Centre for Geo-Information

Thesis GRS-2012-24

Soil texture mapping on a regional scale with remote sensing data

Submitted by

Anton Bakker

October 2012

WAGENINGEN UNIVERSITY
WAGENINGEN UR

Centre for Geo-Information

Thesis GRS-2012-24

Thesis title:	Soil texture mapping on a regional scale with remote sensing data
Supervisor:	Titia Mulder
Date:	October 2012
Course code:	GRS-80436
Author:	Anton Bakker
Reg. Nr:	870115026040

ACKNOWLEDGEMENTS

Foremost I have to thank my supervisor Titia Mulder. Always when I came by with a question, she made time to help me and really took the time to talk it through. I was amazed by her quick replies on emails and requests for feedback, especially in the last couple of weeks. Considering she is/was finishing up her PhD. Furthermore I would like to thank my partial co-supervisor Sytze de Bruin. He was involved at the start and the end of the thesis and supplied a few essential ideas for this study. I am also thankful for his incredible quick response on emails and requests for feedback, even from abroad.

Big thanks go to my fellow students, for discussing relevant and irrelevant topics in the thesis room and off course also for helping me solve a multitude of technical problems. Writing a thesis without them would not be half as much fun as it was now the last few months. I will miss this the coming months, while doing my internship.

I could not have written my thesis without the support of my friends and family, whom I might have neglected a bit the last couple of weeks (my sincerest apologies for this, I will make up for it). Despite my neglecting everyone kept on supporting me, which I am really thankful for. I would especially like to thank my sister Jenneke Bakker, my brother in law Roel Jansen and my father Jan Bakker for giving me feedback on my early works. This was a great help in getting me started in the writing process and keeping me motivated.

Finally, thanks to everyone I did not mention so far.

ABSTRACT

The current availability of quantitative soil information does not meet the requirements with respect to quality, cost and coverage for environmental monitoring and modelling for regional to global scale studies. More specifically, for countries where adaptation to global climate change and improvement of production systems is crucial for human wellbeing, such as most parts of Africa, general qualitative soil information is available at reconnaissance scale only. Fortunately, remote sensing, an efficient data source proven by its long track record, can provide relevant data over vast areas. However, for regional scale studies in development and transitional countries, the overall problem is how to use remote sensing for soil property mapping with few existing soil data available. Current methods focus on property mapping at the plot scale. However, they do not work beyond the plot due to local calibration of models. For operational use, these methods have to be extended beyond the plot level. Therefore, in this study we aim to correlate local soil samples to an exhaustive covariate dataset in order to predict soil texture on a regional scale. Based on the soil-landscape paradigm physical meaningful covariates are being used such as ASTER imagery, DEM-derived terrain descriptors, and additional climate data (precipitation, surface temperature and soil moisture). Three different methods have been tested for the prediction of soil texture; (1) multiple linear regression, (2) stepwise regression and (3) regression trees. Results show that stepwise regression is the best performing modelling method in this study; explained variance of the soil sample dataset for clay, silt and sand is 33%, 54% and 59% respectively. The variables that are included in the model showed that mineral indices, terrain descriptors and climate data are important variables for modelling soil texture. Tree based modelling proves not to be feasible on the soil sample dataset. The relationships resulting from the three different stepwise models have been applied to the entire study area, resulting in a soil texture map of the study area. This study demonstrates that remote sensing has a fundamental role for delivering detailed soil data on global and regional scale which is required for research focussing on environmental monitoring and modelling for regional to global scale studies.

TABLE OF CONTENT

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS

CHAPTER 1: INTRODUCTION

An increasing global need for good quality inexpensive spatial soil data exists. This is mainly caused by the world's expanding human population and the associated pressures on resources. More specifically, soil data is necessary for policy-making, managing of land resources and monitoring the environmental impact of development. Lack of comprehensive information about land resources will lead to uninformed policies, continuing degradation of land and water resources, unnecessary carbon emissions and ultimately no likelihood of achieving the UN's Millennium Development Goals. Soils do not only affect the food and water security, but also the viability and cost of vital infrastructure and the response to environmental change (e-SOTER, 2008).

At the moment the soil data supply does not meet the demand. A reason for this is that supply is insufficient because traditional methods, for acquiring spatial soil data, rely heavily on soil samples, which are costly to collect for regional scale prediction (Scull et al., 2003). Another reason is that demand is increasing because soil data are increasingly used in environmental monitoring and modelling (Rossel et al., 2006).

When time and budget are the limiting factors to collect soil data, the traditional methods are restricted to small scale studies. This is often the case in developing countries and nowadays these countries lack the necessary soil data to answer the many questions that act on a regional[1] or national scale (McKenzie and Ryan, 1999). This makes the need for a new methodology that is suitable beyond the plot scale critical.

A recent advancement in delivering soil survey information is the concept of digital soil mapping (DSM). DSM allows us to map soil properties quantitatively with the help of a range of different data sources and on a range of different spatial scale levels (Boettinger et al., 2010).

DSM requires a smaller soil sampling effort, compared to more traditional mapping methods. This smaller amount of sampled soil data is complemented by environmental data from different sources, which are related to the soil. The different environmental data sources used for DSM are expected to be related to soil through the soil forming function. This function regards the soil forming process as a function the soil forming factors; Climate, Organisms, Relief, Parent material and Time (CLORPT) (Jenny, 1941). For quantitative prediction purposes McBratney et al. (2003) have called this the CLORPT equation and Mckenzie and Austin (1993) have termed this approach 'environmental correlation'. The soil-landscape model forms the conceptual framework for relating environmental data sources (from now on environmental covariates) to soil properties as described by the soil forming function. Soil properties can be predicted based on the use of environmental covariates describing the soil forming factors (Mulder et al., 2011).

Remote sensing (RS) can be an inexpensive and exhaustive (covering the entire region of interest) source of information for mapping soils. Studies have shown that RS data can be used for soil spatial prediction with DSM (Ben-Dor, 2002; Mcbratney et al., 1991; Odeh et al., 1994), but also other spatial data sources have potential, such as global climate data sets.

[1] Different scale levels: Plot < 10^2 km^2, Local=>10^2 km^2 & < 10^4 km^2 , Regional => 10^4 km^2 & <10^7 km^2, Global => 10^7 km^2 (Mulder, V.L., de Bruin, S., Schaepman, M.E., 2012. Representing major soil variability at regional scale by constrained Latin Hypercube Sampling of remote sensing data. Int J Appl Earth Obs.)

Future studies should focus on the integrated use of RS methods for spatial prediction of soil properties, especially since future (airborne and space born) RS instruments will be launched that support these methods at larger spatial scales, enhancing the perspective of DSM (Mulder et al., 2011).

In short, a lack of methods for mapping soils on regional scale exists. The concept of DSM is a promising concept for this study to facilitate the mapping of soils on regional scale. Increasing availability of RS data helps to acquire data relevant for mapping soils at low costs. To study the potential of the combination of DSM with RS data, soil texture will be mapped in this study. This will be done for a study area of ~40.000 km^2 which is located in Morocco. Elevation ranges between ~0 and 3000 m above sea level, the climate is warm temperate with dry and hot summer (Mulder et al., 2012).

Soil texture was selected as the soil property of interest for two reasons. (1) Soil texture is a principal soil property, affecting many of the physical characteristics and behaviour of the soil, such as soil water retention, nutrient holding capacity and susceptibility to erosion (Greve et al., 2012) . Thus, soil texture maps have a wide range of applications. (2) Several studies have predicted soil texture with the help of RS (Breunig et al., 2008; Salisbury and Daria, 1992), this makes the use of RS data in combination with DSM promising.

Two different modelling methods will be tested and compared for prediction of soil texture; multiple linear regression and regression tree. A sparse soil sample set will be combined with an exhaustive RS based environmental covariate dataset. The resulting model for soil texture will be applied to the entire study area. The performance of the different modelling methods will be compared and validation of the models will be done with cross-validation.

1.1 Research objective and questions

The main objective of this study is to develop a method for mapping soil texture on regional scale by integrating remote sensing and environmental data. In order to meet this objective four research questions have been formulated.

1) Which methods do exist and are most suitable to predict/map soil texture with the help of RS data?
2) What candidate covariates exist that can be used to predict/map soil texture?
3) What methods and measures are available for determining the accuracy of the prediction?
4) How do the results fit in the soil-landscape paradigm?

1.2 Outline of report

Chapter two presents the literature study conducted for the selection of methods to model soil properties and for the selection of candidate covariates the can be used for modelling soil properties. Chapter three provides a theoretical background on digital soil mapping and the use of remote sensing. It also discusses how the different methods work for predicting soil properties and discusses the sampling method which was used for collecting the soil samples. Chapter four explains the methodology used in this study. Chapter five provides results and a brief description of the results. Chapter six gives an in-depth discussion of the results. Chapter seven provides conclusions and recommendations for future research.

CHAPTER 2: LITERATURE STUDY

2.1 Selection of modelling methods

Soil properties can be modelled with the soil sample data as target variables and the exhaustive RS covariate dataset explanatory variables. Different linear statistical methods have been used to model the relationship between the exhaustive covariate data and soil properties, such as multiple linear regression, stepwise regression, principal component regression and correlation analysis.

Multiple Linear Regression (MLR) is capable of deriving soil properties with multi- and hyper spectral reflectance data. Several examples can be found in literature. Ben-Dor and Banin (1995) used MLR on laboratory reflectance data, which was processed to simulate six TM bands in the VIS-NIR-SWIR region, to predict carbon concentration, organic matter content, specific surface area and total silica content. MLR was also used to estimate topsoil organic matter on a local scale with the use of high spatial resolution and hyper spectral RS data, acquired by the HyMap scanner (Selige et al., 2006). Another study used MLR for predicting soil properties using the relationships of terrain attributes; such as plan curvature, TWI, upslope area and elevation, with soil attributes; such as A horizon depth, Solum depth and E horizon presence (Gessler et al., 1995).

For MLR stepwise regression can be used as an explanatory variable selection procedure. The advantage of stepwise regression is that it utilizes interdependencies between covariates, while this is not the case if only relationships are tested between individual covariates and the target variable. Although stepwise regression is a greedy algorithm, once a covariate is selected, it cannot be dropped anymore. This is a disadvantage, not all possible interactions between the covariates are taken into account. A risk of applying stepwise regression is that it might lead to data dredging; uncovering misleading relationships in data sets (Neter et al., 1996).

Soil texture is compositional data; the fractions of clay, silt and sand always add up to 100% (Aitchison, 1981). This compositional relationship can be utilized with the right tools. For this purpose the functionalities of the R package 'compositions' (Van den Boogaard et al., 2011; van den Boogaard et al., 2012) was explored. Unfortunately the functionalities of the R package were too limited to use for compositional model building.

According to Ziadat (2005) non-parametric classification algorithms are preferred, when multisource data are used; such as classification and regression trees (CART), artificial neural networks (ANN), evidential reasoning, or knowledge based approaches. The underlying assumption is that the relation between soil types and the additional attributes is expected to be non-linear and such methods can better deal with non-linearity's (Mulder et al., 2011). However ANN's are not suitable for this study, typically ANN's are difficult to interpret (McKenzie and Ryan, 1999). Kriging methods are not an option for this study since the soil sample dataset used in this study does not express spatial correlation (Mulder et al., 2012).

Tree-based models can be used for classification and regression purposes. However a regression tree (RT) can only be used for regression purposes; RTs are a subclass of CART. Regression trees (RT) are capable of predicting soil properties with the help of RS imagery and or variables derived from a DEM (Digital Elevation Model). Greve et al. (2012) predicted soil texture, with the help of RTs and LIDAR derived terrain descriptors. Also McKenzie and Ryan (1999) were able to map soil properties with RS and topographical data. Besides the fact that RTs are good at dealing with non-linear relationships, they can also help discover interactions between covariates and the target variables; RTs are easy to interpret. A drawback of RTs is that the predictions are being composed of a small number of discrete values, while quantitative soil properties are in general expected to have a continuous variation. Therefore a loss of information occurs when applying a regression tree for prediction (McBratney et al., 2000).

Because of the considerations above three modelling methods were selected. (1) MLR was selected, because it has been used and tested in many different studies. (2) Stepwise regression was selected, because it will help to select a better set of explanatory variables to model the target variable and (3) RT was selected, because this modelling method is better at dealing with non-linearity's; non-linear relationships are expected between soil and soil forming factors. The notion that RTs are easy to interpret was also considered to be important; expected interactions between covariates and target variables can be determined.

2.2 Selection of potential covariates

The study is focused on areas with little or no detailed soil data available. Therefore no existing soil maps will be included in the method. The criterion used for selection is that covariates have to represent the soil forming factors and have to be RS based.

A one to one relationship between most covariates and the soil forming factors does not exist. Most soil forming factors have multiple related environmental covariates. Here, different environmental covariates that have been used in previous studies will be presented. These environmental covariates are proxies for the soil forming factor. The relationships between environmental covariates and soil forming factors are complex, due to the complex interrelationships. For instance organisms are influenced by climate. Climate can be influenced by relief and relief can influence organisms. Cause and effect are intertwined, therefore only the correlations are discussed.

Problematic is the fact that climate cannot be observed directly with remote sensing, considering the scope of this study; RS based covariates. Datasets on climate variables do exist, but they are not only based on RS data. WorldClim provides interpolated climate surfaces for global land areas at a spatial resolution of 30 arc s (~1 km resolution). These climate datasets are based on weather stations records, which are spatially interpolated with the help of a DEM. Apart from areas with a low station density, the highest uncertainty in climate data is in areas with high variation in elevation. The climate variables provided by WorldClim are monthly precipitation and mean, minimum and maximum temperature (Hijmans et al., 2005). Even though they are not entirely RS based, these covariates will be used in further analysis.

RS proxy variables exist that contain information on climate. The normalized difference vegetation index (NDVI) is a measure for vegetation intensity and will contain information on the spatial climate variation; vegetation intensity will be partly influenced by precipitation and temperature (Dobos et al., 2000). Soil moisture will contain some information on climate; precipitation will play a role in soil moisture variability. Soil moisture can be observed by RS, with the help of radar RS. However, soil moisture is largely influenced by relief. This means that the DEM and derived covariates Topographical Wetness Index[2] (TWI), slope and curvature will also give information on climate.

Organisms that influence soil formation are vegetation and organisms that live in the soil. However, only vegetation can directly be observed with remote sensing. Spatial and temporal variations in vegetation indices have been found to be linked to physical soil properties. The normalized difference vegetation index is one of the most common indicators of crop growth characteristics and, indirectly also an indicator of site quality. The soil background reflectance is influencing NDVI in partly vegetated areas, which is problematic if the soil spectral response is required. This produces lower NDVI values, with increasing soil brightness under otherwise identical conditions. Therefore several other vegetation indices have been developed, based on NDVI, such as the soil adjusted vegetation index (SAVI). Mono-temporal NDVI data have been linked in local scale studies to a range of soil properties, including soil texture (Mulder et al., 2011). NDVI has also been linked with temperature en precipitation regimes. It was also found that the use of spectral indices in combination with a DEM often produced soil patterns comparable to existing regional scale soil and terrain data (Dobos et al., 2000).

Relief is cardinal factor, first of all because soil development often is governed by the way in which water moves through and over the landscape. This in turn is determined by the local relief (Scull et al., 2003). Second, variables relating to relief can compensate for topographical distortion in the remote sensing data. Therefore, relief will be most useful in landscapes where topographic shape is strongly influencing the soil formation process (McKenzie and Ryan, 1999). Different relief variables have been proposed to be linked with soil characteristics. Slope and topographic wetness index (TWI) have been found to be highly correlated with surface soil attributes. TWI is an integrative topographic variable that is an indicator for water and sediment movement in landscapes. It quantifies the position of a site in the landscape and has been successfully used for predicting soil properties (Moore et al., 1993). Another study found that terrain descriptor data, such as elevation, slope, aspect and curvature, were essential for classifying soil types in combination with multispectral data (Dobos et al., 2000).

Sand, silt, and clay content have been found to be depended on slope and profile and tangential curvature. Profile curvature is parallel to the direction of maximum slope, a negative value indicates an upwardly convex surface, a value of zero indicates a linear surface and a positive value indicates an upwardly concave surface. It is a measure of acceleration and deceleration of flow across the surface. Tangential curvature or planform curvature is perpendicular to the direction of maximum slope, a positive value indicates a sideward convex surface, a value of zero indicates a linear surface and a negative value indicates a sideward concave surface. Tangential curvature is a measure for convergence and divergence of flow across the surface (Pachepsky et al., 2001).

Soil moisture content is an environmental covariate that is related to relief. The spatial variability of soil moisture mainly depends on relief, because the spatial distribution of characteristics of the topsoil layer as water-absorbing and water-retaining abilities are determined mostly by the relief (Svetlichnyi et al., 2003). Microwave remote sensing of soil moisture content is based on the difference in dielectric behaviour between dry soil and water. Currently the most advanced index is SWI (Mulder et al., 2011). SWI is based

[2] TWI is also known as the Compound Topographic Index (CTI)

METOP ASCAT (Meteorological Operation, Advanced Scatterometer) and ENVISAT ASAR GM (Environmental Satellite, Advanced Synthetic Aperture Radar, Global Mode) data and has a resolution of 1 km and gives relative values of soil water content over the rooting depth. The temporal resolution of SWI is 2 weeks.

Soil moisture content is also directly related with to soil texture; coarse sandy soils are usually well drained, resulting in low moisture content and relatively high reflectance; poorly drained fine-texture soils will generally have lower reflectance for panchromatic images (Lillesand et al., 2008).

Parent material can be determined with remote sensing. Studies show that different parent materials could be distinguished with the help of shortwave infrared (SWIR) and thermal infrared (TIR) ASTER (Advanced Spaceborn Thermal Emission and Reflection Radiometer) bands (Boettinger et al., 2010). ASTER-TIR is the first multispectral TIR remote-sensing satellite system with sufficient spectral, spatial and radiometric resolutions for geological applications. Several mineralogical indices have been proposed, based on analysis of TIR spectral properties of typical rocks on the earth. These indices include the Quartz Index (QI), Carbonate Index (CI) and Mafic Index (MI) for detecting mineralogical or chemical composition of quartzose, carbonate and silicate rocks with ASTER TIR bands (Ninomiya et al., 2005).

The time factor refers to the age of the soil surface and is mainly a function of the age of the deposit. The age of the deposit could significantly influence the kind and condition of the vegetation, so NDVI might also reveal some information related to time. However the soil forming factor Time is difficult to characterize (Dobos et al., 2000).

Other potential covariates that are not directly linked to a specific soil forming factor are RS multi-spectral reflectance data and the PC's of the multi-spectral reflectance data; several soil forming factors have been found to be expressed in remote sensed multi-spectral reflectance data (Mulder et al., 2011). The PC's have been found to capture >90% of the spectral variability (Lillesand et al., 2008); it also represents variation in covariates describing the soil forming factors such as landscape, geology and vegetation.

Principal component analysis (PCA) can be used for data reduction of a data matrix; in this case the data matrix was the stack of ASTER reflectance bands. PCA estimates the correlation structure of all the reflectance bands, thereby compressing most of the information of the reflectance bands in a few bands (Wold et al., 1987).

Due to these findings of previous studies the potential covariates were selected. Two different categories of potential covariates can be distinguished; (1) covariates acquired from the source and (2) covariates derived from the source data. Potential covariates of the first category are: ASTER VNIR, SWIR and TIR bands, ASTER DEM, SWI, temperature, precipitation and seasonality. Potential covariates of the second category are: the first three principal components of ASTER VNIR and SWIR bands, slope, TWI, curvature, profile and planform curvature, NDVI, CI, MI and QI.

CHAPTER 3: METHODOLOGICAL BACKGROUND

This chapter provides a background on how to utilize the relationship between soil forming factors and soils (3.1), how to observe the soil forming factors with RS (0), the object of interest; soil and soil texture (3.3) and the methods used for modelling of soil texture (3.4).

3.1 DSM and CLORPT

This study will use the concept of DSM. DSM is the creation of soil maps, based on digital data layers from a geographic information system (GIS). Digital layers can be raster data or vector data of environmental and topographical properties. Conceptually DSM is the art of fitting quantitative relationships between soil properties or classes and 'their environment' (McBratney et al., 2003). In DSM, soil is a sparsely available target variable and the 'environment' of the soil is a set of exhaustively available explanatory variables, which will be called covariates. The explanatory variables are stored in digital data layers.

The basic premise in this study is that the soil formation process is governed by the State Factor Equation of soil formation, where soil (*S*) is a function (*f*) of the soil-forming factors Climate, Organisms, Relief, Parent material and Time; CLORPT (Jenny, 1941) (Equation 1).

$$S = f(cl, o, r, p, t) \qquad \text{Equation 1}$$

Organisms are the natural vegetation as well as soil life. Relief is topography, which is described by terrain attributes, such as elevation, slope and curvature. Parent material is the underlying geological material (generally bedrock or drift deposit) in which soil horizons form. Time is the age of the soil. The function $f()$ is some form of empirical quantitative function linking the CLORPT factors to the soil S.

Equation 1 describes that soil in general is dependent of the soil forming factors. However, the soil forming function can also be used in a spatial framework. In which the soil forming function is valid for a location or area with little or no variation in soil; the choice of area size is dependent on the study area and soil property of interest. Soils have spatial patterns that act on different scale levels for different soil properties (Gessler et al., 1995). In DSM the soil forming factors are observed at a certain point or area for which the soil properties need to be determined.

In the traditional soil sciences CLORPT has been used as qualitative list for understanding the soil formation process. Often soil surveyors have been mapping soils by visually interpreting the implicit relationship with landscape characteristics. This results in soil maps that show polygons in which changes in soil properties are considered to be abrupt, although in most cases changes in soil properties are gradual (Ziadat, 2005). The challenge is to quantify the soil-landscape relationships, such that they can be used to map soils in a quantitative fashion. The increased availability of data describing the soil forming factors and advances in information technology has made it possible to use CLORPT for quantitative soil prediction and mapping.

3.2 RS and DSM

Remote sensing data are an important component of DSM because they provide a spatially exhaustive, quantitative measure of surface reflectance, which is related to several properties of the topsoil. Physical factors; such as particle size and surface roughness, and chemical factors; such as mineralogy, organic matter content and soil moisture, determine soil spectral reflectance (Scull et al., 2003).

Three purposes of remote and proximal sensing in (digital) soil mapping can be distinguished; the first purpose is stratifying the landscape into large relatively homogeneous soil-landscape units. The soil compositions of the stratified units can be determined by sampling or used as a source of secondary information. The second purpose is supporting spatial interpolation of sparsely sampled soil property data; when measurements of soil are sparse or poorly correlated in space, the estimation of the soil property of interest is in general improved by accounting for secondary information from other related variables, such as a DEM or RS data. The third purpose is allowing prediction of soil properties by means of physically-based and empirical methods; when spatial interpolation of soil properties is not feasible, a statistical relation between the sampled soil and a RS based covariate dataset can be modelled. This modelled relationship can be applied to the entire study area (Mulder et al., 2011).

Different studies have shown that soil texture can be mapped with RS data and DSM. Soil texture has also been predicted and mapped with proximal sensing. In proximal sensing, soil texture is typically determined by partial least-square regression and multiple linear regression (MLR) with the measured spectral responses of multiple bands as explanatory variables. Calibration of these models is done using data from a sample. It has been shown that these methods are useful for predicting soil texture, but calibration of the models is based on local conditions and therefore these models will in general not work outside the local scope (Mulder et al., 2011). The same holds true for soil mapping with remote sensing of the spectral response of the soil.

Salisbury and Daria (1992) showed that the ratio of ASTER bands 10/14 can be used to estimate particle size in soils, if other ASTER bands were used to minimize the confusion factors provided by soil moisture, vegetation cover, soil organic content, and the presence of abundant minerals other than quartz.

In a related study, Breunig et al. (2008) showed that the combination of the shortwave infrared (SWIR) ASTER bands 5 and 6 and thermal infrared (TIR) bands 10 and 14 discriminated dark red clayey soils and bright sandy soils from non-photosynthetic vegetation. ASTER band five and six coincide with the hydroxyl absorption band, which is associated with clayey soils. ASTER band 10 and 14 coincide with the reststrahlen feature, which is produced by the presence of quartz. Quartz is the main constituent of sand. It is known that the magnitude of resstrahlen feature varies with particle size.

However, it is known that predicting soil properties with RS data has its challenges. For instance, due to the heterogeneity of landscapes and the spatial resolution of the RS imagery, it is difficult to find pure pixels representing soil or bare rock. Vegetation can obscure part of the soil spectral response. Estimation of soil properties gets inaccurate if pixels have a vegetation cover over 20% (Bartholomeus et al., 2007).

Part of the soil is covered by vegetation, but part of the landscape is not covered with soil. Mapping of soil properties of the topsoil does not make sense for these areas. Areas that are not covered by soil can be built up area, roads, water surfaces and outcrops. The spectral reflectance of these objects does not contain information relevant for soil property mapping; therefore these areas should be excluded from the mapping procedure.

Research has shown that it is difficult to relate RS soil spectral response directly to soil properties. Soil spectral response is governed by complex interactions between constituents, such as soil moisture content, organic matter content, soil texture, surface roughness and presence of iron oxide (Lillesand et al., 2008). Under laboratory conditions most of these factors can be controlled for. This is not the case for large scale soil property mapping with RS. Also other factors complicates measurements of soil properties with RS; such as atmospheric influences, structural effects, lower spectral and spatial resolution, geometric distortions and spectral mixture of features (Mulder et al., 2011).

3.3 Soils and soil texture

In general soil is the product of physical and chemical weathering of rocks, soils and minerals. More specifically, soils are formed out of unconsolidated earth materials that develop distinct layers over time, called soil horizons. Physical and chemical weathering happens through the natural processes of weathering, caused by the effects of climate and plant and animal activity. The top layer is called the A horizon and has typically a thickness of 30 cm. The A horizon is the most weathered horizon. The second layer is the B horizon and called the subsoil, with a typical thickness of 45 to 60 cm. The C horizon is the underlying geological material from which the A and B horizon have developed. The concept of soil profile development is crucial for agricultural soil mapping and productivity mapping, as well as for development uses of the landscapes (Lillesand et al., 2008).

Soils can be differentiated by their soil particle size and classified into the relative fraction of the soil texture classes; clay, silt and sand. Particle size terminology is not standardized, the definition used in engineering is different from the one used in soil science. For this study the soil science definition will be used (Table 1). Soil texture is a principal soil property, affecting many of the physical characteristics and behaviour of the soil, such as soil water retention, nutrient holding capacity and susceptibility to erosion (Greve et al., 2012).

Table 1: Soil particle size definition used in this study (Locher and de Bakker, 1990)

Soil particle size class name	**Soil science definition (mm)**
Sand	0.05-2.0
Silt	0.002–0.05
Clay	<0.002

3.4 Modelling methods explained

Multiple Linear Regression (MLR)

MLR is capable of predicting soil properties with multi-spectral data. An advantage of MLR is that it can also handle nominal covariates. The idea is that the environmental covariates are used as explanatory variables and the soil texture variables are used as target variables (response variables). The regression model with *n* predictor variables is:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \epsilon \qquad \text{Equation 2}$$

In which Y is the target variable, β_0 is the intercept and $X_{1/n}$are explanatory variables, $\beta_{1/n}$ are the associated regression coefficients and ϵ is the error term. Regression estimates the mean of the probability distribution of the target variable. The response function is the overall solution that minimizes the sum of the squares of the errors for each data point.

The predicted value by the regression model contains an error component, for a data point this error is the difference between fitted value and the known value of the data point and is called the residual. The average value of all the residuals for the entire model is assumed to be zero, due to the requirement of normal distributed residuals.

The parameter β_n indicates the change in the mean response E with a unit increase of the predictor variable X_n .The parameter β_0 is the intercept, when the scope of the model includes $X_{n-1} = 0$, β_0 gives the mean of the probability distribution of Y at $X_{n-1} = 0$ (Neter et al., 1996).

Training data which will be used for MLR has to meet several criteria. First, the predictor variables should be independent. Second, the distribution of the residuals should have a normal distribution and third, the residuals should be homoscedastic.

The first requirement can be tested by calculating the correlation between the independent variables; there should not be any obvious correlation between independent variables. Second, the residuals should be normally distributed, because the error term is assumed to be random; a normal distribution can be used to approximate a random variable that clusters around a single mean value. This can be tested by the Shapiro-Wilk test, which is based on the coefficient of correlation between the ordered residuals and their expected values under normality (Neter et al., 1996). If the error term is not random, than a trend is present in the dataset that is not accounted for by the model. Homoscedasticity, of the residuals, means that the residuals should exhibit constant variance. For homoscedasticity constant variance of residuals is required over the predicted values and over any of the independent variables. This is required since non constant variance of residuals means there is an unexplained trend in the model.

Homoscedasticity of the residuals can be tested by means of a Breusch-Pagan test. This test assumes that the error terms are independent and normal distributed and the variance of the residuals, denoted by σ_i^2, is related to the level of the explanatory variable x_i , according to Equation 3.

$$\log_e \sigma_i^2 = \gamma_0 + \gamma_1 x_i \qquad \text{Equation 3}$$

Constancy of the residual variance corresponds to γ_1=0. The null hypothesis is that γ_1=0; the residual variance is constant over the independent variables (Neter et al., 1996).

Covariate selection

Different methods to select the explanatory variables to include in the regression model can be used. One method is to test the relationship between the individual covariates and the target variable by calculating the correlation coefficient. Another method is to apply stepwise regression, in which the choice of covariates is carried out by an automated procedure.

In the case of forward stepwise regression the procedure starts with an initial model with only an intercept. A selected model that will be used in a next step will be called base model. In the next step a linear model is fitted for each of the explanatory variables available. For each model in this step a test statistic can be calculated to assess the performance of the model. The model that is best performing will be used as the base model for the next step. This step is repeated with the remaining explanatory variables that are not added to the model yet, until the test statistic cannot be improved (sufficiently) anymore.

Stepwise regression can also be performed backwards. In the case of backwards stepwise regression the procedure starts with a linear model that has all available explanatory variables included in the model. Each iteration, a new model is fitted for each of the explanatory variables included in the model without this explanatory variable. The best performing model, based on the calculated test statistic for each model, will be used as the base model for the next step. This step is repeated until the test statistic cannot be (sufficiently) improved anymore.

Stepwise regression can also be done both ways; forward and backward. In this case, at each iteration, models are fitted for all variables that can be included in the model and models are fitted for all variables that can be excluded in the model. Also here the best performing model of all fitted models will be selected as the base model for the next step (Neter et al., 1996).

Regression tree

RT analysis is a data mining technique which iteratively splits a dataset into two increasingly homogenously subsets on explanatory variables, aiming for nodes with the smallest variance. In each terminal node the predicted target value $y(t_n)$ is constant, therefore the tree can be thought of as histogram estimate of the regression surface.

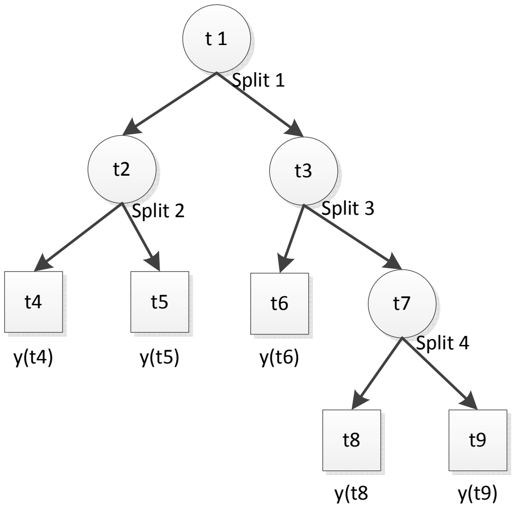


Figure 1: Example of a regression tree, nodes are reprented by circles, end nodes by squares, node t and predicted target value $y(t)$ (Breiman, 1984).

In order to build a parsimonious tree, first a maximum tree is grown until no splits can be made anymore due to lack of data. After growing the maximum tree a cross-validated error is calculated for all the possible trees that are contained by the maximum tree. If the maximum tree has 10 splits, a cross-validated error will be calculated for all trees within the maximum tree for 1 to 10 splits. Subsequently the maximum tree is reduced in a process of pruning where least important splits, based upon the cross-validated error estimates, are removed.

The tree can be pruned on the smallest cross-validated error. However, a more cautious approach is to select the smallest tree which is still accurate. For this, it is suggested use the 1 SE rule, which means to choose the smallest number of splits with a cross-validate error rate smaller than the minimum cross-validated error rate plus 1 standard error of the minimum cross-validated error rate. In general this results in trees with fewer splits and good predictability thereby reducing the chance of over fitting and increasing the model stability (Breiman, 1984).

CHAPTER 4: MATERIALS AND METHODS

4.1 Overview of methodology

This study aimed to develop a method for mapping soil texture on regional scale with the use of remote sensed data. For this objective four research questions have been formulated. To answer the research questions, seven steps were defined in the research proposal phase (Figure 2). Step 1 is to find suitable methods for modelling soil texture with remote sensing data in literature. Step 2 is to make an inventory of potential covariates that can be included in the models. These environmental variables/datasets that describe the CLORPT factors will be called covariates in this study. Therefore the potential covariates will be selected and acquired, based on findings in literature. Step 3 is pre-processing of the potential covariates, so they can be used in the modelling environment. Step 4 is an exploratory data analysis on the acquired data, to see whether the data is suitable to be used in the modelling methods proposed in step 1 and to see how the data set looks like. This is important for interpreting the results of the modelling methods. Step 5 is the modelling of soil-texture with the selected modelling methods. Step 6 is the interpretation of the results and cross-validation of the models to get a measure of the accuracy of the prediction. Step 7 will provide a spatial measure of the error of the prediction. This will give an additional measure of accuracy, independent of cross-validation. A more detailed description of the activities carried out will be provided in this chapter, each step is described in a section.

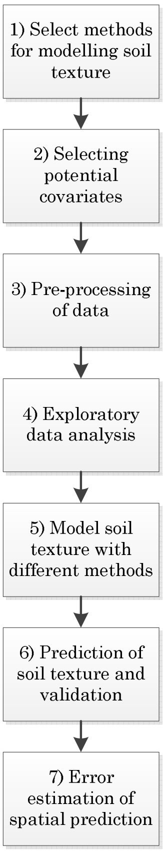

1) Select methods for modelling soil texture
2) Selecting potential covariates
3) Pre-processing of data
4) Exploratory data analysis
5) Model soil texture with different methods
6) Prediction of soil texture and validation
7) Error estimation of spatial prediction

Figure 2: Overview of the different steps in the methodology used in this study

Conceptually the methodology of this study will use DSM with the State Factor Equation. The RS based environmental covariates, describing the CLORPT factors, will be used as primary data source for predicting soil texture, by modelling the relationship between the covariate dataset and the soil sample data set available for the study area.

4.2 Select modelling methods

The first step of the methodology of this study is to select the modelling methods which the soil properties will be predicted with. Based on a literature study (2.1 Selection of modelling methods), three modelling methods were selected; multiple linear regression (MLR), MLR with stepwise variable selection (note; in this study we use the term stepwise regression for MLR with stepwise variable selection) and RT.

4.3 Select potential covariates

A literature study was carried out to select a number of potential covariates (2.2 Selection of potential covariates). Based on findings in literature multiple potential covariates were selected. The criterion for selection of the covariates was that covariates have to represent the soil forming factors. Two different categories of selected covariates can be distinguished; (1) covariates acquired from the source (Table 2) and (2) covariates derived from the source data (Table 3).

The covariate dataset used comes from three data sources; ASTER, a multispectral space born sensor (Abrams et al., 2002) , ASCAT (Advanced Scatterometer), a space born real aperture radar sensor (Brocca et al., 2010) and WorldClim, a source for climate datasets, based on spatially interpolated measurements of weather stations (Hijmans et al., 2005).

Table 2: Selected covariates from source data describing the source, resolution, spectral range and CLORPT factor of each covariate.

Covariate (# of bands)	**Source**	**Resolution (m)**	**Spectral range (µm)**	**CLORPT factor**
VNIR reflectance (3)	ASTER	15	0.52 - 0.86	-
SWIR reflectance (6)	ASTER	30	1.60 – 2.43	-
TIR emissivity (5)	ASTER	90	8.125 – 11.65	Parent Material
DEM	ASTER	30	-	Relief
SWI	ASCAT	12500	-	Relief, Climate
Temperature	WorldClim	760	-	Climate
Seasonality	WorldClim	760	-	Climate
Precipitation	WorldClim	760	-	Climate

Table 3: Selected covariates derived from source data describing the source, spatial resolution and CLORPT factor of each covariate.

Covariate (# of bands)	Resolution (m)	Source (band # used)	CLORPT factor
PC reflectance (3)	30	ASTER VNIR & SWIR	-
Slope	30	ASTER DEM	Relief
TWI	30	ASTER DEM	Relief, Climate
Curvature	30	ASTER DEM	Relief
Profile curvature	30	ASTER DEM	Relief
Plan form curvature	30	ASTER DEM	Relief
NDVI	30	ASTER VNIR (bands 2, 3)	Organisms
CI	90	ASTER TIR (bands 13, 14)	Parent material
MI	90	ASTER TIR (bands 12, 13)	Parent material
QI	90	ASTER TIR (bands 10 , 11, 12)	Parent material

4.4 Data pre-processing

This section explains where the data comes from, how the data is pre-processed and which information the data provides. An overview of the data pre-processing is supplied in .

Over 50% of the selected covariates have a resolution of 30 m; therefore all covariates were resampled to a resolution of 30 m. Consequently, a coarser resolution would result in a loss of information; a finer resolution would result in large file sizes without much gain of information. A more detailed resolution would also result in longer processing times for applying the models. The ASTER DEM was chosen as the template raster, other raster were snapped to the DEM so all the pixels of all the layers overlap were aligned.

Not all covariates share the same coverage. The ASTER dataset does not cover the entire study area (Figure 3). Therefore the actual study area is the area which is covered by all covariates exhaustively. This area is equal to the area covered by the ASTER data. All the other covariates have a full coverage of the study area. The ASTER data was projected in UTM zone 30N (WGS1984). All other data were reprojected to this projection while resampling. Clipping and resampling of the datasets was performed in ArcGIS (ESRI, 2011).

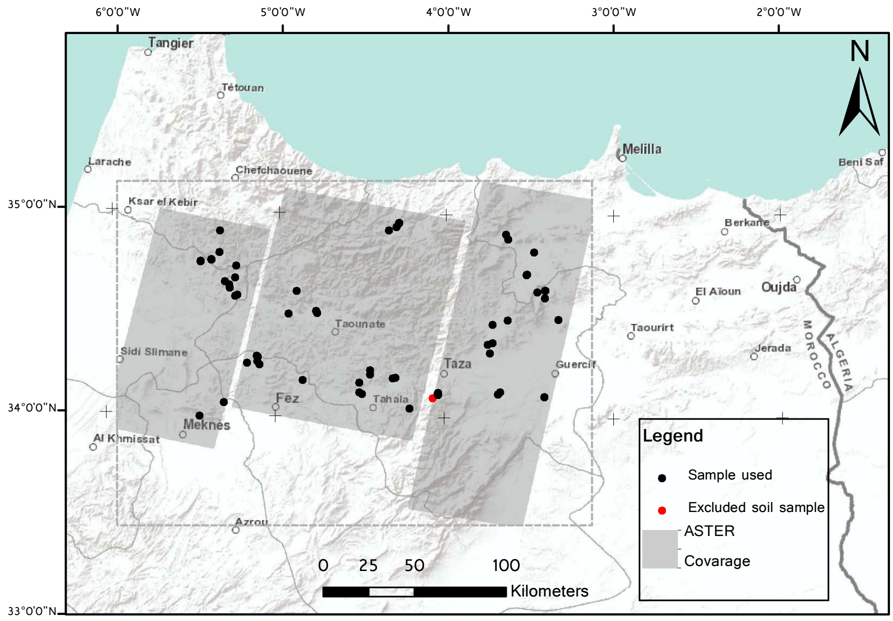


Figure 3: Map of the study area with location of soil samples and coverage of ASTER data indicated. Grey dotted line indicates the extent of the study area, light grey lines show administrative boundaries, background shows shaded relief base map (source: ESRI).

ASTER reflectance and emissivity data

The ASTER instrument has three VNIR bands, six SWIR bands and five TIR bands (Table 2). The ASTER products that were used are: ASTER Surface Reflectance: VNIR & Crosstalk Corrected SWIR & Surface Emissivity. With ASTER data several soil properties can be determined by means of spectral analysis, therefore all ASTER bands of them are seen as potential covariates for soil texture modelling (Mulder et al., 2011).

The ASTER data is a mosaic of images was which were taken in the same season over the years 2005 and 2007. Different images were used to prevent cloud cover in the final image. The ASTER source data came in separate files, one file for each tile. Each wavelength domain (VNIR, SWIR and TIR) has separate files. For image processing ENVI version 4.8 was used (Exelis Visual Information Solutions, Boulder, Colorado). ENVI is software for analysis and visualisation of different types of digital imagery. First, a mosaic was made out all the files, after this the files were stacked in one raster file. The next step was to resample the data to a resolution of 30 m. From the pre-processed ASTER image the first three principal components were calculated. Only the reflectance bands were used for the principal components. The soil sampling scheme was based on the first three PC bands of the ASTER reflectance data.

The three mineral indices CI (Equation 4), MI (Equation 5), QI (Equation 6) and NDVI (Equation 7) were calculated from the ASTER reflectance and emissivity data in R (equations are expressed in band numbers).

$$\mathrm{CI} = \frac{\mathrm{B\ 13}}{\mathrm{B\ 14}}$$ Equation 4

$$MI = \frac{\mathrm{B\ 12}}{\mathrm{B\ 13}}$$ Equation 5

$$\mathrm{QI} = \frac{\mathrm{B\ 11} * \mathrm{B\ 11}}{\mathrm{B\ 10} * \mathrm{B\ 12}}$$ Equation 6

$$\mathrm{NDVI} = \frac{\mathrm{B\ 3} - \mathrm{B2}}{\mathrm{B\ 3} + \mathrm{B\ 2}}$$ Equation *7*

ASTER DEM

The ASTER DEM used is version 1, also called GDEM (Global DEM). The GDEM is created by compiling both VNIR images of band 3, because band 3 has two telescopes; one backward and one nadir looking and has a spatial resolution of 30 m (Abrams et al., 2002). However, research has shown that the resolution of the GDEM is overly optimistic; a resolution of 90 m would be a better fit for the data. The GDEM reports canopy elevations and a large portion of the GDEM (20%) contains anomalies (Hengl and Reuters, 2011).

The following covariates were derived from the ASTER DEM: slope, curvature, profile curvature, planform curvature, Topographic Wetness Index (TWI). Slope, curvature, profile curvature, planform curvature are local terrain parameters, which are calculated by using a 3x3 grid cells window in ArcGIS. Curvature is the slope of the slope, profile curvature is the curvature in the direction of the maximum slope and the planform curvature is perpendicular to the direction of maximum slope (ESRI, 2011).

TWI is a complex terrain parameter, which establishes a broader spatial relationship by combining local upslope contributing area and slope to quantify topographic influence on hydrological processes (Sorensen et al., 2006). TWI was calculated with R, by using the 'RSAGA' package (Brenning, 2011). This package provides access to spatial and terrain analysis functions of SAGA, which is an open source GIS. ArcGIS desktop 10.0 does not have a function to calculate TWI.

ASCAT time series

The ASCAT (Advanced Scatterometer) soil water index (SWI) is a Level 3 soil moisture product and obtained by a modelling approach using the ASCAT Level 2 Surface Soil Moisture data (SSM) and other data sources as input. SSM data are retrieved from the radar backscattering coefficients measured by the ASCAT on board the MetOp satellite using a change detection method. Long-term scatterometer data are used to model the incidence angle dependency of the radar backscattering signal.

Knowing this dependency, the backscattering coefficients are normalised to a reference incidence angle. The relative soil moisture data, ranging between 0% and 100% are then derived by scaling the normalised backscattering coefficients between the lowest/highest values corresponding to the driest/wettest soil conditions. The Soil Water Index is then derived by using a two-layer water model representing the topmost layer and the reservoir below with an exponential function resulting in the SWI. The product represents the soil moisture content in the first meter of the soil in relative units, ranging between wilting level and field capacity. SWI is available at 8 different temporal resolutions; 1, 5, 10, 15, 20, 40, 60, 100 days. The model integrates all measurements for the different temporal resolutions supplied and weights the most recent measurements to contribute strongest to the SWI sample. The period for which daily SWI data is available is from January 2007 till August 2011, data are stored for each day at 12:00 UTC (I.P.F., 2011).

For this study SWI with a temporal solution of 100 days was used, because the coarsest temporal resolution is less sensitive for short term fluctuations. Therefore it provides more information on the long term changes in soil moisture. The data is delivered in a grid format; the grid spacing is approximately 12.5 km. For each grid point, a separate CSV file was supplied containing the time series. Information on the geographical location of the grid points is stored in a separate file, which is the grid point index.

Pre-processing of the SWI data was done with several Python scripts and ArcGIS. The first step in pre-processing was to filter out the grid points which were lying outside the study area, using the grid point index. The next step was to reduce the time series to one variable. In order to reduce the time series of SWI100 to one variable the total average and the average per period of three consecutive months have been calculated (calculated for each pixel). For each of these means the variance in the entire study area was calculated. The average with highest variance in the study area was for the period of August, September and October. Therefore the average of August, September and October of the SWI100 was chosen as the best descriptor in soil moisture variability in the study area. Points with this average were extracted from this dataset and transformed to a raster dataset in ArcGIS. The resulting raster was resampled to a spatial resolution of 30 m.

WorldClim data

The WorldClim dataset has a resolution of 30 arc-seconds, which is ~760 m at the latitude of the study area. From the WorldClim dataset the following variables were selected: annual mean temperature, annual precipitation and temperature seasonality (standard deviation of yearly temperature * 100). This is climatic data for current conditions (~1950-2000) (Hijmans et al., 2005). These grids were clipped to the size of the study area and resampled to a spatial resolution of 30 m.

Combining all covariate rasters

After pre-processing of the individual covariates all the layers were stacked and saved as one GeoTIFF file. The major water surfaces and built-up areas were masked out from the covariate raster stack. Stacking covariate rasters and masking the covariate raster stack were done in ArcGIS. Data for the major cities were derived from a topographic base map, which was provided through ArcGIS desktop. Data on the major water surfaces were digitized from a topographic base map (ANWB, 2010).

Soil data

The soil samples used for this study were collected with constrained Latin Hypercube Sampling (LHS), which is a model-based sampling scheme; for this study LHS selected sampling locations based on multiple predetermined remote sensed covariates and constraints. Constrained LHS aims for allocating sites such that the range of all the predetermined covariates are sampled while at the same time honouring constraints and minimizing the cost of sampling. The cost of sampling is related to sample size, time spent on sampling and accessibility. The covariates that were used are the first three PCs of the

ASTER VNIR and SWIR bands data and elevation (ASTER DEM). It was assumed that these covariates represent the spatial variability of the soil in the defined study area. The relation between the covariates and soil properties is based on the soil-landscape paradigm. The constraints used were: (1) the cost of travel off-road increases linearly with increasing distance from the road and (2) sites with slopes steeper than 45° or rivers and lakes are not selectable. Comparison of the samples taken with legacy data showed that the sampling approach was successful in representing major soil variability.

The first three PC's and elevation showed strong spatial correlation, which supports the premise that once the relation between the exhaustive covariate data and soil properties is modelled, the remote sensed data can provide the spatial basis for mapping soil properties. However, the resulting soil sample set did not show spatial correlation, because the sampling method favoured high variability for soil properties at short distances between sample points (Mulder et al., 2012).

The main reason for using a constrained LHS soil data set was the goal of this study; mapping soils on a regional scale, with limited budget. The constrained LHS method allows for sampling the soil variability on regional scale with a relatively small sampling effort due to the set constraints. Also, a design-based sampling scheme may fail to sample the tails of a distribution of the exhaustive covariate dataset, so that a soil with specific soil properties that depend on less frequent occurrences of specific covariate values may be missed.

At the sample sites square shaped plots of 15 by 15 meters were defined in which a mixed soil sample was taken from each corner and the middle of the plot. Analysis of the soil samples were done in the laboratory; the samples were dried at 70º C, sieved at 2 mm and analysed on several soil properties; including soil texture (Mulder et al., 2012).

Soil sample pre-processing and covariate extracting was done in R (R Development Core Team, 2011), with the contributing package 'raster' for geographic analysis and modelling with raster data (Hijmans and van Etten, 2011). The soil sample data were delivered in a table format containing 64 samples with the corresponding soil property information. With the corresponding x and y coordinates, site numbers and textures classes of clay, silt and sand. One sample did not coincide with the ASTER data and was excluded for further analysis (Figure 3). Next, the table was converted to a spatial point format and the information from the raster file containing all the covariates was extracted at these points (Figure 4). After extraction of the covariates the point data set contained the data of all the covariates of the location of the points. This enabled modelling the relationship between soil texture and the covariate dataset.

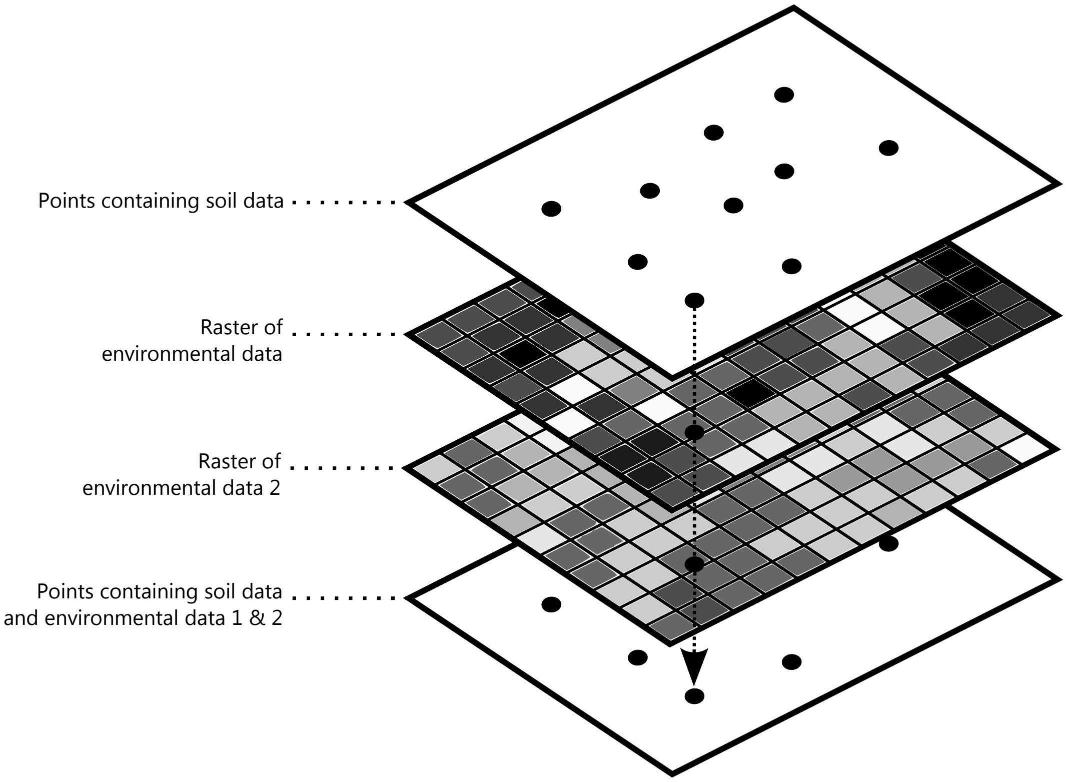

Points containing soil data
Raster of environmental data
Raster of environmental data 2
Points containing soil data and environmental data 1 & 2

Figure 4: Schematic representation of covariate data extraction to soil sample points:

4.5 Exploratory data analysis

The exploratory data analysis consisted of four steps and was performed in R with contributing packages ‘ggplot2’; for plotting of histograms and boxplots (Wickham, 2009), ‘ellipse’; for plotting of correlation matrices (Murdoch and Chow, 2007) and ‘raster’; for handling the raster datasets.

Step 1

Check the thematic distribution of the target variables for the sampled locations. For this purpose histograms were computed.

Step 2

This study is interested in the relationship between the covariates and soil texture. Therefor a first step is looking at the relationships between (a) the collected sample (clay, silt and sand content) and (b) of the individual covariates with each other.

a) Check the individual relationships between the target variables and the covariates by plotting a correlation matrix of all the covariates with the function *plotcorr* from the R package ‘ellipse’. For each relationship the overall shape of the scatterplot is plotted. The overall shape will be round if there is no relationship and ellipse if there is a relationship. The ellipse is shaped to be contours of a bivariate normal distribution with unit variances and correlation coefficient r (Murdoch and Chow, 1996). The correlation coefficient (also known as Pearson’s r) for each of these relationships is calculated as well. The correlation coefficient is a measure of strength and direction of the linear relationship between two variables (Neter et al., 1996).

b) Check the individual relationships between all the covariates. This is done by plotting a correlation matrix of all the covariates with the function *plotcorr*. The relationship of a covariate with another covariate is of interest for modelling the relationship between soil texture and the covariates. It might help uncover unknown relationships in the dataset.

Step 3

Check whether the soil sample locations cover the thematic range of the exhaustive covariate data. For this purpose two types of boxplots were created with the help of R.

a) Boxplots of the soil sample covariate values.
b) Boxplots of a random sample (n=50000) of the corresponding exhaustive covariate data. This helps to check whether the covariate values of the soil samples are representative for the entire covariate raster.

4.6 Modelling soil texture

Multiple linear regression, stepwise regression and regression trees were used for modelling of clay, sand and silt content. Based on the cross validation the most accurate models were selected and used for prediction of soil texture. Modelling was performed in R with contributing packages 'MASS' (Venables and Ripley, 2002) for stepwise regression, 'rpart' (Therneau et al., 2010) RTs and 'ggplot2' (Wickham, 2009) for creating diagnostic plots of MLR models.

Multiple linear regression

The first modelling method used all the variables having a correlation coefficient with the target variable >0.2|<-0.2. For each target variable (clay, silt and sand) a linear model was fitted. Assessment of the performance of the models consisted of 4 different steps:

1) The output of the summary method for the generated linear models in R was inspected. The summary of a linear model returns the list of coefficients and the t-statistic with corresponding p-value, the p-value of the F-test and the adjusted R^2 (Neter et al., 1996).
2) To get a measure of the error of the model on the training data, the root mean square error (RMSE) of the estimate was calculated for each model. This is calculated by taking the mean of the root of the squared difference of the fitted values and the sampled values.
3) Check for normal distribution of the residuals by means of a Shapiro-Wilk test and a standard quantile-quantile (QQ) plot, which shows the standard distribution on the X-axis against the standardized residuals on the Y-axis. Distribution of the residuals is a measure of performance for linear regression.
4) Check for homoscedasticity of the residuals, by plotting the fitted values versus the residuals; this is for checking of constancy of residual variance over the fitted values. Also a Breusch-Pagan test was conducted to test for constancy of residual variance over the independent variables.

The test for normal distribution of the residuals indicated it was necessary to perform a natural logarithmic transformation on a target variable (Equation 8). In this equation is Y the predicted value, b_0 the intercept, b_1 regression coefficient, x the explanatory variable and ϵ the random error term of the model (fitted values minus predicted values).

$$\log(Y) = b_0 + b_1 x + \epsilon \qquad \text{Equation 8}$$

The predictions of this model are log-transformed. To get untransformed predictions of the model it was necessary to back-transform the predictions. However, it is not statistically sound to simply back-transform by taking the exponent of the prediction. The relationship between the untransformed response variable and linear model is described in

Equation 9. In which b_{0a} is the antilog of b_0 and e is the base of the natural logarithm. If the regression residuals were normally distributed e^{ϵ} can be estimated by $e^{MSE/2}$, where MSE is the mean squared error from the regression (Newman, 1993).

$$Y = b_{0a} e^{b_1 x} e^{\epsilon} \qquad \text{Equation 9}$$

Stepwise regression

The second method used for creating models was stepwise regression; a variable selection algorithm for multiple linear regression. All potential covariates were included in the stepwise regression procedure, for the three target variables clay, sand and silt. In R the function *stepAIC* chooses a model by the Aikaike information criterion (AIC) in a stepwise algorithm and returns this model (Venables and Ripley, 2002). In the case of clay content this was used in the following way:

stepAIC(lm(clay~allcovariates), direction="both", k=log(63))

The first argument of the function step contains the linear model with clay as target variable and all the covariates as independent variables. The second argument is the mode of the stepwise search, whether the stepwise procedure add and drops or only adds or only drops variables. Different values have been tested for direction, resulting in different models. Forward selection resulted in models with lower predicted values than the MLR models. Best results were obtained with starting with a model with all covariates included and allowing the procedure to add and drop variables; therefore direction was set at 'both'. The third argument determines whether the stepwise procedure uses the AIC (in the case of k=2) or the Bayesian information criterion (BIC) (in the case of k=log(n)) (Venables and Ripley, 2002). In this case was chosen for BIC, because it resulted in better models. BIC reduces the risk of overfitting by introducing a penalty term for the number of parameters included in the model; this penalty term is greater in BIC than in AIC (Schwarz, 1978). Results of stepwise regression confirm this; AIC resulted in models with more variables which were not all contributing significantly (p-value>0.5). Using BIC resulted in models with fewer variables that were all contributing significantly; p-value<0.5.

After the stepwise selection procedure the models were investigated by using the function *drop1* from the 'stats' package. The function *drop1* is applied to an existing model. For all each variable that can be dropped from the existing model, a new model is fitted. The changes in fit for dropping each variable are calculated; showing the best option for removing a variable in the model. The purpose of the investigation was to see whether models with comparable predictive power with fewer variables would be possible.

Regression tree

The third method used for creating models is RT. Modelling a regression tree consists of 3 steps:

1) The initial tree was grown with the following control parameters; minsplit=2, cp=0 and xval=63. These parameters are mainly dependent on the size of the training dataset. The minsplit parameter is the minimum number of observations that must exist in a node in order for a split to be attempted. The cp parameter specifies that any split that does not decrease the overall lack of fit by a factor of cp is not attempted. The minsplit and cp parameter were set at 2 and 0 respectively in order to not restrain the growth of the initial tree.
2) Create diagnostic plots of the initial tree; plot the relative error of cross validations against the size of the tree.
3) Prune the initial full tree by using the 1 SE rule (3.4 Regression Tree).

Considering the fact that RT is suitable for non-linear relationships (Mulder et al., 2011), there is no need for transforming any of the explanatory variables. Therefore three trees were constructed; for clay, silt and sand.

4.7 Prediction of soil texture

The best performing models were used to predict the target variables for the study area. This was done by applying the relationships found between the covariates and the target variables to the complete extent of the study area ( ). Pixels with NoData in one covariate dataset will result in a NoData value for the prediction raster.

All covariate rasters have the same extent and resolution; for each pixel in all of the covariate rasters a target variable value can be calculated (Figure 5). However not all covariate rasters completely cover the area; the ASTER data contains NoData values. The package 'raster' of R was used. This package has an implementation for the function *predict*; meaning that model predictions can be performed on rasters.

The stepwise models predicted values smaller than 0 and bigger than 100. Therefore the predictions of the stepwise models were scaled back to the range of 0-100%. Two methods for scaling back were looked into.

a) Adding up the individual predictions of the stepwise models of clay, silt and sand to a total. After which the individual predictions were divided by the total and multiplied by 100.
b) Adding up the predictions of the two best performing models (called here x and y) to a total. Subtracting this total from 100; the result of this subtraction is the predicted fraction for the soil texture class with the worst performing model (called z), provided that this result is positive. If the result is negative, than the prediction for z is set at zero. This also means that the total for the predictions for x and y is bigger than 100; the predictions of x and y have to be scaled back. This is done by dividing the prediction of the individual models of x and y by the total of x and y and multiplying the result by 100.

Note, the unprocessed predictions of the stepwise models will be referred to as unscaled predictions; the scaled predictions will be referred to as scaled predictions. The stepwise models with the subsequent scaling of the predictions will be referred to as the scaled models.

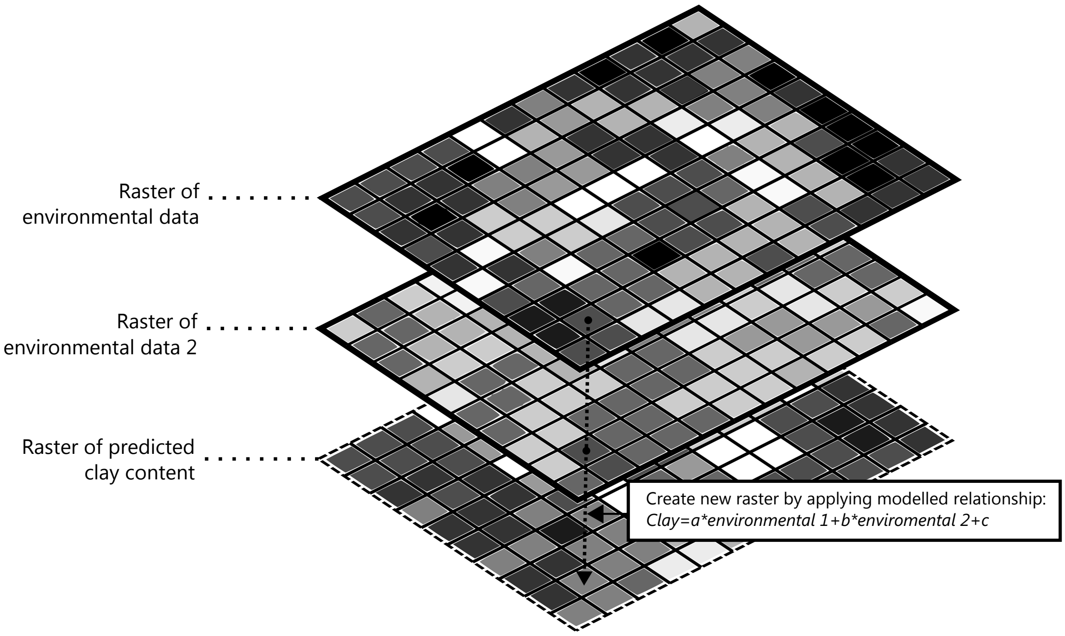

Raster of environmental data
Raster of environmental data 2
Raster of predicted clay content
Create new raster by applying modelled relationship:
*Clay=a*environmental 1+b*enviromental 2+c*

Figure 5: Schematic representation of prediction of clay content with modelled relationship in the case of two covariate rasters (more covariate rasters are used in actual models).

Prediction consisted of seven steps:

1) Applying best performing models to the entire extent of the study area; this resulted in three rasters.
2) Calculating a total predicted fraction raster by adding up the fractions of the three rasters at step 1.
3) Scaling back the fractions calculated at step 1 to a range of 0-100% by dividing each of these three rasters by the total fraction raster and multiplying it by 100.
4) Calculating the RMSE of the scaled predictions.
5) Taking a random sample of the unscaled prediction rasters of step 1 of n=50000 and taking a random sample of the scaled prediction rasters of step 3 of n=50000.
6) Creating histograms of the samples taken at step 4.
7) Creating maps of the predictions:
    - Maps of the scaled predictions for clay, silt and sand.
    - A RGB (red, green and blue) map of the scaled predictions, by stacking the scaled prediction rasters of clay, silt and sand.

4.8 Cross-validation

Only one option for the validation of the models was available; cross-validation (CV). This is due to the unavailability of an independent validation data set and limited size of the sample to validate the models. The applied cross validation methods was leave-one-out cross validation. This is equivalent to n-fold cross validation; n is the number of data points on which the model is trained (Alfons, 2012).

Cross-validation was performed for the best performing models. Both the unscaled and scaled models were cross validated.

1) Cross validation of the unscaled models was performed in R with the package 'cvTools' for the linear models (Alfons, 2012). The function *CVlm* is used to obtain measures of predictive accuracy of the models. The number of folds is set at 63 (this is equal to leave-one-out CV). The function returns the cross-validated root mean squared error (RMSE). The function *CVlm* was used on all the models.
2) Leave-on-out cross-validation for the scaled was performed in R with the help of custom script. This consisted out of 63 iterations, each iteration one data point was taken out the dataset as test data and the remaining data would be used as training data. Unscaled predictions for clay, silt and sand were done with the models fitted on the training dataset. These predictions were then scaled to the range of 0-100%. The cross-validated scaled predictions for clay, silt and sand were subtracted from the value of the test dataset; from this difference the squared root was taken. After all cross-validation iterations were finished, the mean of all the squared root differences were taken for clay, sand and silt.

4.9 Error estimation of prediction

An estimation of the error of the prediction was made by creating a 95% confidence interval map of the prediction for each of the stepwise regression models. These maps were made by predicting a raster of the lower bound of the 95% confidence interval and a raster of the upper bound of the 95% confidence interval in R with function *predict* of the 'raster' package. After which the lower bound raster was subtracted from the upper bound, resulting in a raster of the 95% confidence interval.

CHAPTER 5: RESULTS

5.1 Exploratory data analysis results

Results for the exploratory data analysis will be presented step by step as described in 4.5.

Step 1

Inspection of the sampling distribution shows that the sampled clay content is not normally distributed and skewed to the right (figure 5). Clay content with a percentage between 0% and 10% occurs most frequent. The distribution for sand and silt content seems normally distributed, with the mean of the distribution around 50%. The distribution of silt content is skewed to the left and the distribution of sand content is skewed to the right.

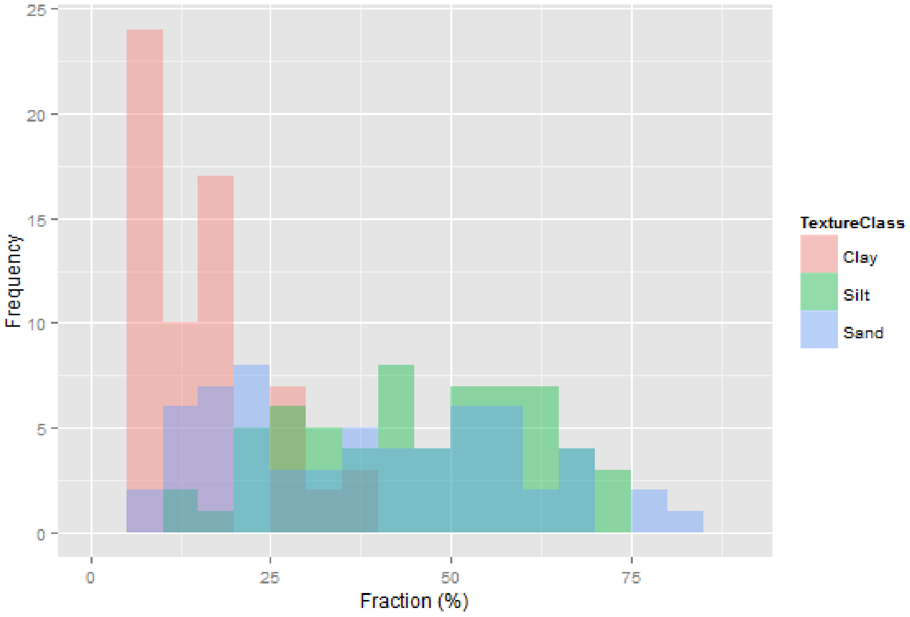


Figure 6: Distribution of target variables of sample points for clay, silt and sand content.

Step 2

Table 4 lists the correlation coefficient between each target variable and each covariate. Correlation<0.2|correlation>-0.2 is considered to be very weak or negligible. Covariate (or variable) selection for the MLR models is based on this criterion (see also).

Table 4: Correlation coefficients between target variables and covariates. Coefficient>0.2 bold and green; corresponding covariate selected for multiple linear regression model

Covariate	Clay	Log(Clay)	Silt	Sand	Covariate	Clay	Log(Clay)	Silt	Sand
A1	-0.11	-0.12	**-0.37**	**0.34**	PC3	0.05	0.07	**0.40**	**-0.34**
A2	-0.13	-0.15	**-0.38**	**0.36**	NDVI	-0.10	-0.12	**0.34**	**-0.22**
A3	**-0.22**	-0.25	**-0.33**	**0.36**	DEM	**-0.21**	**-0.21**	-0.13	0.19
A4	-0.17	-0.16	**-0.35**	**0.35**	Slope	0.02	-0.01	**0.21**	-0.18
A5	**-0.21**	**-0.20**	**-0.38**	**0.39**	TWI	0.16	**0.23**	0.07	-0.13
A6	**-0.22**	**-0.20**	**-0.38**	**0.40**	Curvature	0.08	0.08	0.11	-0.12
A7	**-0.24**	**-0.22**	**-0.39**	**0.41**	Profile	-0.05	-0.05	0.02	0.00
A8	**-0.22**	**-0.20**	**-0.37**	**0.39**	Plan form	0.09	0.10	**0.24**	**-0.23**
A9	-0.17	-0.16	**-0.32**	**0.33**	Temperature	**0.21**	**0.21**	0.15	**-0.21**
A10	0.17	0.10	**0.30**	**-0.31**	Precipitation	**0.30**	**0.29**	**0.30**	**-0.37**
A11	**0.24**	0.17	**0.32**	**-0.36**	Seasonal	0.04	0.01	**0.32**	**-0.27**
A12	0.10	0.03	**0.25**	**-0.24**	SWI	0.10	0.10	0.13	-0.15
A13	0.11	0.09	0.06	-0.10	CI	-0.16	-0.18	-0.19	**0.22**
A14	**0.26**	**0.25**	**0.23**	**-0.30**	MI	0.08	0.01	**0.28**	**-0.26**
PC1	**0.20**	0.19	**0.37**	**-0.38**	QI	**0.34**	**0.34**	0.11	**-0.23**
PC2	-0.05	-0.11	0.04	-0.01					

b) Strong correlations between covariates are present in the covariate dataset ( ). Some correlations are caused by deriving covariates from covariates; MI has for instance a strong correlation with ASTER TIR bands. MI is calculated with these TIR bands. Correlations that cannot be explained by derivations are:

- The strong correlation between temperature and DEM. Temperature is negatively correlated to elevation.
- Precipitation is not well correlated with NDVI. However, precipitation is well correlated with the VNIR and SWIR ASTER bands, just like NDVI.
- Temperature is moderately negatively correlated with NDVI.
- SWI has a strong negative correlation with DEM.

Step 3

Boxplots for checking whether the soil sample locations sampled the thematic range of the covariate rasters ( ) show that:

- The range of the ASTER VNIR and SWIR bands and PC1 are well represented by the soil sample points.
- NDVI is not well represented by the soil sample set.
- The range of the DEM and derived covariates are not well represented by the soil sample points.
- The range of ASTER TIR bands is not well represented by the soil sample points. In all the boxplots of the ASTER TIR bands it is visible that the soil samples did not capture the lower values of the ASTER TIR rasters.
- The mineral indices are not well represented by the sample points either.
- Representation of the range of temperature by the sample points is poor too, inspection of the boxplot learned that the sample points did not sample the lower tail of the raster temperature distribution.
- The sample points failed to account for the upper tail of the SWI raster.

5.2 Modelling results

Different methods have been used for modelling soil texture. Methods that can be used are multiple linear regression (MLR), stepwise regression and regression trees. Results of the different models will be presented below.

MLR based models

Results of the models ( ) show that it was necessary to log-transform the target variable clay; the residuals of the MLR model of clay were not normal distributed. This was indicated by the $p_{\text{SW-test}}$ ($p_{\text{SW-test}}$ for clay 0.02, silt 0.27, sand 0.21, log(clay) 0.53) and visual inspection of the standard QQ plots.

The model for log(clay) shows that even though the residuals are normal distributed, the model is still not performing well (Figure 7). The $p_{\text{F-test}}$ for the model of log(clay) is 0.06; this means there is no significant linear relationship between the included explanatory variables and log(clay). This is also visible in the fitted vs. sampled values plot (sampled values are the values from the soil sample dataset); the linear model cannot predict the sampled values. The models for sand and silt have normal distributed residuals and a statistically significant relationship between explanatory variables and target variable.

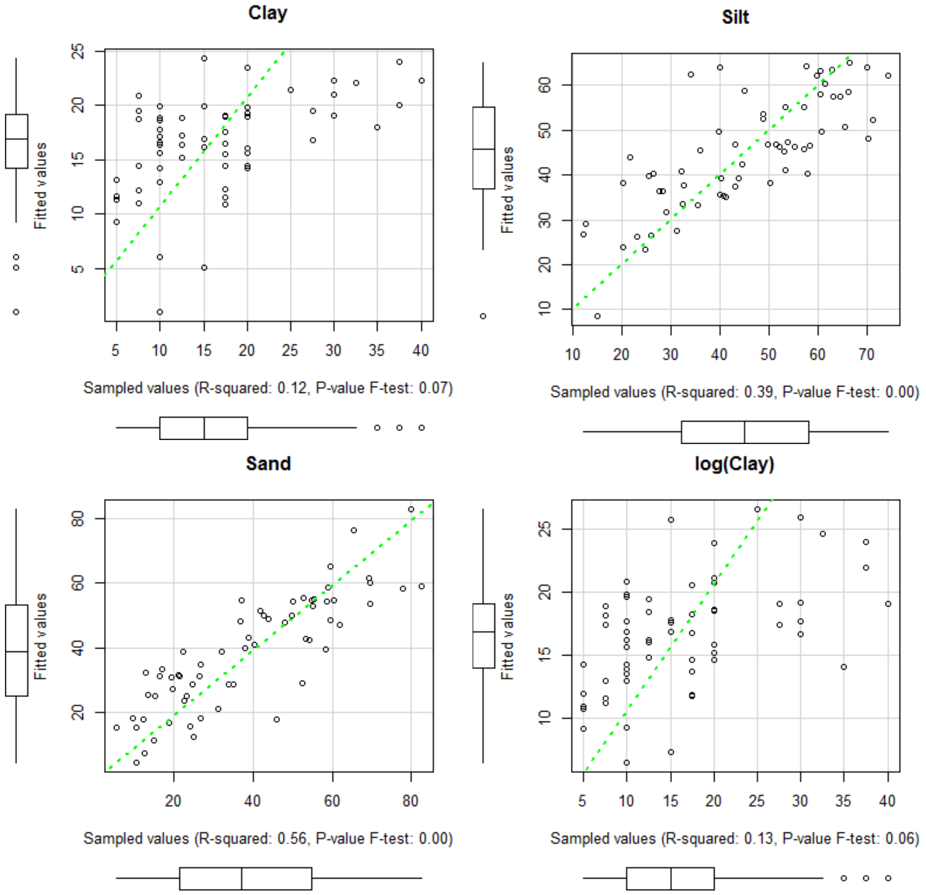


Figure 7: Fitted versus sampled values for MLR models. Dotted green line is y=x. Distributions of the fitted and sampled values are indicate by boxplots on the axis. Reported R^2 is the adjusted R^2.

Another measure for model performance is the $\text{RMSE}_{\text{Fitted}}$ (Table 5). Note the relatively low values for the models of clay. The models for sand and silt are better performing, but have a higher $\text{RMSE}_{\text{Fitted}}$. The MLR models for silt and sand have all of the CLORPT factors incorporated in the models (Table 5). Clay does not have the CLORPT factor Organism included in the model. However, most of the explanatory variables do not contribute significantly to the response.

Table 5: Covariates included as independent variables in MLR models grouped per soil forming factor per model with the RMSE of the fitted values. Covariates that did not contributed significantly ($p_{T\text{-}test}$>0.05) to the response are underlined.

	Soil forming factors					
Model	**Cl**	**O**	**R**	**P**	**-**	**RMSE Fitted (%)**
Clay	temp+precip		DEM	QI	A5+A6+A7+A8+A14	6.5
Log (Clay)	temp+precip		DEM+TWI	QI	A3+A6+A7+A8+A14	6.1
Silt	precip+seas	NDVI	slope+pla_curv	MI	A1+A2+A3+A4+A5+A6+A7+A8+A9+A10+A11+A12+A14+PC1+PC3	8.0
Sand	precip+seas	NDVI	DEM+pla_curv	CI+MI+QI	A1+A2+A3+A4+A5+A6+A7+A8+A9+A10+A11+A12+A14+PC1+PC3	8.6

Besides normal distribution of the residuals, linear models are required to have homoscedastic residuals; constant variance of residuals. Plots for checking this are fitted values versus the residuals plots ( ). These plots show that the residuals are not dependent on the fitted values for the models of silt and sand. The residuals for clay and log(clay) seem to increase with decreasing fitted values.

The conducted Breusch-Pagan for constancy of residual variance over the independent variables indicate that residuals are not correlated with any of the explanatory variables ($p_{BP\text{-}test}$ clay 0.44, silt 0.98, sand 0.83, log(clay) 0.06). However the $p_{BP\text{-}test}$ for log(clay) is low, indicating that the model for log(clay) is not performing well.

Stepwise regression based models

These models are MLR models of which the explanatory variables have been selected by a stepwise regression procedure (from now on stepwise models). QQ-plots ( ) show that the residuals of all the models were normal distributed. This is also indicted by the $p_{SW\text{-}test}$; clay 0.16, silt 0.36 and sand 0.82.

The stepwise models perform better than the MLR models (Figure 8). All adjusted R^2 are higher than the stepwise models. However the plot for clay shows that the points are not scattered around the line y=x. Points with a high value for clay content are underestimated by the stepwise model for clay, which indicates that the linear model is not performing well. For sand and silt the points are scattered around the line y=x. The $p_{F\text{-}test}$ of all the stepwise models indicates that there is significant linear relation between the explanatory variables and the target variable.

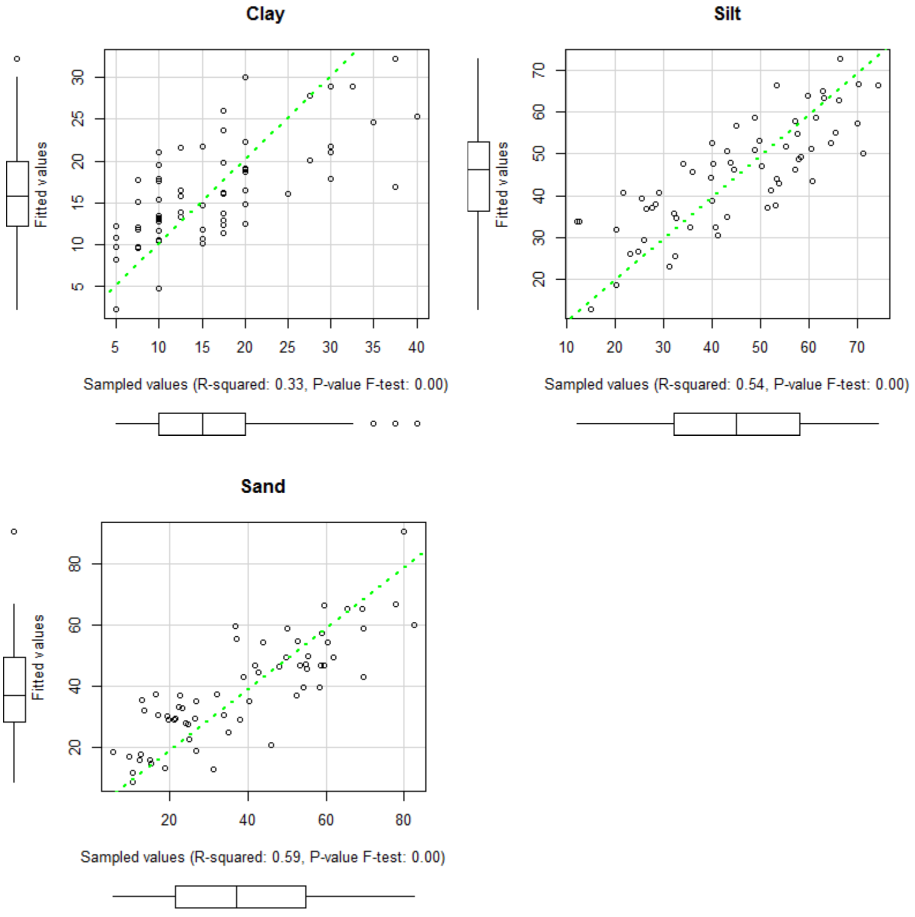


Figure 8: Fitted versus sampled values for stepwise models. Dotted green line is y=x. Distributions of the fitted and sampled values are indicate by boxplots on the axis. Reported R^2 is the adjusted R^2.

The $RMSE_{Fotted}$ for the stepwise model of clay shows a relative low value (Table 6). The models for sand and silt are better performing judging by the adjusted R^2, but have a higher $RMSE_{Fitted}$. The stepwise model for silt has all of the CLORPT factors incorporated in the model. In the model for clay and sand the CLORPT factor Organisms is not included. Only QI is not contributing significantly to the response, however the $p_{T\text{-}test}$~0.05 for QI in the stepwise model for clay.

Table 6: Covariates included as independent variables in stepwise models grouped per soil forming factor per model with the RMSE of the fitted values. Covariates that did not contributed significantly ($p_{T\text{-test}}$>0.05) to the response are underlined.

	Soil forming factors					
Model	**Cl**	**O**	**R**	**P**	**-**	**RMSE fitted (%)**
Clay	precip+ seas		pro_curv	CI+ MI+ QI	A8+A9+A10+ A11+A12+A14+ PC1	5.2
Silt	temp+ precip	NDVI	slope+ pla_curv	CI+ MI	A3+A4+A6+A9+ A12+A14	8.2
Sand	temp+ precip+ seas		slope+ pla_curv	CI+ MI	A3+A4+A5+ A9+A10+A11+ A12+A14	9.2

The plots of the residuals versus the plotted values ( ) show that the residuals have constant variance over the predicted values. This is also the case for the model of clay, even though the plot of clay (Figure 8) seems to suggest the opposite. The conducted Breusch-Pagan test for each of the models shows that the residuals also have constant variance over the explanatory variables ($p_{BP\text{-test}}$ clay 0.49, silt 0.89 and sand 0.73).

Investigation of the stepwise models by using the function *drop1* showed that removing variables from the stepwise models did not result in better models. Dropping the variable QI in the stepwise model for clay resulted in a model with other variables not significantly contributing to the response. For sand and silt removing variables resulted in a lower adjusted R^2 (a decrease of ~0.05 in adjusted R^2 per dropped variable).

RT based models

Plots of the initial trees for the target variables clay, sand and silt (Figure 9) show the cross-validated error versus the size of the tree. The horizontal dotted line marks the point of the minimum cross-validated error rate plus 1 standard error of the minimum cross-validated error rate, for using the 1 SE rule. According to this rule the best tree is the smallest tree with a cross-validated error smaller than the horizontal dotted line. For sand and silt the best tree has only one split. For clay the best tree has, according to 1 SE rule, five splits. However, the cross-validated error for this tree is higher than for a tree with zero splits. This means that the mean is a better predictor for clay than the tree with 5 splits. Therefore regression trees do not have any predictive power for modelling soil texture with this soil sample dataset.

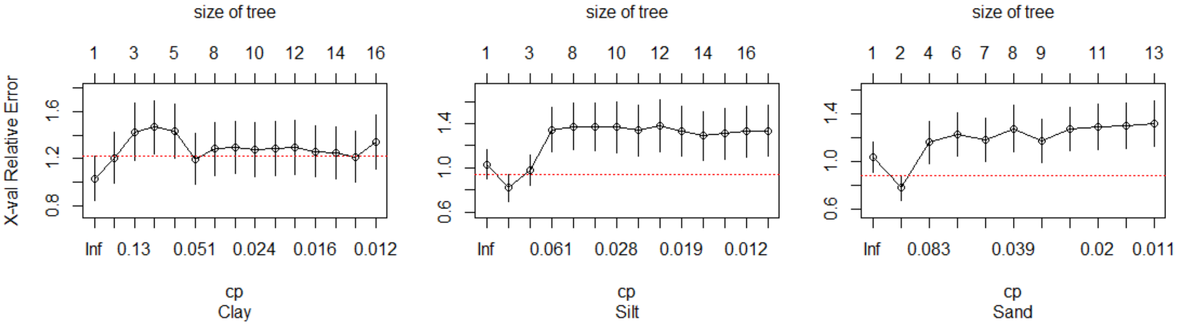


Figure 9: Plots of size of tree (x-axis, expressed in complexity parameter) against cross-validated relative error (y-axis) for full regression trees for the response variables clay, sand and silt. Horizontal dotted line is for applying 1 SE rule.

5.3 Prediction for study area

Prediction of soil texture was performed with the stepwise regression models, because these models were best performing. The scaling method *a* was selected (4.7 Prediction of soil texture), due to the lower RMSE$_{\text{Fitted}}$ for this scaling method.

Table 7: RMSE of fitted value of scaled predictions for stepwise models for the different scaling methods

Scaling method	Clay	Silt	Sand
a	5.0%	7.8%	8.9%
b	6.4%	8.2%	9.2%

The RMSE$_{\text{Fitted}}$ of the scaled predictions by scaling method *a* (Table 7) is lower than the RMSE$_{\text{Fitted}}$ of the unscaled predictions (Table 6). This means that scaled models by scaling method *a* are better performing than the unscaled models for the fitted values. Histograms of scaled and unscaled prediction (Figure 10) show that the scaling of the predictions does not influence the overall shape of the distribution much.

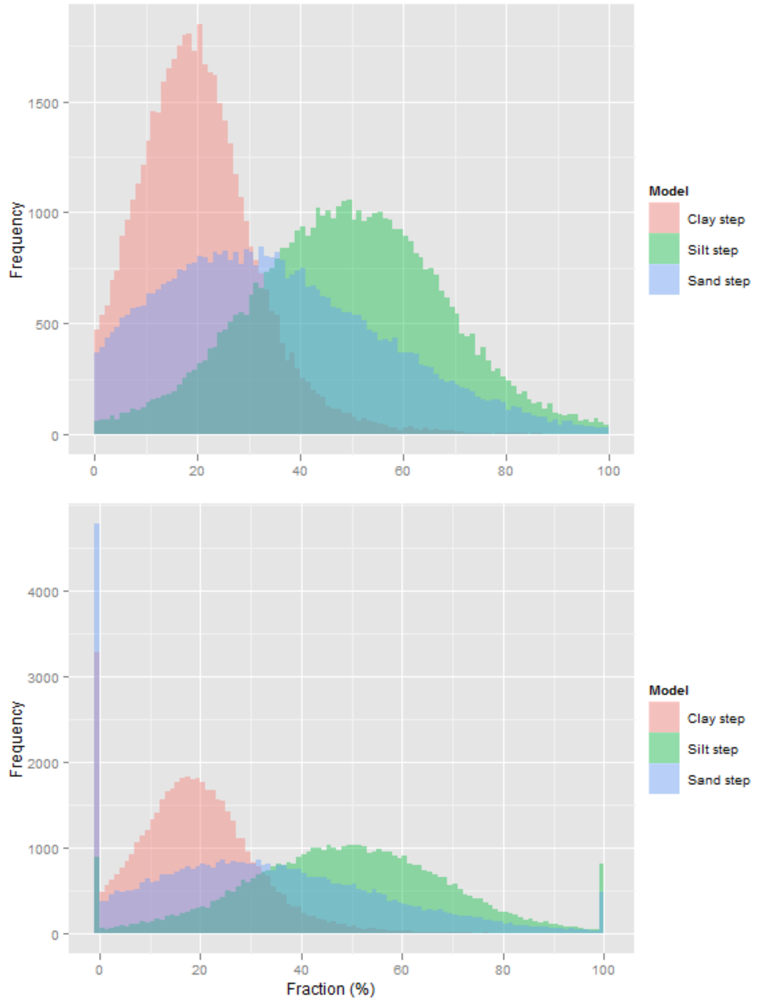


Figure 10: Histogram of the unscaled (top) and scaled (bottom) predictions for clay, silt and sand (random samples taken from prediction raster with n=50000). Unscaled histogram does not show all data, x-axis constrained to range 0-100%.

The histograms do show that at the extremes of the range the distribution is heavily influenced, because all the unscaled values that were too low or high were set at 0 and 100% respectively.

If we compare the distribution of the scaled prediction with the distribution of the sampled values (Figure 6) we see similar patterns for silt and sand. For both distributions the peak of sand is ~25% and the peak of silt is ~50%. For clay the patterns are slightly different, this is caused by the model for clay, which does not predict enough low values in the range of 0-10%. The low number of samples makes it difficult to compare, but the major features of the distributions are clearly visible in both plots.

Prediction maps

The predicted maps consist out of three contiguous sections; one on the left, one in the centre and one on the right. The scaled predictions for clay, silt and sand (Figure 11) and the RBG map of the scaled predictions (Figure 12) show that:

- High values of clay (30-50%) are mainly predicted in the southeast of the study area and for locations that are part of the river courses. Note the estuary of the river in the bottom of the middle section; this clearly stands out with high predicted values for clay content.
- High values for silt (50-80%) are mainly predicted in the west of the study area. Note the very low values for silt in the northern part of the middle section and in south of the right section.
- Very high values for sand (80-100%) are predicted in the northern part of the middle section. High values are predicted in the northern part of the right section. Low values for sand are predicted in the left section

The maps also show several data artefacts:

- The diagonal stripe with abnormal values running SW-NE in the centre section is visible in all prediction rasters and is caused by very low values in the ASTER TIR bands and is a data artefact.
- The thin line with NoData values is a data artefact in ASTER bands 7 till 9 (SWIR) caused by the mosaicking procedure.
- At a close zoom level diagonal lines running SW-NE are visible as deviating values in the prediction rasters. This is caused by the artefacts from the DEM derived rasters.
- Noise from the ASTER TIR bands is clearly visible in the prediction rasters, especially in the prediction raster of clay. This contributes to the grainy appearance of the prediction rasters.

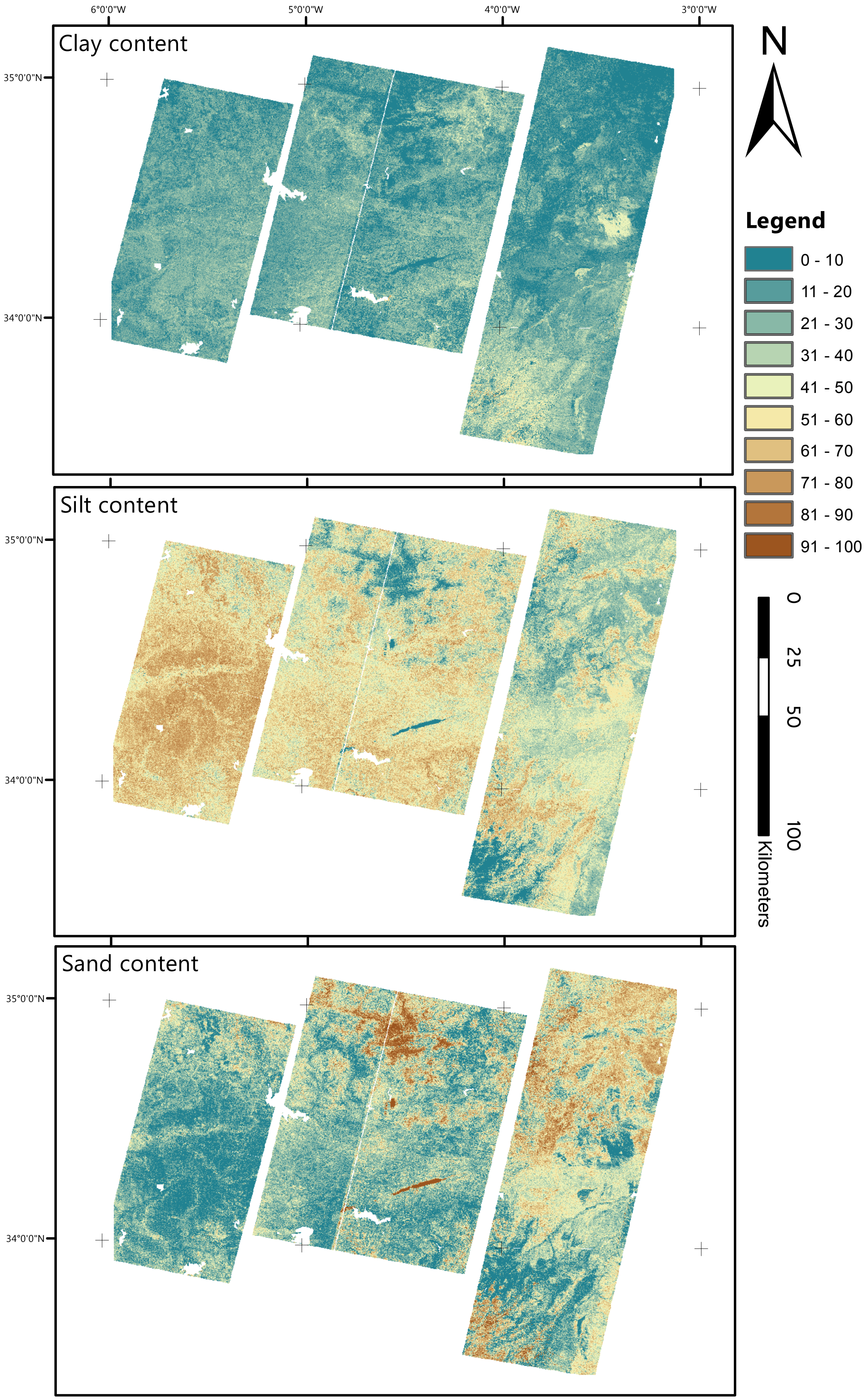


Figure 11: Maps of scaled predictions of clay, sand and silt content.

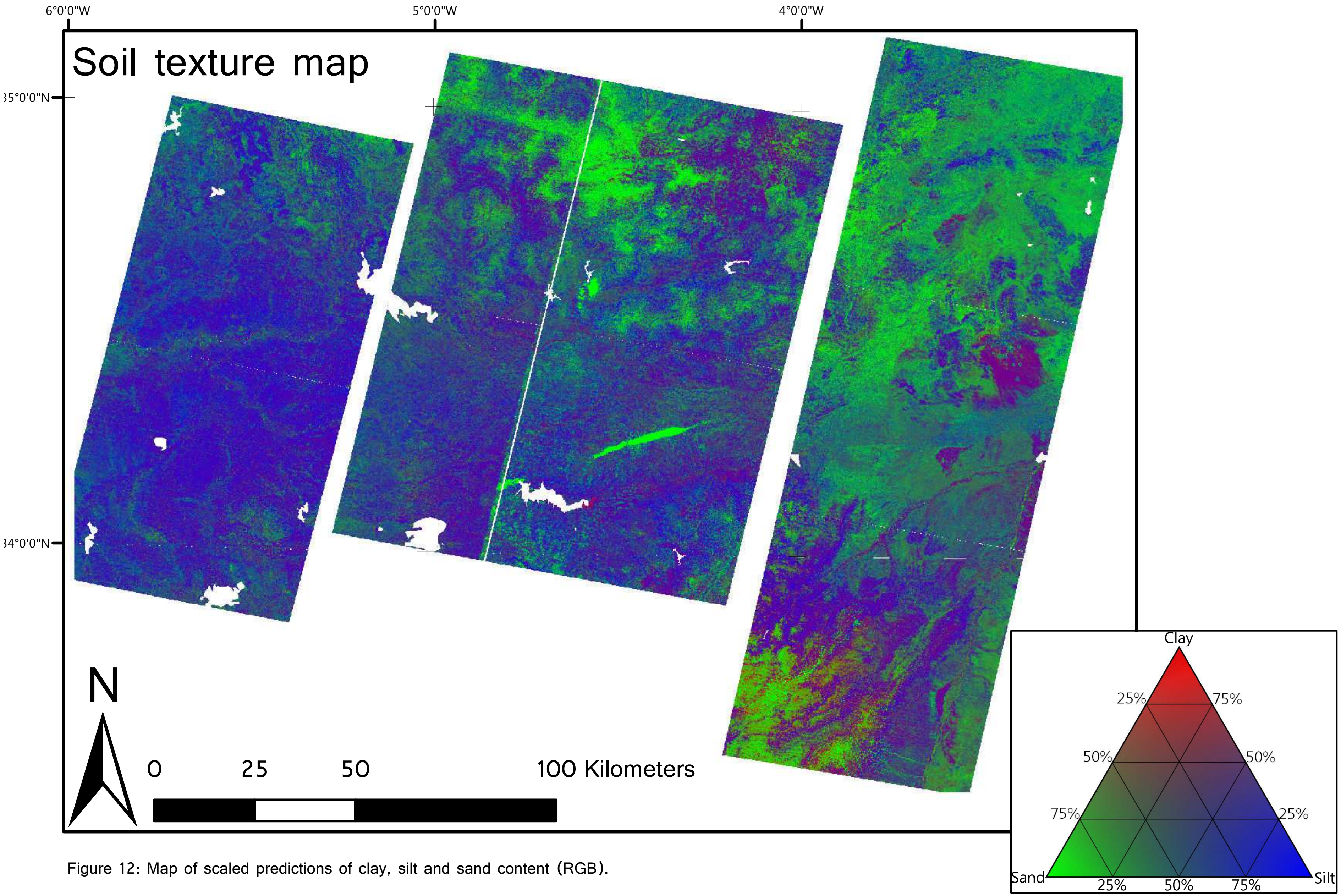


Figure 12: Map of scaled predictions of clay, silt and sand content (RGB).

5.4 Cross-validation of models

Cross-validation was performed for the unscaled models and the scaled models (Table 8). Cross-validation results shows that scaling the prediction does influence the accuracy of the prediction negatively (increased $RMSE_{CV}$ of ~2%).

Table 8: Cross-validated root mean square predicted error of the different models.

	Target variable	**$RMSE_{CV}$**
Unscaled models	Clay	7.9%
	Silt	12.3%
	Sand	15.4%
Scaled models	Clay	10.0%
	Silt	14.3%
	Sand	15.5%

5.5 Error estimation of predictions

The maps of the 95% confidence interval indicate that the unscaled predictions have quite a wide confidence interval. This is also shown by the mean of the confidence interval rasters; for clay ~20%, silt ~30% and sand ~40%. Locations with the biggest confidence interval are areas with a pronounced topography. Especially the northern part of the centre section of the image shows big confidence interval values. Also in the south-east of the image big confidence interval values for silt and sand prevail.

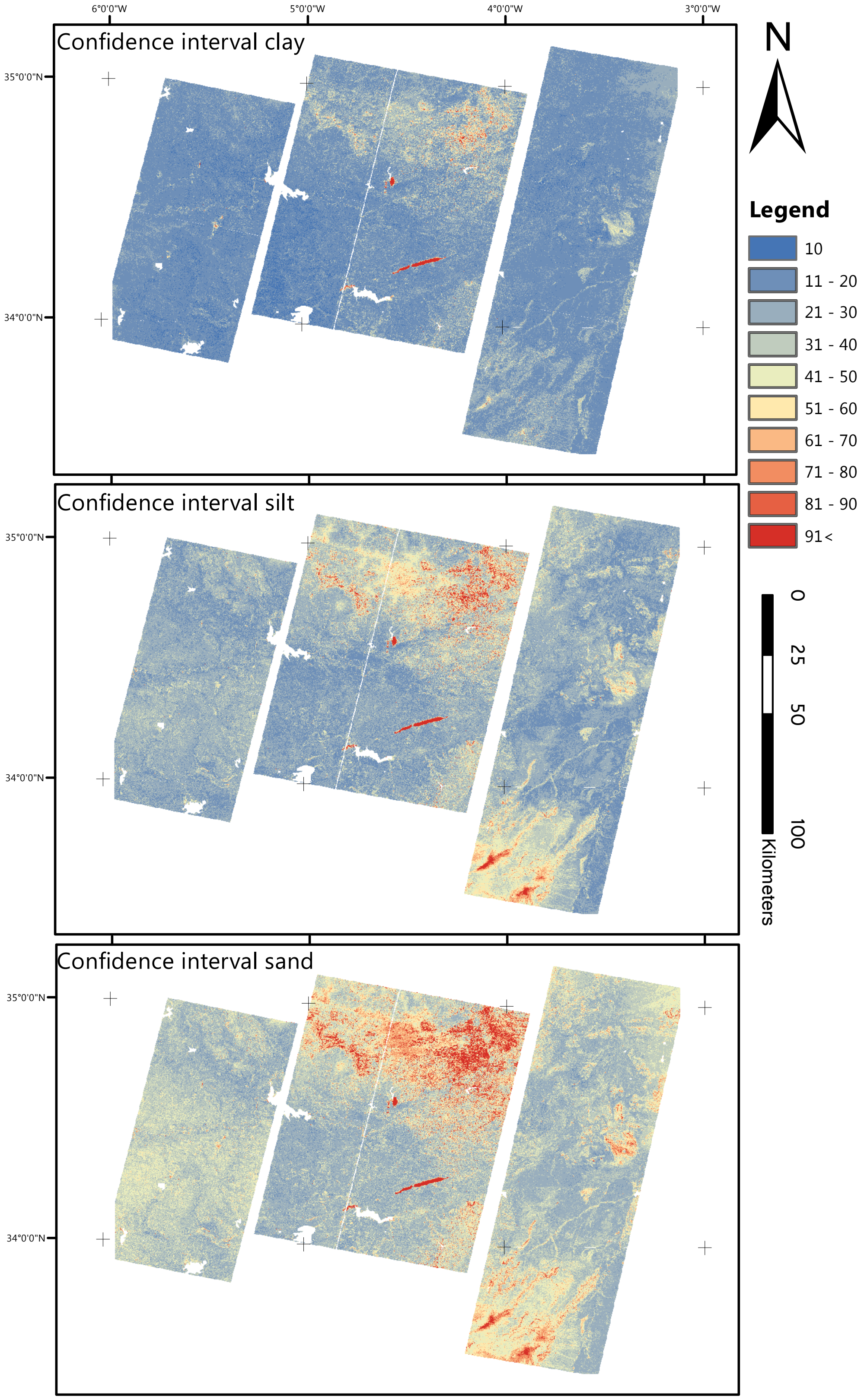


Figure 13: Confidence interval (95%) for stepwise regression model predictions (unscaled)

CHAPTER 6: DISCUSSION

6.1 Multiple linear regression and stepwise regression

The MLR models are poorly fitted models, only a few of the explanatory variables contribute significantly to the response (Table 5). This is the case for all fitted models; clay, silt, sand and log(clay). This shows in the plots of the fitted values versus the sampled values (Figure 7); the point are not neatly scattered around the line y=x. However, the diagnostic tests (Shapiro-Wilk and Breusch-Pagan test) to assess the performance of the models did not signify that the models were poorly fitted.

Including variables in a regression model based solely on their individual relationship (correlation coefficient>0.2 & <0.2) with the target variable appears not a good idea. The reason for this is, that this method does not acknowledge interaction between the explanatory variables. The stepwise variable selection procedure, on the other hand, does take interactions between variables into account (Neter et al., 1996). This results in better performing models (Table 6); almost all explanatory variables are significantly contributing to the response and the residual standard error is lower for all models, compared with the MLR models.

That the stepwise models are better performing is also illustrated by Figure 8; these plots show that the points of sand and silt are centred more around the line y=x compared with the plots of the MLR models. The explained variance of the soil sample dataset is higher for the stepwise models as well; for clay 33%, silt 54% and sand 59%.

The plot for clay shows that the stepwise model for clay does have a better fit for the sampled values, compared with MLR models of clay and log(clay). Despite this, the model seems to underestimates the value for clay for higher sampled values. Whether this is really the case, cannot be put to a test, because only a limited number of samples with high values for clay are taken. Diagnostic plots of the residuals of the stepwise models ( ) indicate that the residuals are normal distributed; this was also indicated by the Shapiro-Wilk test of the residuals. The stepwise models of sand and silt seem to be a good fit on the data. The fit of the model of clay is not as good as the fit of the model on the other variables (Figure 8), although diagnostic measures (SW_{test} and BP_{test}) do not suggest this.

A problem for regression models in general is that regression extrapolates target variable values for locations in the feature space that are not present in the training data. This is evident in the prediction rasters for silt and sand content; for areas with pronounced topography (steep slopes) extreme values are predicted (Figure 11 and Figure 12). This is due to soil sample dataset. The locations of the soil sample dataset do not properly represent the variability of slope, TWI and curvature of the study area ( ). Pixels that have values for these covariates that fall outside the sampled ranged will have extrapolated values fitted to by the regression models.

6.2 Performance of regression tree

The expectation was that RTs would perform better in modelling soil texture than the MLR model; it was assumed that the soil-landscape relations would be non-linear (Mulder et al., 2011). The results (Figure 9)show that RT based modelling is not feasible. The cross-validated error keeps on increasing with increasing tree size; a tree without any splits and therefore the mean, is the best predictor. However, this result does not mean that RTs are not suitable for modelling soil-landscape relationships, it only means that RTs might not be suitable for this soil sample set.

An advantage of a RT is that the tree does not extrapolate any values. The tree predicts values that fall in the range of the training dataset; in that sense the prediction of a RT is more realistic, compared with a regression model. The tree seems to predict discrete values, but the prediction of the tree can be thought of as histogram of the regression surface (Breiman, 1984). The predicted values do not represent actual values, but a histogram bin with a certain size. This can be seen as a disadvantage of the regression tree; the discrete predictions at each terminal node results in a non-smooth of the prediction surface (McKenzie and Ryan, 1999).

The bad performance of the RT could be due to (1) the nature of the regression tree. RT it is a greedy algorithm; at each node an optimum decision is made based on the best partition of data by a specific variable. The optimum decision for a node might not be the optimum decision for the entire tree.

The concept of greedy algorithms might be best explained with the help of an analogy. Choosing an optimal subset of explanatory variables in decision space can be compared with determining a route from point a to point b in geographic space. At each cross-road you can choose which way to go. You can take the route that gets you to the next cross-road the fastest. However this might not be the overall fastest route to point b. The narrow inaccessible path leading the other way might be a shortcut; getting you to point b the fastest. This example illustrates the weakness of growing regression tree; seemingly unsuccessful paths are not explored. This might mean that predictive trees can be grown on this dataset, but RT based modelling is not capable of growing the trees.

Another reason (2) for the bad performance of the regression based models might be the inherent structure of the dataset. Perhaps the dataset is simply not fertile (read: unsuitable) for growing RTs. Non-greedy methods for variable selection for RTs exist. Genuer et al. (2010) proposed a variable selection method for CART based models; random forests. It is recommended to use random forests on this dataset, for checking whether the dataset really is unsuitable for growing RTs.

6.3 Covariates used in the models

The main premise of this study was that if the soil forming factors are known, soil properties can be determined. However, the stepwise models presented in this study do not take all CLORPT factors in to account (Table 6). The soil forming factor Organisms, which is represented by NDVI, is not explicitly present in all stepwise models. The reason for this might be that the soil spectral response is gradually blocked with increasing value of NDVI, until a certain threshold value. This means that with increasing value of NDVI, the information on the soil contained by the spectral reflectance decreases. If NDVI itself contains information on the soil, this effect can be compensated for by including in NDVI in the model. This is illustrated by the stepwise model of silt; NDVI is only included in this model and silt has the strongest correlation with NDVI.

The soil forming factor Time is also not included in any of the models. The reason for this is that it is difficult to characterize the age of soils. There are methods to characterize the age of the soil, such as radiometric dating, but these methods are not capable of scanning and producing full coverage maps in way RS does. The soil forming factor Time can be derived with export knowledge, but it remains difficult to characterize well (McBratney et al., 2003).

All three stepwise models also have covariates included which are not explicitly linked to any of the soil forming factors. These covariates are the ASTER reflectance and emissivity bands and PCs of the ASTER reflectance. It might not be possible to link these covariates explicitly to any of the soil forming factors. However these bands do contain information the soil forming environment, because all of these variables are contributing significantly to the response for the stepwise models (Table 6).

If we look at the covariates that have been included in the stepwise models, we can distinguish five important groups that help to characterize soil texture:

1) Covariates on climate
2) Slope and curvature related variables.
3) Mineral indices.
4) ASTER SWIR bands.
5) ASTER TIR bands.

Surprising in the first group is that temperature is included in the models for silt and sand. It was not expected for temperature to be explicitly related to soil texture. Temperature in itself does not directly govern the processes that lead to differentiation in soil properties. The inclusion of covariate temperature in the models could be explained by the fact that temperature might help to distinguish major climatological areas at the scale level of the study area; these areas might differ in soil development. Temperature is off course related to the topography.

Another explanation for the inclusion of temperature in the models, is that the WorldClim temperature surface is generated with weather stations record and a DEM. This is shown by the strong correlation between temperature and elevation ( ). However, the presence of climatic variables in all stepwise models shows that climatic variables are important for soil property prediction.

Slope and curvature related variables help to characterize soil texture, because soil development occurs in response to the way water moves through and over the landscape (Moore et al., 1993). It was expected that TWI would also help to explain soil texture variation. However, this was not the case. Why TWI was not included in any of the models remains unclear. Perhaps this might be due to the poor quality of the DEM. Elevation was not included in any of the models, since elevation in itself does not determine how the water moves through and over the landscape.

The mineral indices, derived from the ASTER TIR bands, help to explain the variation in parent material (Ninomiya et al., 2005). Mineral indices appear to be an important factor for determining soil texture. Part of the variation in parent material is also explained by the presence of ASTER TIR bands in all of the models. However, it was shown that the ASTER

TIR bands are not well represented by the soil samples (5.1 Step 3). This might be due to covariates used for the constrained LHS method; for this the PCs of only the ASTER VNIR and SWIR bands were used.

Covariates that were absent in the models are SWI and the PC bands. Soil moisture (and therefore SWI) was expected to be well correlated with different soil texture classes. Perhaps SWI was not correlated with the different soil texture classes due to the characteristics of the data. The big difference in spatial resolution might be problematic (12.5 km vs. 90 m ASTER TIR); this means that soil moisture is averaged out over a large area. The absence in the model and large spatial resolution of SWI indicates that the scale of the SWI raster does not match the soil forming process scale. This makes SWI unsuitable for soil property modelling.

The absence of the PCs in the models is surprising, since the first three PCs and the DEM were used for determining the sample locations with the use of constrained LHS sampling (Mulder et al., 2012). It was shown that the sample points chosen with this sampling method did represent the variability of soil texture of the study area. Therefore it was expected that the first three PC's would be included in the models for determining soil texture.

So the absence of the PCs was surprising. Perhaps the absence of the PCs in the models might be explained by the presence of ASTER VNIR and SWIR bands in the models; adding the PCs would lead to data redundancy.

6.4 Spatial prediction

The major geological features of the study area are known, this will help to validate the soil texture maps with this expert knowledge. The eastern part of the study area contains a spur of the Atlas Mountains. The weathering product of the bedrock of the Atlas is sandy; therefore it is expected that the eastern part of the study area is predominantly sandy. The western part of the study area contains a flysch basin. Flysch is a sequence of sedimentary rocks deposited in a marine environment, and the weathering product is expected to have high a fraction of silt. The north of the study area is part of the Rif Mountains; this a relatively young mountain range consisting mainly out of schists. The weathering product of schists is typically clayey, but the Rif Mountains are relatively young; therefore not too many clayey soils are expected.

Overall, these expectations are met by the generated soil texture map (Figure 11 and Figure 12), which show silty soils in the west and sandy soils in the east. Also high values for clay are predicted in the Rif Mountains; these are located in the north of the middle section of the study area. However, in the Rif Mountains very low values (0-10%) for clay were predicted as well, which coincides with very low values (0-10%) for silt and very high values (91-100%) for sand. This indicates that the models have difficulties predicting soil texture for this area. After inspection of the different covariate rasters used it appears that this is a densely vegetated area, which explains the extreme predictions; the soil sample dataset does not contain soil sample points with a high NDVI ( ). The model is simply not trained for regions with dense vegetation; the dense vegetation is blocking part of the soil spectral response on which the model is trained. The pronounced topography of this area also contributes to the extreme predictions. Elevation and derived covariates are not represented well by the soil sample dataset as well.

Another problematic region is the spur of the Atlas Mountains; located in the south-east of the study area. Here very high values for sand are predicted as well. This region does not have notable high values for NDVI, but does have a pronounced topography. Inspection of the DEM shows that this region has elevations over 2,500 m; this explains the extreme predictions for this region. The ASTER VNIR bands showed that even some of the mountain tops are covered by snow. These observations are confirmed by the maps of the 95% confidence interval of the unscaled predictions (Figure 13). These maps show that the predictions for sand and silt have a very wide (>100%) confidence interval for both of these regions. For clay this is only the case for the region in the Rif Mountains.

The confidence interval raster also show that predictions are problematic for locations with a pronounced topography and with a high NDVI. Inspection of the covariate rasters show that locations with a pronounced topography coincide with locations with a high NDVI. A positive correlation between elevation and NDVI is also shown by the correlation matrix ( ). This is can be explained by the climate of the study area, the mountainous regions of the study area receive the most precipitation. Therefor most vegetation is present in regions with a pronounced topography; resulting in high values for regions with a pronounced topography. This shows that it is relatively easy to improve the prediction by the models, by simply masking out the regions that display pronounced topographical features and high values for NDVI.

Other landscape elements that are clearly visible in the soil texture maps are the courses of several rivers; these are the thin lines with a high value of clay content meandering though the image. High clay values are mainly found in areas with a low elevation; this is expected, in general clay is deposited by slow flowing water. However, also high values for clay are found for locations with high elevation, as seen in the Rif Mountains.

The observation that elevation and derived covariates are not well represented by the soil sample dataset can be explained. This due to constrains used in the LHS sampling method (Mulder et al., 2012). High elevations were excluded from the possible sample locations to minimize the sampling effort.

An additional possible reason for the poor representation of elevation by the soil sample dataset might be data quality. As mentioned (4.4 ASTER DEM), the DEM contains anomalies for 20% of the DEM. These anomalies are visible in the DEM derived covariates (slope and curvature) as repeating diagonal lines of 1 pixel wide. These anomalies are still visible in the soil texture maps.

The 95% confidence intervals of the unscaled stepwise predictions (Figure 13) helped to determine which regions have uncertain prediction. In general the confidence interval maps show that it is quite certain that prediction is quite uncertain. The mean of the confidence interval rasters is substantial (clay, silt and sand; 20%, 30% and 40% respectively). However the mean does not have to be representative of the spatial distribution of the uncertainty. The maps show that very high values are predicted for just a few locations, while the majority of locations have quite an acceptable uncertainty.

From a statistical perspective, these high values for the confidence intervals have two causes. (1) The models for predicting soil texture do not contain a strong relationship between explanatory variables and the target variable. (2) The dataset on which the models are trained is quite small; a bigger soil sample dataset will result in predictions with a smaller confidence interval.

6.5 Cross-validation of the models

Validating the models posed a challenge for this study; there was no validation dataset available and the training dataset was relatively small (n=63), considering typical soil sample sizes used in other digital soil mapping studies (McBratney et al., 2003).

Therefore cross-validation was the only option. The training dataset theoretically describes all of the soil variability in the study area. Constrained LHS takes care that the entire thematic space of the variables that are related to soil properties are covered. Removing a part of the soil sample dataset for cross validation would result in a training dataset that does not represent all the soil variability in the study area. By using leave-one-out cross-validation this effect is minimized and an accuracy of the prediction can be determined. The RMSE_{CV} determined by leave-one-out cross-validation of the scaled stepwise model for clay is 10%, for silt 14% and for sand 16% (Table 8). A validation with an independent dataset would probably result in a lower RMSE_{CV}, because the original model is slightly different from the cross-validated models. The lower RMSE_{CV} of the model for clay can be explained by the range of clay values of the soil sample dataset (Figure 6). The range of clay is smaller than range for sand and silt. Therefore the performance of the different stepwise models cannot be compared with one another.

The predictions of the stepwise models had to be cut off and scaled back (4.7 Prediction of soil texture), because unrealistic values were predicted by the models (predicted values>100 & <0). Cutting of and scaling back the predictions might seem artificial, but it is a pragmatic solution. It results in fractions of the three soil texture classes that add up to 100%. Scaling the models did not affect the accuracy much. The RMSE_{CV} for the unscaled models is slightly lower (~2% for clay and silt and 0.01% for sand) for the unscaled models, showing that the accuracy of the models decreases by scaling the models to 0-100%. The $\text{RMSE}_{\text{fitted}}$ for the unscaled predictions is even higher than for the scaled predictions (Table 6 and Table 7). The applied scaling method was selected based on its $\text{RMSE}_{\text{Fitted}}$. However, it is not certain whether the RMSE_{CV} will also be lower for this method. Also other methods for ensuring that the predictions add up to 100% can be thought of. Compositional data analysis was shortly investigated, however at the time this did not seem feasible and promising (it looked like the narrow inaccessible path leading the other way).

Considering the number of variables used in the stepwise models (13, 13 and 15 for clay, silt and sand respectively) and the number of observations in the training data (n=63), there is a risk that the models are overfitting. However, the magnitude of the model error ($\text{RMSE}_{\text{Fitted}}$) corresponds with the expected level of error in the data. The RMSE_{CV} of the different models is comparable to the $\text{RMSE}_{\text{Fitted}}$, indicating that the stepwise models do reasonable predictions for a new dataset. Although cross-validation tends to report higher accuracy for models, than validation with an independent dataset does.

6.6 Results of other studies

Moore et al. (1993) have used MLR with stepwise regression for explanatory variable election for determining different soil properties, including sand and silt content. Explanatory variables used for modelling were terrain descriptor. The explained variance in the soil sample data set was 52% for sand and 64% for silt content. However, this study was conducted on a much smaller scale (0.054 km^2), with soil samples taken on a regular 15 m grid.

A study conducted by Odeh and McBratney (2000) used AVHRR images for determining clay content. Independent validation of the model for the prediction of clay content showed that the RMSE of the prediction was 13.29%. This study was conducted for a study area of approximately the same size (12,000 km^2) and with a soil sample data set in the same order of magnitude.

Predicting soil texture with RTs has been done by different studies. Greve et al. (2012) statistically explained soil texture fractions clay, silt and sand (60%, 60% and ~53% respectively of the variance in the soil sample data set) with seven terrain parameters derived from airborne LIDAR data by using RTs. However, the size of the used soil data set is in a different order of magnitude; the data set for this study consisted out of 45,224 sampling sites. The study area is in the same order of magnitude; 43,000 km^2.

McKenzie and Ryan (1999) used RT based modelling for determining soil depth, phosphor content and carbon content (accounting for 42%, 78% and 54% of the variance in soil sample data set). In this case the size of the study area was of a different order of magnitude; 500 km^2. The soil sample data set used was also somewhat bigger in size (165). These successful studies with regression trees suggest that the soil sample dataset in this study might be too small for effective RT based modelling.

As proposed, random forests might be better suitable modelling method for this study. Wiesmeier et al. (2011) predicted soil organic matter with random forest for study area of about 4,300 km^2, with a relatively small soil sample set (120 observations). Covariates used in this study were land use units, geological units and terrain attributes derived from a DEM; this study explained variances in the soil sample data set from 50-75%.

Comparing the results of this study with the results of previous studies remains difficult, due to many different aspects of soil property mapping. Digital soil mapping studies can differ in scale levels, sampling methods, soil properties of interest, geographic locations of the study area and covariates used for modelling. An infinite number of combination of aspects is possible for soil property mapping; choosing the right aspects is an problem of optimization in itself.

Despite this, the proposed modelling method in this study for soil texture prediction performed quite well, compared with other studies. It should be noted that this study was conducted on a regional scale, with a limited soil sample set collected by constrained LHS method. No other studies have been conducted with a combination of a study area of this size and a soil sample dataset of this size (McBratney et al., 2003). Considering this scope, the explained variance in the soil sample set by the unscaled models is good; 33%, 54% and 59%, for clay, silt and sand respectively. The RMSE$_{CV}$ for the scaled predictions is quite acceptable as well, compared with similar study in terms of study area and soil sample size by (Odeh and McBratney, 2000); 10%, 14% and 16% for clay, silt and sand respectively. The RMSE found by Odeh and McBratney (2000) was 13.29%; note this was an independent validation, these tend to be predict lower RMSE than cross-validations.

CHAPTER 7: CONCLUSION AND RECOMMENDATIONS

7.1 Conclusion

This study presents a method for prediction of soil texture with the help of RS based covariates representing the soil forming factors of the soil-landscape paradigm. Multiple linear regression, stepwise regression and regression trees were used to model the relationship between soil texture and the covariates. The best method was selected to predict soil texture for the entire study area; best was determined based on reported adjusted R^2 and $RMSE_{Fitted}$ of the models. It was required to scale the predictions, the individual models predicted values bigger than 0 and smaller than 100%; the total fraction of clay, silt and sand has to add up to 100% for all locations. The accuracy of the prediction was determined with the help of cross-validation and the uncertainty of the prediction was determined with a 95% confidence interval of the unscaled stepwise models.

It was found that the best modelling method for this study was stepwise regression (read: multiple linear regression with a stepwise regression variable selection procedure). The unscaled stepwise models accounted for clay, silt and sand accounted for 33%, 54% and 59% respectively of the variation found in the soil sample data set. Regression tree based models were not suitable for predicting soil texture with this soil sample dataset. This could be due to two reasons. (1) The soil sample data set might not be suitable for RTs; this might be due to size of the soil sample set or the inherent structures present in the soil-landscape relationships for the study area. (2) The algorithm used for growing the RT is not suitable for the training data.

All the foil forming factors were included in the stepwise models for soil texture prediction; modelling the variation in the soil sample dataset reasonably well. Covariates related to Climate, Parent material and Relief were important predictor for all texture classes. Covariates related explicitly to Organisms were not present in all models, but Organisms might be accounted for by the VNIR and SWIR ASTER bands. Variables that were not included in the stepwise models were SWI and the PCs of the reflectance. SWI was not suitable for determining soil texture due to the scale of the SWI raster, it was suggested that SWI might not match the soil forming process scale. A reason for the absence of the PCs is proposed; it could be due to the presence of ASTER VNIR and SWIR bands in the models, addition of the PCs would lead to data redundancy.

However, the stepwise models do show that the soil-landscape paradigm is useful for modelling soil properties and that RS based covariates are essential for soil texture prediction. RS based covariates provide cheap, exhaustive and quantifiable information on the soil forming factors in the study area.

Three different methods have been used to assess the accuracy of the model:

a) Expert knowledge validation of the scaled prediction
b) Leave-one-out cross validation of the scaled prediction
c) Determining the 95% confidence interval of the unscaled stepwise models for the entire study area

Expert knowledge validation indicated that the scale models succeeded in the large scale patterns of soil variability in the study area, which supports the soil-landscape paradigm. Soil-landscape relationships are useful for the prediction of soil properties in general and soil texture in specific. Leave-one-out cross validation of the scaled predictions showed that the scaled stepwise models have a $RMSE_{CV}$ of 10%, 14% and 16% for clay, silt and sand respectively. This is an acceptable result, considering results of other studies. However, it should be noted that an independent validation of the predictions will probably lead to higher reported RMSE's.

The maps of the 95% confidence intervals show that we are quite certain the prediction is uncertain. Besides the overall indication of uncertainty, the maps also show the spatial distribution of the uncertainty, the uncertainty of the predictions is high for regions with a pronounced topography and high values for NDVI.

7.2 Recommendations

An obvious improvement for this study is to mask out the regions that have a combination of pronounced topography and high values for NDVI. Predictions for these regions are very uncertain; not having any predictions might be better than having really inaccurate predictions. An alternative approach would be to mask out area's that contain confidence interval values above a certain threshold. Perhaps the resulting "holes" in the prediction could be patched up with kriging.

Another point of improvement for this study is to include the TIR bands in the principal component analysis used for constrained LHS method. The stepwise models showed that mineral indices are important covariates for predicting soil texture. The mineral indices are derived from the ASTER TIR bands and the ASTER TIR bands were not well represented by the soil sample dataset. A reason for this might be that the principal components of only the ASTER VNIR and SWIR bands were used for constrained LHS. Including the ASTER TIR bands in the principal component analysis and using this result for constrained LHS, might improve the representation of ASTER TIR bands in the soil sample dataset. However, this might lead to decreased representation of ASTER VNIR and SWIR bands; there probably is a limit to the amount of multidimensional variability 63 sample points can cover.

Compositional analysis was shortly investigated. The tools and the knowledge available at that time did not make it look a promising statistical technique. Now I would recommend the use of compositional analysis, it makes full use of the ontology of texture; which in theory would lead to better modelled relationships.

The scope of this study is to map soils on a regionals scale with a limited soil data set. This study indicated that the sample size is too small; this was concluded from several observations. It was difficult to assess linear models, due to the limited number of data points. Trends in diagnostic plots could not be established properly; a 'guesstimate' had to suffice, but an estimate is desirable. Second, validation possibilities for this study were quite limited. The 95% confidence interval is substantial; the prediction is uncertain for all three texture variables. Other studies were able to model soil-landscape relationships with tree based models; most of these studies used a bigger sample size. This suggests that a bigger sample size can enable tree based modelling for this study area. More data would solve these issues and possibly also other issues.

This raises the question: how much more data is required? Looking at other studies is not going to solve this question; what is done in practice is not necessarily ideal. Therefore I suggest a sensitivity analysis for soil property mapping on a regional scale. The goal of this sensitivity analysis is to determine the amount of soil samples required for a desired accuracy of the prediction for different modelling methods. For this an extensive soil data set is required on a regional scale, perhaps this is rather ambitious. However, if answers on ideal soil sample size for soil property mapping studies are required; this is the way to go.

REFERENCES

Abrams, M., Hook, S., Ramachandran, B., 2002. ASTER User Handbook, Jet Propulsion Laboratory / California Institute of Technology.

Aitchison, J., 1981. A New Approach to Null Correlations of Proportions. J Int Ass Math Geol 13(2), 175-189.

Alfons, A., 2012. cvTools: Cross-validation tools for regression models.

ANWB, 2010. ANWB Wegenkaart Marokko. ANWB Media.

Bartholomeus, H., Epema, G., Schaepman, M., 2007. Determining iron content in Mediterranean soils in partly vegetated areas, using spectral reflectance and imaging spectroscopy. Int J Appl Earth Obs 9(2), 194-203.

Ben-Dor, E., 2002. Quantitative remote sensing of soil properties. Advances in Agronomy, Vol 75 75, 173-243.

Ben-Dor, E., Banin, A., 1995. Near-Infrared Analysis as a Rapid Method to Simultaneously Evaluate Several Soil Properties. Soil Sci Soc Am J 59(2), 364-372.

Boettinger, J.L., Howell, D., Moore, A.C., Hartemink, A.E., Kienast-Brown, S., 2010. Digital soil mapping : bridging research, environmental application, and operation. Progress in soil science. 1st ed. Springer, New York.

Breiman, L., 1984. Classification and regression trees. The Wadsworth statistics/probability series. Wadsworth International Group, Belmont, Calif.

Brenning, A., 2011. RSAGA: SAGA Geoprocessing and Terrain Analysis in R.

Breunig, F.M., Galvao, L.S., Formaggio, A.R., 2008. Detection of sandy soil surfaces using ASTER-derived reflectance, emissivity and elevation data: potential for the identification of land degradation. Int J Remote Sens 29(6), 1833-1840.

Brocca, L., Melone, F., Moramarco, T., Wagner, W., Hasenauer, S., 2010. ASCAT soil wetness index validation through in situ and modeled soil moisture data in central Italy. Remote Sens Environ 114(11), 2745-2755.

Dobos, E., Micheli, E., Baumgardner, M.F., Biehl, L., Helt, T., 2000. Use of combined digital elevation model and satellite radiometric data for regional soil mapping. Geoderma 97(3-4), 367-391.

e-SOTER, www.esoter.net, 2008, Acces date: 11/07/2012

ESRI, 2011. ArcGIS Desktop. Environmental Systems Research Institute, Redlands, CA.

Genuer, R., Poggi, J.-M., Tuleau-Malot, C., 2010. Variable selection using random forests. Pattern Recogn (31).

Gessler, P.E., Moore, I.D., Mckenzie, N.J., Ryan, P.J., 1995. Soil-Landscape Modeling and Spatial Prediction of Soil Attributes. Int J Geogr Inf Syst 9(4), 421-432.

Greve, M.H., Kheir, R.B., Greve, M.B., Bocher, P.K., 2012. Quantifying the ability of environmental parameters to predict soil texture fractions using regression-tree model with GIS and LIDAR data: The case study of Denmark. Ecol Indic 18, 1-10.

Hengl, T., Reuters, H., 2011. How accurate and usable is GDEM? A statistical assessment of GDEM using LiDAR data, International Society for Geomorphometry.

Hijmans, R.J., Cameron, S.E., Parra, J.L., Jones, P.G., Jarvis, A., 2005. Very high resolution interpolated climate surfaces for global land areas. Int J Climatol 25(15), 1965-1978.

Hijmans, R.J., van Etten, J., 2011. raster: Geographic analysis and modeling with raster data.

I.P.F., 2011. Product Sheet: ASCAT Soil Water Index 25 km Time Series, Institute of Photogrammetry and Remote Sensing Vienna, Austria.

Jenny, H., 1941. Factors of soil formation; a system of quantitative pedology. McGraw-Hill publications in the agricultural sciences L J Cole, consulting ed. 1st ed. McGraw-Hill, New York, London,.

Lillesand, T.M., Kiefer, R.W., Chipman, J.W., 2008. Remote sensing and image interpretation. John Wiley & Sons, Inc., Hoboken.

Locher, W.P., de Bakker, H., 1990. Algemene Bodemkunde. Bodemkunde van Nederland. Malmberg, Den Bosch.

Mcbratney, A.B., Hart, G.A., Mcgarry, D., 1991. The Use of Region Partitioning to Improve the Representation of Geostatically Mapped Soil Attributes. J Soil Sci 42(3), 513-531.

McBratney, A.B., Odeh, I.O.A., Bishop, T.F.A., Dunbar, M.S., Shatar, T.M., 2000. An overview of pedometric techniques for use in soil survey. Geoderma 97(3-4), 293-327.

McBratney, A.B., Santos, M.L.M., Minasny, B., 2003. On digital soil mapping. Geoderma 117(1-2), 3-52.

Mckenzie, N.J., Austin, M.P., 1993. A Quantitative Australian Approach to Medium and Small-Scale Surveys Based on Soil Stratigraphy and Environmental Correlation. Geoderma 57(4), 329-355.

McKenzie, N.J., Ryan, P.J., 1999. Spatial prediction of soil properties using environmental correlation. Geoderma 89(1-2), 67-94.

Moore, I.D., Gessler, P.E., Nielsen, G.A., Peterson, G.A., 1993. Soil Attribute Prediction Using Terrain Analysis (Vol 57, Pg 443, 1993). Soil Sci Soc Am J 57(6), 1548-1548.

Mulder, V.L., de Bruin, S., Schaepman, M.E., 2012. Representing major soil variability at regional scale by constrained Latin Hypercube Sampling of remote sensing data. Int J Appl Earth Obs.

Mulder, V.L., de Bruin, S., Schaepman, M.E., Mayr, T.R., 2011. The use of remote sensing in soil and terrain mapping - A review. Geoderma 162(1-2), 1-19.

Murdoch, D., Chow, E.D., 2007. ellipse: Functions for drawing ellipses and ellipse-like confidence regions.

Murdoch, D.J., Chow, E.D., 1996. A graphical display of large correlation matrices. Am Stat 50(2), 178-180.

Neter, J., Kutner, M.H., Nachtsheim, C.J., Wasserman, W., 1996. Applied linear statistical models. Tom Casson.

Newman, M.C., 1993. Regression-Analysis of Log-Transformed Data - Statistical Bias and Its Correction. Environ Toxicol Chem 12(6), 1129-1133.

Ninomiya, Y., Fu, B.H., Cudahy, T.J., 2005. Detecting lithology with Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER) multispectral thermal infrared "radiance-at-sensor" data. Remote Sens Environ 99(1-2), 127-139.

Odeh, I.O.A., McBratney, A.B., 2000. Using AVHRR images for spatial prediction of clay content in the lower Namoi Valley of eastern Australia. Geoderma 97(3-4), 237-254.

Odeh, I.O.A., Mcbratney, A.B., Chittleborough, D.J., 1994. Spatial Prediction of Soil Properties from Landform Attributes Derived from a Digital Elevation Model. Geoderma 63(3-4), 197-214.

Pachepsky, Y.A., Timlin, D.J., Rawls, W.J., 2001. Soil water retention as related to topographic variables. Soil Sci Soc Am J 65(6), 1787-1795.

R Development Core Team, 2011. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Rossel, R.A.V., Walvoort, D.J.J., McBratney, A.B., Janik, L.J., Skjemstad, J.O., 2006. Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties. Geoderma 131(1-2), 59-75.

Salisbury, J.W., Daria, D.M., 1992. Infrared (8-14 Mu-M) Remote-Sensing of Soil Particle-Size. Remote Sens Environ 42(2), 157-165.

Schwarz, G., 1978. Estimating Dimension of a Model. Ann Stat 6(2), 461-464.

Scull, P., Franklin, J., Chadwick, O.A., McArthur, D., 2003. Predictive soil mapping: a review. Prog Phys Geog 27(2), 171-197.

Selige, T., Bohner, J., Schmidhalter, U., 2006. High resolution topsoil mapping using hyperspectral image and field data in multivariate regression modeling procedures. Geoderma 136(1-2), 235-244.

Sorensen, R., Zinko, U., Seibert, J., 2006. On the calculation of the topographic wetness index: evaluation of different methods based on field observations. Hydrol Earth Syst Sc 10(1), 101-112.

Svetlichnyi, A.A., Plotnitskiy, S.V., Stepovaya, O.Y., 2003. Spatial distribution of soil moisture content within catchments and its modelling on the basis of topographic data. J Hydrol 277(1-2), 50-60.

Therneau, T.M., Atkinson, B., Ripley, B., 2010. rpart: Recursive Partitioning.

Van den Boogaard, K.G., Tolosnana, R., Bren, M., 2011. compositions: Compositional Data Analysis.

van den Boogaard, K.G., Tolosnana, R., Bren, M., 2012. Package 'compositions', CRAN.

Venables, B., Ripley, B., 2002. Modern Applied Statistics with S. Fourth ed. Springer, New York.

Wickham, H., 2009. gplot2: elegant graphics for data analysis. Springer, New York.

Wiesmeier, M., Barthold, F., Blank, B., Kogel-Knabner, I., 2011. Digital mapping of soil organic matter stocks using Random Forest modeling in a semi-arid steppe ecosystem. Plant Soil 340(1-2), 7-24.

Wold, S., Esbensen, K., Geladi, P., 1987. Principal Component Analysis. Chemometr Intell Lab 2(1-3), 37-52.

Ziadat, F.M., 2005. Analyzing digital terrain attributes to predict soil attributes for a relatively large area. Soil Sci Soc Am J 69(5), 1590-1599.

APPENDIX 1: DATA ACTION MODEL

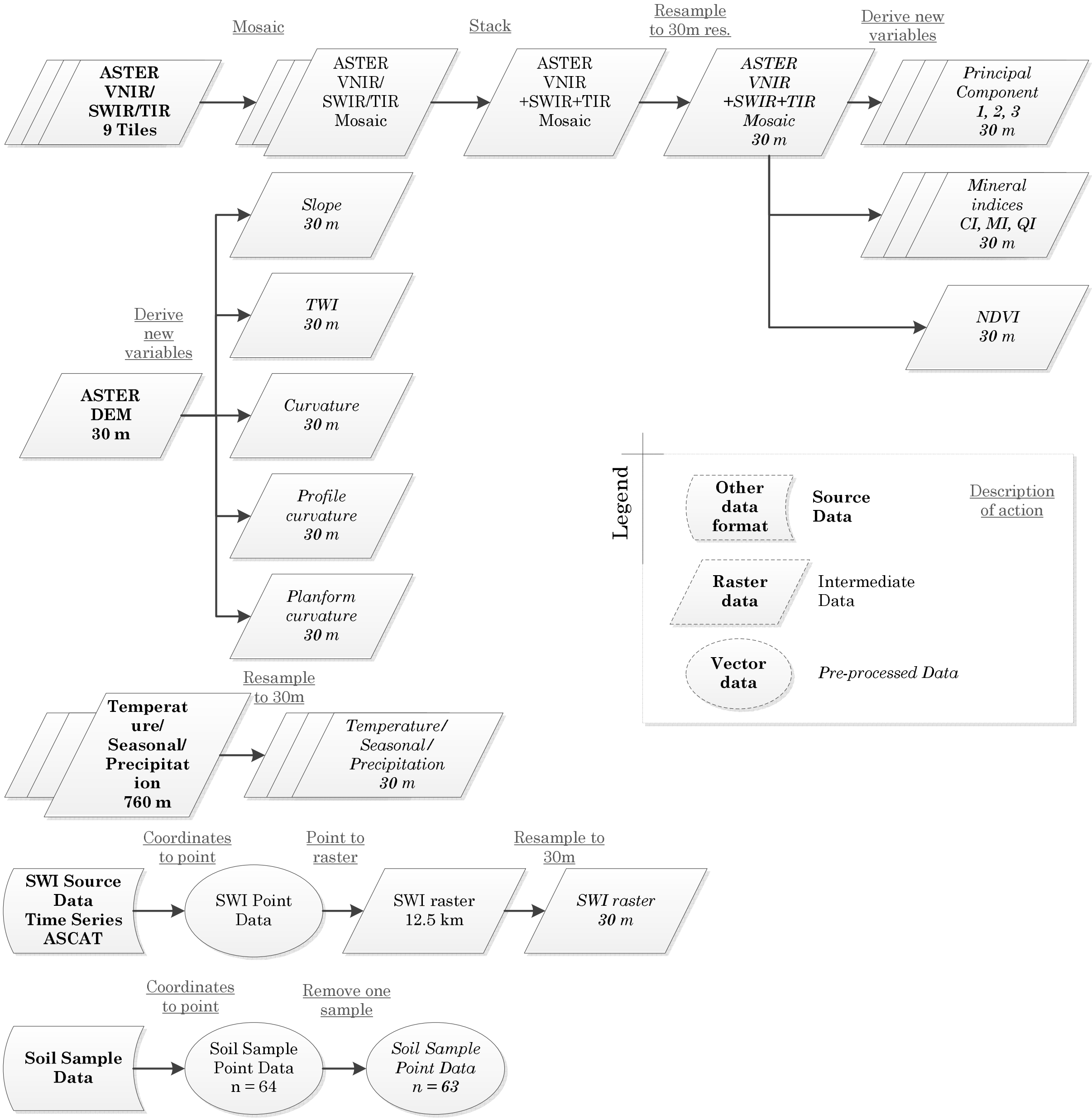


Figure 1: Data action model of pre-processing of data used in this study (1/3)

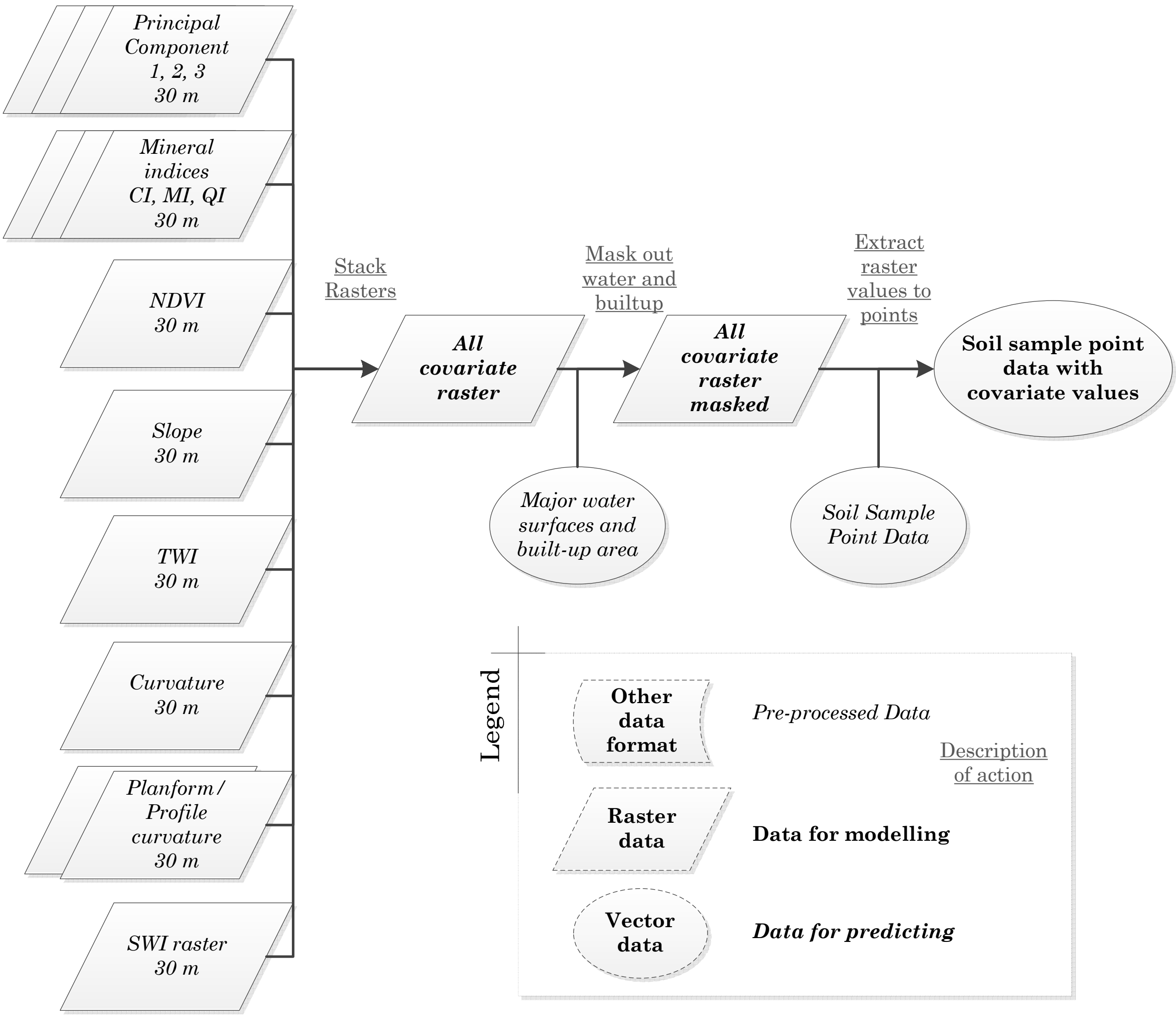

Principal Component 1, 2, 3 30 m
Mineral indices CI, MI, QI 30 m
NDVI 30 m
Slope 30 m
TWI 30 m
Curvature 30 m
Planform / Profile curvature 30 m
SWI raster 30 m
Stack Rasters
All covariate raster
Mask out water and builtup
Major water surfaces and built-up area
All covariate raster masked
Extract raster values to points
Soil Sample Point Data
Soil sample point data with covariate values
Legend
Other data format
Pre-processed Data
Description of action
Raster data
Data for modelling
Vector data
Data for predicting

Figure 2: Data action model of pre-processing of data used in this study (2/3)

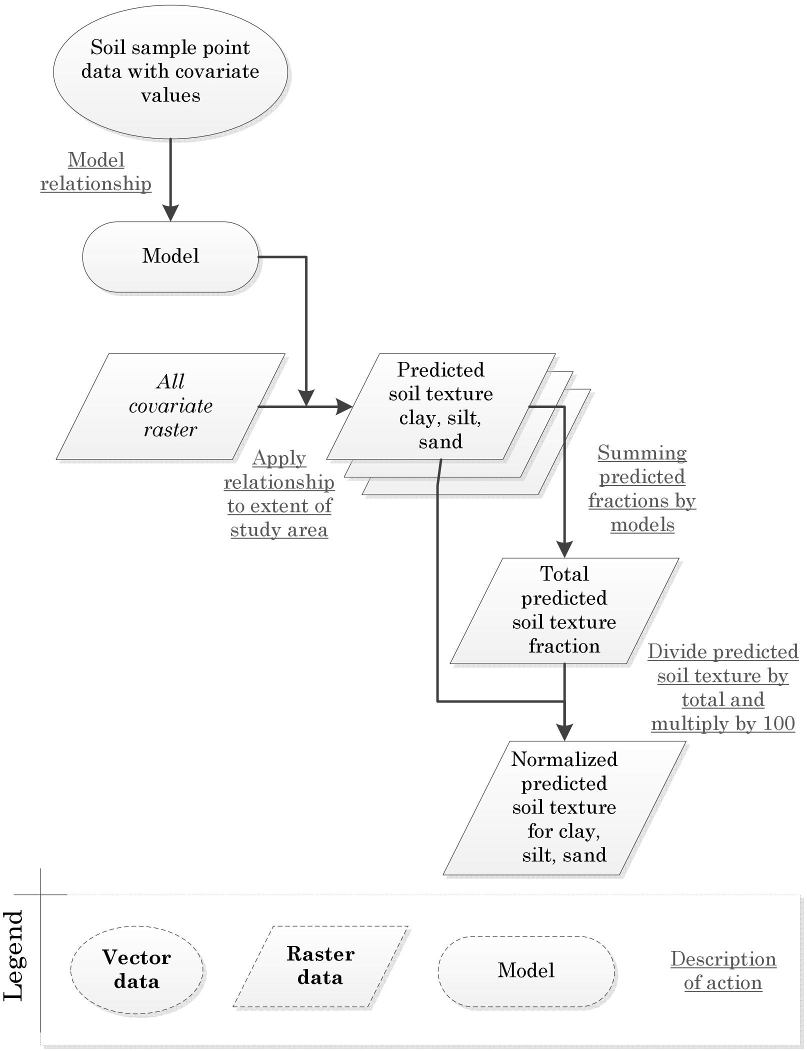
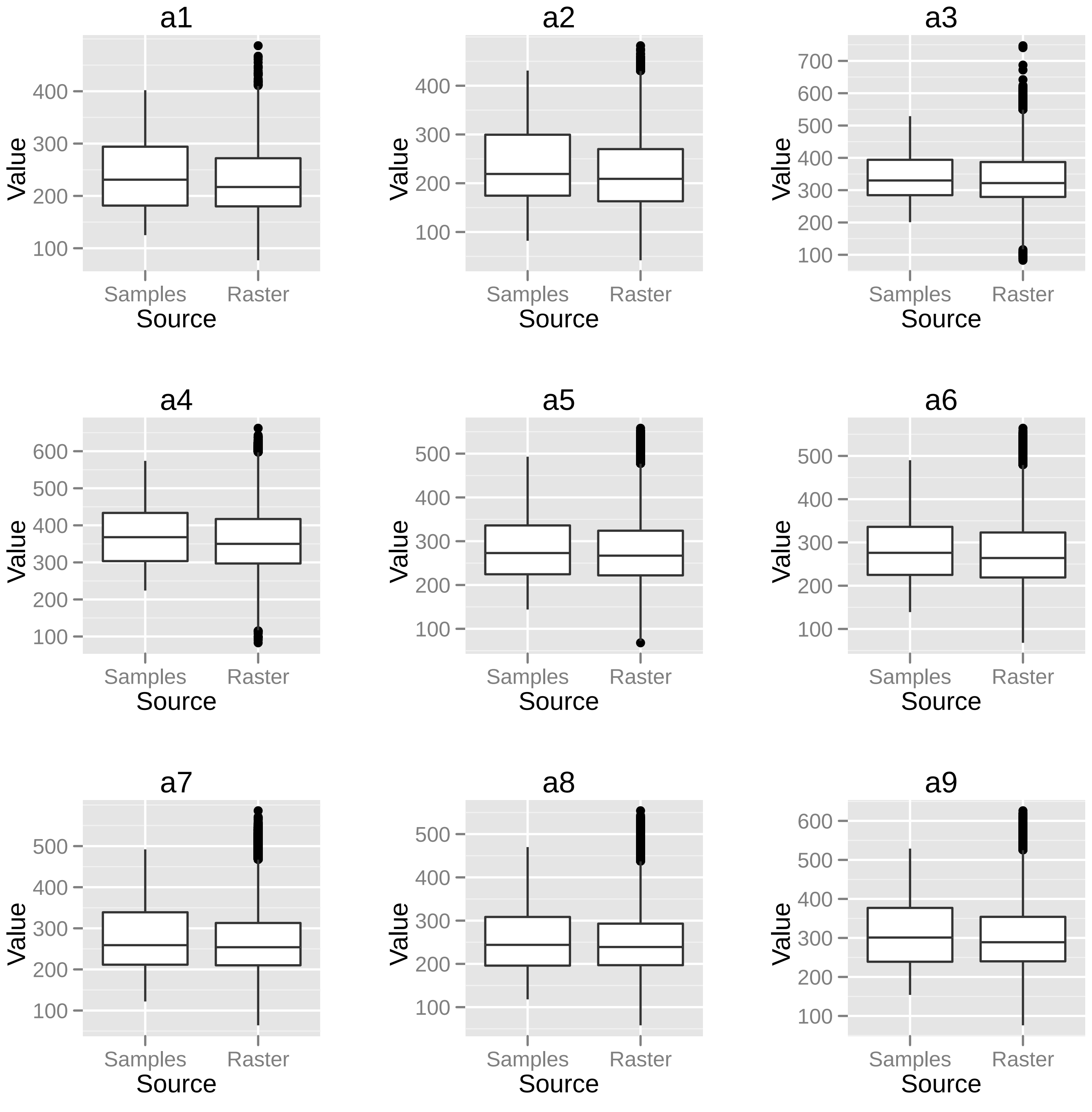
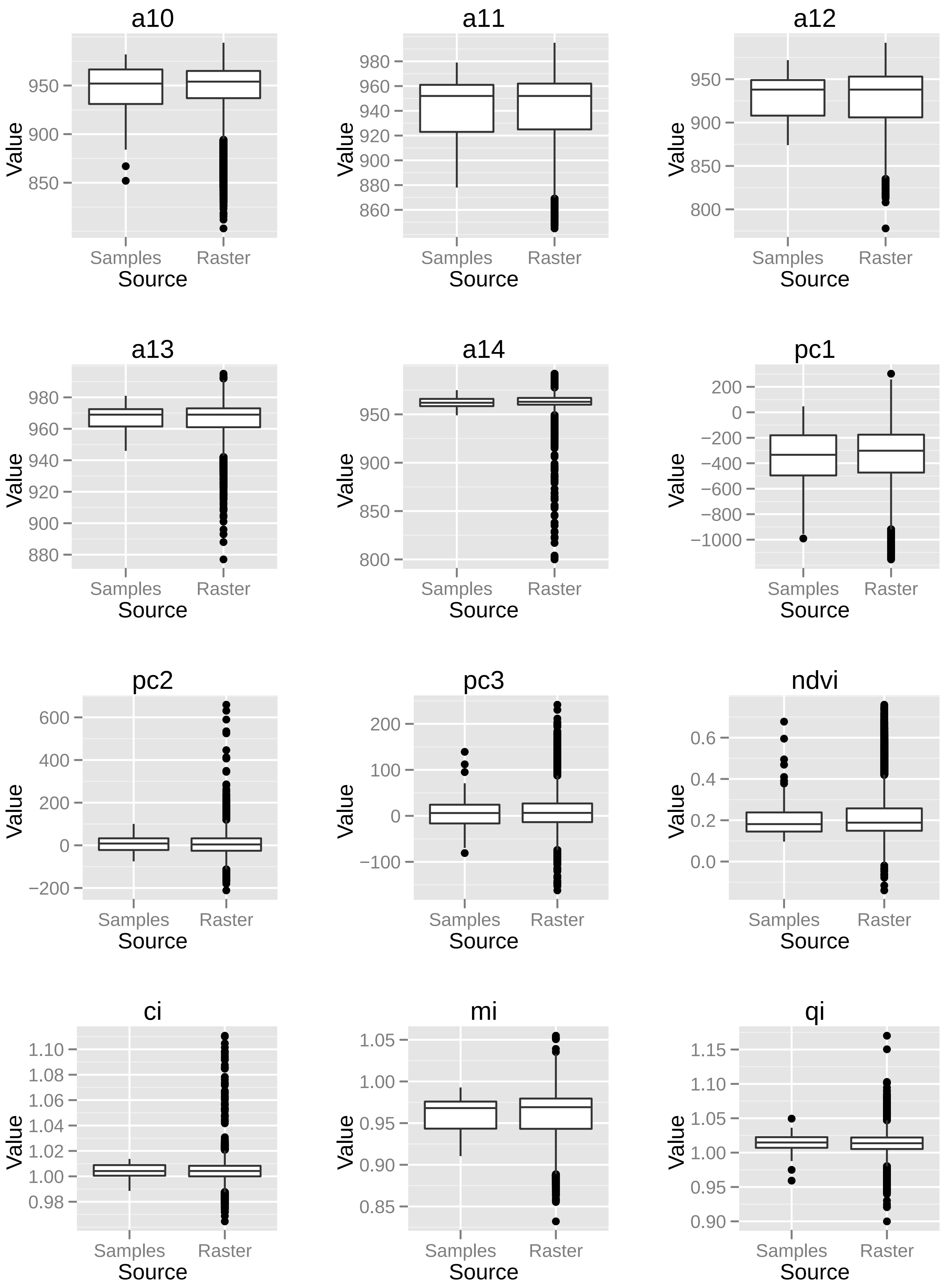
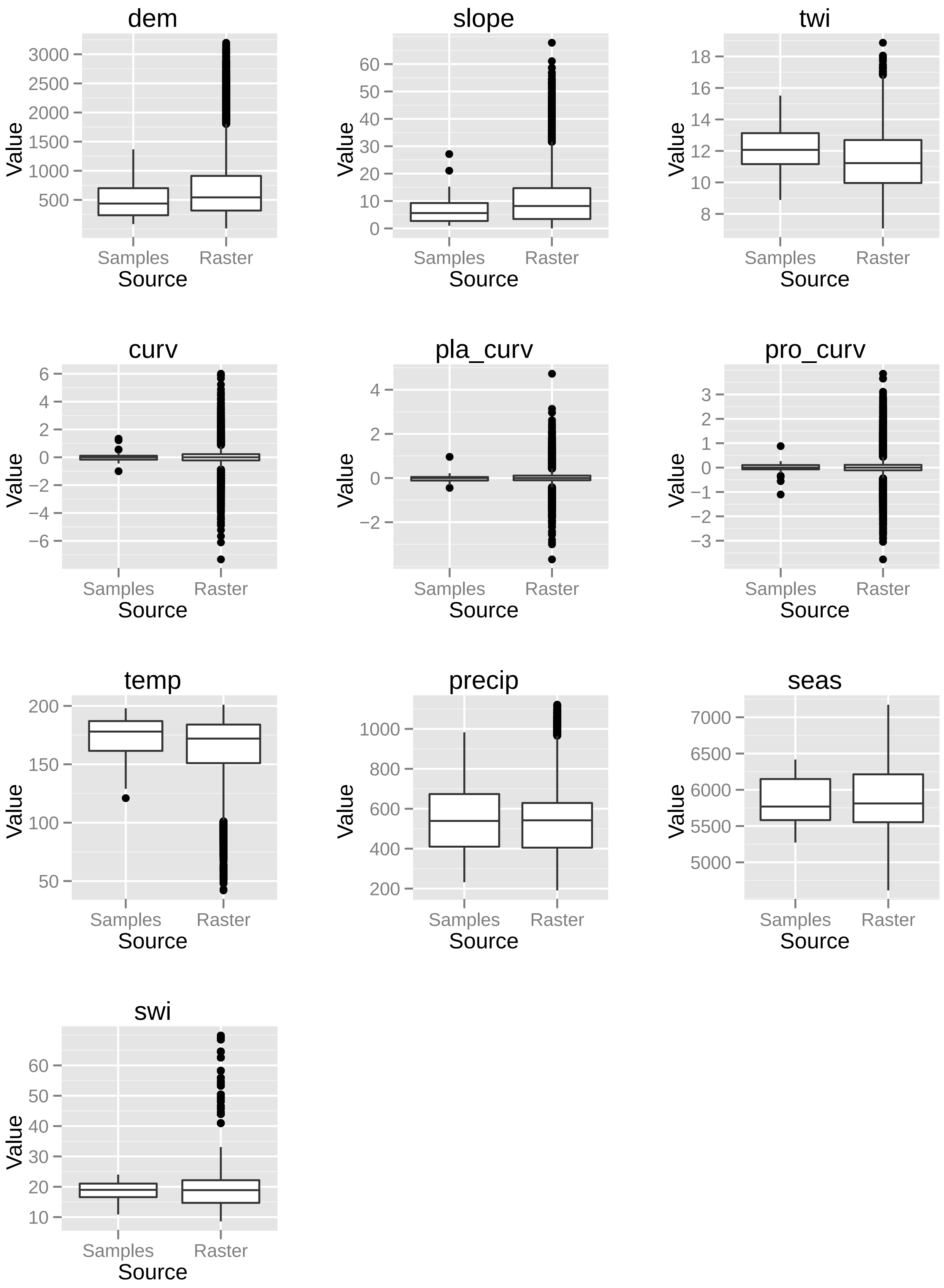

Soil sample point data with covariate values
Model relationship
Model
All covariate raster
Apply relationship to extent of study area
Predicted soil texture clay, silt, sand
Summing predicted fractions by models
Total predicted soil texture fraction
Divide predicted soil texture by total and multiply by 100
Normalized predicted soil texture for clay, silt, sand
Legend
Vector data
Raster data
Model
Description of action

Figure 3: Data action model of soil texture prediction (3/3).

APPENDIX 2: BOXPLOTS SAMPLED/RASTER COVARIATES




a10
950
900
850
Value
Samples
Raster
Source
a11
980
960
940
920
900
880
860
Value
Samples
Raster
Source
a12
950
900
850
800
Value
Samples
Raster
Source
a13
980
960
940
920
900
880
Value
Samples
Raster
Source
a14
950
900
850
800
Value
Samples
Raster
Source
pc1
200
0
−200
−400
−600
−800
−1000
Value
Samples
Raster
Source
pc2
600
400
200
0
−200
Value
Samples
Raster
Source
pc3
200
100
0
−100
Value
Samples
Raster
Source
ndvi
0.6
0.4
0.2
0.0
Value
Samples
Raster
Source
ci
1.10
1.08
1.06
1.04
1.02
1.00
0.98
Value
Samples
Raster
Source
mi
1.05
1.00
0.95
0.90
0.85
Value
Samples
Raster
Source
qi
1.15
1.10
1.05
1.00
0.95
0.90
Value
Samples
Raster
Source


dem
Value
3000
2500
2000
1500
1000
500
Samples
Raster
Source
slope
Value
60
50
40
30
20
10
0
Samples
Raster
Source
twi
Value
18
16
14
12
10
8
Samples
Raster
Source
curv
Value
6
4
2
0
−2
−4
−6
Samples
Raster
Source
pla_curv
Value
4
2
0
−2
Samples
Raster
Source
pro_curv
Value
3
2
1
0
−1
−2
−3
Samples
Raster
Source
temp
Value
200
150
100
50
Samples
Raster
Source
precip
Value
1000
800
600
400
200
Samples
Raster
Source
seas
Value
7000
6500
6000
5500
5000
Samples
Raster
Source
swi
Value
60
50
40
30
20
10
Samples
Raster
Source

APPENDIX 3: DIAGNOSTIC PLOTS OF MLR MODELS

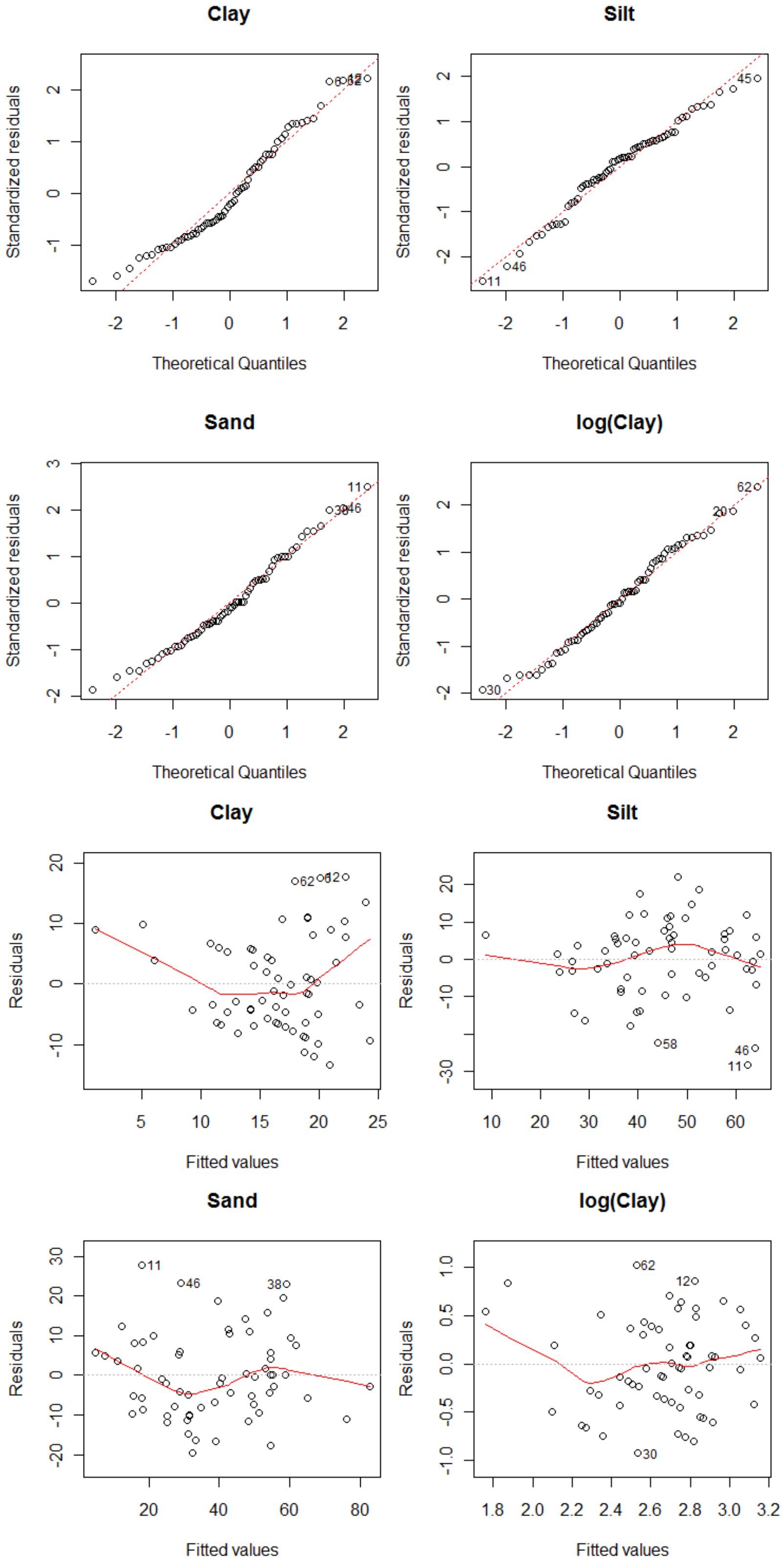


Figure 4: Diagnostic plots of MLR models. Top: Standard QQ-plot for clay, silt, sand and log(clay). Bottom: Residuals versus fitted values for clay, silt, sand and log(clay).

APPENDIX 4: DIAGNOSTIC PLOTS OF STEPWISE MODELS

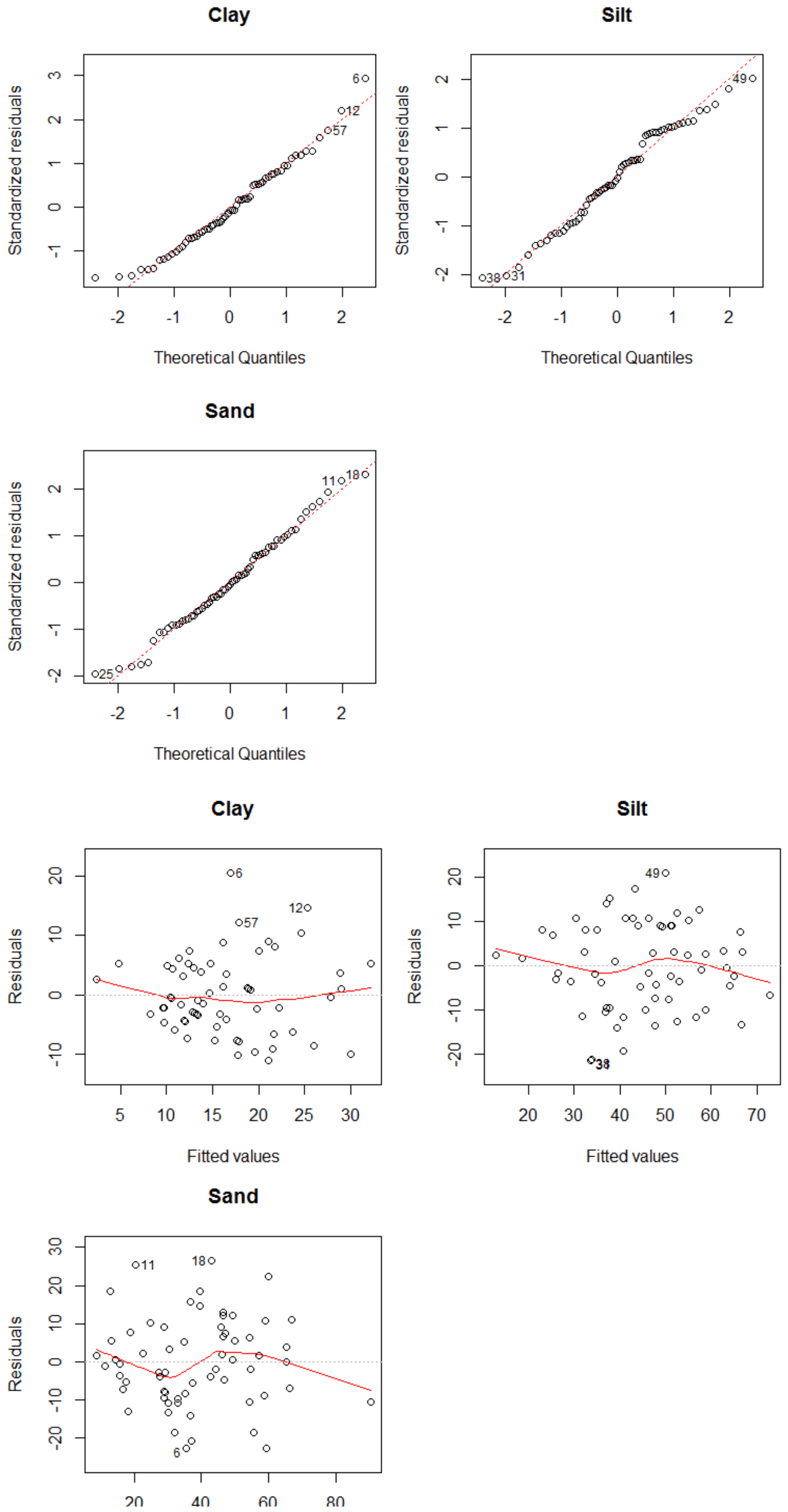


Figure 5: Diagnostic plots of stepwise models. Top: Standard QQ-plot for clay, silt and sand. Bottom: Residuals versus fitted values for clay, silt and sand.

APPENDIX 5: CORRELATION MATRICES

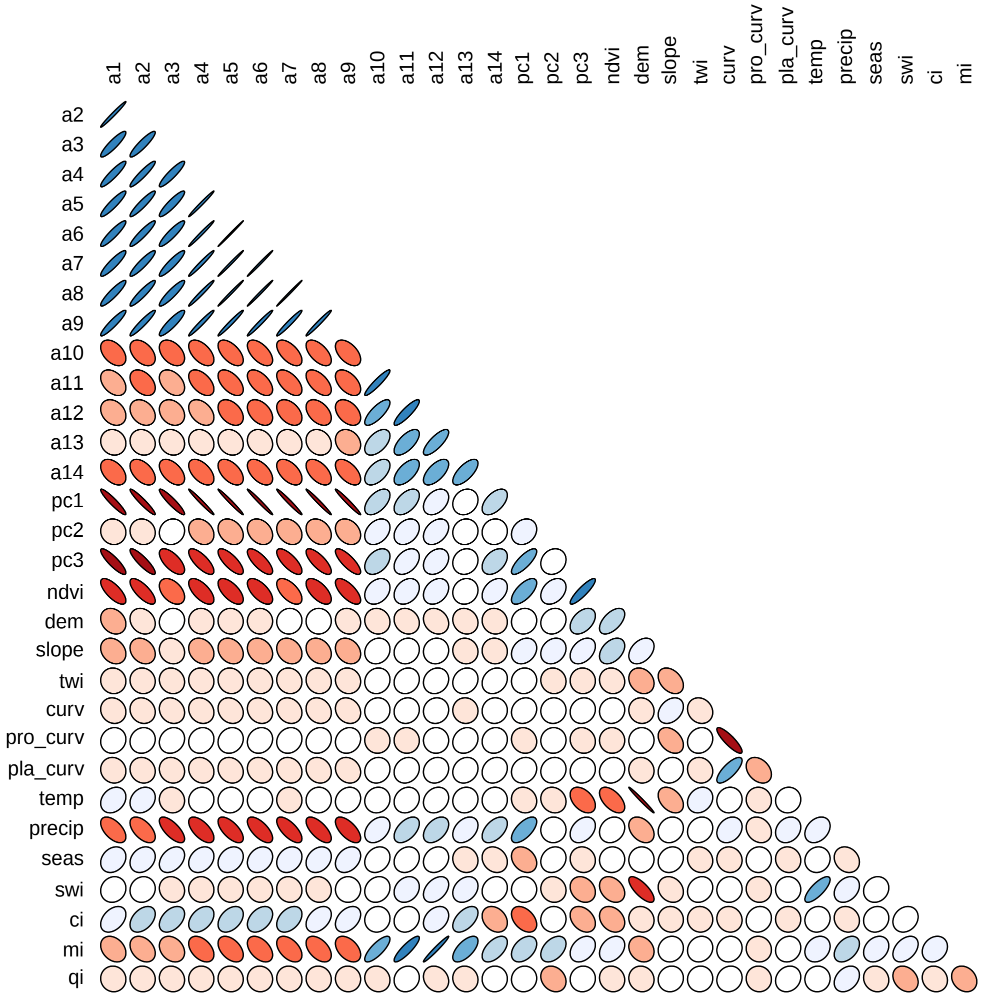


Figure 6: Graphic representation of correlation matrix of the covariates as sampled by the soil sample locations, shape indicates strength of relationship (round=low correlation, ellipsoid=high correlation), colour and orientation indicate direction of relationship.

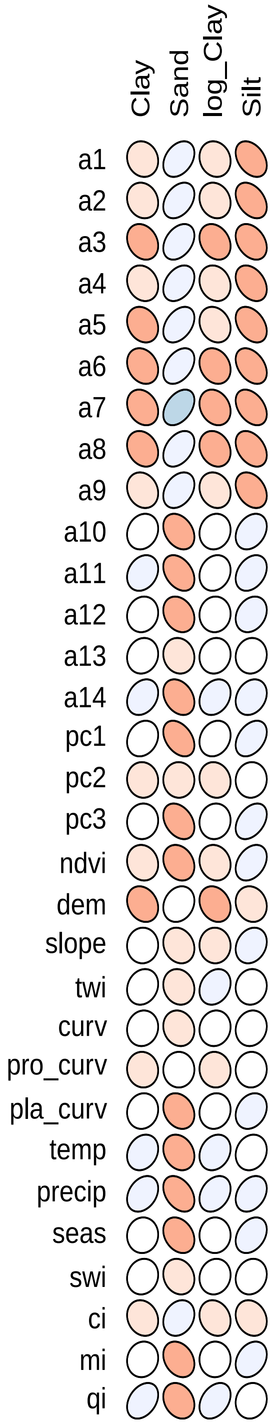


Figure 7: Graphic representation of correlation between response variables log(clay), silt and sand and all covariates