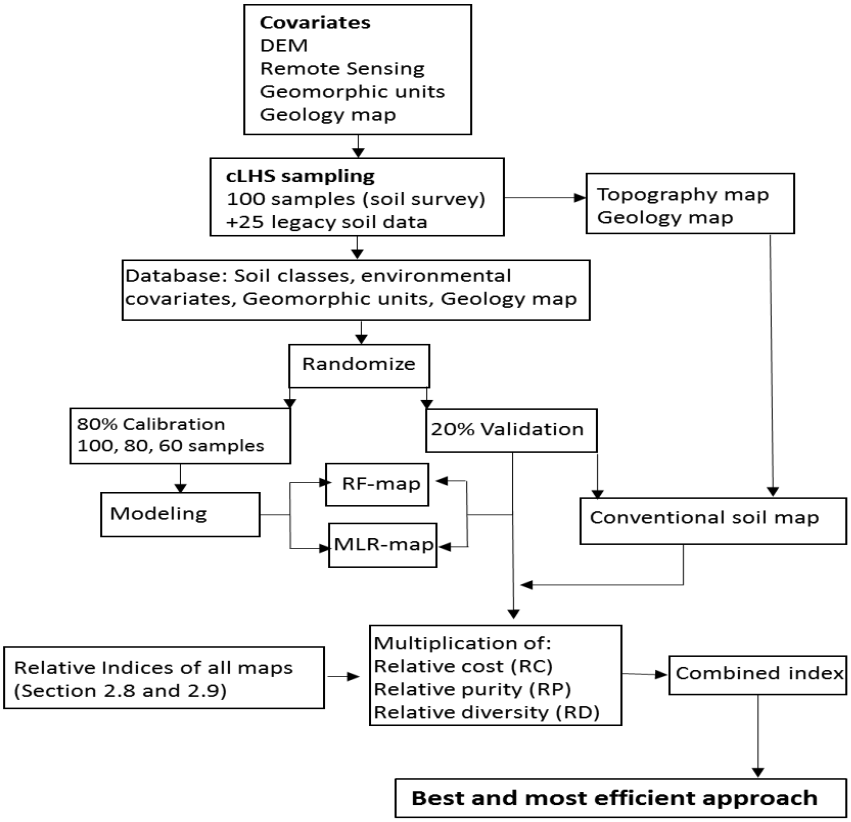


Figure 3-2 Flowchart of procedures, modelling, and validation used for DSM and CSM approaches in this study.

3.3 Results and Discussion

In the study area, 31 geomorphic units were identified (by air photo interpretation, Figure 3-2 and Table 3-1), and Inceptisols, Entisols and Mollisols (Soil Survey Staff, 2014) were the main soil orders observed and described by field sampling. The majority of Mollisols occurred in geomorphic units Pl111 and Pl211 (low land landscape), where the shallow ground water table leads to organic carbon accumulation (melanization) to form a mollic epipedon (Bockheim, 2015). Inceptisols were mostly concentrated in piedmont landscapes (Pi) and all Entisols were located in the mountain (Mo), piedmont (Pi), and hill (Hi) landscapes (Table 3-1, Figure 3-2).

3.3.1 Digital soil mapping using MLR and RF modelling

As shown in Table 3, three orders, five suborders, seven great groups, and twelve subgroup classes were identified in the study area. The results of MLR and RF modelling indicated that scenarios with the geomorphology map as a covariate had in most cases better performance in terms of map quality than the scenarios without geomorphology map (Table 3-3). The spatial patterns of soil types by MLR and RF followed those of the geomorphic units (Figure 3-2). The positive effect of using the geomorphology map on map quality was consistent with that reported in previous studies (Behrens et al., 2005; Jafari et al., 2013; Scull et al., 2005). Generally, in arid and semi-arid regions, soil types and properties are mostly controlled by parent materials and topographic positions (Florinsky et al., 2002; Mehnatkesh et al., 2013; Tajik et al., 2012), which are represented well by the geomorphology map. As expected, map purity and the Kappa index decreased from the order to subgroup levels for both models, except for the suborder level predicted by RF. This is presumably related to the better relationship between environmental covariates and the most observed suborder level (Xerepts, see Table 2-5; chapter 2). From the order to subgroup level, the Shannon index increased as the map intricacy increased towards the

lower taxonomic levels (Table 3-3). This result is in agreement with the findings of Toomanian et al. (2006) and Jafari et al. (2013) who found the highest diversity indices at the subgroup level.

Table 3-3 Map purity, Kappa and Shannon index for all scenarios and based on four Soil Taxonomic levels in the study area

Soil Taxonomy levels	**Multinomial Logistic Regression (MLR)**																	
	Map purity						Kappa						Shannon Index					
	With Geomorphic units			Without Geomorphic units			With Geomorphic units			Without Geomorphic units			With Geomorphic units			Without Geomorphic units		
	100	80	60	100	80	60	100	80	60	100	80	60	100	80	60	100	80	60
Order	0.80	0.80	0.68	0.68	0.64	0.72	0.67	0.67	0.46	0.48	0.41	0.53	0.55	0.55	0.57	0.54	0.54	0.60
Suborder	0.68	0.64	0.56	0.64	0.68	0.64	0.51	0.47	0.35	0.42	0.48	0.46	0.56	0.59	0.58	0.58	0.60	0.60
Great Group	0.52	0.48	0.44	0.56	0.56	0.32	0.36	0.30	0.27	0.40	0.38	0.09	0.61	0.59	0.61	0.63	0.63	0.66
Subgroup	0.40	0.44	0.28	0.32	0.44	0.32	0.31	0.34	0.17	0.19	0.34	0.20	0.61	0.60	0.66	0.65	0.69	0.64
Soil Taxonomy levels	**Random Forest (RF)**																	
	With Geomorphic units			Without Geomorphic units			With Geomorphic units			Without Geomorphic units			With Geomorphic units			Without Geomorphic units		
	100	80	60	100	80	60	100	80	60	100	80	60	100	80	60	100	80	60
Order	0.68	0.64	0.60	0.60	0.60	0.60	0.46	0.39	0.33	0.33	0.34	0.33	0.53	0.53	0.56	0.53	0.53	0.53
Suborder	0.72	0.60	0.56	0.68	0.60	0.48	0.53	0.33	0.27	0.47	0.33	0.14	0.54	0.54	0.54	0.54	0.55	0.55
Great Group	0.60	0.52	0.48	0.52	0.48	0.44	0.43	0.31	0.24	0.31	0.24	0.17	0.60	0.58	0.59	0.59	0.61	0.61
Subgroup	0.44	0.36	0.40	0.44	0.44	0.36	0.33	0.23	0.28	0.32	0.32	0.22	0.65	0.63	0.64	0.66	0.64	0.67

Contrary to our expectation, producing more detailed maps by MLR from the order to subgroup levels did not increase the number of predictors (Table 3-4), showing that the environmental covariates used at the order level to represent soil types are able to explain and capture the complexity of the lower Soil Taxonomic levels as well.

Table 3- 4 Selected covariates by stepwise Akaike information criterion for MLR modelling (N=100).

Soil Taxonomy levels	*Environmental covariates in MLR
Order	SAVI+CI+NDVI+GEO+GEM+TWI+WI+PVI
Suborder	SAVI+CI+NDVI+GEO+GEM+TWI+WI+PVI
Great group	SAVI+CI+NDVI+GEO+GEM+TWI+WI+PVI
Subgroup	CI+SAVI+NDVI+GEO+GEM+TWI+PVI

*Variables abbreviations explained in Table 1.

The importance of the selected variables for MLR is presented in Figure 3-3. The soil adjusted vegetation index (SAVI), clay index, and normalized difference vegetation index (NDVI) were the most important covariates across order, suborder and great group levels (Table 3-4 and Figure 3-3). At the subgroup level, the clay index had the largest contribution. At all levels, SAVI contributed more importance than NDVI which may be attributed to the lack of dense vegetation in this arid and semi-arid region. NVDI, which reflects the influence of the soil on vegetation, was weaker than the SAVI index (Huete, 1988). This was also observed by (Reza Pahlavan Rad et al., 2014) when updating a soil map in northern Iran.

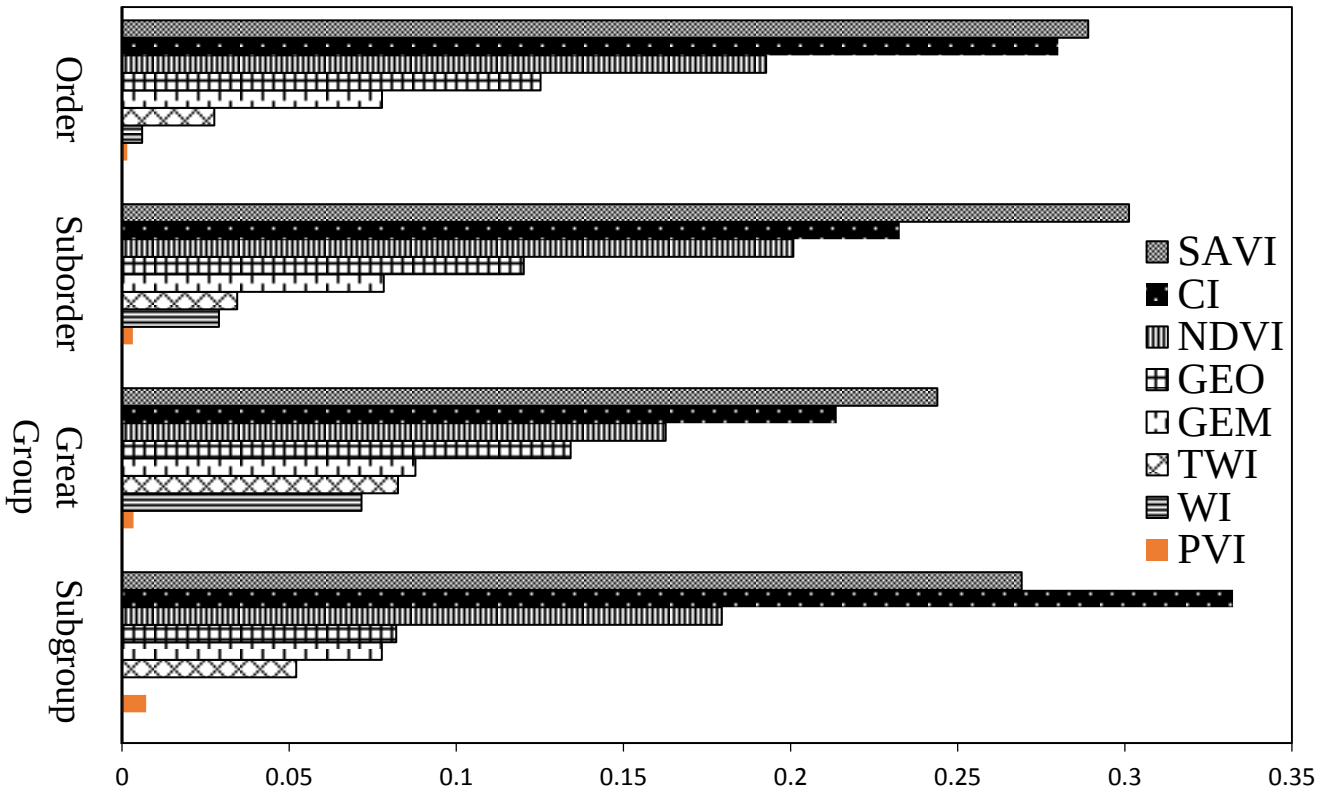


Figure 3-3 Variable importance for the MLR approach at four taxonomic levels.

All remote sensing indices were involved in the soil prediction maps at four taxonomic levels, showing that the spatial distribution of soils has a strong relationship with surface reflectance. Water flow and soil water content had strong impact on the soils formed in the study area, since both the SAGA wetness index and topographic wetness index participated in the prediction model by MLR (Table 3-4). Wang and Laffan (2009) and Whiteway et al. (2004) showed that the topographic wetness index defined flow intensity and potential sediment accumulation, and that the SAGA wetness index also indicated the potential degree of wetness. A higher SAGA wetness index corresponds with higher potential wetness and organic matter accumulation in lowland positions. Debella-Gilo and Etzelmüller (2009) showed that the high probability areas for each great group occurred simultaneously with the known landscapes, therefore, the recognition of soil–landform relationships could provide a powerful tool to aid soil mapping activities (Holliday, 2006).

According to the variable importance measurement, two measures for RF modelling are shown in Figure 3-4; including mean decrease of Gini and mean decrease of accuracy at the order and subgroup levels. Amongst all covariates, at both taxonomic levels (order and subgroup levels), geomorphology and geology maps were identified as the most important variables. In fact, at the order and subgroup levels, the variables that increased the model accuracy the most was the geology map followed by the geomorphology map. In terms of decreasing node impurity and increasing map purity for the predicted maps, geomorphology played the most effective role (Figure 3-4 a and b). The results of the present work suggest that the geomorphology map had the most important impact for both RF and MLR models.

In comparison with the results of Roecker et al. (2010) (51% for subgroup level); Reza Pahlavan Rad et al. (2014) (48.5% for great group level, 51.5% for subgroup levels); and Barthold et al. (2013) (51.61% for Reference Soil Groups (FAO classification)), the accuracy and performance of RF modelling showed reasonable accuracy in spatial prediction. Reza Pahlavan Rad et al. (2014) and Jafari et al. (2013) also observed a reduction in the prediction accuracy while there was an increase in taxonomic detail.

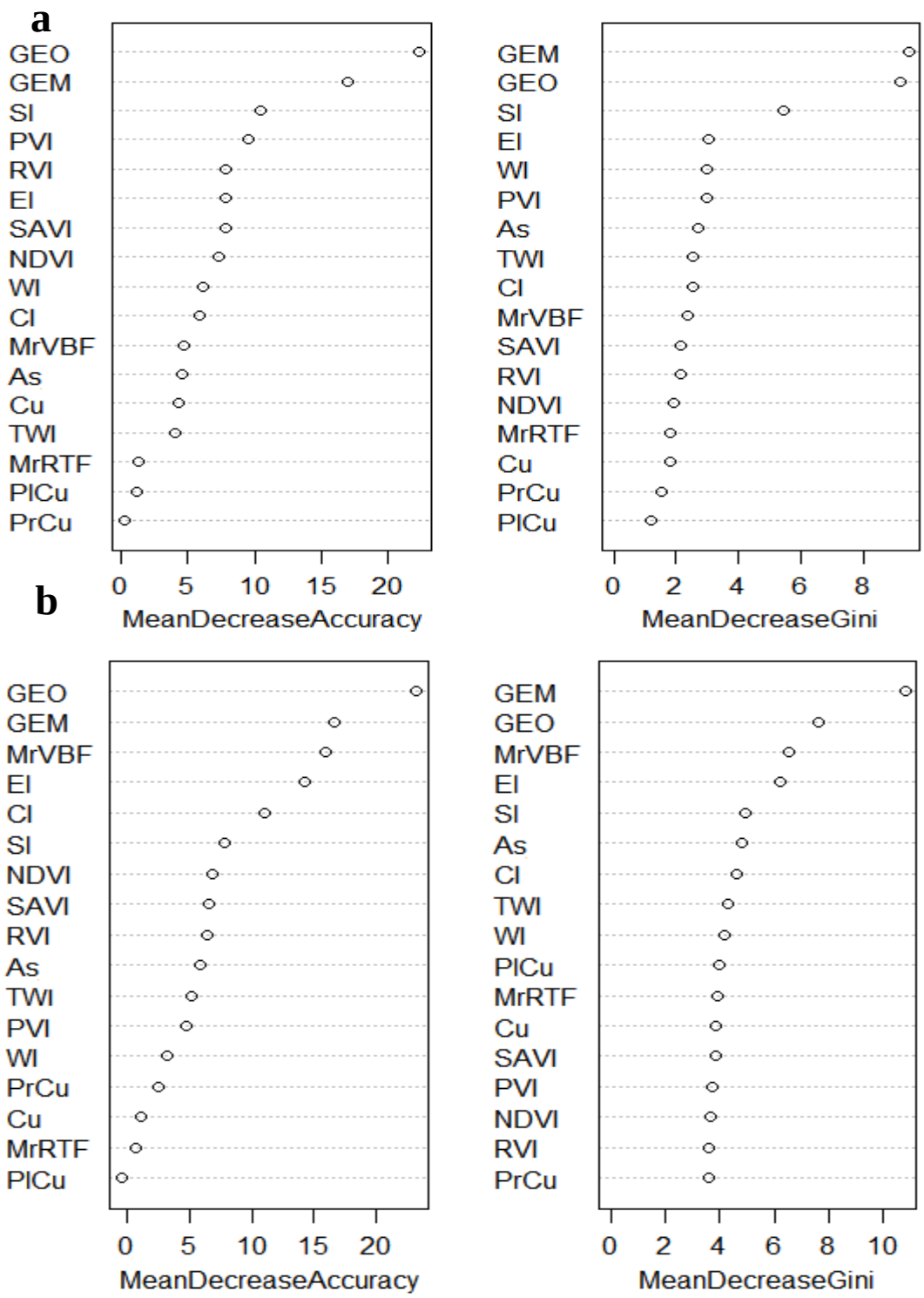


Figure 3-4 Mean decrease of accuracy, mean decrease of Gini, and covariate importance for RF at the (a) order level and (b) subgroup level. Abbreviations of covariates are explained in Table 3-1.

3.3.2 Conventional soil survey

In this study, the CSM was based on air photo interpretation. Soil patterns were mapped using knowledge of physiographic units and homogeneous areas (elevation and geology). By the CSM approach, two orders, five suborders, seven great groups and twelve subgroups were mapped over the ten landforms identified in the study area (Table 3-2). Topography plays a vital role in arid and semi-arid regions on the spatial distribution of soils (Cantón et al., 2003). Based upon the high variability of topography in the study area, it is believed that this is the major factor out of the five soil forming factors proposed by (Jenny, 1941).

The estimated overall map purity of the CSM map varied from 0.48 to 0.72 at different taxonomic levels (Table 3-5). Map purity decreased from the order to subgroup level, but increased in the suborder level, likely because the CSM represented well the dominant suborder (Xerepts) in the validation set. At the order level, due to map scale (cartographic) limitations, it was impossible to delineate Mollisols because this order was not dominant in any map polygon, and they were associated with Inceptisols. Zhu (1997) Such limitations are well known in CSM, as described by Zhu (1997) and Menezes et al. (2014). The Kappa index ranged from 0.29 to 0.55. The lowest Kappa occurred at the order level (0.29, Table 3-5), indicating a high probability of chance classification in spite of good purity (Girard and Girard, 1999). The Shannon index for map intricacy was constant for the CSM approach at all taxonomic levels. This is probably due to the cartographic criteria that were followed in CSM, which prevented the delineation of small polygons (Figure 3-5e). In contrast, DSM methods might show polygons as small as the resolution of the covariate maps.

Table 3- 5 Validation criteria for CSM-approach at four level of Soil Taxonomy.

Soil Taxonomy levels	Map Purity	Kappa	Shannon index
Order	0.60	0.29	0.53

Suborder	0.72	0.55	0.53
Great Group	0.56	0.39	0.53
Subgroup	0.48	0.39	0.54

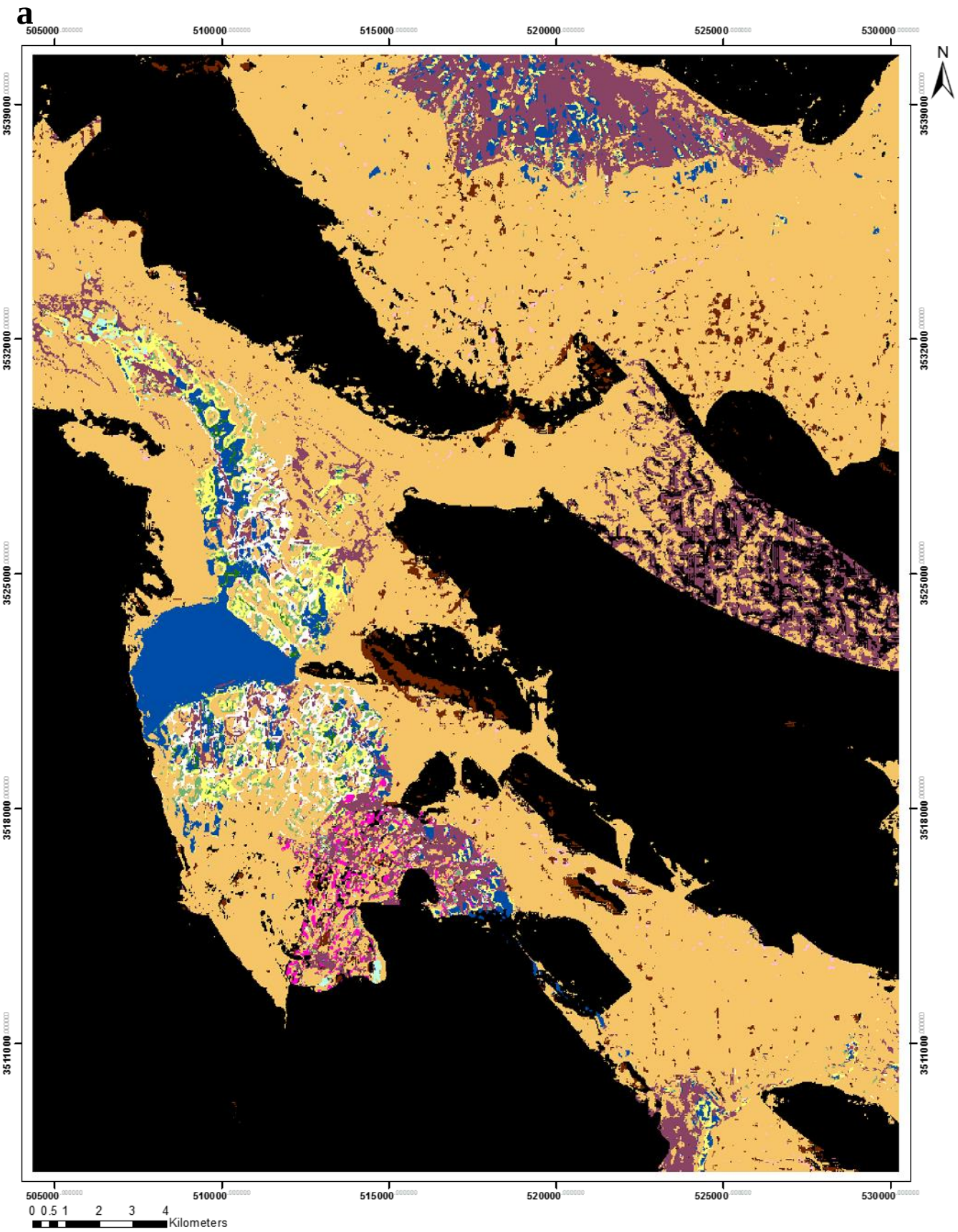
Table 6. Validation criteria for CSM approach at four levels of Soil Taxonomy.

3.3.3 Statistical comparison

Both RF and MLR models were trained with the same number of observations. Comparison of the results of the two mapping approaches (Table 3-3) clearly showed that both methods had higher performance with the geomorphology map as a predictor.

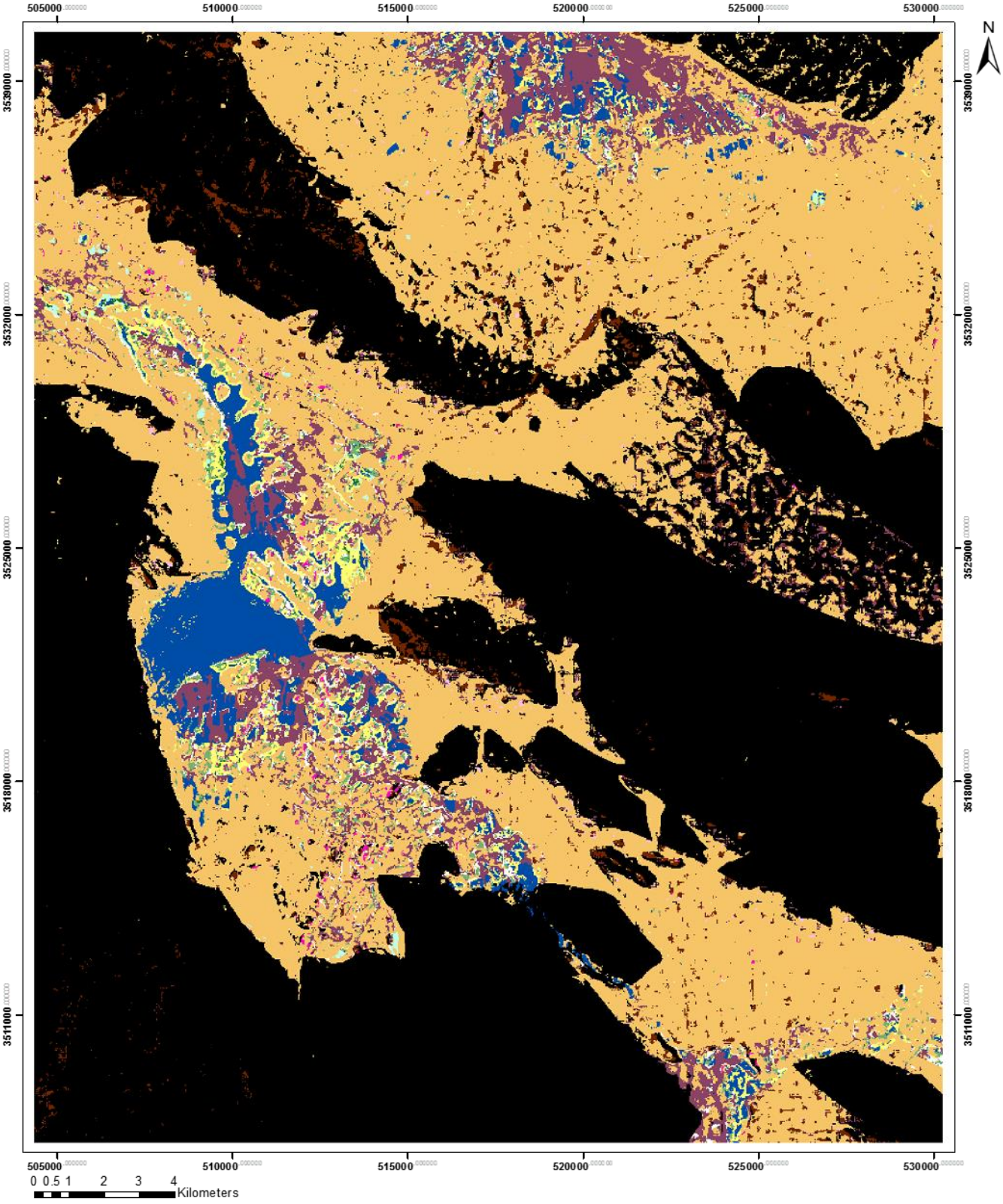
In terms of map purity and Kappa, MLR showed a better performance at the higher taxonomic levels (order and suborder) than RF. On the other hand, RF was more powerful at the lower taxonomic levels (great group and subgroup). MLR was less sensitive (more efficient) than RF when the number of training observations decreased from 100 to 60 locations. The Kappa index for MLR approach was higher than that for RF, which suggested that chance classifications are less likely for MLR methods. Shannon index, in most cases, was lower for RF than for MLR. RF produced soil maps with more homogeneous units (i.e., map units in Figure 3-5 a and b and Table 3-3, lower Shannon index). Kempen et al. (2009) concluded that areas with high heterogeneity showed high entropy and low purity, and their prediction was associated with high uncertainty. As we found the maps with the highest Shannon index (obtained using MLR) also showed the highest purity, this seems a contrasting result. Although the entropy in Kempen et al. (2009) refers to the pixel scale (the probability of occurrence of each taxonomic class at each pixel is used to calculate Shannon index), which makes it a map precision estimator, we are calculating the average Shannon index for the final map using the local entropy of predicted soil types in a moving window as a proxy for map intricacy, which makes it a map information content indicator.

To evaluate the performance of DSM approaches and the CSM approach, we compared the maps based on 100 samples. Figure 3-5 illustrates soil classes at the subgroup level for the DSM approaches (with and without geomorphology map as covariate) and the CSM approach. In both DSM modelling approaches, the soil patterns with and without usage of the geomorphology map were similar. In the scenarios without the geomorphology map, most of the soil patterns were explained by the geology map. When using the geomorphology map as covariate, a slightly higher purity was obtained. These results showed the importance of the parent material, represented in the geology map, on the training of DSM methods to predict soil types. The number of geomorphologic surfaces (31) is fairly large relative to the sample size (100) and is incompletely verified with the validation data (25 samples), but does result in a slightly higher map purity.



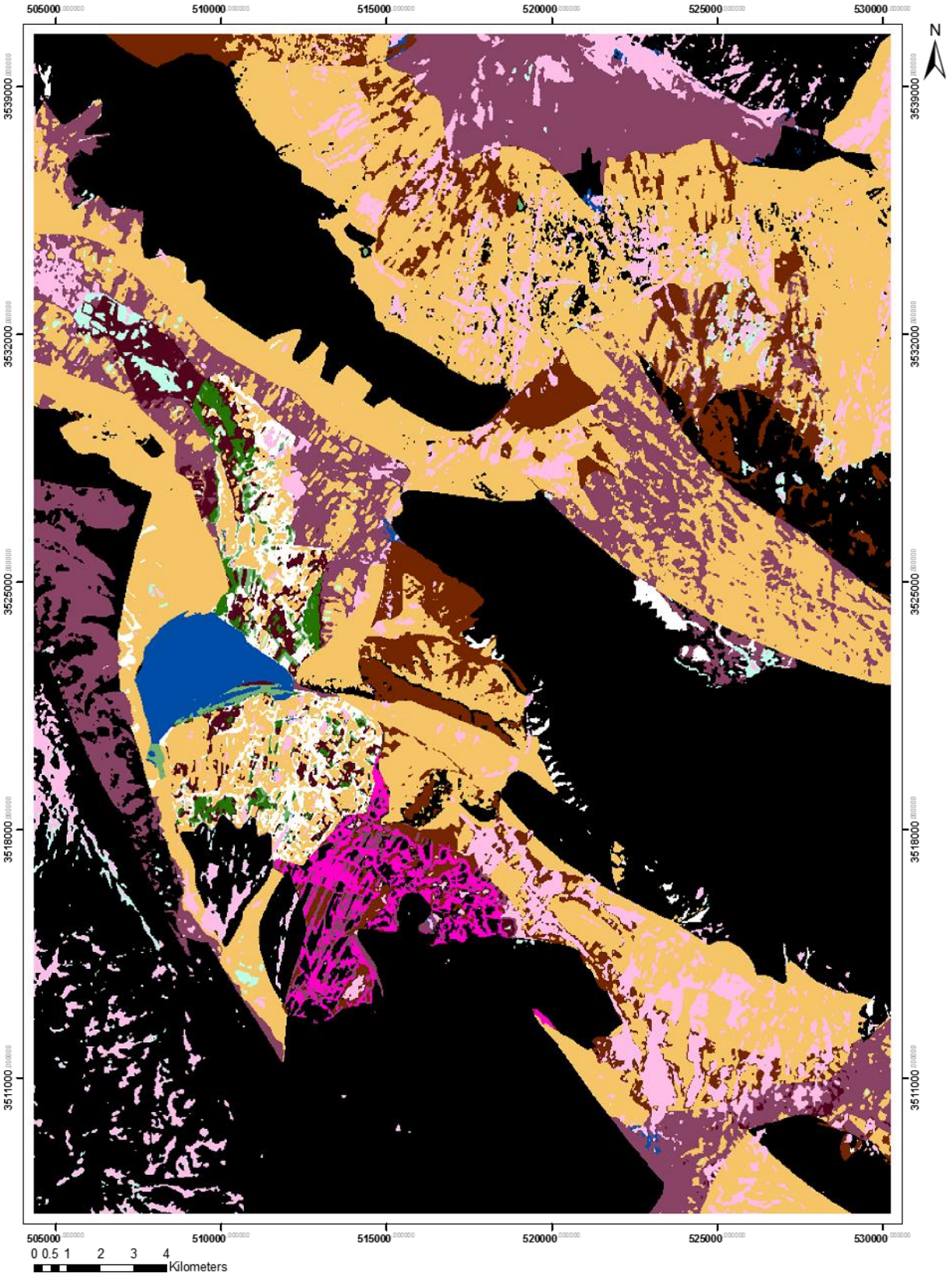
Continue Figure 3-5 Next page

b



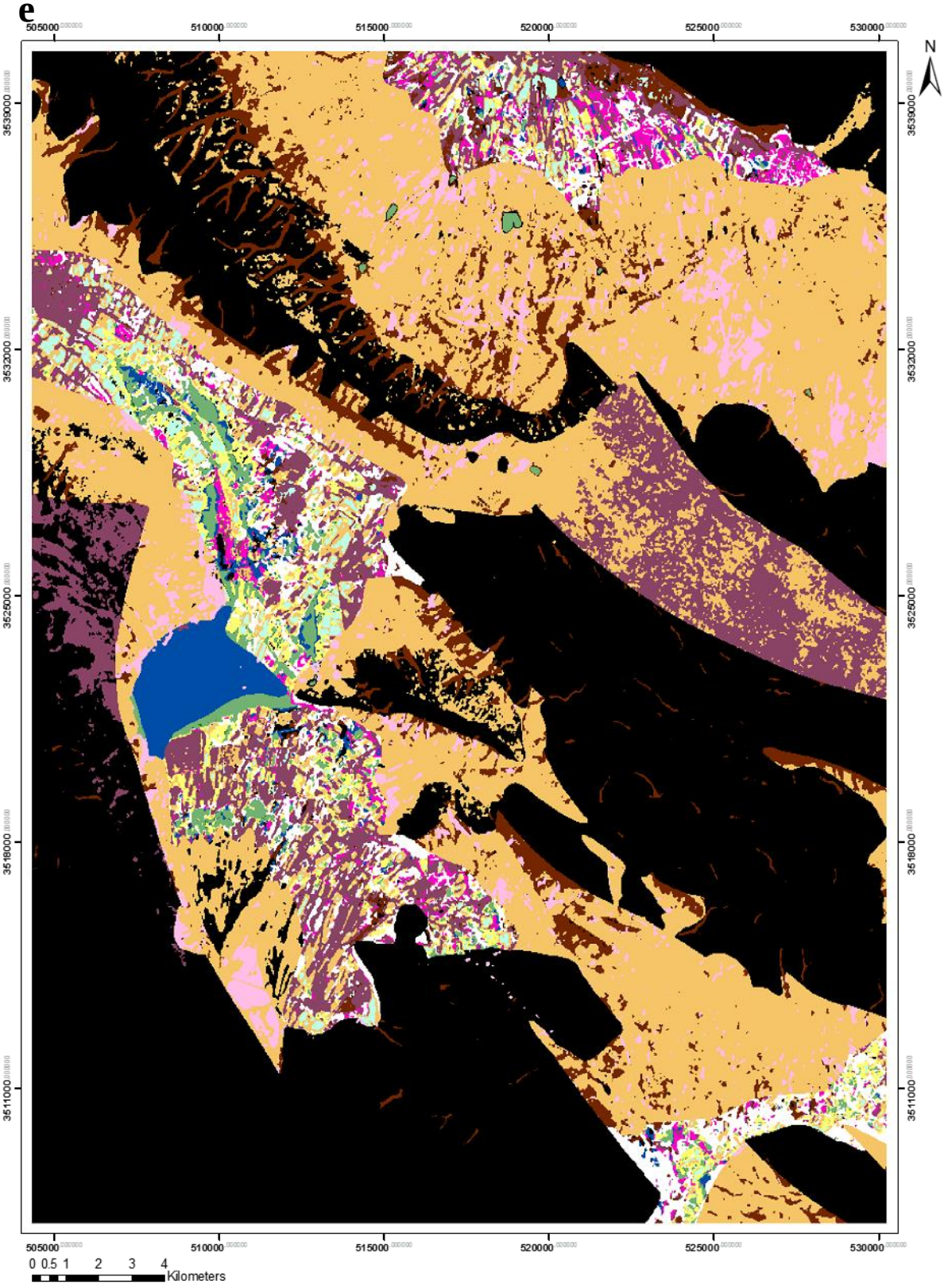
Continue Figure 3-5 Next page

c



Continue Figure 3-5 Next page

e



Continue Figure 3-5 Next page

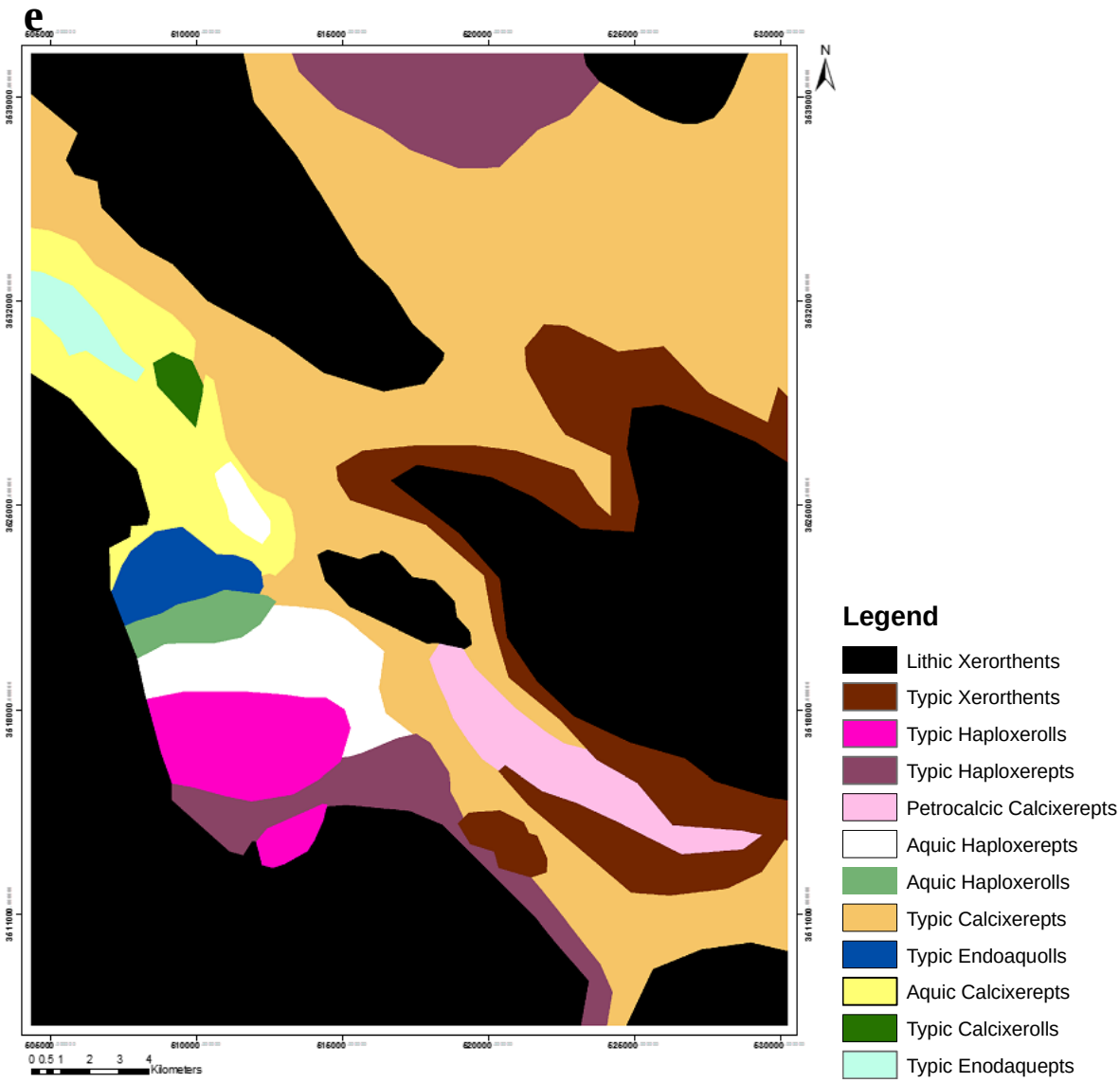


Figure 3-5 Predicted subgroup soils using the same legend for DSM approaches and the CSM approach. (a) RF with the geomorphology map, (b) RF without the geomorphology map, (c) MLR with the geomorphology map, (d) MLR without the geomorphology map, (e) the CSM approach, and (f) the digitized geology map (geology codes explained in Table 3-2).

RF using the geomorphology map wrongly predicted Typic Xerorthents in the Mo111 geomorphic unit (Rock outcrop, Table 3-2 and Figure 3-5 a and b) instead of Lithic Xerorthents. RF also did not recognize Typic Endoaquepts in the river plain landform (Pi411, Table 3-2 and Figure 3-5 a and b), while MLR with the geomorphology map correctly predicted it (Figure 3-5 a

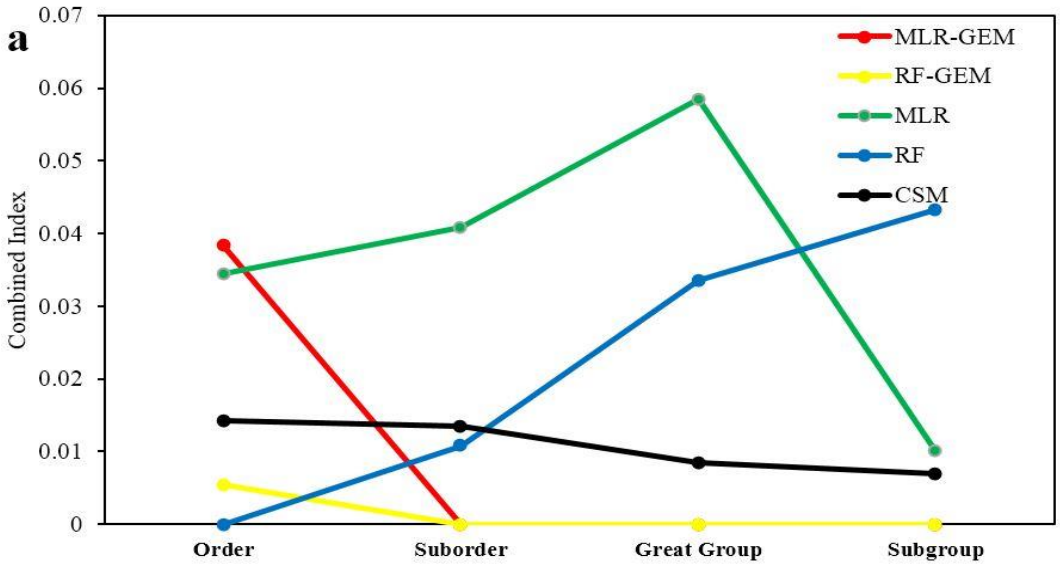
and c). On the other hand, MLR over-estimated the occurrence of Petrocalcic Calcixerepts (three observations in the piedmont landscape). Both approaches (using the geomorphology map) correctly predicted Typic Calcixerolls in the alluvial fan landform (Figure 3-5 a and c). MLR using the geomorphology map predicted Typic Endoaquolls better than RF. These soil types only occur in the lake and wetland landforms (Pl111 and Pl211, Figure 3-5, Table 3-2). In accordance with the findings of Brungard (2009), Hengl et al. (2007), Jafari et al. (2012), Kempen et al. (2009) and (Barthold et al., 2013), our results suggest that soil types with higher numbers of observations had higher prediction accuracy. The conventional approach resulted in a map with larger polygons on average, and also a larger minimal polygon size (Figure 3-5 e). The large minimal polygon size is related to the classical cartographic criteria applied in creating the geology map: the smallest polygon should approximately equal 0.25 cm^2 printed map, which corresponds to 250×250 m^2 on the terrain at 1:100,000 scale. This is much larger than the pixel size of the DSM maps (30×30 m^2). The larger average polygon size reflects the larger geological map as well. As a consequence, the diversity index mapped with a moving window would encounter less map unit transitions with CSM than with DSM maps, and thus the averaged diversity index would be lower and map intricacy is thus lower. Roecker et al. (2010) concluded that map units developed by a conventional approach were over-specified to the soil observations.

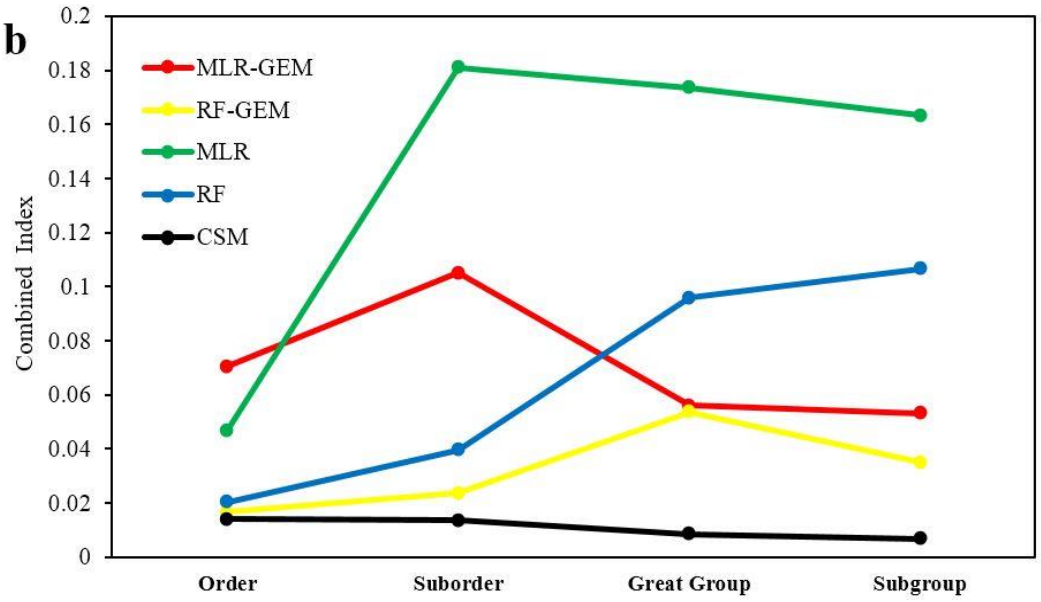
With respect to map purity, MLR at the order level (0.80), RF at the suborder (0.72) and great group level (0.60), and the CSM approach at the subgroup level (0.48) produced the most accurate maps in the study area (Tables 3-4 and 3-6). Zhu et al. (2001), in two case studies in different terrains, showed that soil maps produced by CSM were less accurate (61% and 67% accuracy) than soil maps produced by the soil-land inference model (81% and 84% accuracy).

Menezes et al. (2014) utilized ArcSIE (ArcMap extension) to generate a soil type map and then a solum depth map in Brazil, and found a greater accuracy using DSM compared to CSM.

3.3.4 Performance indicators

With respect to efficiency in terms of cost (map purity and diversity), it would be optimal to produce soil maps with high purity and at low cost, which also well represents soil diversity. We developed a combined index to represent all those three criteria for all scenarios (Figure 3-6). All the relative indices have the optimum value equal to 1 and the least optimum value equal to 0 (Figure 3-6). Amongst the two DSM approaches and CSM, MLR was the most effective in terms of the combined index. CSM was not as effective as the other methods because while the cost of both DSM and CSM were almost equal, CSM did not produce maps with high purity, Kappa, and diversity. Therefore, the combined index for CSM was lower than for the DSM methods. In DSM, the geomorphology map was applied as the main variant (section 3.2.5). Since most of the variability of soil types was explained by the SAVI, the clay index, NDVI, and the geology map (section 3.3.3), inclusion of the geomorphology map was not cost-effective (Figure 3-6). Values of the combined index were higher (better) at smaller sample sizes (Figure 3-6 c) because the decrease in cost outweighs the poorer quality. The combined indices revealed a preference for the order level at small sample size, while a larger sample size led to a preference for the great group level (Figure 3-6).





Continue Figure 3-6 Next page

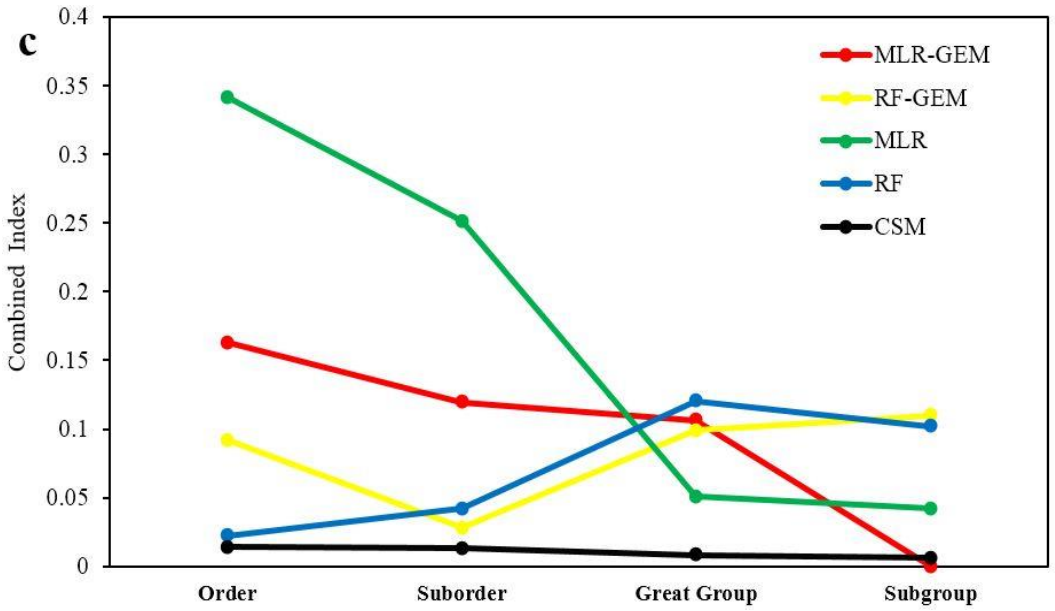


Figure 3-6 Variation of the combined index, (a)100, (b) 80, and (c) 60 points data training along with the CSM approach at the four hierarchic taxonomic levels.

When the cost is considered as a decisive factor for soil surveys, decreasing the operating cost is an acceptable strategy. In this study, we decreased the number of training observations (from 100, to 80, and to 60 observations) for the best scenarios (DSM approaches with the geomorphology map). The results revealed that by decreasing the number of observations points from 100 to 60, the cost decreased 20-40%, while map purity and Kappa did not decrease with the same rates (e.g., map purity decreased from 100 to 60 points for DSM-MLR, 0 and 15% at the order level, 5.9 and 17.6% at the suborder level, 7.7 and 15.7% at the great group level, and -10 and 30% at the subgroup level). However, the lower sampling densities can produce DSM maps which are cost effective in comparison of map purity and Kappa. Therefore, we concluded that the results at the high sampling density (100) were not persistent at lower sample densities. If funds are the limiting factor, decreasing the sample size and mapping at a higher taxonomic level is acceptable, although lower accuracy would be obtained.

3.4 Conclusions

Validation results of soil maps produced by DSM approaches with and without using a geomorphology map were similar because the geology map (as indicator of parent material) already explained the most of the variability and distribution of soil types at different soil taxonomic levels. Inclusion of the geomorphology map did slightly improve map purity and Kappa, and decreased the noise in soil maps, but the number of geomorphic units is large compared to the sample size and conclusions are uncertain. In general, models which were developed using all covariates in the most scenarios showed higher performance. One of the questions addressed in this study was which DSM approach had the highest performance and which approach was the most cost efficient method of soil mapping. The results showed that MLR had a higher performance at the higher taxonomic levels (order and suborder levels) while RF showed higher performance at the lower taxonomic levels (great group and subgroup levels). MLR was less sensitive than RF to a decrease in the number of training observations, thus the rate of decrease in map purity and the Kappa index for MLR was lower than for RF.

Using the combined index, the MLR-DSM was the most effective approach for soil mapping. CSM was not as effective as the other methods: costs were similar but Kappa and depicted intricacy were lower. Because of the additional cost of the geomorphology map and the predictive strength of the geology map, inclusion of geomorphology was not cost-effective in this study, but it helped to improve map quality. Since preparation of the geomorphology map is costly, air photo interpretation with consultation of a geology map as a covariate (the same procedure as CSM approach) might be used as the alternative method in the DSM approach to reduce costs. The values of the combined index were higher at smaller sample sizes; a small sample size led to a preference for the order level, while a larger sample size led to a preference

for great group level. We conclude that in a data-poor region such as Iran, where little soil information is available, DSM and low density sampling with high resolution ancillary data can be an attractive approach for soil mapping at the large scale.

Chapter 4

Spatial Prediction of Some Soil Properties

Abstract

The distribution of soil properties over the landscape is required for a variety of land management applications, modelling, monitoring of landscapes and land recourses. This chapter considers the ability of different digital soil mapping (DSM) approaches (linear and non-linear models) to predict some of the topsoil properties including soil organic carbon (SOC), clay content (Cl) and calcium carbonate equivalent (CCE) in Borujen region, Chaharmahal-Va-Bakhtiari Province, Central Iran. Three non-linear models: Cubist (Cu), Random Forest (RF), Regression Tree (RT) and two linear models: Multiple Linear Regression (MLR) and Stepwise Multiple Linear Regression (SMLR) were evaluated to predict and map soil properties. 304 soil samples in the study area were analyzed and fed to the models to identify the relationship between soil properties and ancillary variables (terrain attributes and remote sensing indices). Model training and validation was done by the 10-fold cross-validation approach. Root mean square error (RMSE), the coefficient of determination (R^2), adjusted coefficient of determination (Adj R^2) and mean error (ME) were considered as the performance indicators of the models. Based on 10-fold cross-validation, the SMLR and RF models showed the highest performance to predict CCE, Cl and SOC content respectively. Also, results revealed that all models could not predict the spatial distributions of clay content properly.

The terrain attribute “elevation” is the most important variable among all studied models. As a conclusion, we recommend that more observations and denser sampling should be carried out in the whole study area. Alternatively, the study area could better be divided into homogeneous

sub-areas, stratifying the area by elevation, and then sampled. The stratified sampling and applied models in this study, probably will increase the performance of soil property predictions.

4.1 Introduction

Accurate and detailed spatial soil information are essential for sustainable land use and management as well as environmental modelling and risk assessment. The distribution of soil properties over the landscape is required for a variety of land management applications, modelling, monitoring of landscapes and land resources. In precision monitoring of land recourses, hydroecological and other environmental modelling applications, high-resolution spatial information on soils can assist decision makers to better understand the variability of soil properties over an area (Forkuor et al., 2017). Therefore, this is crucial for the sustainable use of the soil resources particularly in the context of sustainable land use and climate change.

Traditional soil mapping approaches have mostly depended on ground-based surveys. On the other hand, unfortunately, these approaches rarely provide information about the spatial distribution of soil properties at the desired resolution over the landscape. Acquiring soil information by traditional field surveys including soil sampling and laboratory analyses are time-consuming, expensive, especially when the mapping is being done at national, regional or global scales. So mapping of continuous spatial variation of soil properties is an almost impossible effort (Forkuor et al., 2017; Jafarisirizi, 2012). Obviously, it is practically impossible and hard to measure soil properties continuously, therefore, it is necessary to have robust systems and models that can predict soil properties at a given location or scale. Consequently, fast and accurate prediction of soil properties is a necessity for soil scientists and decision-makers to overcome the lack of measured soil property information. Considerable progress in the past three decades has been made in such predictions following the development of geostatistics and

modelling whereby predictions were made with calculated levels of accuracy and error. Subsequent advances in a range of sensing techniques (aircraft, satellite, on-the-ground spectroscopy etc.) allowed that soil properties can now be accurately predicted with new tools and approaches like digital soil mapping (McBratney et al., 2003; Minasny and Hartemink, 2011). By applying digital soil mapping (DSM), soil properties that are expensive to measure, time-consuming or unavailable can be predicted from the point observations and ancillary data such as remote sensing data and terrain attributes (McBratney et al., 2003).

Some soil properties have a high spatial variability, especially in agricultural areas. For better soil management practices, it is necessary to know the spatial distribution of soil properties. The spatial variability of soil properties is influenced by parent material characteristics, topography, climate, vegetation, time and anthropogenic activities (Fenton and Larterbach, 1999; Mulder et al., 2011).

In recent years, several studies have investigated the spatial variability of different soil properties such as soil pH (Behera and Shukla, 2015), soil organic matter (Byrne and Yang, 2016), electrical conductivity (EC) (Ranjbar and Jalali, 2016), phosphorus (Wilson et al., 2016), potassium (Behera and Shukla, 2015) and soil texture (Barnes and Baker, 2000).

Digital soil mapping methods are widely employed to assess the spatial distribution of soil properties in agricultural areas and other land resources (Taghizadeh-Mehrjardi et al., 2016; Forkuor et al., 2017; Minasny and Hartemink, 2011). The soil properties maps produced by DSM approaches show the spatial variation of the desire variables in the area of interest.

Numerous prediction models have been developed and introduced to correlate ancillary variables and soil properties through the DSM framework suggested by McBratney et al. (2003). For example, Minasny et al. (2013) introduced a comprehensive review of SOC modelling.

Hengl et al. (2015) stated that multiple and linear regression have been used commonly for relating SOC to ancillary variables. Comparing to popular approaches, fewer studies used generalized linear models (Karunaratne et al., 2014), regression tree models (Martin et al., 2010; Taghizadeh-Mehrjardi et al., 2014), random forest (Hengl et al., 2015; Were et al., 2015), artificial neural networks (Dai et al., 2014), support vector regression (Forkuor et al., 2017; Were et al., 2015), random forest regression (Forkuor et al., 2017), to build the relationships between soil properties and ancillary data. The advantage of these modelling techniques is that they have the potential for detecting non-linear relationships and might therefore prove more powerful for digital soil properties mapping. Soil organic carbon (SOC), soil calcium carbonate equivalent (CCE) and clay content are some of the most important soil properties that can define soil quality and can also be an indicator of soil fertility. These soil parameters can be highly variable in space and time, especially in agricultural areas, with implications for crop production (Bogunovic et al., 2017).

The objectives of this chapter were to predict and map the spatial distribution and variation of soil properties (SOC, CCE and clay content) using different digital soil mapping techniques in a semi-arid region of Iran. To investigate soil properties variation, five statistical approaches comprising multiple linear regression (MLR), random forest (RF), cubist (Cu), stepwise multiple linear regression (SMLR) and regression tree (RT) were explored to identify the most suitable method for prediction and mapping of soil properties in the study region. The research questions of this chapter tried to address: (1) Which DSM approaches offer the best accuracy for predicting soil properties? (2) Which ancillary data is the most important for predicting soil properties?

4.2 Materials and Methods

4.2.1 Description of the study area

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

4.2.2 Soil sampling scheme

In addition to the samples collected in chapter 2, a number of 209 surface samples (0-30 cm) were collected by contributing of quantitative covariates (Table 3-1) and applying cLHS approach. In total, 334 observations were collected for prediction and mapping of some soil properties (Figure 4-1).

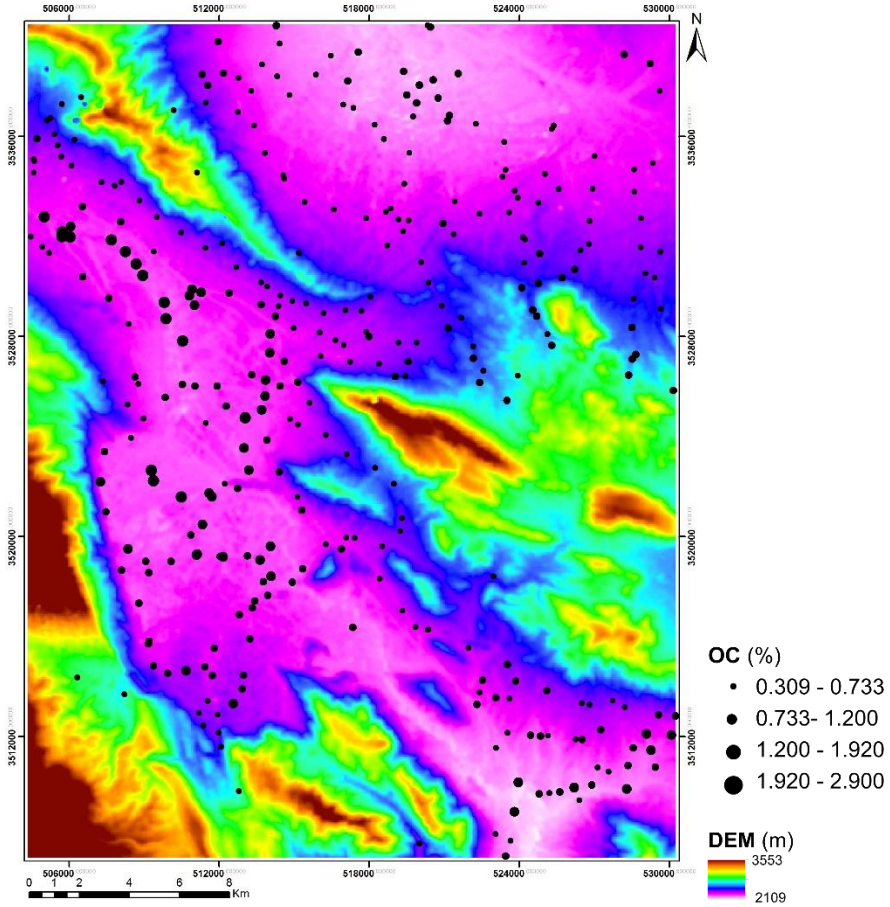


Figure 4-1 Digital elevation model and locations of sampling points with distribution of soil organic carbon contents in the study area.

4.2.3 Soil properties modelling

In this chapter, we also used the scorpan spatial soil prediction function framework in the mapping procedure (McBratney et al., 2003), where a soil property can be mapped as an empirical function of its environmental covariates: soil, climate, organisms, relief, parent material, age, and spatial position. In this study, the ancillary covariates presented in chapter 3 were also applied for the modelling of soil properties. In addition, based on important predictors found in the models, the best covariates were presented for each soil property.

Three machine learning models including Cubist (Cu), Random Forests (RF), Regression Tree (RT) and two linear models comprising; Multiple Linear Regression (MLR) and Stepwise Multiple Linear Regression (SMLR) were evaluated and applied for predicting, mapping and estimate soil organic carbon (SOC), calcium carbonate equivalent (CCE) and clay contents. These techniques are described briefly in the following subsections.

3.2.3.1 Cubist

Cubist is an advanced version of the regression tree algorithm and was performed using the "cubist" R package (Kuhn et al., 2013; R Development Core Team, 2013). Cubist is an extension of the Quinlan's M5 model tree (Quinlan, 1993; Quinlan, 1992). Cubist is similar to common regression trees, except that the leaves are in the form of a linear regression of the covariates (Malone et al., 2014; Taghizadeh-Mehrjardi et al., 2014). In Cubist, the prediction is based on linear regression models instead of discrete values. Cubist constructs different models from training data. Each model consists of several rules and each rule has one or several conditions. Whenever a case satisfies all conditions of a rule the linear form is applied to predict the value. Unlike regression trees, which predict a rigid value for each "leaf", regression rules build a multivariate linear function. The rules of one model are sorted in descending order of importance

by cubist. This means that the first rule has the greatest contribution to model accuracy of the training data; the last rule has the least impact. The Cubist model has been used effectively in various soil prediction and mapping procedures (e.g., Bui et al., 2009; Henderson et al., 2005; Kidd et al., 2014; Minasny et al., 2008; Viscarra Rossel et al., 2014).

3.2.3.2 Random Forest

Random Forest (RF) is a classifier or regression model which consists of many decision or regression trees where each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the data. The output of the model is an average of all the regression or decision trees. The RF was applied by "randomforest" R package (Liaw and Wiener, 2002; Brewer et al., 2015). Breiman (2001) firstly introduced the Random Forest, which is a tree-based ensemble method. RF contains many regression trees instead of a single standard regression tree, like a forest. RF operates on two subsets: (1) on predictor variables at each node and (2) on individual data by a bootstrapping technique. Three parameters control the fitting of RF models: (i) the number of trees (*ntree*), (ii) the minimum number of samples in the terminal node n_{min}, and (iii) the number of predictors to be used for the fitting of each tree (*Mtry*) (Grimm et al., 2008). The importance of the predictive variables is determined by two methods in RF, the first one is standard measure and second is the Gini measure. In the first method, variable importance replaces the true values of the variable with randomly generated (likely incorrect) values for each tree in the grove and assesses the impact on classification (Salford Systems, 2004). Then, if there is no impact on the error of the tree the significance of the variable decreases. On the other hand, if the tree's ability to predict the out-of-bag error observations is diminished, the variable is considered important. The Gini

importance ranks variables according to how clearly the variable separated classes when selected at a node (Salford Systems, 2004).

The number of trees is a compromise between the accuracy and processing time (Oshiro et al., 2012). The range of the number of trees was set between 100 and 1000 at intervals of 10, and in this study we used 520 trees. The relative importance of the predictor variables in modelling of soil properties were assessed using the "importance" function in the "randomforest" R package (R Development Core Team, 2013).

3.2.3.3 Regression Tree

The Regression Tree (RT) approach is used to characterize the relationship between soil properties and covariates. RT analysis is a parametric data mining technique that can handle both non-linear and linear relationships (Breiman et al., 1984; Myles et al., 2004), and it is reported that this technique has been widely used in the field of DSM (Taghizadeh-Mehrjardi et al., 2014). The response as well as the covariates are numeric. The dataset of the response variable is split in a tree like manner into successively smaller groups on the basis of the ancillary covariates that maximizes the homogeneity of the groups (De'ath and Fabricius, 2000). The RT approach was run by "rpart" R package (R Development Core Team, 2013).

3.2.3.4 Multiple Linear Regression

The general purpose of MLR is to analyze the relationship between several independent or predictor variables and a dependent or predicted variable. Multiple regression analysis fits a straight line (or plane in an n-dimensional space, where n is the number of independent variables) to the data (Mbagwu and Abeh, 1998). So, the Multiple Linear Regression model is used in order to provide quantitative estimation for each soil property. The linear regression model can be expressed as,

$$\boldsymbol{f(X) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j} \qquad \text{(Eq. 4.1)}$$

where β_0 is the interception of the linear model, X_j represents the auxiliary or secondary variables or covariates and β_j are the unknown coefficients for the auxiliary covariates and p is the number of auxiliary covariates (Hastie and Tibshirani, 2008). Regression methods explore a possible functional relationship between the primary variable (soil properties) and explanatory variables (SCORPAN factors). The "rgdal" R package was used to run the Multiple Linear Regression model (R Development Core Team, 2013).

3.2.3.5 Stepwise Multiple Linear Regression

Both direction (Backward and Forward combination) Stepwise Multiple Linear Regression (SMLR) was performed using the 'MASS' package in R (Venables and Ripley, 2003) and covariates selection were determined by minimizing the Akaike Information Criterion (AIC; Akaike, 1974). The AIC is presented as follow:

$$\boldsymbol{\mathrm{AIC} = 2K + Nln(L)} \qquad \text{(Eq. 4.2)}$$

where K is the number of the estimated parameters included in the model, L is maximized the value of the likelihood function for the estimated model which is readily available in the statistical output and reflects the overall fit of the model. In itself, the AIC value for a given data set has no meaning. It becomes interesting when it is compared to the AIC of a series of models, here different models constructed based on combination of the different covariate, then one with the lowest AIC being the best model. If many models have similarly low AICs, the one with the fewest predictor variables should be chosen.

4.2.4 Models Validation

Ten-fold cross-validation was used to evaluate the prediction performance of five models (Hengl et al., 2015). Cross-validation provides a structure for creating several train/test splits in

the dataset and guaranteeing that each data point is in the test set at least once. The procedure is simple, at first, the main data is split into 10 equal-sized groups. Then, for 1 to 10, select group 1 to be the test set and all other (9) groups to be the training set. After that, train the model on the training set and evaluate on the test set. Each iteration is called a fold. The advantage of this method is that it performs reliably and is unbiased on smaller data sets. This approach requires much more computational effort than simple trained-and-tested (hold out) procedures (Taghizadeh-Mehrjardi et al., 2014).

The performances of the models during training and testing in the ten-fold cross-validation procedure were evaluated using four different error parameters which are popularly used in digital soil mapping by applying the R statistical software package (R Development Core Team, 2013).

1. Root Mean Square Error (RMSE): The RMSE represents the accuracy of the model predictions

$$\mathbf{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}[O_i - P_i]^2}{n}} \qquad \text{(Eq. 4.3)}$$

2. Coefficient of determination (R^2): to determine and evaluate the model performance R^2 was used for each model according to the following equation:

$$R^2 = \left[\frac{\sum_{i=1}^{n}(O_i - O_{avg})(P_i - P_{avg})}{\sqrt{\sum_{i=1}^{n}(O_i - O_{avg})^2 (P_i - P_{avg})^2}}\right]^2 \qquad \text{(Eq. 4.4)}$$

3. Mean error (ME): a measure of model's prediction bias:

$$ME = \frac{1}{n}\sum_{i=1}^{n}(O_i - P_i) \qquad \text{(Eq. 4.5)}$$

4. Adjusted coefficient of determination (Adj R^2):

$$\boldsymbol{Adj\ R^2 = \sqrt{R^2}} \qquad \text{(Eq. 4.6)}$$

where P_i, O_i, O_{ave}, P_{ave},and *n* are the predicted and observed values, average of observed and predicted soil properties values at the *i*th point, and number of data, respectively.

The results of these criteria allowed comparison of the effectiveness of the different models to predict the spatial distribution of soil properties and final maps. Ideally, the best model would maximize the R^2, Adj R^2 which is close to 1 and minimize RMSE and ME which is close to 0.

For RT, Cu, MLR and SMLR the variable importance values were converted to relative importance by scaling model importance values from 0 to 100%. Consequently, variables with higher relative importance values are more important.

4.3 Results and Discussion

4.3.1 Soil calcium carbonate equivalent (CCE) prediction and mapping

The average validation criteria for soil calcium carbonate equivalent (CCE) prediction in the studied area are presented in Table 4-1. Among the applied models in the validation data set, the most accurate model for predicting of soil CCE was SMLR and the least accurate model for CCE was RT. Contrary to the validation results, in the calibration data set the most precise model was RF and the least accurate model was Cu. These results indicate that although some models had higher performance in the calibration data set, they showed lower performance in the validation data. The performance of RF is inferior with RMSE= 9.49 percent and R^2=0.24 in the validation subset comparing to the performance in the calibration subset with RMSE= 3.83 percent and R^2= 0.93.

SMLR and MLR models resulted in similar performances in the both validation and calibration data set: RMSE about (≈) 8.2 percent and R^2≈0.45 for SMLR, RMSE≈8.3 percent and

R^2≈0.43 for MLR, RMSE≈ 8.4 percent and R^2≈ 0.29 for SMLR, RMSE≈ 8.4 percent and R^2≈ 0.30 for MLR. Nevertheless, linear prediction models showed a higher accuracy compared to other non-linear prediction models were used.

According to the dominant parent material and geology (chapter 3; Table 3-2, lithology) (Borujen Geology Map, 1990), high soil calcium carbonate concentrations occurred in most of the study area. Additionally, because of insufficient precipitation rate and high temperature in the arid and semi-arid region (Borujeni et al., 2010), calcium carbonates accumulated in the soil.

Wilford et al. (2015) concluded that minimal leaching of soil calcium carbonate associated with low rainfall, high temperature and evaporation leads to the retention of atmospherically sourced calcium including that from rainfall, sea-spray, aerosols and aeolian dust in the soil. Therefore, the distribution of CCE in the study area due to rich soil calcium carbonate did not show effect by different landform or other soil forming factors (ancillary data). Eventually, the models could not build the proper relationship between CCE and the predictors, so the predictions led to low performance in terms of accuracy for most models. As a conclusion, the main explanation of the soil calcium carbonate distribution is the parent material and the climate condition, specifically rainfall and temperature. Parent material is illustrated by high soil calcium carbonate concentration (chapter 2, table 2-4) over the highly calcareous bedrock (chapter 3; Table 3-2, lithology).

Table 4-1- Average validation criteria for prediction of calcium carbonate content (CCE %) from10-fold cross validation. The most accurate method is shown in bold.

	Validation				Calibration			
	RMSE	R^2	Adj R^2	ME	RMSE	R2	Adj R2	ME
RF	9.492±0.611	0.237±0.039	0.226±0.039	0.949±0.088	**3.832±0.059**	**0.933±0.003**	**0.932±0.003**	**0.092±0.009**
Cu	8.828±0.997	0.431±0.067	0.422±0.068	0	8.535±0.152	0.273±0.008	0.270±0.008	0
MLR	8.275±0.687	0.430±0.020	0.422±0.020	0.768±0.524	8.374±0.188	0.300±0.001	0.297±0.001	0
SMLR	**8.178±0.716**	**0.447±0.040**	**0.437±0.040**	**0.768±0.524**	8.432±0.197	0.290±0.032	0.288±0.003	0
RT	9.975±0.226	0.195±0.062	0.191±0.058	-0.139±0.914	7.410±0.443	0.452±0.042	0.450±0.042	0

RF: Random Forest; Cu: Cubist; MLR: Multiple Linear Regression; SMLR: Stepwise Multiple Linear Regression; RT: Regression Tree; RMSE: Root Mean Square Error; Adj R^2: Adjusted R^2; ME: Mean Error

Table 4- 2- Average validation criteria for prediction of clay content (CI %) from10-fold cross validation. The most accurate method is shown in bold.

	Validation				Calibration			
	RMSE	R^2	Adj R^2	ME	RMSE	R2	Adj R2	ME
RF	**7.452±0.101**	**0.185±0.101**	**0.173±0.102**	**-0.741±1.554**	**3.041±0.058**	**0.944±0.001**	**0.944±0.001**	**-0.044±0.024**
Cu	7.976±0.032	0.047±0.040	0.032±0.041	0	6.752±0.163	0.206±0.066	0.203±0.066	0
MLR	7.863±0.208	0.081±0.091	0.067±0.093	-0.802±1.489	6.979±0.072	0.130±0.034	0.126±0.034	0
SMLR	7.857±0.104	0.079±0.076	0.065±0.077	-0.802±1.489	7.042±0.118	0.114±0.047	0.110±0.047	0
RT	7.918±0.771	0.134±0.141	0.121±0.143	-0.618±0.827	5.709±0.215	0.417±0.056	0.415±0.056	0

RF: Random Forest; Cu: Cubist; MLR: Multiple Linear Regression; SMLR: Stepwise Multiple Linear Regression; RT: Regression Tree; RMSE: Root Mean Square Error; Adj R^2: Adjusted R^2; ME: Mean Error

Table 4- 3- Average validation criteria for prediction of organic carbon content (OC %) from10-fold cross validation. The most accurate method is shown in bold.

	Validation				Calibration			
	RMSE	R^2	Adj R^2	ME	RMSE	R2	Adj R2	ME
RF	**0.319±0.042**	**0.627±0.053**	**0.622±0.053**	**0.023±0.037**	**0.150±0.006**	**0.928±0.004**	**0.928±0.004**	**0.003±0.001**
Cu	0.321±0.034	0.616±0.068	0.611±0.068	0	0.284±0.018	0.679±0.032	0.678±0.032	0
MLR	0.345±0.039	0.560±0.072	0.554±0.074	-0.014±0.045	0.335±0.009	0.531±0.021	0.529±0.021	0
SMLR	0.344±0.039	0.561±0.070	0.554±0.071	-0.014±0.045	0.338±0.010	0.523±0.018	0.521±0.018	0
RT	0.357±0.040	0.544±0.072	0.538±0.073	-0.002±0.054	0.273±0.041	0.683±0.086	0.681±0.087	0

RF: Random Forest; Cu: Cubist; MLR: Multiple Linear Regression; SMLR: Stepwise Multiple Linear Regression; RT: Regression Tree; RMSE: Root Mean Square Error; Adj R^2: Adjusted R^2; ME: Mean Error

4.3.2 Clay content (Cl) prediction and mapping

Table 4- 2 shows the summary of models performance for prediction of clay content (Cl). The comparison of models were conducted using 10-fold cross-validation. Results suggested that linear models (MLR and SMLR) had similar performance and are not appropriate for the prediction of clay content.

According to the RMSE, R^2 and ME values in table 4- 2, results indicated that among the investigated models in the validation and calibration data set, the RF technique had the highest performance to predict clay content. Among the studied models, the cubist model showed the highest RMSE and lowest R^2 for clay prediction and could not explain the variability of this property. In spite of reasonable performance in the calibration subset, all of the models did not have appropriate performance in the validation data set.

Validation statistics indicated that the RMSE and R^2 values were high and low respectively for clay content which means none of the models could predict clay content accurately (table 4- 2).

4.3.3 Soil organic carbon (SOC) prediction and mapping

The summary of models performance for soil organic carbon content are shown in Table 4- 3. The results of soil organic carbon prediction (SOC) are excellent with RMSE≈ 0.32 percent and R^2≈ 0.63 for RF and RMSE≈ 0.32 percent and R^2≈ 0.62 for Cu techniques (Table 4- 3). Both of these models also had high performance in the calibration data set. These results confirm the good correspondence between modeled and measured soil organic carbon. Prasad et al. (2006) reported that superior capability of RF is its prediction performance. Our results support this assessment with high prediction accuracies in model performance in the validation and calibration subsets. The linear models had the same performance in the validation and calibration

data set. The correlation between predicted and observed SOC (R^2) ranged from 0.54 for RT to 0.63 for RF (Table 4- 3). Overall, all these approaches showed acceptable performance.

Summary statistics of SOC in the predicted maps and soil dataset are shown in Table 4-4. According to the table, the average SOC levels ranged from 0.309 to 2.90 percent. The linear models underestimated SOC with high values of standard deviation. Cubist approaches predicted SOC content near to minimum and maximum in the observed data.

Table 4- 4- Descriptive statistics of soil organic carbon content (OC %) in the predicted maps and soil data set

	Min	1st Qu	Median	Mean	3rd Qu	Max
Soil data set	0.309	0.538	0.703	0.855±0.497	0.958	2.900
RF	0.412±0.011	0.627±0.013	0.712±0.016	0.824±0.011	0.913±0.018	2.435±0.067
Cu	0.324±0.063	0.611±0.022	0.709±0.024	0.806±0.021	0.895±0.026	2.530±0.138
MLR	-1.810±0.724	0.539±0.024	0.693±0.029	0.787±0.027	0.934±0.037	3.092±0.254
SMLR	-1.520±1.324	0.558±0.042	0.700±0.031	0.798±0.034	0.924±0.021	2.954±0.239
RT	0.511±0.043	0.642±0.017	0.675±0.051	0.808±0.007	0.836±0.077	2.331±0.291

Figure 4- 2 showed the plot between observed and predicted SOC content for the calibration and validation subsets. The validation subset showed more scattering compared to the calibration subset. Excellent agreements were found in both calibration and validation subsets between the observed and predicted SOC especially for Cubist and Random Forest. Although the basis of the investigated non-linear models are regression trees, the comparison between RF and Cu with RT showed that advanced models (i. e. RF and Cu) had better performance than simple RT. RT used discrete values in the terminal nodes for splitting the trees but RF and Cu approaches could use a regression model in the terminal nodes, and therefore produce a range of predictions (Malone, 2013). In Figure 4-2, the RT approach predicted discrete values for SOC content.

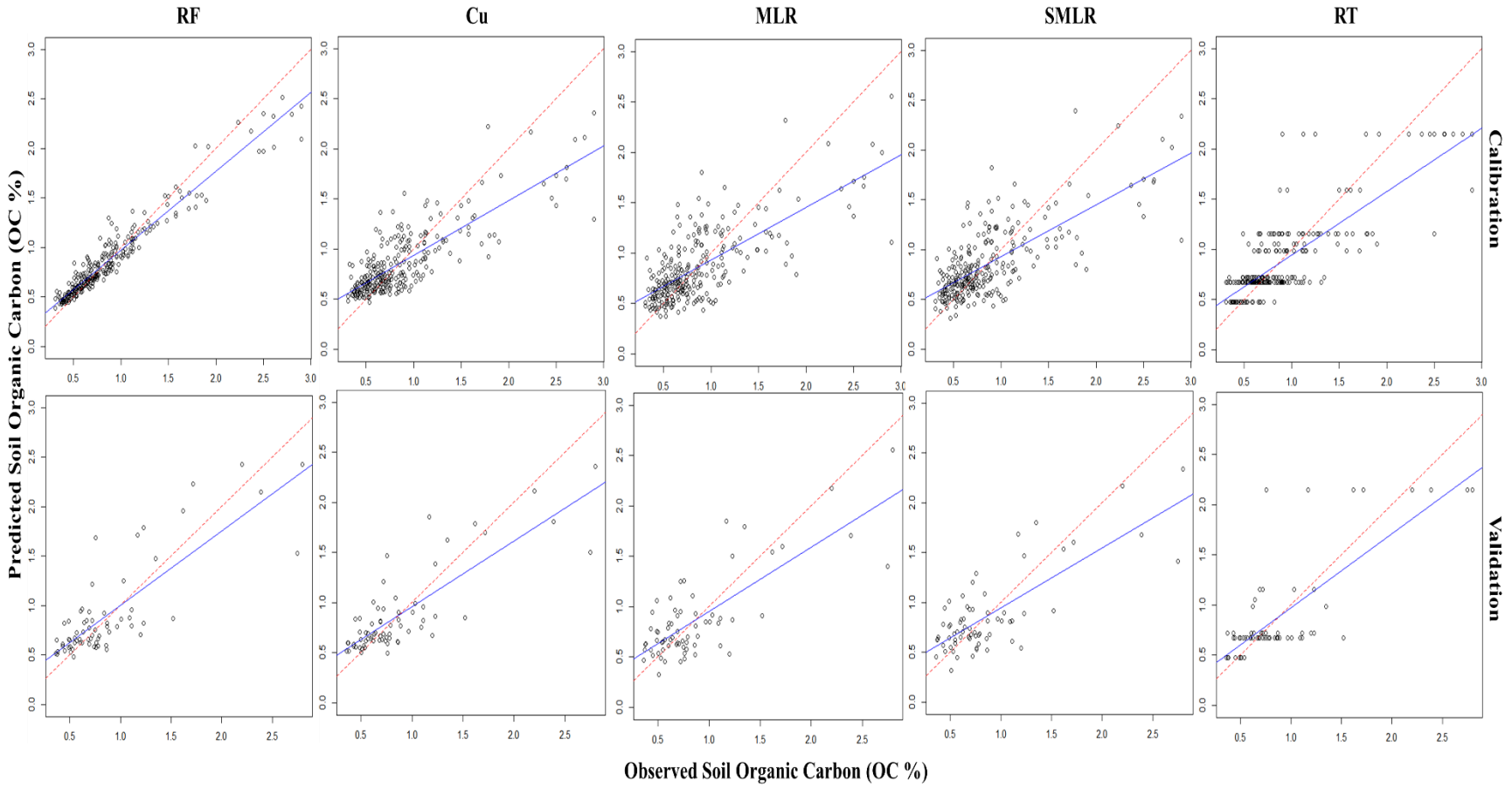


Figure 4- 2- Agreement between observed and predicted of soil organic carbon content (OC %) from five models: Cubist (Cu), Random Forest (RF), Multiple Linear Regression (MLR), Stepwise Multiple Linear Regression (SMLR) and Regression Tree (RT). Linear fit between observed and predicted (solid blue lines) and concordance 1:1 line (dash red lines).

Figures 4-3 and 4-4 showed the variables of importance for the applied models. The variable of importance for RF 'importance plots', presented by two criterions, comprise an increase in accuracy and increase in node purity (Figure 4-3). Among all of covariates DEM, clay index and PVI come out to be the most important predictors for predicting SOC measured by the mean decrease in prediction accuracy (Figures 4-3, left; Figure 4-5 RF SOC prediction map). When the Gini measure considered MRVBF, clay index and DEM are the strongest predictive variables. These variables helped to decrease the noise in the selected at a node of trees.

Genuer et al. (2010) expressed that the Gini measurement of importance is not fair in favor of predictor variables when there are many categories and variables (Strobl et al., 2007), while the importance measurement of RF by mean decrease in accuracy is a more reliable indicator. However, DEM and clay index both contributed as the most importance variables in random forest approach.

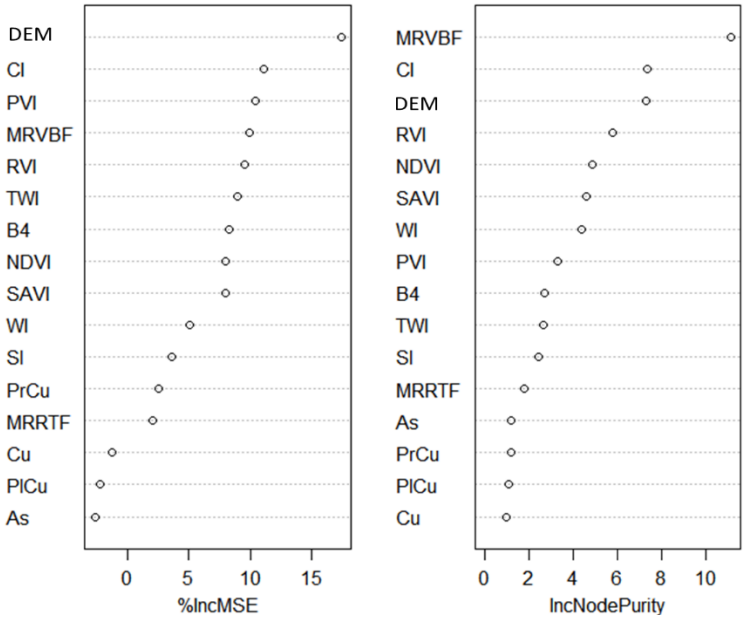


Figure 4- 3- Variables of importance for predicting soil organic carbon content (OC %) according to Random Forests (RF), increase in accuracy (left) and increase in node purity (right). Variables abbreviation presented in table 3- 1.

For RT, Cu, MLR and SMLR models, the indices of relative importance are calculated for each variable, which are the importance of each variable in the 10-fold cross validation for prediction SOC content. Clay index, SAVI and MRVBF were considered as the most important predictors (Figure 4-4).

According to Figure (4-4) in the cubist model the clay index, MRVBF and DEM were the most important variables which explained more than 50 % of SOC variability (Figure 4-5, Cu). Three top variables in the RT approach were included SAVI, RVI and DEM that explain about 35 % of SOC variability in the studied area (Figure 4-4 and 4-5, RT).

Three top most variables for linear models were aspect, clay index and MRVBF, however in the SLMR model explained more than 80 % of SOC variability while those three most important variables in MLR model about 80 % of SOC variability (Figures 4- 4 and 4- 5, SMLR and MLR). Based on the results obtained and presented in Figures 4- 3 and 4- 4, it can be concluded that the terrain attributes can be the main predictors and on the second level remote sensing attributes by relative contribution for SOC content prediction. The results also confirm that other used ancillary covariates could not properly capture the variation of SOC in this study area, but using those improved the accuracy and purity of predicted maps. (Nath, 2006) reported that the terrain attribute “curvature” was the most important factor for prediction of the SOC with multiple regression model and described 31 % of its variability. According to the relative importance in Figure (4- 4) our results are in line with previous work. Regarding terrain attributes Kheir et al. (2010) described the influence of elevation, soil types, and slope for prediction of SOC in wet cultivated lands in Denmark. The vertical distribution of SOC content was modeled and mapped at five standard soil depth intervals using environmental variables (soil map, geology, precipitation, wetness index, land use, aspect, and elevation) in a relatively flat

area. Results showed that SOC distribution was influenced by precipitation, land use, soil type, elevation, and wetness index (Adhikari et al., 2014). Mosleh et al. (2016) concluded that terrain attributes were the main predictors for spatial prediction of some soil properties.

In contrast to previous studies, in which Minasny et al. (2013) and Taghizadeh-Mehrjardi et al. (2016) concluded NDVI and the other remotely sensed vegetation parameters are usually good to predict SOC contents, our results demonstrated that in this study area these were not good predictors. Such results might be due to the climate conditions in the arid and semi-arid region that led to poor vegetation development.

.

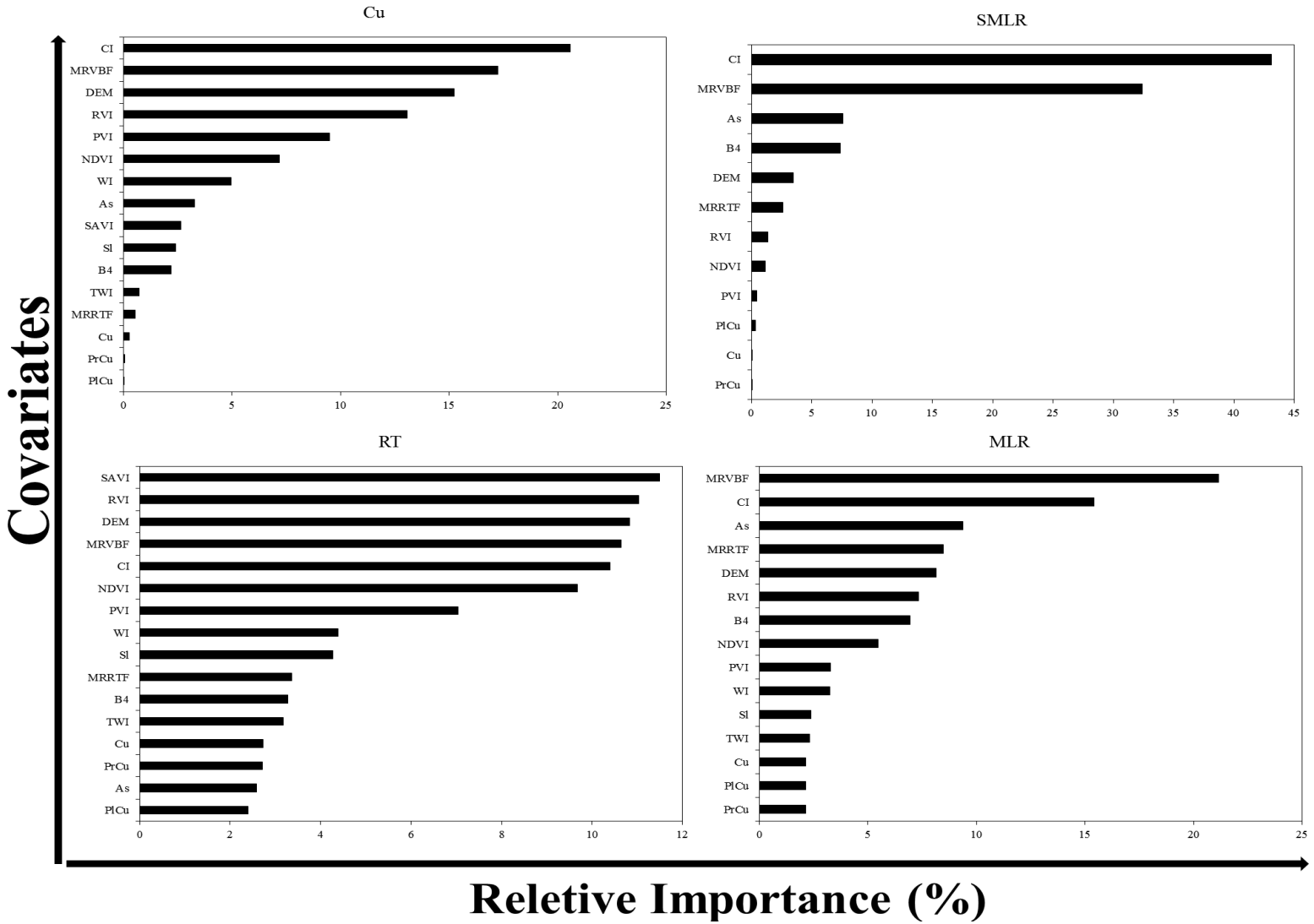
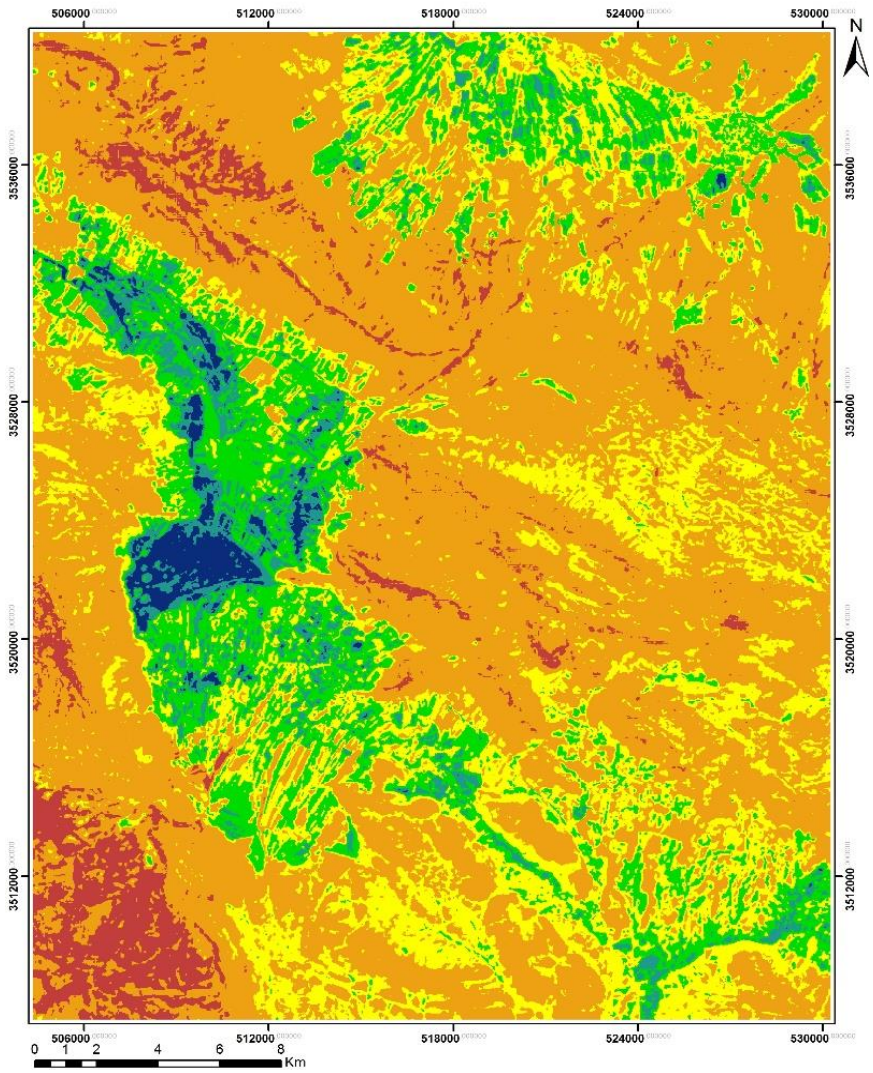
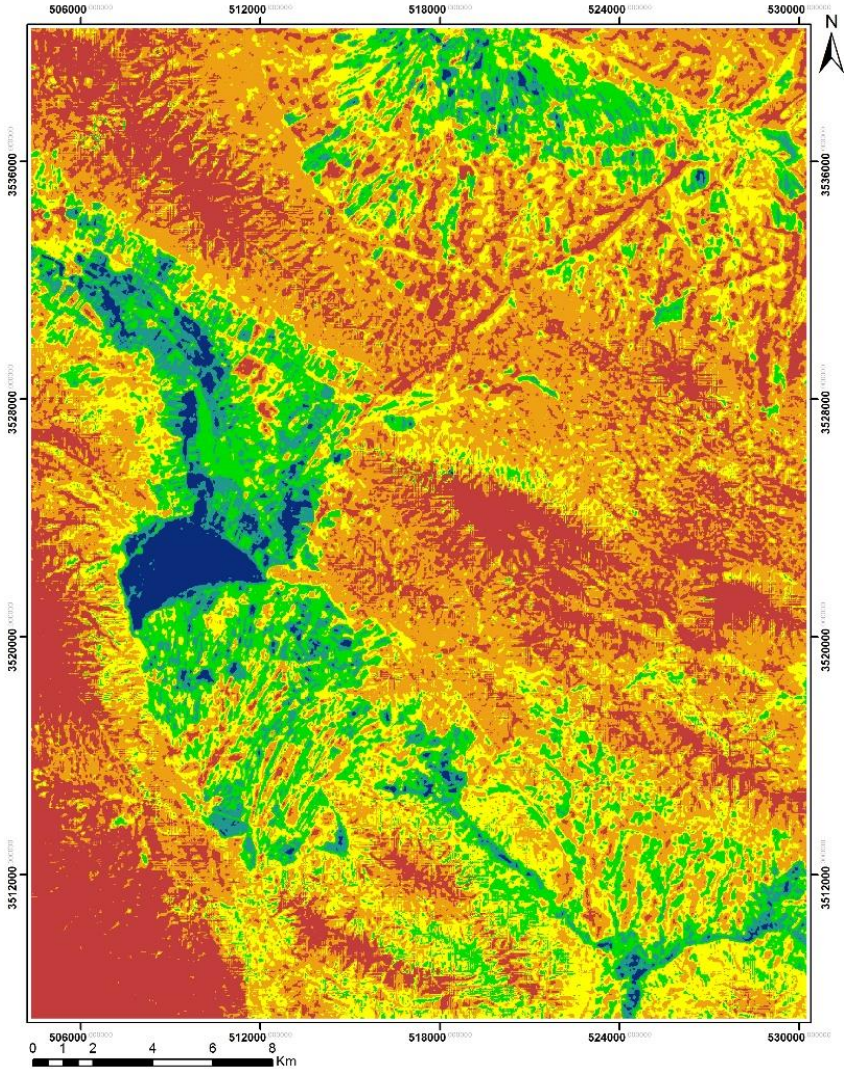


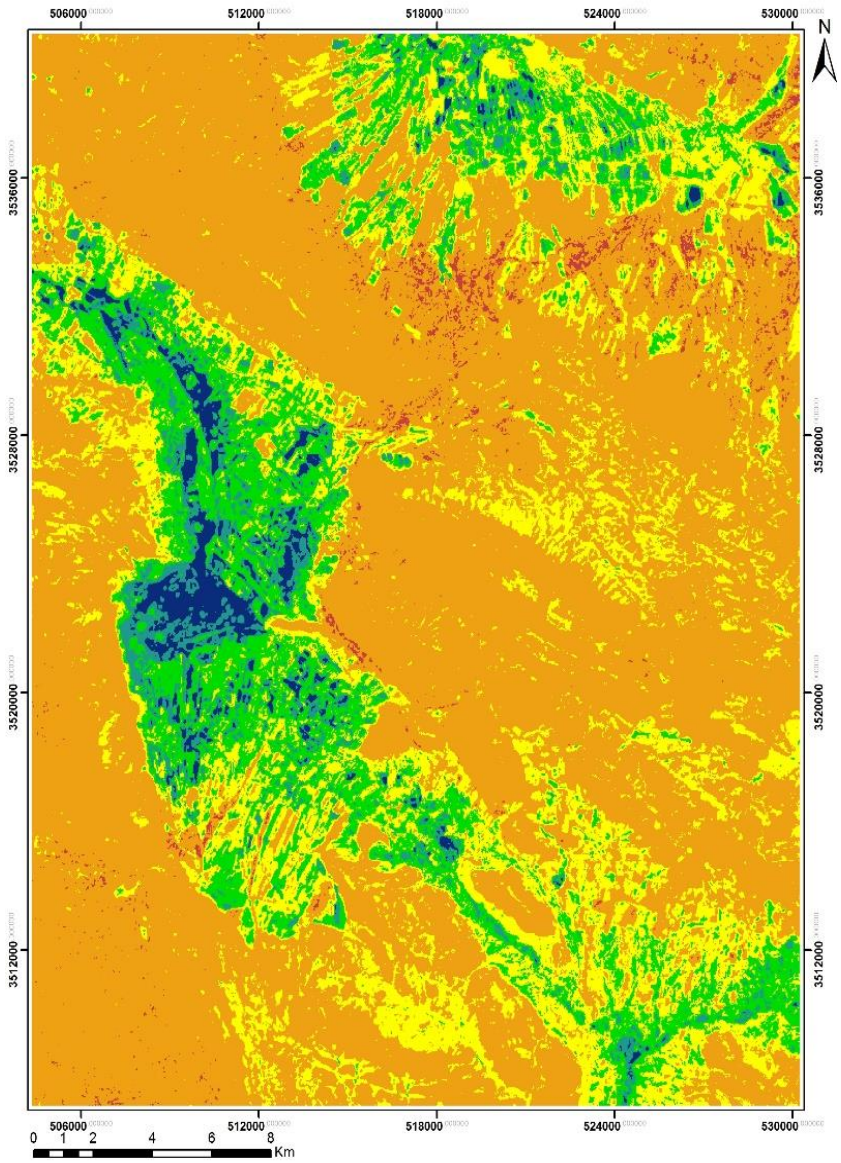
Figure 4- 4- The significance of each auxiliary variable used in four approaches for prediction of soil organic carbon content (OC %) Percentage represents how frequently the auxiliary variable was used in models.


506000
512000
518000
524000
530000
N
3536000
3528000
3520000
3512000
0 1 2 4 6 8
Km

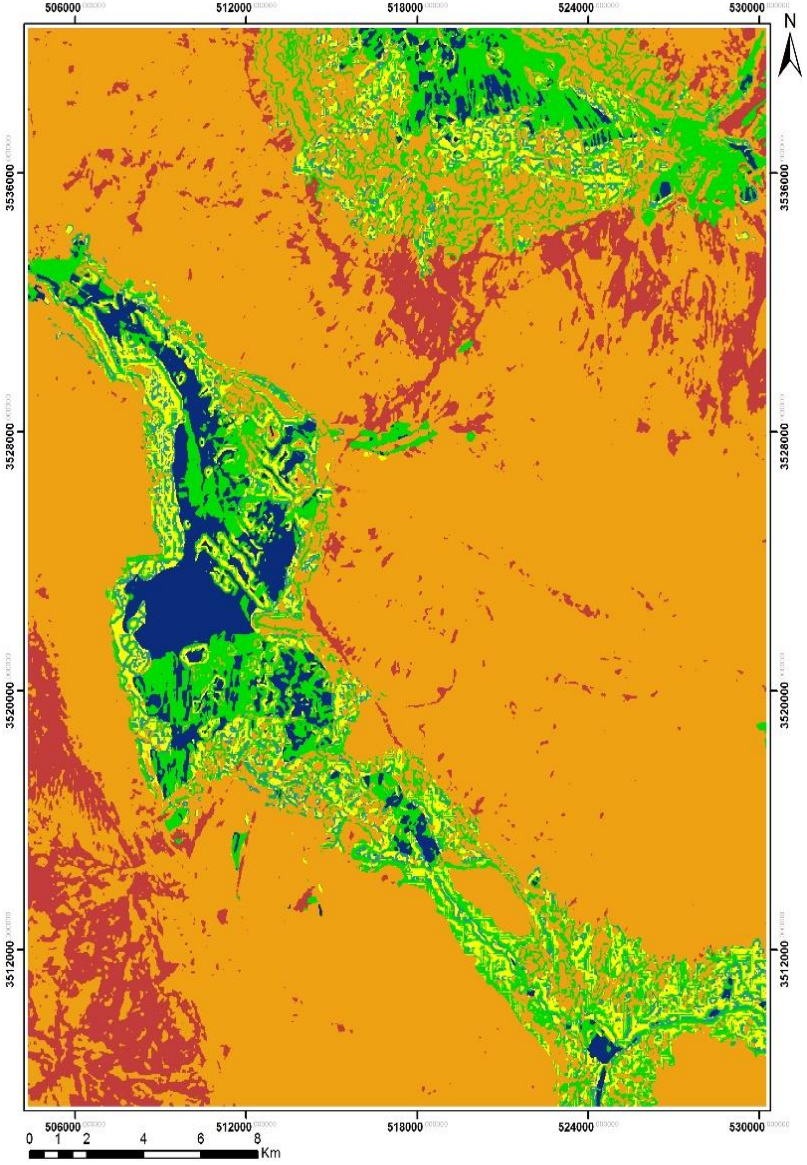
Cu


506000
512000
518000
524000
530000
N
3536000
3528000
3520000
3512000
0 1 2 4 6 8
Km

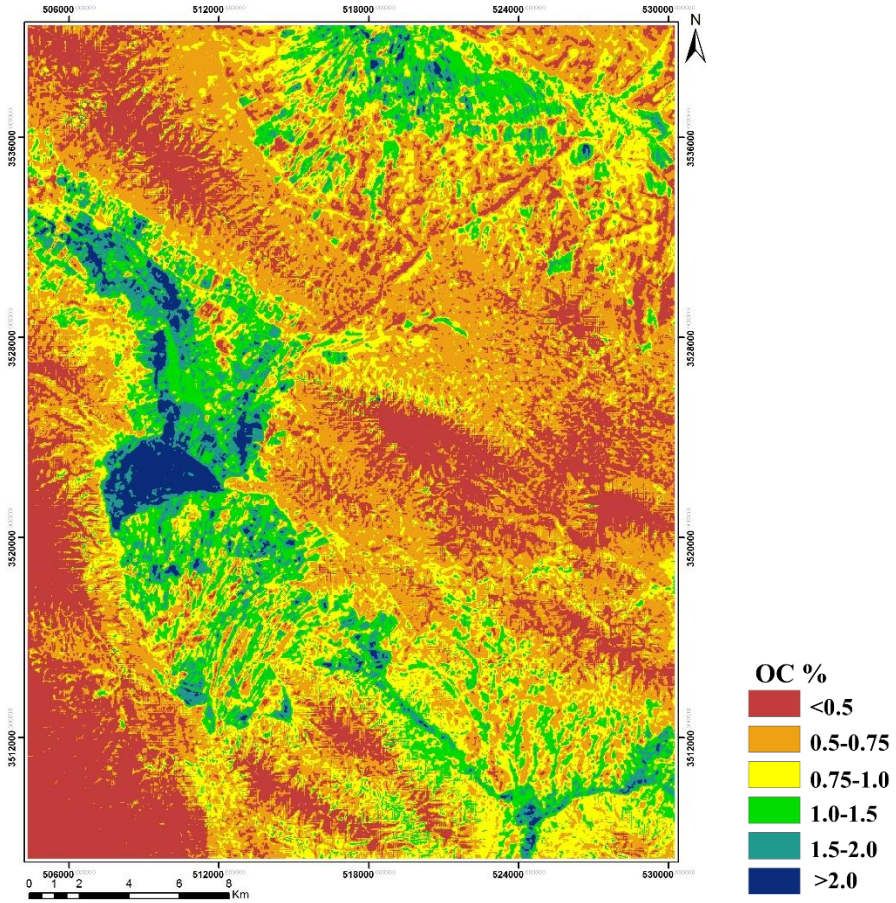
MLR


506000
512000
518000
524000
530000
N
3536000
3528000
3520000
3512000
0 1 2 4 6 8
Km

RF


506000
512000
518000
524000
530000
N
3536000
3528000
3520000
3512000
0 1 2 4 6 8
Km

RT



SMLR

Figure 4- 5- Spatial predicted maps of soil organic carbon content (OC %) across the study area using five data mining approaches.

Prediction maps of SOC content for all five models indicated higher values of SOC in the central to west, northern and southern part of the study area, while lower SOC values were predicted at the higher elevations with mountains and hills (Figure 4-1 and 4-5). This can confirm the influence of elevation and terrain attributes. The previous study by Mishra et al. (2009) proposed the high SOC stocks are located in areas with low gradients (>5%) and high percentages of poorly drained soils.

4.4 Conclusions

This chapter has investigated spatial distributions and variations of soil organic carbon (SOC), calcium carbonate equivalent (CCE) and clay content (Cl) by applying five models including non-linear and linear techniques within the study area in 0-30 cm layer in a semi-arid region of Iran. Among the studied models, the SMLR and RF models had the highest performance to predict CCE, Cl and SOC content respectively. Based on R^2, Random forest is relatively accurate and describes the variation of the SOC content higher than 0.63; while regression tree showed a poor performance. All models could not predict the spatial distributions of clay content properly. The comparison of different models shows that RF can predict relatively spatial distributions of CCE content in this region. The most of models revealed that terrain attributes such as elevation are the most important variable in all studied models. Finally, to improve the accuracy of prediction, we recommend that more observations and denser sampling should be carried out in the whole study area. Alternatively, the study area could better be divided into homogeneous sub-areas, stratifying the area by elevation, and then sampled. The stratified sampling probably will increase the performance of the used models.

Chapter 5

Disaggregation and Updating of Legacy Soil Map

Abstract

In the modern world, the increasing demand for food production, global change and growing population are an enormous challenge that we face. Accurate maps and adequate models are indispensable tools to assist managers, scientists and decision-makers in addressing these challenges. At the level of national or international classification systems, soil polygon maps are present in many locations and area. These valuable legacy soil data need effective methods for disaggregation and incorporation into digital soil mapping approaches at national and regional scales. The objective of this chapter is to disaggregate the legacy soil map of Iran at scale of 1:1,000,000 by three methods of disaggregation, a supervised classification (DSMART algorithm) and two unsupervised classifications (Fuzzy c-means and K-means clustering algorithm). The three approaches and their results at two levels of Soil Taxonomy are discussed and compared with the overall accuracy of the original map sequentially. Field validation indicated that the accuracy of the disaggregated soil maps was lower than that of the conventional soil map at both levels of Soil Taxonomy, but disaggregated approaches produced more detailed maps. The higher overall accuracy of the conventional soil map was due to soil association units which consist of more than one soil class. Fuzzy c-means and DSMART methods produced more accurate and detailed disaggregated soil maps at the great group and subgroup levels respectively. We conclude, that the decision what method to use depends on the map, the level of available information (details of map), expert knowledge and the map's legend such as the composition percentage of soil maps if it is available or not.

5.1 Introduction

In the modern world, increasing demands for food production, global change and growing population are an enormous challenge that we face. At the same time, the priority of land management is to prevent land degradation, restore lands that are already degraded, maintain soil quality, and sustainable use of land resources. Accurate maps and adequate models are indispensable tools to assist managers, scientists and decision-makers in addressing these challenges.

In the context of above challenges and a growing demand for high-resolution spatial soil information for scientists and management, fast and accurate methods for updating legacy and conventional soil maps are essential.

It is known that polygon soil maps or conventional soil maps, in general, are not pure units, which means that the same polygon may contain more than one soil class or component (inclusions) (Soil Survey Division Staff, 1993). Therefore, obtaining information from less detailed scale soil polygon maps is difficult and associated with uncertainty. In these cases, disaggregation of soil polygons could be an alternative to identify soil classes within the same polygon (Holmes et al., 2015; Kerry et al., 2012).

The disaggregation approach produces a refined delineation map in which each soil map unit is assumed ‘pure’, that is containing one and only one soil type or soil class. In consequence, the soil type object and its spatial definition become more consistent with one another. The refined delineation map would therefore be expected to be much more powerful and detailed (Daroussin et al., 2006). Some approaches and methods for disaggregation of soil polygons have been described and explained in Bui and Moran (2001), and also several fundamental ideas and theoretical forms of spatial soil map disaggregation were investigated by McBratney (1998).

The biggest impediment of disaggregation studies is to determine the spatial configuration of the soil classes within each map unit in a quantitative manner. It is often known which soil classes occur in each mapping unit, and sometimes there is also information regarding the relative proportions of each one. In the polygon soil maps, commonly, spatial explicitness and configuration of soil classes within the unit are unknown. This issue is the common difficulty faced in studies seeking the renewal and updating of legacy soil maps (Malone et al., 2017).

At the level of national or international classification systems, soil polygon maps cover many locations and areas. These valuable legacy soil data need effective methods for disaggregation and incorporation into digital soil mapping approaches at national and regional scales. Sampling efforts could reduce by appropriate use of these legacy soil data and utilization of modern digital soil mapping methods (Kerry et al., 2012). Some approaches and advantageous reasons were proposed by De Bruin et al. (1999) and Eagleson et al. (1999) for legacy soil maps and disaggregation of soil polygons.

Some empirical methods accompanied by other methods were investigated by Bui and Moran (2001) for spatial disaggregation of soil polygon maps in the Murray–Darling basin, Australia. To disaggregate soil polygon maps, three approaches were proposed by McBratney (1998): transfer functions, fractal analysis, and pycnophylactic splines.

Several studies proposed clustering methods for spatial disaggregation and prediction of soil types in areas where no soil profiles were available and no detailed information exists on where in the landscape a specific soil type of a complex map unit is located (Häring et al., 2012).

The K-means clustering method was applied by Bui and Moran (2001) to classify soils with Landsat MSS bands, slope position and relief as predictor variables. Fuzzy clustering was used by Yang et al. (2011) to quantify soil–landscape relationships on a 1:20,000 soil map in Canada.

Smith et al. (2010) disaggregated soil maps in British Columbia, Canada using terrain attributes, landform classes, and ecological subzones as predictor variables for fuzzy classification rules.

In cases where soil profiles as training data were not available on the soil polygon maps, unsupervised classifications such as K-means and Fuzzy c-means clustering approaches could be applied. Fuzzy c-means and K-means clustering approaches can be used to generate clusters from environmental covariates and then the soil classes of map legend can be assigned to each cluster (Bui and Moran, 2001; Yang et al., 2011).

On the other hand, when representative soil profiles as training data are available in combination with the legacy soil map, supervised classification is an alternative method for spatial disaggregation and prediction. Comparing to the unsupervised classification, supervised classification has the benefit that is it able to estimate the prediction accuracy and can identify clearly described map units or subunits (Häring et al., 2012). Odgers et al. (2014) present a new algorithm called DSMART which stands for Disaggregation and Harmonisation of Soil Map Units Through Resampled Classification Trees as supervised classification. This approach needs the composition of soil classes in each soil map unit, this information is commonly reported in the map legend. Bui and Moran (2001) reported that the approach to disaggregation relies on the decomposition of map unit descriptions given in legends and companion reports to soil association maps.

As well as in the work of Odgers et al. (2014), the DSMART algorithm has been used in other studies throughout the world. Chaney et al. (2014) used it to disaggregate the entire gridded USA Soil Survey Geographic database. Zund (2014) used DSMART to disaggregate land system mapping units into land units for an area in Central Queensland, Australia.

The national soil map of Iran at scale of 1:1,000,000 is too coarse to show fine-scale patterns or individual soil classes with any degree of legibility, but producing a finer-scale soil map for the entire country is expensive and time-consuming.

On the other hand, the soil observations within the national soil map did not have geographic coordinates. Moreover, this map was made using only 4250 profile observations which were excavated and described in 1990's (Banaei et al., 2005) for the entire country (1.648 million km^2), that means only 2 or 3 observations located in the study area.

Consequently, the supervised classification for disaggregation based on point observations of legacy soil map is impossible, but supervised classification such as DSMART, using the component percentage of map units, is possible.

A few studies have investigated methods to update and disaggregate the national soil map in Iran. The objective of this chapter is to disaggregate the legacy soil map of Iran at scale of 1:1,000,000 by three methods of disaggregation, a supervised classification (DSMART algorithm) and two unsupervised classifications (Fuzzy c-means and K-means clustering algorithm). The three approaches and their results at two levels of Soil Taxonomy are discussed and compared with the overall accuracy of the original map.

5.2 Materials and Methods

5.2.1 Description of the study area

The study area was described in chapter 2. Figure 5-1 demonstrated the map of the study area at a scale of 1:1,000,000 with the soil associations (Banaei et al., 2005; Mohammad, 2000). In this map, the soil polygons contain soil associations and commonly include more than one soil class. The conventional soil map of the country (Iran) at scale of 1:1,000,000 was prepared based on physiography units and did not illustrate soil variation properly (Banaei et al., 2005;

Mohammad, 2000). The lowland landscape (Table 2-1) was delineated as non-soil unit in the conventional map at scale of 1:1,000,000 (Figure 5-1), while the previous study (Gholamzadeh, 2014) and also observed profiles in this study proved that more than 90 percent of the area is Mollisols at the order level and Typic Endoaquolls at subgroup level. Consequently, Typic Endoaquolls were used instead of the non-soil unit.

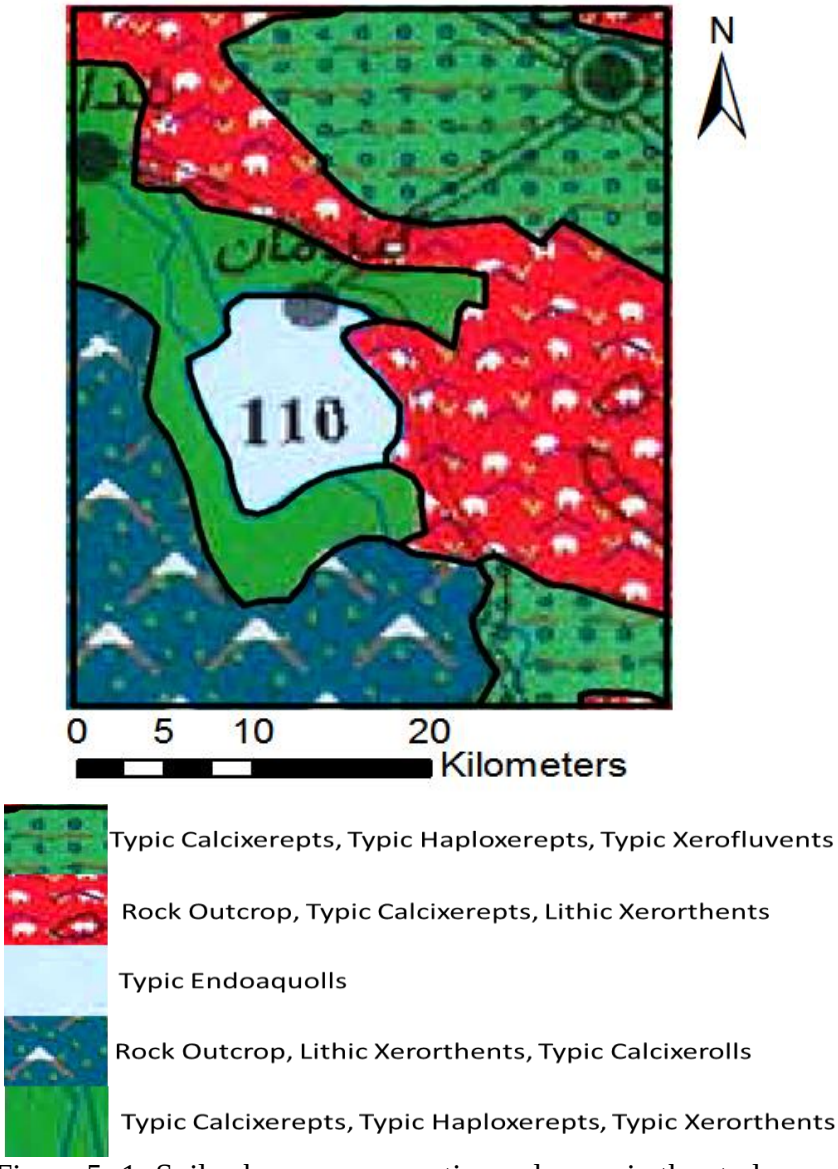


Figure 5- 1- Soil subgroup aggregation polygons in the study area.

In this chapter, we considered the soil map legend (Figure 5-1) to disaggregate soil association polygons. In addition, 15 different quantitative terrain attributes were derived from the DEM and satellite images (Table 3-1) using SAGA GIS software (Olaya, 2004).

5.2.2 Disaggregation approaches

5.2.2.2 K-means disaggregation approach

Hartigan (1975) and McBratney and Gruijter (1992) described the K-means clustering algorithm in detail. The objective of the K-means method is to divide M points in N dimensions (M×N matrix) into K clusters so that the within-cluster sum of squares is minimized. The clustering does not practical to require that the solution has minimal sum of squares against all partitions, except when M, N are small and K=2. Instead, we try to find "local" optima, solutions such that no movement of a point from one cluster to another will reduce the within-cluster sum of squares.

The algorithm requires as input a matrix of M points in N dimensions and a matrix of K initial cluster centres in N dimensions.

K-means clustering runs through the following steps (Hartigan and Wong, 1979):

1- The data points X_i and the number of clusters K are determined. $i=1,2,\ldots,n$

2- The next step could be to continue as below:

   - The data points are randomly allocated into clusters, then each C_j cluster centres are calculated $j= 1,2,\ldots k$
   - Determination of centre clusters (C_j)

3- Within each cluster, the Euclidean distance of each data point to its cluster centre is calculated as the sum of the squared errors. Here, “error” indicates the distance of each

sample to the nearest centre of each cluster. The error sum of squares (ESS) then is calculated as the sum over all clusters by the following formula:

$$\text{ESS} = \text{J} = \sum_{j=1}^{k} \sum_{i=1}^{n} \|Xij - Cj\|^2 \qquad \text{(Eq. 5.5)}$$

where C_j is the j-th cluster centre and X_{ij} is the data point in the j-th cluster.

4- Re-assigned the data points to the nearest cluster and update the center of each cluster after each allocation.

5- Repeat steps 3 and 4 until reassignment of point data to the clusters stop (minimum ESS obtained).

The "cclust" R package was used to run the K-means clustering algorithm (R Development Core Team, 2013).

5.2.2.2 Fuzzy C-means disaggregation approach

The Fuzzy C-means algorithm was described by in detail McBratney and Gruijter (1992) and Bezdek (2013). In the K-means clustering, each sample or variable belongs to one cluster, and the boundary between clusters is distinct, but in the Fuzzy C-means algorithm each sample or variable at least belongs to two clusters.

The centre of clusters determined by minimizing the objective function as follows (Bezdek, 2013):

$$\text{Jm} = \sum_{i=1}^{n} \sum_{j=1}^{c} u_{ij}^{m} \|Xi - Cj\|^2 \qquad 1 \leq m < \infty \qquad \text{(Eq. 5.1)}$$

where m is any number greater than 1, Uij is the membership of Xi in the j cluster, Xi-th sample, Cj is the centre of j-th cluster and $\|Xi - Cj\|$ represents the similarity of samples and the center of each cluster.

In fact, this component ($\|Xi - Cj\|$) shows a function of the distance between samples and the centres of clusters; that distance can be the Euclidean, Mahalanobis or Manhattan distance. The

optimization of the fuzzy C-means clustering algorithm was done by frequent recalculation of the objective function under different data assignments to fuzzy clusters. The membership of U_{ij} and the cluster centre of C_j is determined by the following function:

$$uij = \frac{1}{\sum_{k=1}^{c}\left(\frac{\|Xi-Cj\|}{\|Xi-Ck\|}\right)^{\frac{2}{m-1}}} \quad \text{(Eq. 5.2)}$$

$$cj = \frac{\sum_{i=1}^{n} u_{ij}^{m}.x_i}{\sum_{i=1}^{n} u_{ij}^{m}} \quad \text{(Eq. 5.3)}$$

where the value of ε is between zero and one and k is the iteration. The algorithm includes the following steps:

1- Initial assignment of value to the matrix of $U=[u_{ij}]$ or $U^{(0)}$

2- In the k-th step, calculating the centre of vectors C(k)=[cj] in the matrix $U^{(k)}$ and $U^{(k+1)}$

$$cj = \frac{\sum_{i=1}^{n} u_{ij}^{m}.x_i}{\sum_{i=1}^{n} u_{ij}^{m}} \quad \text{(Eq. 5.4)}$$

3- Updating the matrixes of $U^{(k)}$ and $U^{(k+1)}$

4- If $max_{ij}\left[\left|u_{ij}^{k+1} - u_{ij}^{k}\right|\right] < \varepsilon$ occurs, the algorithm will stop; otherwise the algorithm will re-run from the second step.

The construction of fuzzy membership functions follows the methods described in (Zhu et al., 2010). The "e1017" R package was used to run the fuzzy C-means clustering algorithm (R Development Core Team, 2013).

5.2.2.3 Disaggregating and harmonizing soil map units through resampled classification trees (DSMART) approach

DSMART provides and describes the methodology to disaggregate legacy soil maps and soil map associations into their components for an area of interest. Therefore, by disaggregating soil map units it becomes possible to map individual soil classes and soil types.

A comprehensive explanation of the DSMART algorithm is provided by Odgers et al. (2014) and Malone et al. (2017).

DSMART predicts the spatial distribution of soil classes by disaggregating the soil map units of a soil polygon map. Here soil map units or soil polygons are entities consisting of a defined set of soil classes which occur together in a certain spatial pattern and in an assumed set of proportions. The disaggregated soil class distribution by DSMART approach presents probable soil classes for each one raster cell, so the results contain a set of numerical raster surfaces, with one raster per soil class. The data representation for each soil class is given as the probability of occurrence. According to Malone et al. (2017), in order to generate the probability surfaces, a re-sampling approach is used to generate *n* realizations of the potential soil class distribution within each map unit. Then at each grid cell, the probability of occurrence of each soil class is estimated by the proportion of times the grid cell is predicted as each soil class across the set of realizations. The procedure of the DSMART algorithm can be summarized in 6 main steps (Odgers et al., 2014):

1- Draw *m* random samples (grid cell location) from each soil map polygon.

2- Allocate a soil class to each sampling point, using:

   - Weighted random allocation from soil classes in relevant map unit.

   - Relative proportions of soil classes within map units are used as the weights.

3- Use sampling points and covariate values at these points to build a decision tree to prediction spatial distribution of soil classes.

4- Apply the decision tree across the mapping extent using covariate layers.

5- Repeat steps 1–4 *i* times to produce *i* realizations of soil class distribution.

6- Using *i* realizations generate probability surfaces for each soil class.

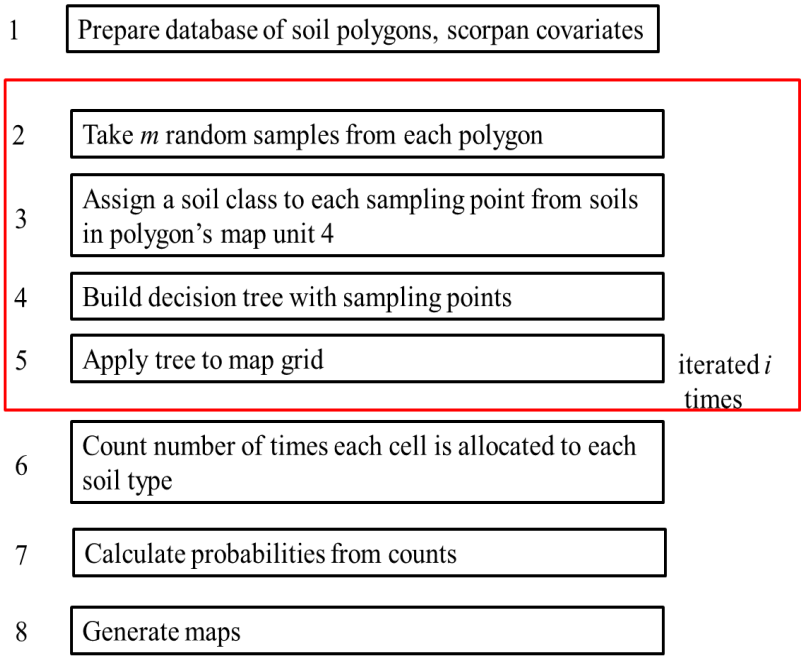


Figure 5-2 Schematic overview of the DSMART algorithm (Odgers et al., 2014).

DSMART allows the user to set project-specific parameters. The random sample size *m* was set to 15. The model type that Odgers et al. (2014) used was the C4.5 decision tree algorithm which was introduced by (Quinlan, 1993a). The type of data mining algorithm implemented in DSMART is not prescriptive; as long as it is robust and, importantly, computationally efficient. For example, Chaney et al. (2014) used Random Forest models (Breiman, 2001) in their implementation of DSMART. In this study, the C4.5 decision tree algorithm was applied.

The DSMART approach has already been written in the C++ and Python computing languages. It is also available in an R package, which was developed at the Soil Security Laboratory. Regardless of computing language preference, DSMART requires three main sources of data:

1- The soil map unit polygons that will be disaggregated (Figure 5-1).

2- Information about the soil class composition of the soil map unit polygons. In this study, for those soil polygons that consist of 3 components, a relative contribution of 50, 30 and 20 percent was considered for each soil class respectively.

3- Geo-referenced raster covariates representing the scorpan factors which have complete and continuous coverage of the mapping extent. For this source data, 15 quantitative terrain attributes were considered as scorpan factors.

Implementation of DSMART algorithm in R program contains two working functions and packages: *dsmart* and *dsmartR* (R Development Core Team, 2013).

5.2.3 Validation of disaggregation approaches

The clusters obtained by unsupervised classification (Fuzzy c-means and K-means clustering approaches) are assigned to each soil classes in the map legend according to expert knowledge and soil type distribution in the landscape.

In all applied disaggregation approaches, Rock Outcrop units in the map of 1:1,000,000 at the subgroup and great group levels have been named as Lithic Xerorthents and Xerorthents, respectively.

To undertake the validation, the disaggregated soil maps predicted by DSMART, fuzzy C-means and K-means clustering were validated by 125 soil profiles samples collected (see chapter 2, section 2.2.3). Overall accuracy was used to evaluate the performance of all disaggregation approach.

To comparison the performance of disaggregation approaches, the overall accuracy was also computed with 125 soil profiles for the legacy soil map at scale of 1:1,000,000. Overall accuracy has a range between zero and one where a good map has a value of map purity close to 1 (Behrens et al., 2010).

The detail depicted on a map informs whether it is useful for subsequent spatial studies. A map with high spatial detail may be useful in the spatial analysis if its accuracy is high as well. We derived a proxy for map intricacy using the Shannon entropy index (S) using a moving

window at great group and subgroup levels for 3 soil disaggregated maps and also for the legacy soil map (more information in chapter 3).

By multiplying Shannon entropy index (S) and overall accuracy, a combined index for final assessment of the best disaggregation approach was calculated. This index was calculated for all disaggregation approaches and the legacy soil map.

5.3 Results and Discussion

The study area covers 86000 ha (860 km^2) of soil map at scale of 1:1,000,000. Based on soil map legend at the subgroup level it consists of 7 soil subgroups: Typic Calcixerepts, Typic Haploxerepts, Typic Xerofluvents, Lithic Xerorthents, Typic Endoaquolls, Typic Calcixerolls, Typic Xerorthents and one non-soil unit including Rock outcrop, which is considered as Lithic Xerorthents in this study. There are six great groups and one non-soil unit including Rock outcrop as well (Figure 5-1).

Typic Xerofluvents presented on the soil map at scale of 1:1,000,000 were not observed in the field observation locations, on the other hand, in addition to the six subgroups included in the legacy map, six extra subgroups were observed in the study area (Table 3-1). Discrepancies and mismatches between the number of soil subgroups in the field observation and the legacy soil map are due to low accuracy and large scale of the legacy soil map.

Because information on soil profiles from which the legacy soil map was produced was not available, the field observations were used as validation points. The validation points (chapter 2) were collected using a cLHS sampling strategy and were intended to cover all the soil types in the area. Consequently, the field observations were also used to evaluate the accuracy of the legacy soil map.

The validation criteria for legacy soil map and disaggregated soil maps are summarized in Table 5-1. The overall accuracy of the legacy soil map was 52% and 65% at the subgroup and great group levels respectively. While the overall accuracy was high, the map was not reliable because each soil polygons consist of more than one soil classes. The inability of this map to present soil variability led to the lowest Shannon entropy index (S) (Table 5-1). Based on Shannon index (S) the original soil map did not show much more spatial detail. Yang et al. (2011) stated that commonly the original map shows less information than the updated map.

5.3.1 Unsupervised classification

The original legacy soil map (Figure 5-1) showed that 7 and 6 soil classes existed in the study area at the subgroup and great group levels respectively. Additionally, one non-soil unit is recognizable that in both unsupervised classifications is assigned to Lithic Xerorthents and Xerorthents at the subgroup and great group levels respectively. Therefore by clustering approaches, 8 clusters for subgroup and 7 clusters for the great group level were generated. Consequently, the disaggregated soil maps were produced with 7 and 6 soil classes at the subgroup and great group levels.

5.3.1.1 Fuzzy c-mean disaggregation approach

By Fuzzy c-means, the environmental covariates, representing the scorpan factors, were clustered. Then 8 and 7 fuzzy membership maps of the environmental clusters based on Euclidean distance, the fuzziness degree of 1.3 and with an iteration of 100 were produced for the soil subgroup and great group respectively. Afterwards, based on the maximum degree of membership from 8 maps of membership (subgroup level) and 7 maps of membership at great group level, two maps with 8 and 7 clusters were produced for soil subgroup and great group respectively. Each cluster is assigned to one soil subgroup or great group by expert knowledge of

the relationships between the soil and the environmental conditions. Zhu et al. (2001) and Qi and Zhu (2003) reported that the relationships between the soil and the environmental conditions are embedded in associations between environmental clusters and mapped soil types. Therefore this knowledge can be extracted and quantified to be useful in digital soil mapping. In this chapter, this knowledge is obtained through the construction of fuzzy membership functions. These fuzzy membership functions describe how similarity between a local soil and the typical case of a given soil type will change as environmental conditions change (Zhu, 1999).

The disaggregated maps at the subgroup and great group levels by fuzzy c-means are presented in Figure (5-3). Estimation of Fuzzy c-means algorithm performance by 125 field validation points gave an overall accuracy of 34% and 43% at the subgroup and great group levels respectively. This approach resulted in a low Shannon index (Table 5-1).

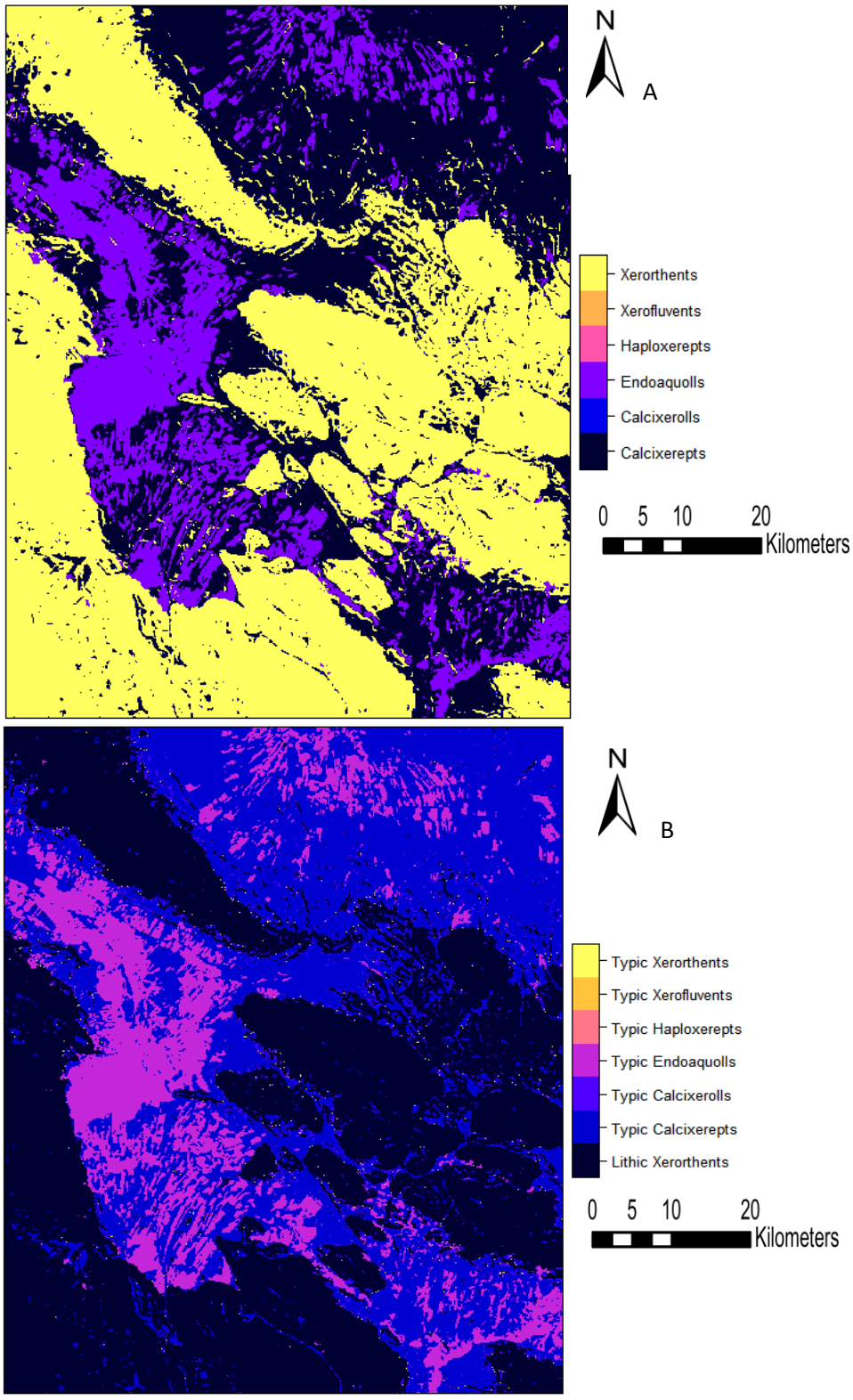


Figure 5- 3- The disaggregated soil map using the fuzzy c-means clustering approach, A) great group level, B) subgroup level.

5.3.1.2 K-mean disaggregation approach

The disaggregated maps produced by this approach have been validated by 125 field observations (Figure 5-4). The results showed that the overall accuracy of the disaggregated maps are 15% and 31% in the subgroup and great group levels respectively.

The Shannon index indicated that this method in both Soil Taxonomy levels produced maps with higher diversity rate (Table 5.1) and this increase in the Shannon index led to a slight increase in the combined index for this method. Comparing to the field observations Figure (5-4) shows that the K-mean disaggregation approach overestimated Xerofluvents and Haploxerepts at the great group level and also Typic Endoaqoulls at the subgroup level.

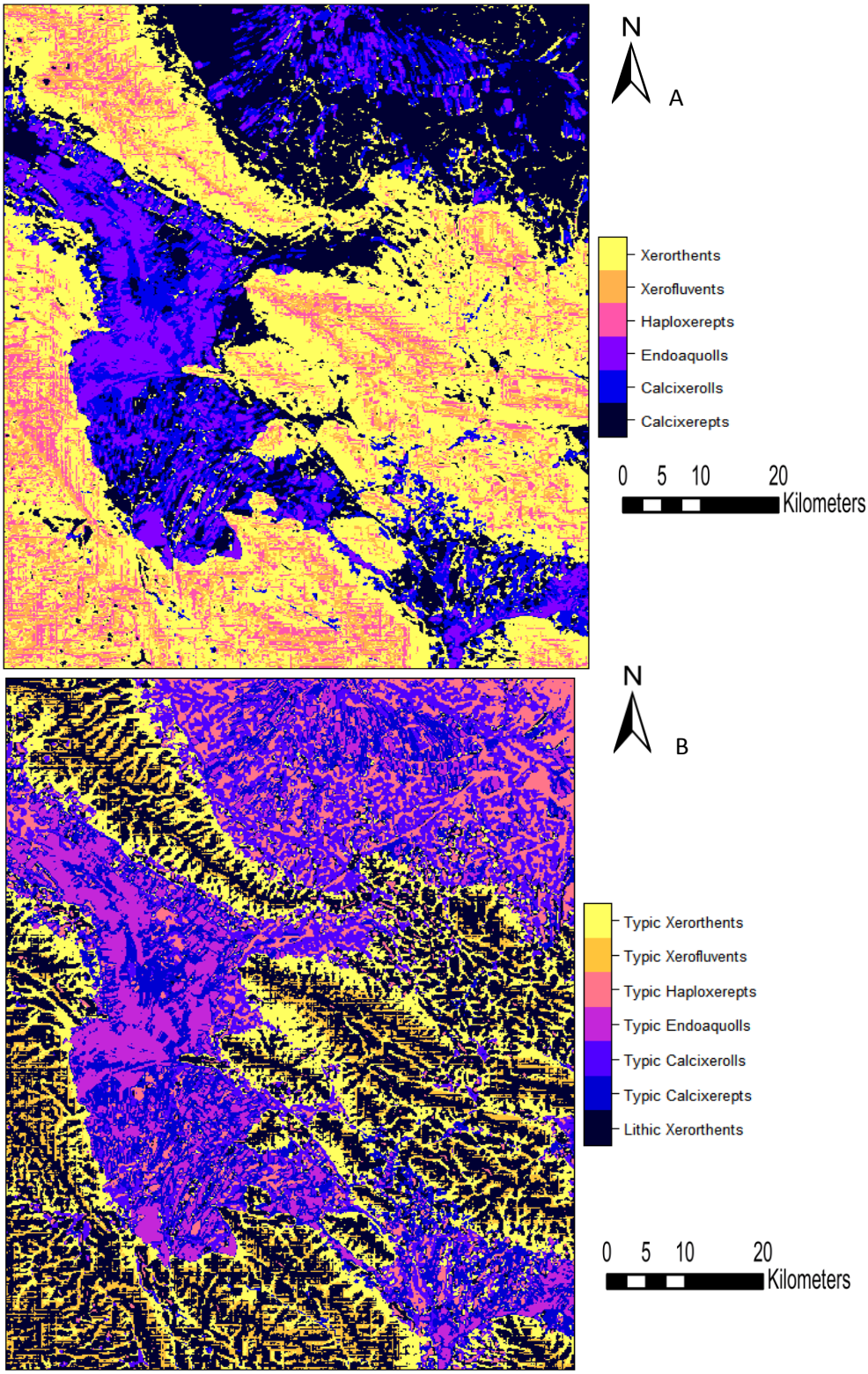


Figure 5- 4- The disaggregated soil map using the K-means clustering approach, A) great group level, B) subgroup level.

5.3.2 Supervised classification

The DSMART was run on the 5 soil polygons (Figure 5-1) for n=10 iterations at m=15 sampling points per polygon and produced probability soil classes for the 6 and 7 soil classes at the great group and subgroup levels respectively. The Rock outcrop unit was also classified and predicted as Lithic Xerorthents subgroup, the same as in the unsupervised classification.

The most probable soil classes and the confusion index maps were also calculated. Validation on 125 observed soil profiles indicated that DSMART approach has 36% and 33% overall accuracy at great group and subgroup levels. This method produced the most detailed disaggregated maps with the highest value of Shannon entropy (Table 5-1).

The study of Chaney et al. (2014) in the USA showed that the DSMART approach can be quite a powerful and successful method for disaggregating a legacy soil map. Though DSMART has many advantages, this method needs specific inputs, particularly in regards to the soil class compositions and their relative proportions within mapping units. In some cases, this kind of information is not easily available and needs to be approximated by some means. The other disadvantage of DSMART is the computational effort to generate disaggregated prediction maps which could be a burden (Malone et al., 2017).

Table 5- 1 Validation criteria for different disaggregation approaches at two levels of Soil Taxonomy in the study area.

	Overall accuracy	Shannon entropy	Combined index
	Great group		
Legacy soil map	0.65	0.55	0.36
DSMART	0.36	0.69	0.25
F c-means	0.43	0.61	0.26
K-means	0.31	0.67	0.21
	Subgroup		
Legacy soil map	0.52	0.55	0.29
DSMART	0.33	0.70	0.23
F c-means	0.34	0.63	0.21
K-means	0.15	0.68	0.10

Each location where predictions are made has a probability series consisting of the most probable soil classes and their probability (Odgers et al., 2014). The most probable soil classes (Figure 5-5) and the confusion index between their probabilities of occurrence (Figure 5-6) were mapped at the subgroup level.

The number of pixels which predicted as Typic Calcixerolls and Typic Xerofluvents almost agreed with the field observations. Typic Xerofluvents had not been observed in the field and a small number of profiles were observed as Typic Calcixerolls. Consequently, according to the expert knowledge, the prediction of DSMART method for these two subgroups are close to reality (Figure 5-5). On the disaggregated map, Lithic Xerorthents, Typic Calcixerepts and Typic Endoaquolls were identified as the three most probable soil classes (Figure 5-5).

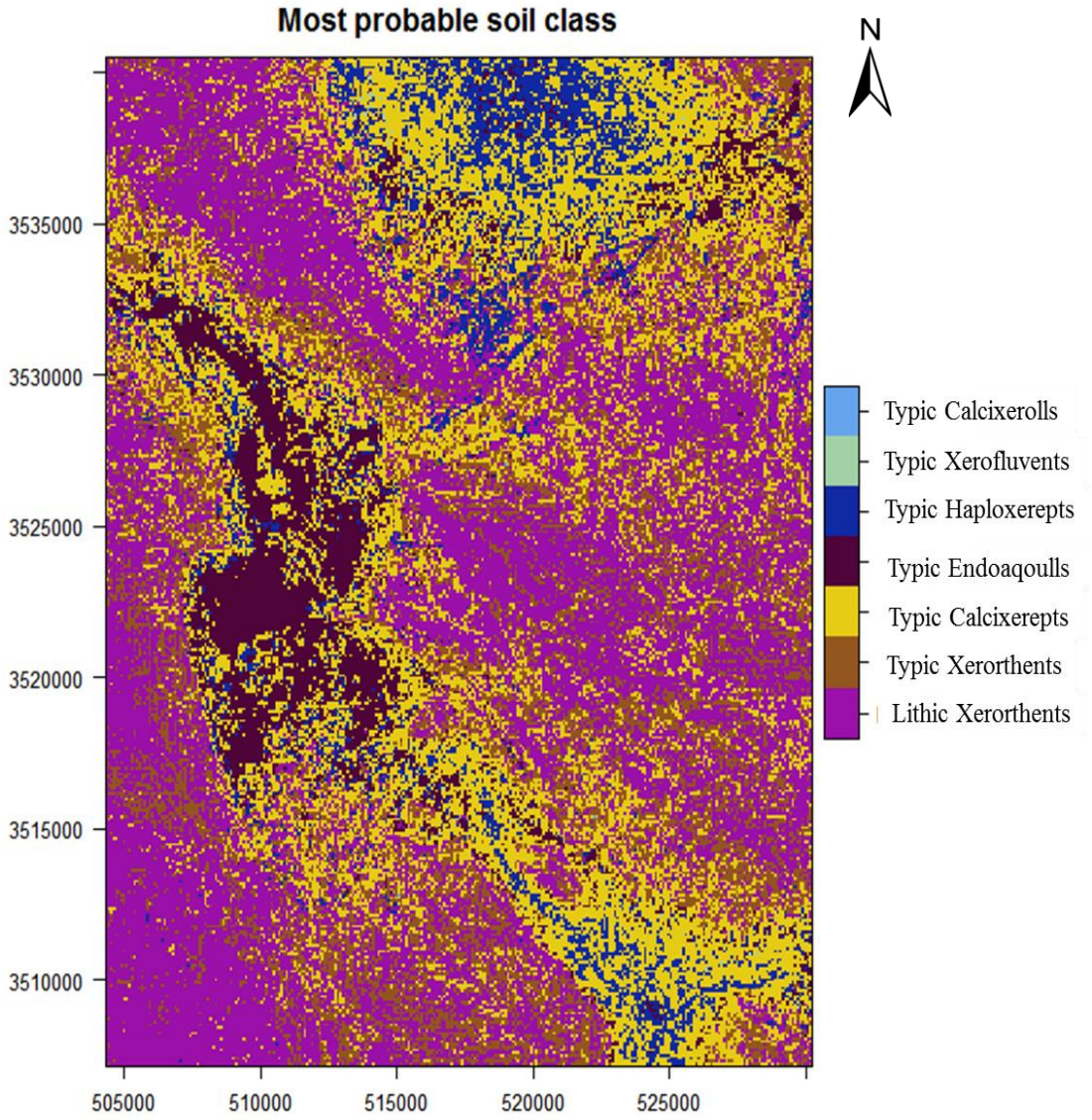


Figure 5- 5- Map of the most probable by DSMART method for soil class at subgroup level.

The confusion index provided an estimate of the uncertainty, therefore this index essentially measures how similar to classification is between (in most cases) most probable and second-most probable soil class predictions at a pixel (Odgers et al., 2014). For instance, if the difference in probability between the most probable and second-most-probable soil class is small (e.g. 0.1), we are more confused about which is the "correct" class than if the difference in probability between the most probable and second-most-probable soil class is large (e.g. 0.9) (Figure 5-6).

In the central and western part of the study area, the probabilities of the most probable soil classes were low and confusion index almost being around 0.40 and occasionally above 0.65; this is the area were the Typic Endoaquolls subgroup was predicted. This means that those soils were predicted with high uncertainty. For the area with confusion index higher than 0.60 (most part of the study area), the probability of the most probable soil was high and those soils were predicted with higher certainty. The confusion index could enhance our understanding of soil-landscape relationships.

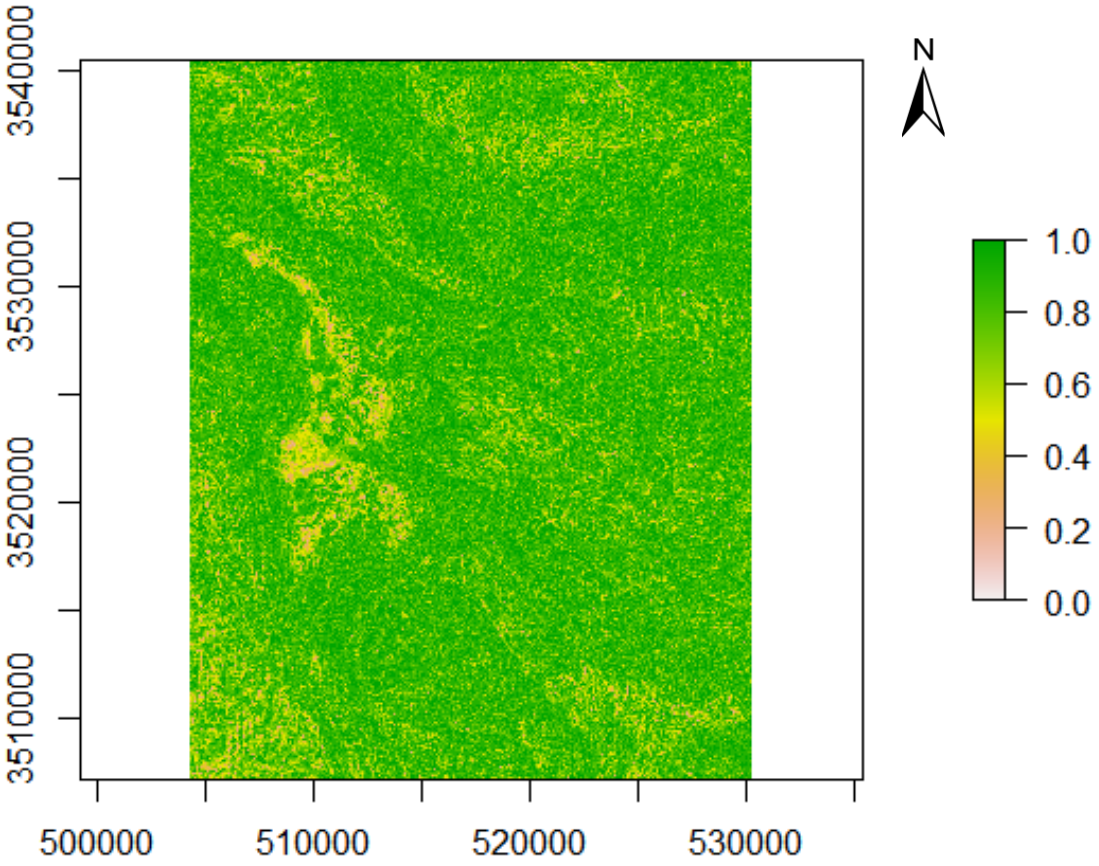


Figure 5- 6- Confusion index of DSMART method for soil subgroup predictions

Tables (5-2 and 5-3) present a summary of the covariate usage across all 10 decision trees. The covariate usage for a given covariate is the proportion of leaves in each decision tree which

require the covariate at some point in one of the if-then rules descending from the root of the tree (Odgers et al 2014). At the subgroup level, elevation, aspect and Multi-resolution Valley Bottom Flatness Index had the highest usage in all the decision trees. Almost all of image satellite covariates were not used frequently in the decision trees. The covariates with high values of standard deviation of usage indicate that some trees used it heavily but others did not (Table 5-2). The mean usage of covariates decreased at great group level. Elevation, aspect and topographic wetness index had the highest usage in all the decision trees (Table 5-3). The importance of covariates in the prediction of soil classes can be concluded by the mean usage of covariates in the decision trees.

Table 5- 2- Summary of covariate usage across 10 decision trees at subgroup level. The covariates are briefly explained in Table 3-1.

Covariate	n	Mean usage (%)	StDev usage (%)	Minimum usage (%)	Maximum usage (%)
El	10	95.00	8.23	80.83	100.00
As	10	48.83	22.05	9.17	78.33
MrVBF	9	43.60	36.70	0.00	100.00
WI	10	37.92	22.22	9.17	70.00
TWI	9	37.50	27.93	0.00	93.33
PrCu	10	34.83	25.21	10.83	100.00
SAVI	9	32.92	19.48	0.00	67.50
MrRTF	10	30.08	21.52	6.67	61.67
CI	10	29.75	19.36	11.67	72.50
PlCu	7	24.42	23.63	0.00	69.17
Sl	8	23.25	25.50	0.00	72.50
Cu	5	21.67	30.26	0.00	87.50
PVI	8	21.50	22.65	0.00	66.67
RVI	6	12.42	19.31	0.00	61.67
NDVI	1	5.08	16.07	0.00	50.83

Table 5- 3- Summary of covariate usage across 10 decision trees at great group level. The covariates are briefly explained in Table 3-1.

Covariate	n	Mean usage (%)	StDev usage (%)	Minimum usage (%)	Maximum usage (%)
El	10.00	71.83	29.41	15.00	100.00
As	10.00	58.17	20.55	30.83	94.17
TWI	9.00	50.90	36.00	0.00	100.00
MrRTF	10.00	48.83	27.85	7.50	100.00

CI	9.00	47.50	27.59	0.00	85.00
PVI	9.00	40.30	40.50	0.00	100.00
WI	9.00	40.20	36.60	0.00	100.00
PlCu	8.00	33.58	31.15	0.00	85.00
MrVBF	9.00	30.40	33.40	0.00	100.00
Cu	7.00	29.40	33.80	0.00	100.00
SAVI	9.00	23.83	17.76	0.00	47.50
PrCu	8.00	21.42	20.34	0.00	50.83
RVI	3.00	21.20	41.70	0.00	100.00
Sl	7.00	13.58	19.73	0.00	61.67
NDVI	1.00	3.42	10.81	0.00	34.17

5.3.3 Validation and models comparison

Generally, the overall accuracy decreased for all approaches because more soil classes were observed in the field observations in comparison to the soil classes of soil map at scale of 1:1,000,000. This led to obtaining low overall accuracy. Among all applied approaches, DSMART produced and generated the most detailed maps of soil classes based on Shannon entropy and the Fuzzy c-means method generated the least detailed maps of soil classes (Table 5-1).

Unlike the unsupervised approaches, DSMART method firstly provides a framework to update and legacy soil maps, to realize and acquire soil property information from existing polygon soil maps (Malone et al., 2017; Odgers et al., 2015). The comparisons between the disaggregated soil maps and conventional soil map showed that all disaggregated soil maps contained much greater spatial detail.

Although Yang et al. (2011) reported that updated soil map by Fuzzy c-means was more detailed and accurate, this study did not achieve more accurate disaggregated soil maps. There is at least one main reason for this: The proposed methods in this study used the soil classes and polygons from the conventional soil map, which showed great difference with field observations (more soil classes in comparison with soil classes in the conventional soil map). None of the

approaches produced a disaggregated map with greater overall accuracy than the conventional soil map but all disaggregated methods generated more detailed soil maps.

Odgers et al. (2014) reported that only 22.5% of field observations were predicted correctly as the most probable soil by the DSMART method. The validation results obtained from all approaches showed that they did not have similar predictive capability. The unsupervised classification, assigning soil classes to clusters, was very dependent on expert knowledge and in the DSMART method, similar predictive capability of models depends on soil information and components of the conventional soil map.

5.4 Conclusions

In this chapter three approaches were evaluated to disaggregate the conventional soil map. It is possible to disaggregate a very large areas of interest. For disaggregation purpose of legacy soil maps by unsupervised classifications, we had to use our expertise and expert knowledge to generate disaggregated maps by their legends. On the other hand, supervised classification like DSMART approach needs more details of legacy soil map that in some cases is not available or has to be estimated by the users. So, deciding what method to use is a function of the map, the level of available information, expert knowledge and the map's legend. The DSMART results showed that topographic attributes have a significant influence on the prediction of disaggregated soil classes at both taxonomic levels of prediction. The proposed methods are appropriate in situations where the study area contains legacy soil sample data that were used to produce the conventional soil map, and after disaggregation the validation is carried out using the legacy soil sample data. In this study area, georeferenced legacy soil samples data were not available, therefore validation was carried out using new field observations, which reduced the overall

accuracy. In a case study like this situation there is much work left to do and also there are great opportunities for applying the outputs of unsupervised and supervised classification.

Chapter 6

Summary and Future Suggestions

6.1 Summary

This study involves mathematical models and expert knowledge to test the capability of these models to predict soil map classes and some soil properties, to disaggregate soil classes of a conventional soil map and to identify the relationship between soils and landscapes. So, the objectives of this thesis were:

(i) to evaluate soil evolution in the landscape and the effect of geomorphic surfaces and landforms on soil;

(ii) to predict soil classes, compare digital soil mapping and conventional soil mapping methods with respect to time, cost and accuracy;

(iii) to predict some soil properties and compare linear and nonlinear prediction models.

Finally, objective (iv) of this study aimed to disaggregate soil associations of conventional soil map to their components.

The following main conclusions can be drawn from different parts of this study:

- Pedodiversity indices increased from the soil order to the soil subgroup level. The relationships between pedodiversity indices and area showed that probably additional soil classes would be observed when the number of observations would increase. So it can be deduced that possibly the number of observations was insufficient in the studied area.
- In most predicted maps using DSM, the best results were obtained using the combination of terrain attributes and the geomorphology map.

- Using the geomorphology map in DSM was not significant but, employing it increased map purity and the Kappa index, and led to a decrease in the ‘noisiness’ of soil maps.
- For prediction of soil classes at higher taxonomic levels, Multinomial Logistic Regression had better performance and Random Forest showed better performance at lower taxonomic levels.
- Multinomial Logistic Regression was less sensitive than Random Forest to a decrease in the number of training observations.
- Based on a combined index of map purity, map information content, and map production cost, Multinomial Logistic Regression was identified as the most effective approach.
- The Stepwise Multiple Linear Regression and Random Forest models showed the highest performance to predict Calcium Carbonate Equivalent, Clay and SOC content. Results revealed that all models could not predict the spatial distributions of clay content properly.
- Elevation is the most importance variable for all studied models to predict soil properties.
- Disaggregated approaches produced more detailed maps compared to the original map.
- Fuzzy c-means and DSMART methods produced more accurate and detailed disaggregated soil maps at the great group and subgroup levels respectively.

6.2 Suggestions for future research

This study presents different methods of digital soil mapping (DSM) for soil classes and properties, and disaggregation approaches. Therefore, there are a few considerations for future research. First, this study did not use high-resolution terrain attributes, so future research could be evaluated for the effect of ancillary data resolutions on model performance. Secondly, for prediction of soil properties, this research recommended that more observations and denser sampling should be carried out in the study area. In the area with high topographic variation, the study area could probably have been better divided into homogeneous sub-area. Third, in map disaggregation studies, data from georeferenced legacy soil samples will increase the validation accuracy, so future research could use legacy soil profiles to validate disaggregated soil maps. DSMART could be the appropriate disaggregated method for continued mapping at the national scale to producing soil classes at the lower taxonomic level (great group and in some cases sub group), for instance in relation to the current nationwide 1:1,000,000 soil map. DSMART has additional functionality which was not tested in this study. For instance, computing second and third most probable soil in each grid cell. Afterwards calculate accuracy for 3 most probable soil in each grid cell which enhance the reliability of final disaggregated soil map.

References

Abdel-Kader, F.H., 2011. Digital soil mapping at pilot sites in the northwest coast of Egypt: A multinomial logistic regression approach. The Egyptian Journal of Remote Sensing and Space Science, 14(1), 29-40.

Adhikari, K., Hartemink, A.E., Minasny, B., Kheir, R.B., Greve, M.B., Greve, M.H., 2014. Digital mapping of soil organic carbon contents and stocks in Denmark. PLoS One, 9(8), e105519.

Akaike, H., 1974. A new lock at the statistical model identification. IEEE T. Automat. Contr, 19(716-723).

Alsharhan, A., Rizk, Z., Nairn, A.E.M., Bakhit, D., Alhajari, S., 2001. Hydrogeology of an arid region: the Arabian Gulf and adjoining areas. Elsevier.

Archer, K.J., Kimes, R.V., 2008. Empirical characterization of random forest variable importance measures. Computational Statistics & Data Analysis, 52(4), 2249-2260.

Banaei, M.H., Bybordi, M., Moameni, A., Malakouti, M.J, 2005. The soil of Iran: New Achievements in Perception, Management and Use. SANA Publishing, Tehran, Iran, 482 pp. (In Persian).

Barnes, E., Baker, M., 2000. Multispectral data for mapping soil texture: possibilities and limitations. Applied Engineering in Agriculture, 16(6), 731-746.

Barthold, F., Wiesmeier, M., Breuer, L., Frede, H., Wu, J., Blank, F., 2013. Land use and climate control the spatial distribution of soil types in the grasslands of Inner Mongolia. Journal of arid environments, 88, 194-205.

Beckett, P.B., SW, 1978. Use of soil and land-system maps to provide soil information in Australia. Commonwealth Scientific and Industrial Research Organization.

Behera, S.K., Shukla, A.K., 2015. Spatial distribution of surface soil acidity, electrical conductivity, soil organic carbon content and exchangeable potassium, calcium and magnesium in some cropped acid soils of India. Land Degradation & Development, 26(1), 71-79.

Behrens, T., Förster, H., Scholten, T., Steinrücken, U., Spies, E.D., Goldschmitt, M., 2005. Digital soil mapping using artificial neural networks. Journal of plant nutrition and soil science, 168(1), 21-33.

Behrens, T., Schneider, O., Lösel, G., Scholten, T., Hennings, V., Felix-Henningsen, P., Hartwich, R., 2009. Analysis on pedodiversity and spatial subset representativity—the German soil map 1: 1,000,000. Journal of Plant Nutrition and Soil Science, 172(1), 91-100.

Behrens, T., Zhu, A.-X., Schmidt, K., Scholten, T., 2010. Multi-scale digital terrain analysis and feature selection for digital soil mapping. Geoderma, 155(3), 175-185.

Brewer, S., Liaw, A., Wiener, M., Liaw, M.A., 2015. Package 'randomForest'.

Bezdek, J.C., 2013. Pattern recognition with fuzzy objective function algorithms. Springer Science & Business Media.

Bockheim, J.G., 2015. Classification and development of shallow soils (< 50cm) in the USA. Geoderma Regional, 6, 31-39.

Boettinger, J., Ramsey, R., Bodily, J., Cole, N., Kienast-Brown, S., Nield, S., Saunders, A., Stum, A., 2008. Landsat spectral data for digital soil mapping, Digital Soil Mapping with limited data. Springer, pp. 193-202.

Bogunovic, I., Pereira, P., Brevik, E.C., 2017. Spatial distribution of soil chemical properties in an organic farm in Croatia. Science of the Total Environment, 584, 535-545.

Borujen Geology Map. 1990. Borujen geology map 1:100,000, http://www.ngdir.ir/Downloads/Downloads.asp.

Borujeni, I.E., Mohammadi, J., Salehi, M., Toomanian, N., Poch, R., 2010. Assessing geopedological soil mapping approach by statistical and geostatistical methods: a case study in the Borujen region, Central Iran. Catena, 82(1), 1-14.

Bouma, J., 2006. Soil functions and land use. In: Certini, G. and Scalenghe, R., 2006. Soils: Basic Concepts and Future Challenges. Cambridge University Press, U.K.

Breiman, L., 2001. Random forests. Machine learning, 45(1), 5-32.

Breiman, L., Friedman, J., Stone, C.J., Olshen, R.A., 1984. Classification and regression trees. CRC press.

Brungard, C.W., 2009. Alternative Sampling and Analysis Methods for Digital Soil Mapping in Southwestern Utah2.

Brungard, C.W., Boettinger, J.L., Duniway, M.C., Wills, S.A., Edwards, T.C., 2015. Machine learning for predicting soil classes in three semi-arid landscapes. Geoderma, 239, 68-83.

Brus, D., Kempen, B., Heuvelink, G., 2011. Sampling for validation of digital soil maps. European Journal of Soil Science, 62(3), 394-407.

Bui, E., Henderson, B., Viergever, K., 2009. Using knowledge discovery with data mining from the Australian Soil Resource Information System database to inform soil carbon mapping in Australia. Global biogeochemical cycles, 23(4).

Bui, E.N., Moran, C.J., 2001. Disaggregation of polygons of surficial geology and soil maps using spatial modelling and legacy data. Geoderma, 103(1), 79-94.

Byrne, J.M., Yang, M., 2016. Spatial variability of soil magnetic susceptibility, organic carbon and total nitrogen from farmland in northern China. Catena, 145, 92-98.

Cantón, Y., Solé-Benet, A., Lázaro, R., 2003. Soil–geomorphology relations in gypsiferous materials of the Tabernas Desert (Almería, SE Spain). Geoderma, 115(3), 193-222.

Chaney, N., Hempel, J., Odgers, N., McBratney, A., Wood, E., 2014. Spatial disaggregation and harmonization of gSSURGO, ASA, CSSA and SSSA international annual meeting, Long Beach. ASA, CSSA and SSSA.

Costantini, E.A., L'Abate, G., 2016. Beyond the concept of dominant soil: Preserving pedodiversity in upscaling soil maps. Geoderma, 271, 243-253.

Cutler, D.R., Edwards Jr, T.C., Beard, K.H., Cutler, A., Hess, K.T., Gibson, J., Lawler, J.J., 2007. Random forests for classification in ecology. Ecology, 88(11), 2783-2792.

Dai, F., Zhou, Q., Lv, Z., Wang, X., Liu, G., 2014. Spatial prediction of soil organic matter content integrating artificial neural network and ordinary kriging in Tibetan Plateau. Ecological Indicators, 45, 184-194.

Daroussin, J., King, D., Le Bas, C., Vrščaj, B., Dobos, E., Montanarella, L., 2006. The Soil Geographical Database of Eurasia at Scale 1: 1,000,000: history and perspective in digital soil mapping. Developments in Soil Science, 31, 55-602.

De'ath, G., Fabricius, K.E., 2000. Classification and regression trees: a powerful yet simple technique for ecological data analysis. Ecology, 81(11), 3178-3192.

De Bruin, S., Wielemaker, W., Molenaar, M., 1999. Formalisation of soil-landscape knowledge through interactive hierarchical disaggregation. Geoderma, 91(1), 151-172.

De Vries, F., De Groot, W., Hoogland, T., Denneboom, J., 2003. De bodemkaart van Nederland digitaal; Toelichting bij inhoud, actualiteit en methodiek en korte beschrijving van additionele informatie, Alterra.

Debella-Gilo, M., Etzelmüller, 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(1), 8-18.

Díaz-Uriarte, R., De Andres, S.A., 2006. Gene selection and classification of microarray data using random forest. BMC bioinformatics, 7(1), 1.

Eagleson, S., Escobar, F., Williamson, I., 1999. Spatial hierarchical reasoning applied to administrative boundary design using GIS.

Fenton, T.E., Larterbach, M., 1999. Soil map unit composition and scale of mapping related to interpretations for precision soil and crop management in Iowa. Precision Agriculture(precisionagric4a), 239-251.

Fleiss, J.L., Cohen, J., Everitt, B., 1969. Large sample standard errors of kappa and weighted kappa. Psychological Bulletin, 72(5), 323.
Florinsky, I.V., Eilers, R.G., Manning, G., Fuller, L., 2002. Prediction of soil properties by digital terrain modelling. Environmental Modelling & Software, 17(3), 295-311.
Forkuor, G., Hounkpatin, O.K., Welp, G., Thiel, M., 2017. High Resolution Mapping of Soil Properties Using Remote Sensing Variables in South-Western Burkina Faso: A Comparison of Machine Learning and Multiple Linear Regression Models. PloS one, 12(1), e0170478.
Gallant, J.C., Dowling, T.I., 2003. A multiresolution index of valley bottom flatness for mapping depositional areas. Water Resources Research, 39(12).
Genuer, R., Poggi, J.-M., Tuleau-Malot, C., 2010. Variable selection using random forests. Pattern Recognition Letters, 31(14), 2225-2236.
Gholamzadeh, M., 2014. Feasibility of soil magnetic susceptibility to separate soil drainage classes in Chaharmahal-Va-Bakhtiari Province. MSc Thesis, Isfahan University of Technology.
Girard, M.-C., Girard, C.M., 1999. Traitement des données de télédétection.
Goovaerts, P., 1997. Geostatistics for natural resources evaluation. Oxford University Press on Demand.
Grimm, R., Behrens, T., Märker, M., Elsenbeer, H., 2008. Soil organic carbon concentrations and stocks on Barro Colorado Island—digital soil mapping using Random Forests analysis. Geoderma, 146(1), 102-113.
Guo, Y., Gong, P., Amundson, R., 2003. Pedodiversity in the United States of America. Geoderma, 117(1), 99-115.
Häring, T., Dietz, E., Osenstetter, S., Koschitzki, T., Schröder, B., 2012. Spatial disaggregation of complex soil map units: A decision-tree based approach in Bavarian forest soils. Geoderma, 185, 37-47.
Hartigan, J.A., 1975. Clustering algorithms: John Wiley and Sons, New York, 351 p.
Hartigan, J.A., Wong, M.A., 1979. Algorithm AS 136: A k-means clustering algorithm. Journal of the Royal Statistical Society. Series C (Applied Statistics), 28(1), 100-108.
Hastie, T., Tibshirani, R., and Friedman, J. 2008. The Elements of Statistical Learning; Data Mining, Inference and Prediction. Springer, New York.
Henderson, B.L., Bui, E.N., Moran, C.J., Simon, D., 2005. Australia-wide predictions of soil properties using decision trees. Geoderma, 124(3), 383-398.
Hengl, T., Heuvelink, G.B., Kempen, B., Leenaars, J.G., Walsh, M.G., Shepherd, K.D., Sila, A., MacMillan, R.A., de Jesus, J.M., Tamene, L., 2015. Mapping soil properties of Africa at 250 m resolution: Random forests significantly improve current predictions. PloS one, 10(6), e0125814.
Hengl, T., Rossiter, D.G., 2003. Supervised landform classification to enhance and replace photo-interpretation in semi-detailed soil survey. Soil Science Society of America Journal, 67(6), 1810-1822.
Hengl, T., Toomanian, N., Reuter, H.I., Malakouti, M.J., 2007. Methods to interpolate soil categorical variables from profile observations: lessons from Iran. Geoderma, 140(4), 417-427.
Heung, B., Bulmer, C.E., Schmidt, M.G., 2014. Predictive soil parent material mapping at a regional-scale: a random forest approach. Geoderma, 214, 141-154.

Hojati, S., Khademi, H., 2011. Factors Affecting Palygorskite Distribution and Genesis in Selected Soils Developed on Tertiary Parent Materials in the Isfahan Province.
Holliday, V.T., 2006. A history of soil geomorphology in the United States. Ariel, 150, 238.250.
Holmes, K., Griffin, E., Odgers, N., 2015. Large-area spatial disaggregation of a mosaic of conventional soil maps: evaluation over Western Australia. Soil Research, 53(8), 865-880.
Hosmer Jr, D.W., Lemeshow, S., 2004. Applied logistic regression. John Wiley & Sons.
Hua, J., Xiong, Z., Lowey, J., Suh, E., Dougherty, E.R., 2005. Optimal number of features as a function of sample size for various classification rules. Bioinformatics, 21(8), 1509-1515.
Huete, A.R., 1988. A soil-adjusted vegetation index (SAVI). Remote sensing of environment, 25(3), 295-309.
Ibáñez, J.-J., Effland, W.R., 2011. Toward a Theory of Island Pedogeography: Testing the driving forces for pedological assemblages in archipelagos of different origins. Geomorphology, 135(3), 215-223.
Ibáñez, J., Caniego, J., San Jose, F., Carrera, C., 2005. Pedodiversity-area relationships for islands. Ecological Modelling, 182(3), 257-269.
Ibañez, J.J., De-Alba, S., Lobo, A., Zucarello, V., 1998. Pedodiversity and global soil patterns at coarse scales (with discussion). Geoderma, 83(3), 171-192.
Ibañez, J.J., De-Albs, S., Bermúdez, F., García-Álvarez, A., 1995. Pedodiversity: concepts and measures. Catena, 24(3), 215-232.
Ibáñez, J.J., Vargas, R.J., Vázquez-Hoehne, A., 2013. Pedodiversity state of the art and future challenges. Pedodiversity, CRC Press (Taylor and Francis Group) Boca Ratón. California, 1-28.
Jafari, A., Ayoubi, S., Khademi, H., Finke, P., Toomanian, N., 2013. Selection of a taxonomic level for soil mapping using diversity and map purity indices: a case study from an Iranian arid region. Geomorphology, 201, 86-97.
Jafari, A., Finke, P., Vande Wauw, J., Ayoubi, S., Khademi, H., 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(2), 284-298.
Jafarisirizi, A., 2012. Digital soil mapping in a selected arid landscape in southeastern Iran. PhD-thesis, Ghent University and Isfahan University of Technology.
Jalalian, A., Mohmmadi, M., 1989. Comprehensive study of reclamation and development of agriculture and natural resources in the Northern Karoon river basin (In Persian).
Jenny, H., 1941. Factors of soil formation—-A sytem of quantitative pedology, 281. McGraw-Hill, New York.
Karunaratne, S., Bishop, T., Baldock, J., Odeh, I., 2014. Catchment scale mapping of measureable soil organic carbon fractions. Geoderma, 219, 14-23.
Kelishadi, H., Mosaddeghi, M., Hajabbasi, M., Ayoubi, S., 2014. Near-saturated soil hydraulic properties as influenced by land use management systems in Koohrang region of central Zagros, Iran. Geoderma, 213, 426-434.
Kempen, B., Brus, D.J., de Vries, F., 2015. Operationalizing digital soil mapping for nationwide updating of the 1: 50,000 soil map of the Netherlands. Geoderma, 241, 313-329.
Kempen, B., Brus, D.J., Heuvelink, G.B., Stoorvogel, J.J., 2009. Updating the 1: 50,000 Dutch soil map using legacy soil data: A multinomial logistic regression approach. Geoderma, 151(3), 311-326.

Kempen, B., Brus, D.J., Stoorvogel, J.J., Heuvelink, G., de Vries, F., 2012. Efficiency comparison of conventional and digital soil mapping for updating soil maps. Soil Science Society of America Journal, 76(6), 2097-2115.

Kerry, R., Goovaerts, P., Rawlins, B.G., Marchant, B.P., 2012. Disaggregation of legacy soil data using area to point kriging for mapping soil organic carbon at the regional scale. Geoderma, 170, 347-358.

Kheir, R.B., Greve, M.H., Bøcher, P.K., Greve, M.B., Larsen, R., McCloy, K., 2010. Predictive mapping of soil organic carbon in wet cultivated lands using classification-tree based models: The case study of Denmark. Journal of Environmental Management, 91(5), 1150-1160.

Kidd, D., Malone, B., McBratney, A., Minasny, B., Webb, M., 2014. Digital mapping of a soil drainage index for irrigated enterprise suitability in Tasmania, Australia. Soil Research, 52(2), 107-119.

King, D., Jamagne, M., Arrouays, D., Bornand, M., Favrot, J., Hardy, R., Le Bas, C., Stengel, P., 1999. Inventaire cartographique et surveillance des sols en France. Etat d'advancement et exemples d'utilisation. Étude et Gestion des Sols, 6(4), 215-228.

Kovačević, M., Bajat, B., Gajić, B., 2010. Soil type classification and estimation of soil properties using support vector machines. Geoderma, 154(3), 340-347.

Krasilnikov, P., García-Calderón, N., Ibáñez-Huerta, A., Bazán-Mateos, M., Hernández-Santana, J., 2011. Soilscapes in the dynamic tropical environments: the case of Sierra Madre del Sur. Geomorphology, 135(3), 262-270.

Kuhn, M., Weston, S., Keefer, C., Coulter, N., Quinlan, R., 2013. Cubist: Rule-and instance-based regression modeling. R package version 0.0, 13.

Lagacherie, P., McBratney, A., Voltz, M., 2006. Digital soil mapping: an introductory perspective, 31. Elsevier.

Liaw, A., Wiener, M., 2002. Classification and regression by randomForest. R news, 2(3), 18-22.

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 America Journal, 70(4), 1200-1209.

Longuet-Higgins, M., 1971. On the Shannon-Weaver index of diversity, in relation to the distribution of species in bird censuses. Theoretical population biology, 2(3), 271-289.

Lösel, G., 2003. Application of heterogeneity indices to coarsescale soil maps, Abstracts, Pedometrics 2003, International Conference of the IUSS Working Group on Pedometrics, Reading University, Reading, England, September, pp. 11-12.

Magurran, A., 1988. Ecological diversity and its measurement Croom Helm London 179p.

Malone, B., 2013. Use R for digital soil mapping. Soil Security Laboratory, The University of Sydney: Sydney) Available at: www. clw. csiro. au/aclep/documents/DSM_R_manual_2013. pdf (accessed: 1 November 2013).

Malone, B.P., Minasny, B., McBratney, A.B., 2017. Using Digital Soil Mapping to Update, Harmonize and Disaggregate Legacy Soil Maps, Using R for Digital Soil Mapping. Springer, pp. 221-230.

Malone, B.P., Minasny, B., Odgers, N.P., McBratney, A.B., 2014. Using model averaging to combine soil property rasters from legacy soil maps and from point data. Geoderma, 232, 34-44.

Martin, M., Wattenbach, M., Smith, P., Meersmans, J., Jolivet, C., Boulonne, L., Arrouays, D., 2010. Spatial distribution of soil organic carbon stocks in France: Discussion paper. Biogeosciences Discussions.

Martín, M.A., Rey, J.-M., 2000. On the role of Shannon's entropy as a measure of heterogeneity. Geoderma, 98(1), 1-3.

MathWorks, 2009. Matlab. , The MathWorks., Inc., Natick, MA.

May, R., Van Dijk, J., Wabakken, P., Swenson, J.E., Linnell, J.D., Zimmermann, B., Odden, J., Pedersen, H.C., Andersen, R., Landa, A., 2008. Habitat differentiation within the large-carnivore community of Norway's multiple-use landscapes. Journal of Applied Ecology, 45(5), 1382-1391.

Mbagwu, J., Abeh, O., 1998. Prediction of Engineering Properties of Tropical Soils Using Intrinsic Pedological Parameters. Soil Science, 163(2), 93-102.

McBratney, A., 1995. Pedodiversity. Pedometron, 3, 1-3.

McBratney, A., Gruijter, J.d., 1992. A continuum approach to soil classification by modified fuzzy k-means with extragrades. European Journal of Soil Science, 43(1), 159-175.

McBratney, A.B., 1992. On variation, uncertainty and informatics in environmental soil management. Soil Research, 30(6), 913-935.

McBratney, A.B., 1998. Some considerations on methods for spatially aggregating and disaggregating soil information, Soil and water quality at different scales. Springer, pp. 51-62.

McBratney, A.B., Odeh, I.O., Bishop, T.F., Dunbar, M.S., Shatar, T.M., 2000. An overview of pedometric techniques for use in soil survey. Geoderma, 97(3), 293-327.

McBratney, A.B., Santos, M.M., Minasny, B., 2003. On digital soil mapping. Geoderma, 117(1), 3-52.

Mehnatkesh, A., Ayoubi, S., Jalalian, A., Sahrawat, K.L., 2013. Relationships between soil depth and terrain attributes in a semi arid hilly region in western Iran. Journal of Mountain Science, 10(1), 163-172.

Menezes, M.D.d., Silva, S.H.G., Mello, C.R.d., Owens, P.R., Curi, N., 2014. Solum depth spatial prediction comparing conventional with knowledge-based digital soil mapping approaches. Scientia Agricola, 71(4), 316-323.

Minasny, B., Hartemink, A.E., 2011. Predicting soil properties in the tropics. Earth-Science Reviews, 106(1), 52-62.

Minasny, B., McBratney, A.B., 2006. A conditioned Latin hypercube method for sampling in the presence of ancillary information. Computers & Geosciences, 32(9), 1378-1388.

Minasny, B., McBratney, A.B., 2007. Incorporating taxonomic distance into spatial prediction and digital mapping of soil classes. Geoderma, 142(3), 285-293.

Minasny, B., McBratney, A.B., Hartemink, A.E., 2010. Global pedodiversity, taxonomic distance, and the World Reference Base. Geoderma, 155(3), 132-139.

Minasny, B., McBratney, A.B., Malone, B.P., Wheeler, I., 2013. Digital mapping of soil carbon. Advances in Agronomy, 118(3), 4.

Minasny, B., McBratney, A.B., Salvador-Blanes, S., 2008. Quantitative models for pedogenesis—A review. Geoderma, 144(1), 140-157.

Ministry of Economy, Trade and Industry of Japan, National Aeronautics and Space Administration, 2009. http://www.gdem.aster.ersdac.or.jp.

Mishra, U., Lal, R., Slater, B., Calhoun, F., Liu, D., Van Meirvenne, M., 2009. Predicting soil organic carbon stock using profile depth distribution functions and ordinary kriging. Soil Science Society of America Journal, 73(2), 614-621.

Mohammad, H.B., 2000. Soil resources and use potentiality map of Iran. Soil and Water Research Institute, Teheran, Iran.

Moore, I.D., Gessler, P., Nielsen, G., Peterson, G., 1993. Soil attribute prediction using terrain analysis. Soil Science Society of America Journal, 57(2), 443-452.

Mosleh, Z., Salehi, M.H., Jafari, A., Borujeni, I.E., Mehnatkesh, A., 2016. The effectiveness of digital soil mapping to predict soil properties over low-relief areas. Environmental monitoring and assessment, 188(3), 1-13.

Mulder, V., De Bruin, S., Schaepman, M., Mayr, T., 2011. The use of remote sensing in soil and terrain mapping—A review. Geoderma, 162(1), 1-19.

Myles, A.J., Feudale, R.N., Liu, Y., Woody, N.A., Brown, S.D., 2004. An introduction to decision tree modeling. Journal of Chemometrics, 18(6), 275-285.

National Cartographic Center of Iran, 2015. http://www.ncc.org.ir/.

Nath, D.A., 2006. Soil landscape modeling in the Northwest Iowa Plains region of O'Brien County, Iowa.

Nelson, M., Odeh, I., 2009. Digital soil class mapping using legacy soil profile data: a comparison of a genetic algorithm and classification tree approach. Soil Research, 47(6), 632-649.

O'Neill, R.V., Krummel, J.R., Gardner, R.H., Sugihara, G., Jackson, B., Deangelis, D.L., Milne, B.T., Turner, M.G., Zygmunt, B., Christensen, S.W., Dale, V.H., Graham, R.L. , 1988. Indices of landscape pattern. Landscape Ecology 3, 153-162.

Odgers, N.P., McBratney, A.B., Minasny, B., 2015. Digital soil property mapping and uncertainty estimation using soil class probability rasters. Geoderma, 237, 190-198.

Odgers, N.P., Sun, W., McBratney, A.B., Minasny, B., Clifford, D., 2014. Disaggregating and harmonising soil map units through resampled classification trees. Geoderma, 214, 91-100.

Olaya, V., 2004. A gentle introduction to SAGA GIS. The SAGA User Group eV, Gottingen, Germany, 208.

Oshiro, T.M., Perez, P.S., Baranauskas, J.A., 2012. How many trees in a random forest?, International Workshop on Machine Learning and Data Mining in Pattern Recognition. Springer, pp. 154-168.

Pearson, R.L., Miller, L.D., 1972. Remote mapping of standing crop biomass for estimation of the productivity of the shortgrass prairie, Remote Sensing of Environment, VIII, pp. 1355.

Peters, J., De Baets, B., Verhoest, N.E., Samson, R., Degroeve, S., De Becker, P., Huybrechts, W., 2007. Random forests as a tool for ecohydrological distribution modelling. Ecological Modelling, 207(2), 304-318.

Petersen, C., 1991. Precision GPS navigation for improving agricultural productivity. GPS World, 2(1), 38-44.

Petersen, G., Lebed, I., Fohrer, N., 2009. SRTM DEM levels over papyrus swamp vegetation–a correction approach. Advances in Geosciences, 21, 81-84.

Phillips, J.D., 2001. The relative importance of intrinsic and extrinsic factors in pedodiversity. Annals of the Association of American Geographers, 91(4), 609-621.

Pielou, E.C., 1966. The measurement of diversity in different types of biological collections. Journal of theoretical biology, 13, 131-144.
Post, J., Hattermann, F.F., Krysanova, V., Suckow, F., 2008. Parameter and input data uncertainty estimation for the assessment of long-term soil organic carbon dynamics. Environmental Modelling & Software, 23(2), 125-138.
Prasad, A.M., Iverson, L.R., Liaw, A., 2006. Newer classification and regression tree techniques: bagging and random forests for ecological prediction. Ecosystems, 9(2), 181-199.
Qi, F., Zhu, A.-X., 2003. Knowledge discovery from soil maps using inductive learning. International Journal of Geographical Information Science, 17(8), 771-795.
Qi, F., Zhu, A.-X., Pei, T., Qin, C., Burt, J.E., 2008. Knowledge Discovery from Area–Class Resource Maps: Capturing Prototype Effects. Cartography and Geographic Information Science, 35(4), 223-237.
Quinlan, J., 1993. C4. 5: Programs for Empirical Learning Morgan Kaufmann. San Francisco, CA.
Quinlan, J.R., 1992. Learning with continuous classes, 5th Australian joint conference on artificial intelligence. Singapore, pp. 343-348.
Ranjbar, F., Jalali, M., 2016. The combination of geostatistics and geochemical simulation for the site-specific management of soil salinity and sodicity. Computers and Electronics in Agriculture, 121, 301-312.
R Development Core Team, 2013. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna Austria.
Rannik, K., Kõlli, R., Kukk, L., Fullen, M.A., 2016. Pedodiversity of three experimental stations in Estonia. Geoderma Regional, 7(3), 293-299.
Reza Pahlavan Rad, M., Toomanian, N., Khormali, F., Brungard, C.W., Bayram Komaki, C., Bogaert, P., 2014. Updating soil survey maps using random forest and conditioned Latin hypercube sampling in the loess derived soils of northern Iran. Geoderma, 232(97-106), 232-234.
Rhemtulla, J.M., Mladenoff, D.J., Clayton, M.K., 2007. Regional land-cover conversion in the US upper Midwest: magnitude of change and limited recovery (1850–1935–1993). Landscape Ecology, 22(1), 57-75.
Richardson, A.J., Wiegand, C., 1977. Distinguishing vegetation from soil background information.[by gray mapping of Landsat MSS data.
Roecker, S., Howell, D., Haydu-Houdeshell, C., Blinn, C., 2010. A qualitative comparison of conventional soil survey and digital soil mapping approaches, Digital Soil Mapping. Springer, pp. 369-384.
U.S. Geology Survey, 2004. (Geology.com/news/2010/free-lansat-images-from-USGS-2. shtml (http://glovis.usgs.gov)).
Sağlam, M., Dengiz, O., 2015. Similarity analysis of soils formed on limestone/marl-alluvial parent material and different topography using some physical and chemical properties via cluster and multidimensional scaling methods. Environmental monitoring and assessment, 187(3), 1-12.
Saldaña, A., Ibáñez, J., 2004. Pedodiversity analysis at large scales: an example of three fluvial terraces of the Henares River (central Spain). Geomorphology, 62(1), 123-138.
Salford Systems. 2004. Random Forests (software help guide). San Diego, CA.
Scull, P., Franklin, J., Chadwick, O., 2005. The application of classification tree analysis to soil type prediction in a desert landscape. Ecological modelling, 181(1), 1-15.

Smith, S., Bulmer, C., Flager, E., Frank, G., Filatow, D., 2010. Digital soil mapping at multiple scales in British Columbia, Canada, Program and Abstracts, 4th Global Workshop on Digital Soil Mapping Rome, Italy.

Soil and Water Research Institute of Iran, 2015. http://www.swri.ir/.

Soil Survey Staff. 2014. Keys to Soil Taxonomy, 12th ed. USDA-Natural Resources Conservation Service, Washington, DC.

Soil Survey Division Staff. 1993. Soil survey manual. United States Department of Agriculture.

Strobl, C., Boulesteix, A.-L., Zeileis, A., Hothorn, T., 2007. Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC bioinformatics, 8(1), 25.

Suring, L.H., Goldstein, M.I., Howell, S.M., Nations, C.S., 2008. Response of the cover of berry-producing species to ecological factors on the Kenai Peninsula, Alaska, USA. Canadian journal of forest research, 38(5), 1244-1259.

Taghizadeh-Mehrjardi, R., Minasny, B., Sarmadian, F., Malone, B., 2014. Digital mapping of soil salinity in Ardakan region, central Iran. Geoderma, 213, 15-28.

Taghizadeh-Mehrjardi, R., Nabiollahi, K., Kerry, R., 2016. Digital mapping of soil organic carbon at multiple depths using different data mining techniques in Baneh region, Iran. Geoderma, 266, 98-110.

Tajik, S., Ayoubi, S., Nourbakhsh, F., 2012. Prediction of soil enzymes activity by digital terrain analysis: comparing artificial neural network and multiple linear regression models. Environmental Engineering Science, 29(8), 798-806.

Thornbury, W.D., 1969. Principles o f Geomorphology, Second Edition. (John Wiley and Sons Inc.: Toronto).

Toomanian, N., Jalalian, A., Khademi, H., Eghbal, M.K., Papritz, A., 2006. Pedodiversity and pedogenesis in Zayandeh-rud Valley, central Iran. Geomorphology, 81(3), 376-393.

Turner, M.G., Gardner, R.H., 1991. Quantitative methods in landscape ecology: the analysis and interpretation of landscape heterogeneity/Monica G. Turner, Robert H. Gardner, editors. Ecological Studies; v. 82.

Venables, W., Ripley, B., 2003. Modern Applied Statistics with S (Statistics and Computing).

Viloria, J.A., Viloria-Botello, A., Pineda, M.C., Valera, A. 2016. Digital modelling of landscape and soil in a mountainous region: a neuro-fuzzy approach. Geomorphology, 253, 199–207.

Viscarra Rossel, R.A., Webster, R., Bui, E.N., Baldock, J.A., 2014. Baseline map of organic carbon in Australian soil to support national carbon accounting and monitoring under climate change. Global Change Biology, 20(9), 2953-2970.

Wang, D., Laffan, S., 2009. Characterisation of valleys from DEMs, Proceedings of 18th World IMACS/MODSIM Congress. IMACS, MSSANZ. Cairns, pp. 2014-2020.

Were, K., Bui, D.T., Dick, Ø.B., Singh, B.R., 2015. A comparative assessment of support vector regression, artificial neural networks, and random forests for predicting and mapping soil organic carbon stocks across an Afromontane landscape. Ecological Indicators, 52, 394-403.

White, R.E., 2013. Principles and practice of soil science: the soil as a natural resource. John Wiley & Sons.

Whiteway, T.G., Laffan, S.W., Wasson, R.J., 2004. Using sediment budgets to investigate the pathogen flux through catchments. Environmental management, 34(4), 516-527.

Whittaker, R.H., 1977. Evolution of species diversity in land communities [Birds and vascular plants]. Evolutionary biology.
Wiesmeier, M., Barthold, F., Blank, B., Kögel-Knabner, I., 2011. Digital mapping of soil organic matter stocks using Random Forest modeling in a semi-arid steppe ecosystem. Plant and soil, 340(1-2), 7-24.
Wilding, L., 1985. Spatial variability: its documentation, accommodation and implication to soil surveys, Soil spatial variability. Workshop, pp. 166-194.
Wilford, J., De Caritat, P., Bui, E., 2015. Modelling the abundance of soil calcium carbonate across Australia using geochemical survey data and environmental predictors. Geoderma, 259, 81-92.
Wilson, E.O., Forman, R., 1995. Land mosaics: the ecology of landscapes and regions. Cambridge: Cambridge University Press.
Wilson, H.F., Satchithanantham, S., Moulin, A.P., Glenn, A.J., 2016. Soil phosphorus spatial variability due to landform, tillage, and input management: A case study of small watersheds in southwestern Manitoba. Geoderma, 280, 14-21.
Wulf, H., T, M., M, S., A, K., Jörg, P., 2015. Remote Sensing of Soils. Technical report, University of Zurich, 72 pp.
Yang, L., Jiao, Y., Fahmy, S., Zhu, A., Hann, S., Burt, J.E., Qi, F., 2011. Updating conventional soil maps through digital soil mapping. Soil Science Society of America Journal, 75(3), 1044-1053.
Zeraatpishe, M., Khormali, F., 2012. Carbon stock and mineral factors controlling soil organic carbon in a climatic gradient, Golestan province. Journal of soil science and plant nutrition, 12(4), 637-654.
Zhu, A.-X., 1997. A similarity model for representing soil spatial information. Geoderma, 77(2), 217-242.
Zhu, A.-X., 1999. A personal construct-based knowledge acquisition process for natural resource mapping. International Journal of Geographical Information Science, 13(2), 119-141.
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(5), 1463-1472.
Zhu, A.-X., Yang, L., Li, B., Qin, C., Pei, T., Liu, B., 2010. Construction of membership functions for predictive soil mapping under fuzzy logic. Geoderma, 155(3), 164-174.
Zund, P., 2014. Disaggregation of land systems mapping. Technology, Information Division, Science.

Curriculum Vitae
Personal information

Full name:	**Mojtaba Zeraatpisheh**
Date of birth:	24/12/1985, Sepidan, Iran
Nationality:	Iranian
Address:	2/4 alley Soroosh St. AmirKabir Blvd. Shiraz-7178796147, Iran.
Affiliation:	1) Department of Soil Science, Isfahan University of Technology, Isfahan, Iran. 2) Department of Soil Management, Faculty of Bioscience Engineering, Ghent University, 9000 Ghent, Belgium.
Contact details:	Tel : +98-713-8324363 Cell phone: +98-917-7201719 E-mail: mojtaba.zeraatpisheh@ugent.be zeraatpishem@yahoo.com m.zeraatpishe@ag.iut.ac.ir

Education and training

2014- now	**PhD in Geology and Soil Science,** Universiteit Gent, 9000, Gent, Belgium ***Thesis****:* Digital soil mapping, downscaling and updating conventional soil maps using GIS, RS, Statistics and auxiliary data.
2012-2017	**PhD in Soil Genesis, Classification and Land Evaluation** Isfahan University of Technology, Isfahan, Iran ***Thesis****:* Digital soil mapping and scaling of soil classes and some of soil properties in a selected area in Charmahal & Bakhttrairi Province.
2008-2010	**Master in Soil Genesis, Classification and Land Evaluation. 19.70/20** Gorgan University of Agricultural Sciences and Natural Resources, Gorgan, Iran ***Thesis:*** Carbon stock and mineral factors controlling soil organic carbon in a climosequence, Golestan province
2004-2008	**Bachelor in Soil Science, Soil and Water. 17.25 /20** Islamic Azad University, Shiraz, Iran

Awards and fellowships

2015	**6 months scholarship:** Ministry of Science, Research and Technology (MSRT) of Iran

Lab experience

- Diagnosis of nutrient deficiency in plants
- Soil and plant nutrient elements analysis
- Mineralogy (Especially clay minerals in soil)
- Micromorpholoy
- Spectrophotometer
- GPS
- Air photo interpretation
- Cartography

Workshops and training

Year	Title	Description
2015	Advanced Academic English: Conference Skills- Academic Posters	The effective way to prepare and present a poster in academic conference. Gent University.
2015	Advanced Academic English: Conference Skills- Effective Slide Design	How to make a presentation more scientific and effective. Gent University.
2015	Internationalization: Meet people, not cultures	The best and Efficient way to be in touch with a culture and avoid cultural misunderstandings. Gent University.

Computer Skills

Proficient and Familiar with Several Software, Including:	✓ Microsoft office (Word, Excel and PowerPoint, Access) ✓ SAS ✓ Arc GIS ✓ SAGA GIS ✓ ILWIS ✓ Surfer Software ✓ R Programing ✓ MATLAB ✓ ENVI ✓ Spatial analysis

Languages

Persian (native)
English (Advanced)
French (Intermediate)

Teaching experience

Teaching assistant in Soil Science. Shiraz University. Shiraz, Iran
Duration: One semester -year 2013.

Work experience

Work in University of Applied Science and Technology of Jahad-E-University- Gorgan branch. Gorgan, Iran
Duration: 14 months-year 2011.

Publications

International/national peer reviewed

- **M. Zeraatpisheh**, S. Ayoubi, A. Jafari, P. Finke. 2017. A comparison of the efficiency of digital and conventional soil mapping to predict soil classes in a semi-arid region in Iran. Geomorphology, 285: 186-204.
- Khaledian. Y, Ebrahimi. S, Natesan. U, Basatnia. N, Behtari Nejad. B, Bagmohammadi, H., and **Zeraatpisheh, M**. 2017. Assessment of Water Quality Using Multivariate Statistical Analysis in the Gharaso River, Northern Iran. In: Sarma, A.K., Singh, V.P., Kartha, S.A., Bhattacharjya, R.K. (Eds.), 2017 *Urban Hydrology, Watershed Management and Socio-Economic Aspects*. Springer. ISBN 978-3-319-40195-9 (waiting for publishing).
- **M. Zeraatpisheh** and F. Khormali. 2012. Carbon Stock and Mineral Factors Controlling Soil Organic Carbon in a Climatic Gradient, Golestan Province. *Journal of Soil Science and Plant Nutrition*. 12 (4), 637-654.
- **M. Zeraatpisheh** and F. Khormali. 2011. The investigation of Soil Formation and Evolution of Losses-Derived Soils in a Climosequence, Case Study: Eastern of Golestan Province. *Journal of Soil and Water Conversation*.18 (2): 45-65 (In Farsi).
- **M. Zeraatpisheh,** F. Khormali, F. Kiani and M. H. Pahlavani. 2012. Studying Clay Minerals in Soils Formed on Loess Parent Materials in a Climatic Gradient in Golestan Province. *Iranian Journal of Soil Research*, 26(3): 303-316 (In Farsi).
- **M. Zeraatpisheh,** F. Khormali and A. Shahriari. 2011. The Relationship between

Specific Surface Area and Soil Organic Carbon in Loess-Derived Soils of Northern Iran. *J. Dynamic Soil, Dynamic Plant*, 5. Special Issue 1: 87-89.

- **M. Zeraatpisheh** and F. Khormali. 2011. Estimation of Organic Carbon loss Potential in a Climosequence in Golestan Province, Northern Iran. *J. Dynamic Soil, Dynamic Plant*, 5. Special Issue 1: 90-93.
- M. Hossieni, A. R. Movahedi Naeni and **M. Zeraatpisheh**. 2014. Effect of porosity, volumetric water content and soil temperature on the water uptake and dry matter yield of plant in different tillage systems. *Journal of Science and Technology of Agriculture and Natural Resources, Water and Soil Science*, 18(68): 133-146 (In Farsi).

Conference Papers

- **M. Zeraatpisheh**, Y. Khaledian, S. Ebrahimi, H. Sheikhpouri, B. Behtarinejad. 2013. The Effect of Deforestation on Soil Erosion, Sediment and Some Water Quality Indicates. *1th International Conference on Environmental Crisis and its Solutions*. 13-14 Feb, Kish, Iran.
- Y. Khaledian, S. Ebrahimi, Nabee Basatnia, **M. Zeraatpisheh**, A. Ghafarpour. 2013. Evaluation land use changes and anthropogenic influences on surface water quality by principal component analysis, Northern Iran. *1th International Conference on Environmental Crisis and its Solutions*. 13-14 Feb, Kish, Iran (Oral Presentation).
- **M. Zeraatpisheh** and F. Khormali. 2012. The Effect of Different Aggregate Size on Storage of Soil Organic Carbon. *International Soil Science Congress "Land Degradation and Challenges in Sustainable Soil Management"*. İzmir, TURKEY (Abstract).
- **M. Zeraatpisheh,** F. Khormali and A. Ajami. 2011. Forests and Soil evolution; outlook on Development Processes of Soil Evolution in the Presence of Virgin Forest Covers. *12th Soil Science Congers*. Tabriz. Iran (In Farsi).
- **M. Zeraatpisheh** and F. Khormali. 2011. Evaluation of Origin and Micromorpholoy of Calcite and Gypsum in the Loess-Derived Soils, Northen Iran. *12th Soil Science Congers*. Tabriz. Iran.
- **M. Zeraatpisheh,** F. Khormali, A. Shahriari. M. Liaghat and M. Mohammadnezhad. 2011. Using Early Features Found in Soil to Predict the Cation Exchange Capacity of Soils of Golestan Province. *The Second National Symposium on Agriculture and Sustainable Development. Opportunities and Future Challenges*. Shiraz, Iran (In Farsi).
- **M. Zeraatpisheh,** F. Khormali, F. Kiani and M. H. Pahlavani. 2010. Effect of Rain and Temperature on Carbon Stocks and the Rule of That in Sustainable Development at Pasture soils in Golestan Province. *The First National Symposium on Agriculture and Sustainable Development. Opportunities and Future Challenges*. Shiraz, Iran (In Farsi).
- **M. Zeraatpisheh** and F. Khormali. 2010. Evaluation Soil Development in a Climosequence on Losses Parent Material in the Golestan Province. *The 4th Regional Congress on Advances in Agricultural Research (West of Iran),*

Sanandaj, Iran (In Farsi).

- **M. Zeraatpisheh,** F. Khormali, F. Kiani and M. H. Pahlavani. 2010. Effect of Rain and Temperature on Carbon Stocks Rate Total Losses Soils, Golestan Province. *The 4th Regional Congress on Advances in Agricultural Research (West of Iran)*, Sanandaj, Iran (In Farsi).
- E. Bakhshandeh**, M. Zeraatpisheh,** A. A. Baahmani and M, Vali. 2009. Management Nutrient Element and the Role of Conservational Tillage in Agriculture Sustainable. *National Congress Agriculture Sustainable*, Gorgan, Iran (In Farsi).
- **M. Zeraatpisheh**, Y. Khaledian, S. Ebrahimi, H. Sheikhpouri, B. Behtarinejad. 2013. The Effect of Deforestation on Soil Erosion, Sediment and Some Water Quality Indicates. *1th International Conference on Environmental Crisis and its Solutions*. 13-14 Feb, Kish, Iran.
- Y. Khaledian, S. Ebrahimi, Nabee Basatnia, **M. Zeraatpisheh**, A. Ghafarpour. 2013. Evaluation land use changes and anthropogenic influences on surface water quality by principal component analysis, Northern Iran. *1th International Conference on Environmental Crisis and its Solutions*. 13-14 Feb, Kish, Iran (Oral Presentation).
- **M. Zeraatpisheh** and F. Khormali. 2012. The Effect of Different Aggregate Size on Storage of Soil Organic Carbon. *International Soil Science Congress "Land Degradation and Challenges in Sustainable Soil Management"*. İzmir, TURKEY (Abstract).

Appendix 1: List of Tables

Appendix 1: List of Figures