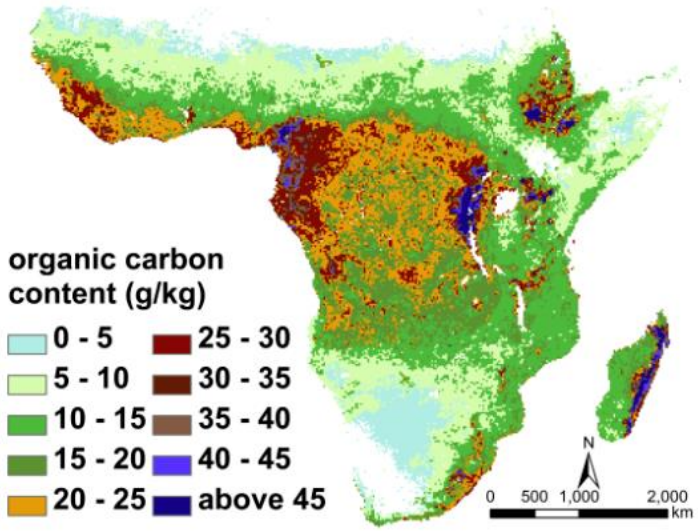
Centre for Geo-Information

Thesis Report GIRS – 2013 -13

Comparison of statistical methods for digital soil mapping of sub-Saharan Africa



Siyu Zhang

Registration No. 880323 987 010

Supervisors:

Dr Gerard Heuvelink (Soil Geography and Landscape / ISRIC)

Dr Ir Sytze de Bruin (Centre for Geo-Information)

Dr Tom Hengl (ISRIC)

Acknowledgement

I am deeply indebted to a number of people without whose encouragement and assistance this thesis would not have been completed. I am profoundly grateful to my supervisors, Gerard Heuvelink and Sytze de Bruin, whose illuminating instruction and expert advice have guided me through every step of my study and writing of this thesis. Gerard continually recommended many valuable papers and books to me. During meetings, Gerard and Sytze gave me lots of inspirations and encouragement. I could always get help immediately when I met problems or got stuck. They read my proposal and thesis draft carefully, helped me to clarify my ideas, clear the thesis structure, and correct content and English mistakes.

I also owe many thanks to Tom Hengl, without whose preparation work for world grids map and GSIF package, I could not start this study easily. He also guided me to start using R to do this study.

I also would like thank Valerio Avitabile to spend his time to share his experience of random forest with me.

I would like to thank every ISRIC staff member. I enjoyed the atmosphere of the excursion, birthday pie breaks, fruitful meetings and any other meeting time.

I would like to thank my friends Yanyan Sun and Huange Wang. When I suffer with some problems, you always have the patience to listen to my complains and try to help and encourage me from both study and life sides.

I finally would like to thank my parents, who always unconditionally support me in my life. Without your financial and emotional support, I would even not have had the opportunity to study abroad.

Abstract

Digital soil mapping is a methodology for finding relationships between known soil data and environmental variables to produce soil maps. These relationships combined with a relatively small sample of measured soil properties can used to predict soil properties at unobserved locations. This study applied five statistical methods, i.e., linear regression without interaction, linear regression with interaction, regression tree, random forest and artificial neural network (ANN) to develop digital soil mapping models for sub-Saharan Africa. The soil variables to be predicted were pH, organic carbon content (SOC) and clay content of the 0-5 cm top layer. The predictor or predictive variables used by the models were related to the soil forming factors relief, climate and organisms. The validity of the statistical models was assessed based on the Root-mean-squared error (RMSE) and explained variance (R^2), both using the calibration data and using independent validation data. The RMSEs were large (around 0.8 for pH, 16g/kg for SOC and 17g/100g for clay content) and R^2 were small (about 22%-44% for pH, around 21%-41% for SOC and approximately 4%-40% for clay content depending on different statistical models) in all models, but the random forest models performed better than the other models for the three soil properties considered. The ANN method could not be configured properly for modelling relationships between the response variables and the predictive variables. All models were built and validated in R, and the digital soil maps were exported by ArcGIS. The results seem that the developed models do not have the enough good quality for predict soil properties.

Key words: statistical models, soil pH, soil organic carbon, soil clay, independent validation, Sub- Saharan Africa

Table of Content

Table List

Figure list

1 Introduction

1.1 Soil data needs in sub-Saharan Africa

A dictionary definition of soil is: “the top layer of the Earth’s surface, consisting of rock and mineral particles mixed with organic matter” (Dictionary, 2011).

Soils are considered of great importance for the environment and mankind. For example, Lal (2004) studied the impacts of soil carbon on global climate change and food security. The paper showed that atmospheric CO_2 can transform to soil carbon which is stored in the soil and hence can mitigate climate change by decreasing the effect of CO_2 emissions from the combustion of fossil fuels. This study also found that an increase of the soil carbon pool in degraded cropland may increase the crops production output. Furthermore, Lal (2001) found that soil degradation and soil erosion impact crop yield, soil quality and productivity, soil carbon dynamics and water pollution. A.S.Kauzeni (1993) reviewed that the necessary of using soil information to plan land use from village scale to regional and national scale, past and present, in Tanzania. The study shown that Tanzania needs a clear comprehensive land use planning guidelines for planners, policy- and decision makers. To arrive this goal, they need adequate natural resources data, including location and administrative frame work, climate, soils and hydrology data, the land tenure and land ownership information. General speaking, all these studies need soil information and this information can help to improve agricultural productivity, land use planning and decision making.

In Sub-Saharan Africa (SSA), the most serious problem is food security. From the 1970s, in several decades SSA has suffered food production shortages. About 180 million African people do not have sufficient food to support their life. Food shortage causes that people suffer from immunological deficiencies, are more prone to infectious diseases, and have a low life expectancy. To improve the quality of life of African people, agricultural development has a high priority . Although soil is as well-known as the “fertile substrate”, not all kinds of soils are suitable for all agricultural production(Parikh, 2012). Therefore, soil properties are usually used to evaluate if the soil suitable for a particular crop. In this thesis three key soil properties, acidity (pH), organic carbon and clay content, will be studied. They are introduced below.

Soil pH reflects the acidity level in soil (Parikh, 2012) , which has interaction with soil properties and effects plant uptake nutrients. In pH optimal range (4-5.6) for tea, pH declines leads to an increase of nitrogenous fertilizer requirements (Owuor, 2012). Furthermore, very low pH (pH<5), the major plant nutrients (Calcium, Magnesium, Phosphorous, Nitrogen and Potassium) may not be sufficiently available for plant growth (Gazey, 2009). Although most agricultural plants grow well in a wide pH range (5.2-8.0) (Lake, 2000), different crops have their own optimizing growing environment. For example, the optimal pH range for potato growing is 5.0-6.0 (Johnston, 2004), soybeans are best grown on soils with pH 6.0-7.0, while corn, wheat and tobacco do best with pH 5.5-7.5 Lake (2000). Using a map of soil pH one can assess the suitability for crops.

Alternatively, liming in combination with appropriate microelement can adjust the soil pH value and hence increase soil fertility.

SOC is the carbon stored within soil; it constitutes part of the soil organic matter. SOC usually comes from the aboveground decaying of natural products, such as fallen leaves and woods as well as decomposition of dead plant roots in the soil (Alvarez and Lavado, 1998). SOC content is an important indicator of soil biological quality and it can also be used in greenhouse gas fluxes estimation (FAO, 2012). SOC content accumulates in the topsoil and reduces exponentially with soil depth. Typically, SOC has high positive correlation with rainfall and clay content (Oades, 1995). Graham Dy and et al. (2002) studied a long-term sugarcane experiment in South Africa and found that protecting SOC content can improve the sustainability of sugarcane production. On the contrary, the SOC content decreases under the intensive agricultural activity and it increases when the agricultural use intensity is reduced(Lugo et al., 1986).

Soil clay is defined as "a very fine-grained material that consists of hydrated aluminium silicate, quartz, and organic fragments and occurs as sedimentary rocks, soils, and other deposits"(Dictionary, 2011). and the mineral materials with a grain size of less than two micrometres (Alvarez and Lavado, 1998). Clay content depends on the geologic conditions, including soil horizons, sediments, volcanic deposits, parent materials and climate (Foley, 1999). In the topsoil, most of the time, clay content is positively associated with organic carbon and precipitation, but negatively correlated with soil depth (Alvarez and Lavado, 1998). In Spain clay content has been found to be negatively correlated with carbon, which can be explained by the free iron oxides which act on organic carbon (Oades, 1995). Soil clay and organic carbon level have effects on many aspects of agriculture, including the soil structure, soil water holding capacity, soil nutrient availability, and contents soil microelement (Owuor, 2012).

The above demonstrates the importance of these three soil properties for crop growth. Accordingly, detailed and accurate soil properties maps for SSA may improve nutrient use efficiencies, prevent and restore degraded soil, and support land use planning, which are ways to improve African food production (Sanchez, 2002). Moreover, soil property maps can also be used in land use planning, soil management, soil degradation and erosion evaluation and in climate change assessments.

Soil mapping in SSA started at the end of the 19th century; initially it focused on commodity crops and soil fertility assessments. Since publication of the first soil map of the world, soil data have been collected occasionally in Africa(Leenaars, 2012). However, because both soil properties and the environment are dynamic, the conventional soil map cannot offer the required information. Data collection for conventional soil mapping has been hampered by the fact that SSA covers a huge area that has many remote and poorly accessible areas. To address these shortcomings and make use of alternative sources of information, digital soil mapping (DSM) has developed since the 1970s (Webster and Burrough, 1972).

1.2 Digital soil mapping

A digital soil map is a spatial database of soil properties (Sanchez et al., 2009), which can be used for land use planning, agricultural management and to support policy decisions, etc. DSM is a tool to produce accurate, up-to-date and spatially explicit soil maps. DSM combines soil observations with auxiliary data (including correlated environmental variables and remote sensing images), using statistical models to predict soil types and properties at unobserved locations in the landscape(Endre Dobos, 2006).

A Geographic Information System (GIS) is a tool for collecting and managing all kinds of spatial information, including maps of the environmental factors that influence the soil. Statistical models can be used to formalize the equation $S = f(CL, O, R, P, T)$ in various ways (Jenny, 1941). The formalized equation can be used to predict the soil at locations where it was not measured. Combing GIS and statistical tools is an approach to find the relationship between soil properties and environmental variables as well as to produce a digital soil map.

There are several different approaches that have been used in DSM, which include linear regression, regression tree, random forest and artificial neural network (ANN).

Xu and Qi (2001) used linear regression to find the relationship between soil moisture and Q10, an indicator of temperature sensitivity that varies depending on geographic location, time, and ecosystem types. Bourennane et al. (2000) used simple linear regression to model the relationship between slope gradient and thickness of a silty-clay-loam horizon in the 'Petite Beauce' in the south-western Parisian Basin. Sheets and Hendrickx (1995) used linear regression analysis to establish the relationship between bulk soil electrical conductivity and soil water content in the top 1.5m soil profile in 40 km northeast of Las Cruces, New Mexico. Jones (1973) used linear regression to find relationships between soil clay content, climate and organic matter in the surface soils of the West African savannah.

Marion Mertens et al. (2001) made a soil texture profiles map by using classification and regression trees in northern Bavaria, Germany. This model used vegetation topographical maps, geologic maps, soil texture with soil profiles data and topographic information. McKenzie and Ryan (1999) established a relationship between soil profile depth and total phosphorus and total carbon in south-eastern Australia using regression trees and generalised linear models.

Wiesmeier et al. (2011) developed random forest methods to show that land use is highly correlated with the soil organic carbon (SOC), total carbon (C_{tot}), total nitrogen (N_{tot}) and total sulphur (S_{tot}). Based on the input variables of land use, reference soil group, geological unit and 12 topography related variables, the SOC, C_{tot}, N_{tot} and S_{tot} are mapped in different landforms in Inner Mongolia, China. Grimm et al. (2008) used the 'random forest' statistical technique to produce a map of spatial concentration and stock estimation of SOC in Barro Colorado Island. McKenzie and Ryan (1999) did similar analyses to predict the spatial distribution of SOC.

In soil science, neural networks have mainly been used to predict soil hydraulic properties (McBratney et al., 2003). Behrens et al. (2005) used an ANN approach with relief, geology and land use map to predict 33 soil units in Rhineland-Palatinate, Germany. The accuracy of the results was high, which indicates the relevance of a digital elevation model and terrain attributes to predict soil properties. Berberoglu et al. (2000) used ANN to integrate spectral and textural information as input neurons to predict land cover in the Mediterranean. Minasny et al. (1999) used the so-called hyperbolic tangent activation with seven input variables and five hidden layers to predict four output variables.

Based on literature, it appears that digital soil mapping of SSA would be possible using the previously mentioned statistical methods. The Africa Soil Information Service project (AFSIS), which is financed by the Bill & Melinda Gates Foundation, aims to develop SSA digital soil maps for small hold famers of Africa. If the project is successful, it will help farmers to manage their land and to decide where, when and what to plant. To achieve these targets, many methods are explored. This study aims to implement and compare several statistical methods for modelling trends of soil properties in sub-Saharan Africa.

General objective:

The objective of this research is to model the relationship between soil properties and environmental covariates for SSA by using (1) multiple statistical methods, (2) the African Soil Profile database and (3) generally available gridded covariate layers, and to compare the results of different statistical methods.

Research questions:

1 Which statistical methods can be used to model the relationship between soil properties and environmental covariates and how do these statistical methods work?

2 Are software implementations in R available for these methods and how can these be used?

3 How can the results of each method be validated?

4 What do the result maps look like and which results are obtained when the methods are applied to soil property prediction in SSA?

5 What can we learn from a case study applying the different statistical methods on data from SSA?

1.3 Outline

This report includes five chapters. The first chapter introduced the purpose of this study as well as the background and problem of SSA and DSM. It also defined the objectives and research questions.

The second chapter describes the study area, the input data (soil profile data and world grids), the methods and the used software (R package and ArcGIS) and model concepts. This chapter also presents the validation methods.

The third chapter presents the results of processed input data, the soil properties maps predicted by five models (linear without interaction, linear with interaction, regression tree, random forest and ANN) and summary statistics (explained variance(R^2), root mean squared error (RMSE), mean, median, min, max) of predicted soil properties value.

The fourth chapter discusses the result of chapter three.

In the last chapter, the report concludes by answering each research question.

2 Materials and methods

This chapter describes the necessary materials and methods for digital soil mapping in SSA. The chapter includes five sections. The first section introduces the geographical, climate and soil information of the study area. The second section describes the soil properties dataset the environmental variables and the used pre-processing methods for the soil dataset. The third section describes the five statistical methods that were used to build relationship between soil properties and environmental variables. In section four, the criteria and method of evaluate the model accuracy and stability are explained. The last section describes how models were developed and predicted in R and presented in ArcGIS Desktop 10.1.

2.1 Study area

The study area SSA lies south of the Sahara and the border is between 18W - 55E of longitude and 18N – 35S of latitude, which corresponds to the area excluding desert area of figure 1. The area occupies approximately 18 million square kilometres and it covers 48 countries that are fully or partially located in SSA. Africa's climate varies generally according to latitude, which can be distinguished as six major zones (Figure 1) that are equatorial, humid tumid tropical, tropical, Sahelian, desert and Mediterranean. The mean monthly precipitation decreases from the equator to north and south (Appendix 1 PREGSM0a), varying from 300 to 3000mm. The annual rainfall has shown a decreasing trend in the last 30 years. The mean soil temperature is 29^{o}C or higher through the whole year in SSA but in the humid areas temperature is lower than in the other areas as a result of cloud cover (Eswaran et al., 1996). The high temperature potential causes evapotranspiration to be relatively high. The major part of the study are except the equatorial area that is largely covered by vegetation suffers different degrees of drought (Barrios et al., 2006).

According to the Food and Agriculture Organization (FAO) land use map (LADA, 2008)and global land cover class map (2000-2005) produced by ISRIC (source), the land use in SSA has 7 major classes that are forest, grassland, shrubs area, cropland area, urban land, bare area and water body. The vegetation cover mainly depends on the amount of precipitation. The most important soils in SSA are Oxisols, Alfisols, Ultisols, Entisols and Aridisols (figure2). This study excluded Aridisols(desert in figure 1) area, because the properties of this soil type are constant and these area do not have much change for land use.

Figure 1 General climate zone of Africa, from *http://www.sc4geography.net/hunckler/internetclass/SubsaharanAfrica/climate.html*, the study area excludes the desert area.

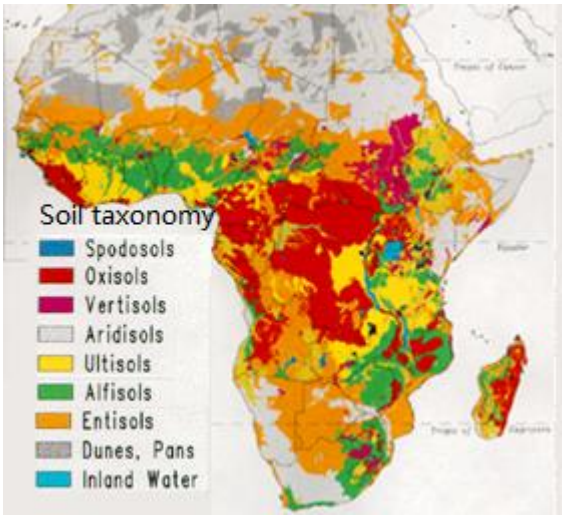


Figure 2 Soil Taxonomy orders for Africa, from *http://soils.usda.gov/use/worldsoils/mapindex/metadata/maps/afrorder.gif* (1996), the study area excludes aridisols area.

2.2 Data description and pre-processing

2.2.1 Soil Data description

In version 1 of the Africa Soil Profiles database (ASPD), 12,574 unique soil profiles are geo-referenced with 50,150 layer data(Leenaars, 2012). The basis 2770 soil profiles of ASPD were derived from the digital profile dataset ISRIC-WISE3 (Batjes, 2008), which are harmonised and screened based on their FAO soil classification. These geo-referenced profiles are distributed over SSA as shown in figure 3. The other legacy datasets come from various sources (over 300) and organizations, such as (ISRIC, FAO, WOSSAC and IRD). The thematic accuracy of the soil attribute data is similar to that of recent soil data, but the positional accuracy of the legacy soil profiles is limited, because most of them are from the pre-GPS era (Leenaars, 2012). However, the legacy data may be inconsistent because:

- They were collected from more than 300 data sources;
- They may be outdated;
- They may have been acquired using different standards for measuring soil properties.

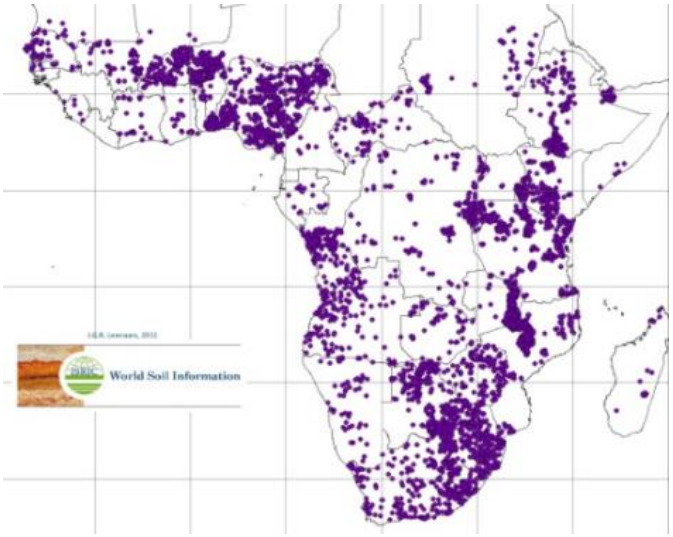


Figure 3 Legacy soil data in Africa Soil Profiles Database version 1.1 (Leenaars, 2012)

This study mainly focused on the soil properties of soil pH (in water), soil organic carbon (SOC) content and clay content of the top 5 cm of the soil that belongs to the top horizon soil. However, the depths of top horizon soil in different location are different, which need to be converted to the fixed depth 5cm. The data density of these three soil properties in ASPD are shown in figure 4. In addition, the soil properties are described by the histogram and box plot separately in figure 5 (pH), figure 6 (SOC), figure 7 (Clay).

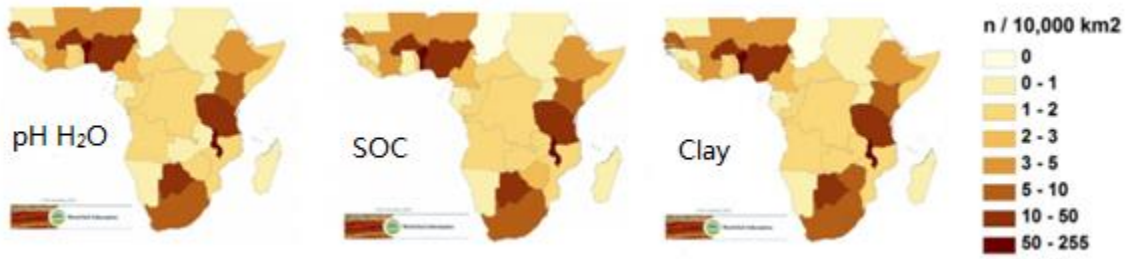


Figure 4 Data density of soil properties in Sub-Saharan Africa (Leenaars, 2012)

Figure 5 shows that the frequency of pH value in first layer distributed like a symmetrical distribution. The pH range in ASPD is from 3.2 to 11, the mean is 6.21 and median is 6.1. In SSA, the range of valid pH values is 2-12.

Figure 6 shows that the frequency of SOC content value in first layer distributed as a skewed distribution. The min value of SOC in ASPD is 0, the max value is 360, the mean is 13.74 and median is 9. In addition, the report states the range of SOC value in SSA is 0-580g/kg(Leenaars, 2012).

Figure 7 shows that the frequency of Clay content value in first layer distributed as a skewed distribution as well. The percentage of clay content in soil is from 0% to 97% , the mean is 23.24 and median is 17.

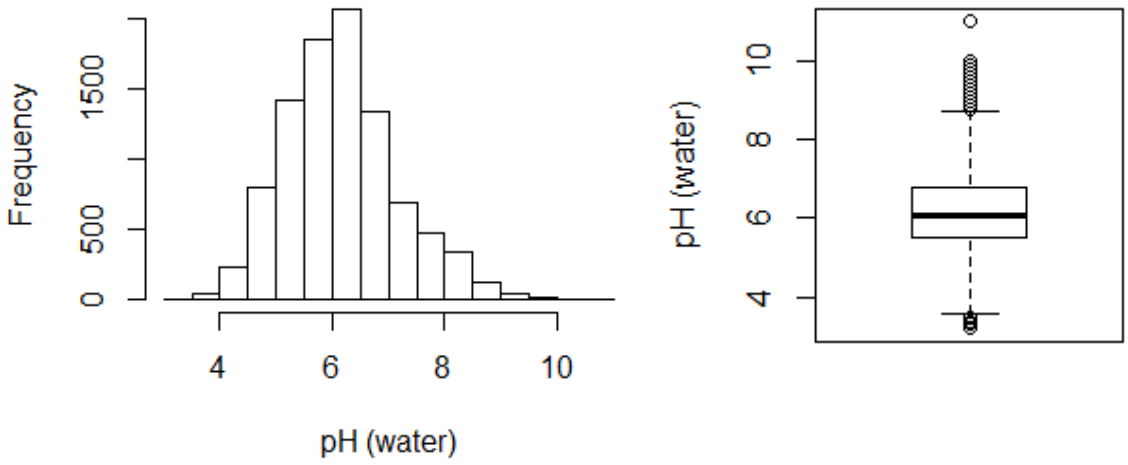


Figure 5 Distribution of pH in histogram and box plot

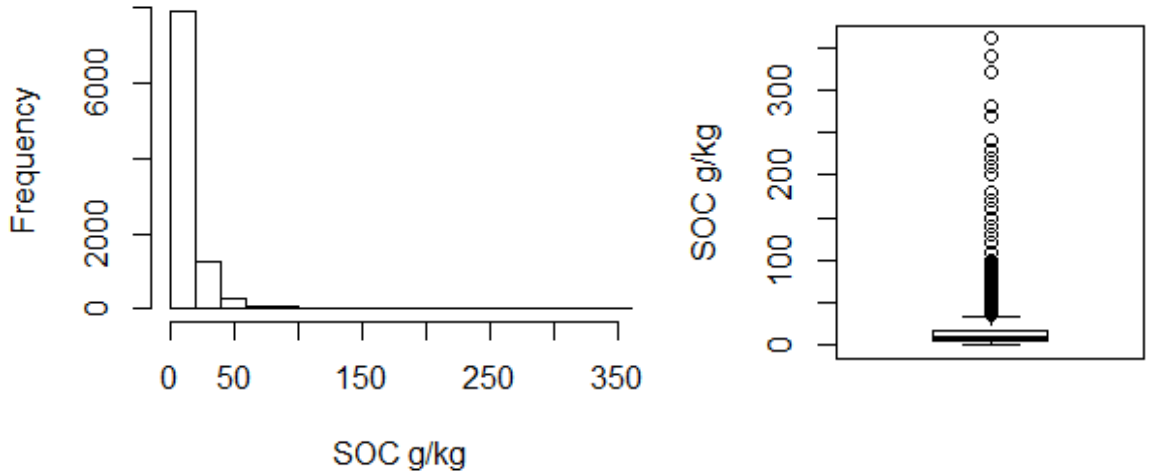


Figure 6 Distribution of soil organic carbon content in histogram and box plot

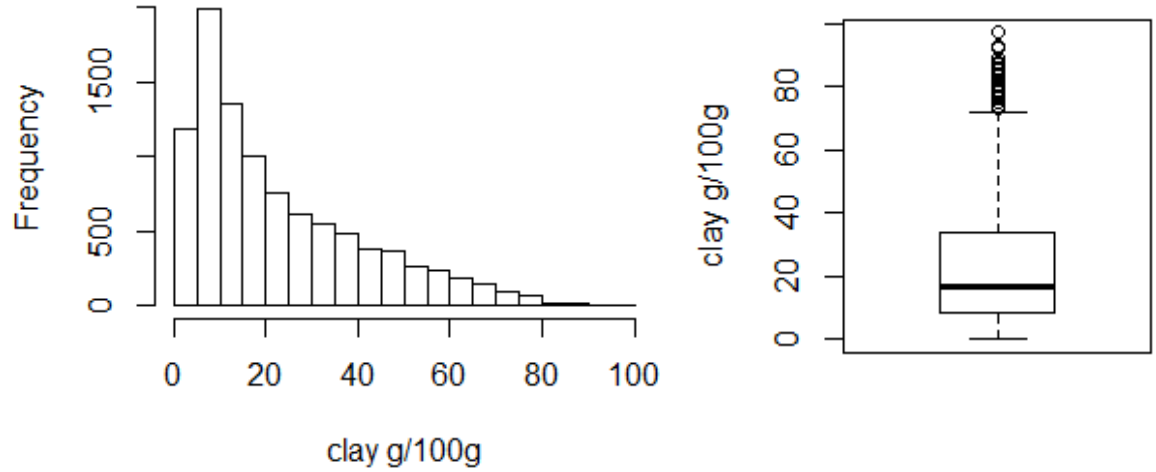


Figure 7 Distribution of clay content in histogram and box plot

2.2.2 Land mask

Because digital soil mapping only produces maps for areas that are covered by soil rather than desert or water body, SMKMOD0a that is a mask file based on MODIS LAI of soil productive was used to select soil area for the other predictor variables.

2.2.3 Environmental variables

The environmental variables, also termed as predictor variables, were downloaded from *www.wordgrids.org*. This website is a public repository and a web processing service for global environmental layers, which is maintained by ISRIC (Hengl and H.I. Reuter, 2013). The downloaded files are related to Jenny's (1941) soil factor equation as follows:

- digital elevation model (DEM) and slope files are used to represent 'relief';
- rainfall and temperature of the earth surface files represent 'climate';
- EVI (explain) products, human activities and land cover files represent the influence of 'organisms'.

The descriptions of predictor variables are shown in Table 1 and maps of these variables are shown in Appendix 1.

Table 1 Environmental covariates

	type	Abbreviated name	Description
1	Climate	PREGSM0a	Mean monthly precipitation
2		TDMMOD0a	Mean value the 8-day MODIS day-time series data
3		TDSMOD0a	Standard deviation of the 8-day MODIS day-time series data
4		TDHMOD0a	Maximum value of the 8-day MODIS day-time series data
5		TNMMOD0a	Mean value the 8-day MODIS night-time series data
6		TNSMOD0a	Standard deviation of the 8-day MODIS night-time series data
7	Relief	DEMSRE0a	Global Relief Model based on SRTM 00+ and ETOPO DEM at 1/5arcdeegres
8		SLPSRT0a	Slope map in present derived using the DEMSRE0a
9	Organisms, vegetation or human activity	EVMMOD0a	Mean value of the monthly MODIS EVI time series data
10		EVSMOD0a	Standard deviation of the monthly MODIS EVI time series data
11		IFLGRE0aa	Intact forest landscapes
12		G01ESA0a	Post-flooding or irrigated croplands
13		G02ESA0a	Rain fed croplands
14		G00ESA0a	Mosaic cropland (50-70%) / vegetation (grassland/shrub land/forest) (20-50%)
15		G04ESA0a	Mosaic vegetation (grassland/shrub land/forest) (50-70%) / cropland (20-50%)
16		G05ESA0a	Closed to open (>15%) broadleaved evergreen or semi-deciduous forest (>5m)
17		G06ESA0a	Closed (>40%) broadleaved deciduous forest (>5m)
18		G07ESA0a	Open (15-40%) broadleaved deciduous forest/woodland (>5m)
19		G09ESA0a	Open (15-40%) needle leaved deciduous or evergreen forest (>5m)
20		G10ESA0a	Closed to open (>15%) mixed broadleaved and needle leaved forest (>5m)
21		G11ESA0a	Mosaic forest or shrub land (50-70%) / grassland (20-50%)
22		G12ESA0a	Mosaic grassland (50-70%) / forest or shrub land (20-50%)
23		G13ESA0a	Closed to open (>15%) herbaceous vegetation (grassland, savannahs or lichens/mosses)
24		G14ESA0a	Closed to open (>15%) (broadleaved or needle leaved, evergreen or deciduous) shrub land (<5m)
25		G15ESA0a	Sparse (<15%) vegetation
26		G16ESA0a	Closed to open (>15%) broadleaved forest regularly flooded (semi-permanently or temporarily) - Fresh or brackish water
27		G17ESA0a	Closed (>40%) broadleaved forest or shrub land permanently flooded - Saline or brackish water
28		G18ESA0a	Closed to open (>15%) grassland or woody vegetation on regularly flooded or waterlogged soil - Fresh, brackish or saline water
29		G19ESA0a	Artificial surfaces and associated areas (Urban areas >50%)
30		G20ESA0a	Bare areas
31		G21ESA0a	Water bodies

2.2.4 Soil Data Pre-processing—— Mass-Preserving-Spline

Because soil attributes vary continually with depth in the soil profile (Russell, 1968), the soil profile data is divided into horizons to record In ASPD. The structure of soil profile data are 1: n, which means one observed location may own several records for different soil layers. However, this study is focused on the top 5cm soil, therefore the response variables require the soil attributes values over a fixed depth interval. Mass – preserving – spline (MPS), also called equal –area spline is a function that can convert soil profile data value to a fixed depth value (Bishop et al., 1999). This function was used for pre-processing the soil data in ASPD.

This study used the MPS algorithm that was developed by (Bishop et al., 1999)(Eq.1). The minimiser of Eq.1 is a quadratic spline. In which n is the number of soil profile horizons, y_i is one horizon soil property value plus the measurement error, $\frac{1}{\mathrm{n}}\ (f_i)$ is the value in one soil horizons, that is measure from the bulk sample of this horizon or mean the value of sum upper and lower boundaries value. The first item means the fit of the spline to the data, the second item defines the spline function $f(x)$ roughness, the parameter λ operates the balance between the fit and spline function $f(x)$ roughness to get the minimum $f(x)$ (Bishop et al., 1999; Odgers et al., 2012).

$$\min(spline) = \sum_1^n (y_n - \frac{1}{n}\ (f_n))^2 + \lambda \int_{x_{n-1}}^{n} [f^{'}(x^2)]dx \qquad \text{(Eq.1)}$$

The function (Eq. 1) assumes the soil attributes value vary smoothly with depth. Therefore a smooth spline is generated to fit the soil attributes value in different horizons, which is demonstrated in Figure 8. In each horizon, the area X_i of the left side of spline is equal to the area Y_i (i = 1-n) at right side of spline, which contributes to the fitting soil properties mean value in each horizon as same as the original horizon value (Ponce-Hernandez et al., 1986). To perform this function, the soil properties should have been measured in at least need 3 horizons, and the top layer soil should be present (Malone and Hengl, 2012a). In addition, the upper and lower boundaries of adjacent horizons should not overlap (Odgers et al., 2012).

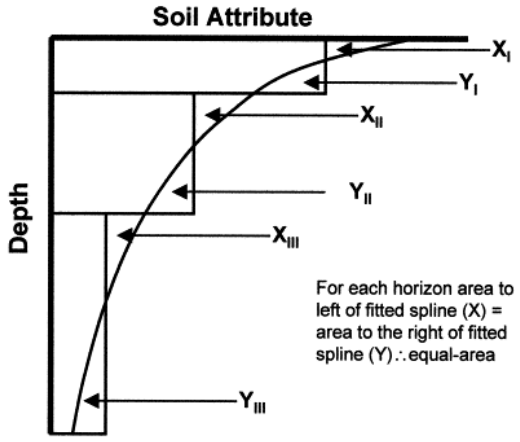


Figure 8 Mass - preserving spline or equal-area quadratic spline (Ponce-Hernandez et al., 1986)

2.3 Modelling statistics concepts

The literature review in the previous chapter showed that there are several statistical methods that can be used for predicting target soil variables using covariates. The concepts of these statistical methods, i.e., linear regression without interaction, linear regression with interaction, regression tree, random forest and ANN are introduced below.

2.3.1 Linear regression without interaction, least squares and step wise regression

Linear regression without interaction (LRW) uses a straight line model the relationship between a response variable and a set of predictive variables, see eq. 2:

$$Y = \beta_0 + \sum_1^n \beta_j x_j + \in \qquad \text{(Eq. 2)}$$

where Y is the dependent variable, $x_{1..n}$ are predictive variables , $\beta_0$stands for intercept, β_j is regression coefficients and $\in$ represents the random error, $\in \sim N(0, \sigma^2)$. In this model, the parameters, $\beta_{0..n}$are estimated based on the training dataset of observed values Y and x_j. If deemed necessary the response variable or predictive variables can be transformed, for example using a log transform.

To fit the linear regression model, the least squares method (eq. 3) is usually to be used, where the coefficients β are considered to minimize the residual sum of squares (RSS):

$$RSS(\beta) = \sum_{\mathrm{i=1}}^{\mathrm{m}} \left(\mathrm{y_i} - \beta_0 - \sum_{\mathrm{j=1}}^{\mathrm{n}} \mathrm{x_{ij}}\beta_\mathrm{j}\right)^2 \qquad (3)$$

Stepwise regression is a well-known method for selecting and predictive variables by an automatic procedure. The main approaches to perform stepwise regression are forward selection, backward elimination and bidirectional elimination. Forward selection starts with no variables in the model and adds variables one by one to the model and testing by some criterion whether this predictor should be included in the model. Backward elimination does the opposite of forward approaches; it starts with all variables in a model. Bidirectional elimination combines the previous two approaches, by adding and testing (Efroymson， MA, 1960).

The criterion for select predictor variables is Akaike information criterion (AIC) that makes decision for predictive variables. AIC tries to find a trade-off between the model complexity and model performance. In general, AIC is defined as

AIC = 2k- 2ln(L) *(Eq. 4)*

Where k is the number of predictive variables, L is maximized value of likelihood function for the developed linear model (Akaike, 1974). Here the likelihood function compares the different between fitting model and predicting model, the minimum difference value has the maximized L.

2.3.2 Linear regression with interaction

Linear regression with interaction model (LRI) invested to study the interaction between each two covariates of Climate, Relief and MODIS EVI covariates affection. The predictor variables that produced from globe land cover map ('G01ESA0a' - 'G21ESA0a') were not used in the interaction calculation, because the data are collected from 2000-2005, but the other covariates Climate, Relief and MODIS EVI data collected from 2010-2012, these two parts data do not have interaction. In addition, the MODIS EVI data have taken part in the Organism part. In LRI, the response variable is not only dependent on individual prediction variables, but that interaction between two variables is also included. In other words, the effect of one variable on the dependent may depend on the value of another variable (Aiken and West, 1991).

The basic fitting strategy of LRI is similar as LRW, the difference is LRI including interaction variables that formed as Eq.5, where x_j the element of eq.2, x_a and x_b are the interaction variables.

$$x_j = x_{(a,\ b)} = \ x_a \times x_b \qquad \text{(Eq. 5)}$$

The interaction model means the relationship between variable x_a and $f(x)$ depends on the value of another variable x_b.

2.3.3 Regression tree

Regression tree is part of CART system, which use recursive partitioning method to fit and predict continues response variables (Leo Breiman, 1984). The tree model is grown as a binary recursive partitioning tree, which has nodes and edges that are used to connect nodes. An example is shown in Figure 9, where all data go in the root node (top) go in the model, the tree model usually works top-down, the input parent nodes with specific rules are split to two children nodes that are put in either left or right directions. At ach node, only one predictive variable is used to decide the direction of children nodes, if the input data meet the node's rule, the child node will go right, otherwise go left. Based on all possible split rules, the process is repeated for each child node until the terminal nodes (red nodes) value are too small or the number of terminal nodes are too few to be split (Ripley, 2013).

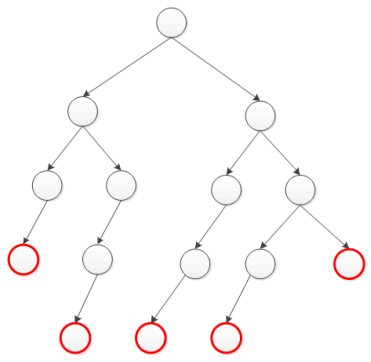


Figure 9 Structure of regression tree

The tree model may fit noise, which means that the model is over-fitting the data and that it will do a bad job if fed with new data. To avoid this problem, pruning of branches is necessary, which remove chapters of a tree model to reduce the tree size and complexity. Leo Breiman (1984) used k -fold cross-validation to decide on pruning branches by minimize the cross-validated sum of squared errors (ErrC(T)). The function are defined as Eq.6, in which "Err(T) is the resubstitution error estimate of tree T; Tn is the cardinality of the set T containing the leaves of the tree T; and cp is the complexity parameter, which defines the cost of each leaf" (Torgo, 1999).

$$\mathrm{EC}_{\mathrm{cp}}(\mathrm{T}) = \mathrm{Err}(\mathrm{T}) + \mathrm{cp} \times \mathrm{Tn} \qquad \text{(Eq.6)}$$

2.3.4 Random forest

Random Forests (RF) is an ensemble classifier of decision trees, which is based on regression trees with random inputs split sub-dataset and predictive variables (Breiman, 2001). Because the input response variables are randomly split to many small dataset and the input explained variables are randomly divided to each small dataset to grow trees, no tree in RF use whole dataset and all predictive variables to fit each RT, therefore the tree can be grow as deep as possible and pruning is not necessary here. The general architecture of RF is shown in Figure 10. In which input dataset are split to three sub-datasets to grow three trees, each tree holds different predictive variables; the terminal red nodes represent each tree predicted value, which will be average of the individual tree outputs to get final prediction *z*. The random forests often grow a crowd of trees to get better results than individual tree and avoid over-fitting.

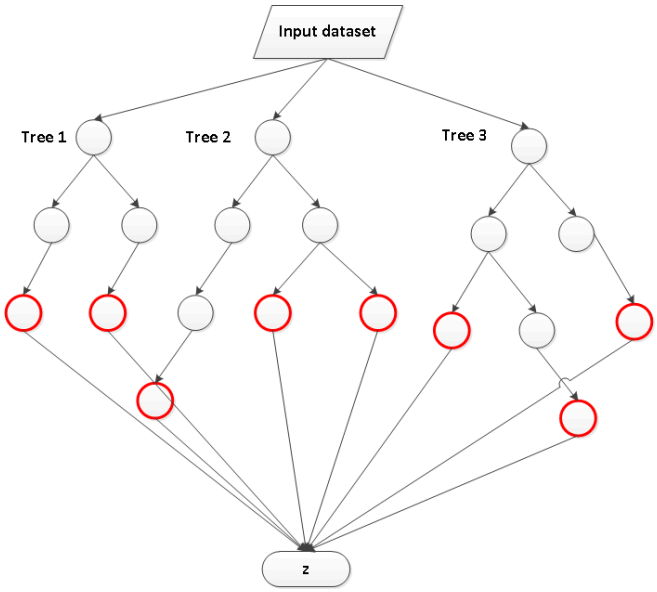


Figure 10 A general architecture of a random forest.

2.3.5 Artificial neural network

Artificial neural networks (ANN) provide a common technique used in artificial and data mining, which is used to find the complex relationship between input and output. In fact, it is a special case of a non-linear regression model. An ANN, which is shown in figure 11, consists of a group of input neurons, a group of output neurons and a group of hidden neurons. The hidden neurons connect input neurons and output neurons, extracting useful information from input neurons and transforming them to predict the output neurons. A set of functions are used to connect them by weight. In most cases, a neural network is an adaptive system, which is able to adapt itself dynamically to complex problems(Ivry and Michal, 2013).

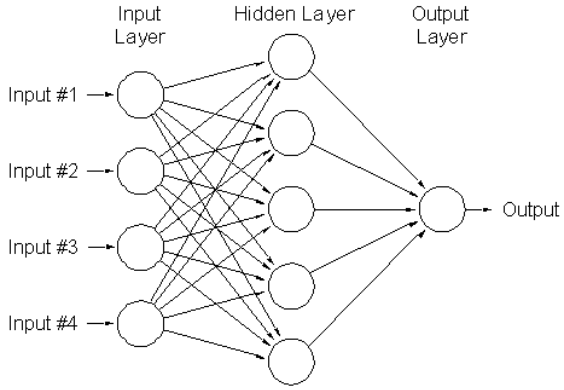


Figure 11 Schematic outline of an artificial neural network (Ivry and Michal, 2013)

The back propagation algorithm (Rumelhart and McClelland, 1986) is used in convergence properties. This means that the hidden layers are organised in hidden layers, and send their weights *w* forward to output layers, and then the errors are propagated backwards from the output layers. The error is the difference between the predicting results and actual results, and ANN back propagation try to find the minimum error. The weights in ANN model are unknown parameters for hidden neurons, which need to be found by fitting the model (Gershenson, 2013). The activation function (A) of the in the hidden layer is sum of the inputs *x* multiplied by themselves weights *w that is same as linear model* and then output function is the sigmoid function (Eq.7).

$$f_k(x_i) = 1/(1 + \mathrm{e}^{\mathrm{A}(\mathrm{x_i},\mathrm{w_i})}) \qquad \text{(Eq.7)}$$

When the ANN model deals with continues value, the model uses sum of squared error (Eq. 8) to measure the fitting, where *k* is the number of hidden neurons and *I* is the number of predictive variables, y_{ki} *is the actual value and* $f_k(x_i)$ is the predicted value

$$\mathrm{RSS}(\theta) = \sum_1^K \sum_1^I (y_{ki} - f_k(x_i))^2 \qquad \text{(Eq. 8)}$$

ANN model is controlled by the model structure and weights. In most of case, a single layer with a large number of neurons can fit the model well. The number of neurons can be set between 5 -100 normally, dependent on the number of input variables (Hastie et al., 2001). Lawrence et al. (1996) found the number of hidden neurons (n) can be tried following three strategies based on back propagation convergence properties:

1) n can be tried start half of input and output variables or

2) n can be tried around two third of input and output variables or

3) n cannot over 2 times numbers of input and output variables.

In regression ANN mode, number of layers is based on the background knowledge and experimentation. Using multiple hidden layers increases the model complexity.

Models fitting with the weight often start near zero, which leads to ANN structure like a roughly linear model. During the fitting calculation, the weights of unknown parameters increase as model becoming non-linear. If the start value set be bigger, it often get poor result(Hastie et al., 2001).

2.4 Assessment of models fitting quality and stability

All models performances were evaluated by root-mean-squared-error (RMSE) (Eq.9) and percentage explained variance (R^2) (Eq.10). RMSE measures the difference between the observed value and predicted value; smaller RMSE means stronger ability of model prediction. R^2 is also called correlation coefficient, which reflects the model regression quality. The larger R^2 indicates the better performance of model fitting.

$$\mathrm{RMSE} = \sqrt{\frac{1}{\mathrm{n}}\sum_1^{\mathrm{n}}(\mathrm{X}_{\mathrm{observed\ n}} - \mathrm{X}_{\mathrm{predicted\ n}})^2} \qquad \text{(Eq.9)}$$

$$\mathrm{R}^2 = 1 - \sum_1^{\mathrm{n}}(\mathrm{X}_{\mathrm{observed\ n}} - \mathrm{X}_{\mathrm{predicted\ n}})^2 / \sum_1^{\mathrm{n}}(\mathrm{X}_{\mathrm{observed\ n}} - \frac{1}{n}\sum(\mathrm{X}_{\mathrm{observed}}))^2 \qquad \text{(Eq.10)}$$

To ensure prediction models stability, all models were tested using independent validation data. The original dataset was randomly split to training dataset (50%, 60%, 70%) and testing dataset (50%, 40%, 30%). The training datasets were used for model development, next the developed models were used for fitting the training dataset and testing dataset. The RMSE and R^2 of the training dataset and test dataset were compared to assess the models stability and accuracy. If the results between training dataset and test dataset similar to each other, that means the developed models were stable. The quality of developed model can be assessment by the size of RMSE and R2, where the smaller RMSE and the larger R^2 indicate better model performance.

2.5 Software implementation

2.5.1 Modelling flow chart

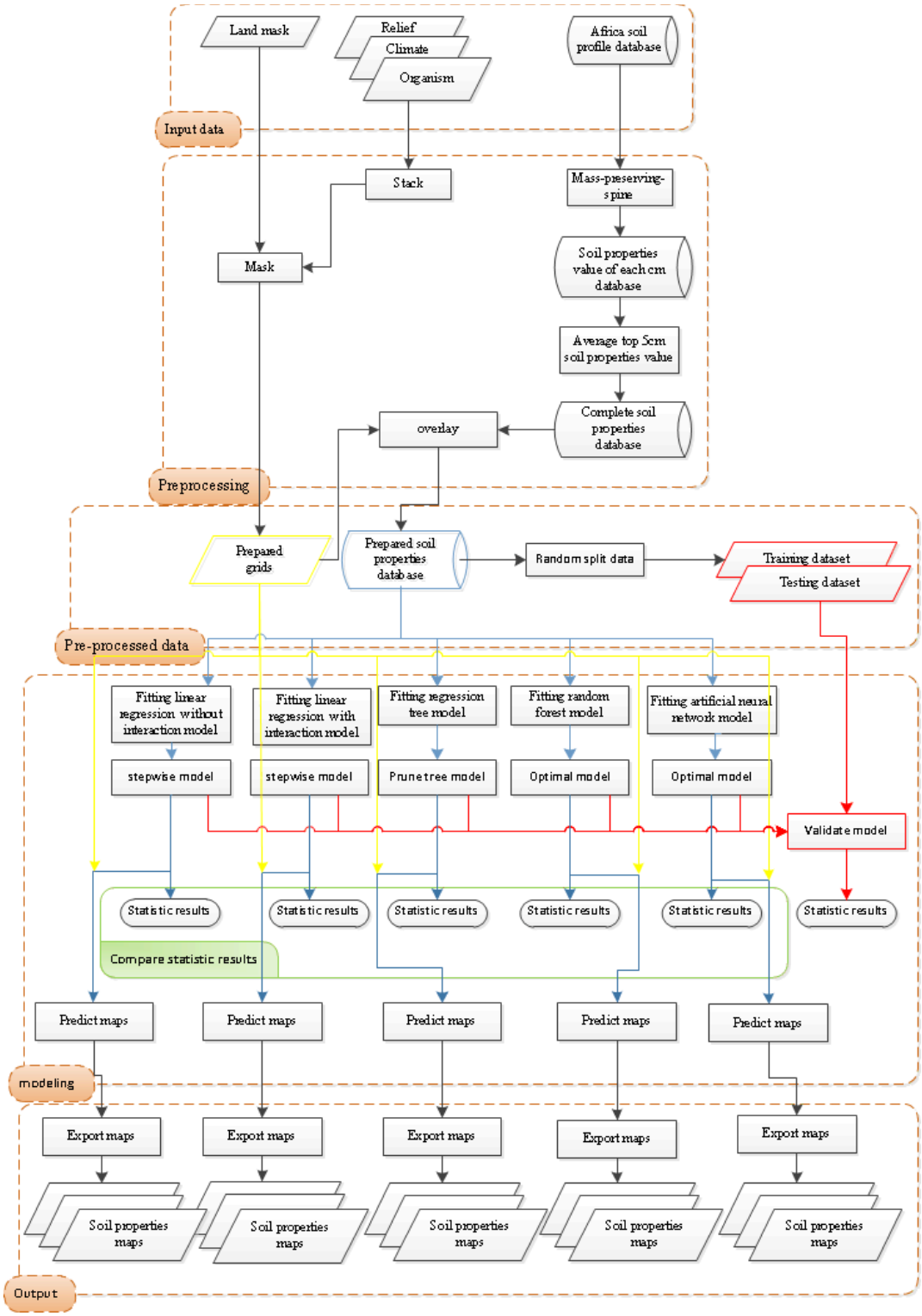


Figure 12 Flow chart of this thesis research

Figure 12 represents the flowchart of this thesis research, which consists of five parts. From top to down, these are: input data (2.2), pre-processing (2.2.2), pre-processed data (3.1.1), modelling development, prediction, comparison and validation (3.2 - 3.5) and output maps (3.5), which are discussed according to the section numbers.

2.5.2 R and R packages

R is a powerful and effective programming language and environment for data operating, statistical computing and graphics displaying. It has been developed and expanded rapidly including a large number of specialized packages, which apply many classical and modern statistical methods and techniques. Most of packages are available through the Comprehensive R Archive Network (CRAN) via *http://cran.r-project.org* (W. N. Venables, 2013), where also statistical methods for digital soil mapping are provided. R studio IDE is a powerful, productive and user friendly interface for R. Both R and RStudio IDE are able to run in Microsoft Windows System and they are free.

All calculation work including data pre-processing, model fitting and predicting and validation were done in R. The necessary libraries, functions and arguments were used and developed as follows:

1. **Prepare soil properties data of top 5 cm (response variables)**

   In this step，the soil profile data were converted to the fixed depth (0 -5 cm) value. Package 'GSIF' (Malone and Hengl, 2012b) offers a function 'mps' to fit soil depths value depending on the soil layers' value. The results of 'mps' present "spline interpolated" soil properties value per cm in a list. The data for the top 5 cm were averaged to get mean values for the top layer. Then, the pre-processed data were added to the original soil database to get the observed location coordinates. Finally, the new database was completed three times with the three response variables separately by deleting the soil data that lack of specific soil property.

2. **Prepare environmental covariates (predictor variables)**

   There were 31 predictors covariates that were introduced in section 2.2.3.They were stacked to one grid file and masked by the mask file that was explained in section 2.2.4. Because all predictor covariates and land mask data were the world data and they had the same spatial reference system, we finally clipped the study area SSA data out of the grids file.

3. **Prepare the modelling database**

   The prepared top soil properties database was spatially overlaid with pre-processed grids file This pre-processed soil database hence included soil properties value in top 5 cm, predictor covariate, observed location coordinates and coordinate system, which are needed for fitting the statistical models.

4. Linear Regression without Interaction

The basic package 'stats' offers a function 'lm' to fit linear models. Because the pH was assumed to follow a Gaussian distribution, the argument 'formula' of pH in all predict models excludes LRI is

pH ~ covariate 1+covariate 2 +·····+ covariate31

Clay content and organic carbon content (SOC) were assumed to be lognormal distributed, therefore the argument 'formula' is

log (1+Clay) ~ covariate 1+covariate 2 +……+ covariate31 and

log (1+SOC)~ covariate 1+covariate 2 +……+ covariate31

Function 'stepwise' was used to select predictors that can improve the model prediction accuracy. Function 'predict' was used to predict the unobserved location soil properties value according to the optimal model and the study area environmental covariates.

5. Linear Regression with Interaction

In linear regression with interaction model, the model fitting and predicting were similar to linear regression without interaction. The only difference was the argument 'formula' setting, here the first eight predictive covariates that means relief, climate and organism (EVI) were assumed to interact with each other, and therefore the formula of pH model was as follows：

pH ~ covariate 1* (covariate 2 +……+ covariate8) + covariate2 *(covariate 3 + covariate 4+ …… + covariate8)+ …… + covariate7* covariate 8+ covariate 1+covariate 2 +·····+ covariate31

The formulas of Clay and SOC model are similar as pH model, only the left side of models had been changed to 'log (1 + Clay)'or log (1 + SOC)'.

6. Regression Tree

Package 'rpart' was used to develop and predict regression trees models. Function 'rpart' was used for fitting a tree model, where the parameters 'cp' was used to control the complexity of trees to avoid model over-fitting. Less important branches of trees are removed when the cp is above a chosen threshold. The 'cp' value setting was based on the result of the model cross-validation. Function 'predict' was used to predict the whole study area map based on the developed RT model (Terry Therneau 2013).

7. Random Forests

Package 'randomForest' has function 'randomForest' to fit the classification and regression tree based on Breiman's algorithm (Breiman. et al., 2012). Three items mtry (splits number), nz (node size) and ntree (tree number) determines the goodness of the model fitting. In which，'mtry' was tried with settings 3, 4 and 5, 'nz' has been set as 5，10，15，20 and 'ntree' has been set as 500, 750, 1000, 1250. Different combinations of these three items were tested to get an optimal model. Function 'predict' was used to predict the whole study area map based on the developed RF model (Breiman. et al., 2012).

8. Artificial Neural Networks

Package 'neuralnet' function 'neuralnet' was used to fit the ANN. The structure of the ANN model is defined in argument 'hidden', which can be set in a flexible way that define the number of hidden layer and the number hidden units in one hidden layer. For example "hidden= c(a, b)", means the ANN model holds 2 hidden layer, the first layer has number a neurons and the second units holds b neurons. The argument "algorithm" was be set as 'backprop' meaning back propagation. Function 'compute' was used to predict the whole study area map based on the developed ANN model (Stefan Fritsch, 2012).

2.5.3 ArcGIS Desktop 10.1

ArcGIS is a geographic information system for mapping, designing, geographic data managing and offer solutions for geographic application, which is developed by ESRI Company. ArcGIS desktop 10.1 standard version is one of ArcGIS production, which runs in a Microsoft Windows environment. Under this version, the software is used to view ESRI format data, to edit the geo database and spatial raster of vector data, to design and develop application and to publish the maps. Here all calculations have been done in R, and ArcGIS 10.1 was used to firstly convert results of point data from R to raster data. Secondly, the colour of result raster maps were reclassified and edited. Finally, ArcGIS 10.1 published maps.

3 Results

3.1 Model input data

After pre-processing the data are prepared for the modelling work. The input data include the soil data and the predictor covariates.

3.1.1 Soil dataset

Appling the mass-preserving spline method, the soil properties in top soil 5 cm are calculated from profile observations at possibly irregular depth intervals at the observation locations. The new soil dataset preserves the observed location attributes, combined with the soil property values in the top soil 5 cm. To check the spline fitting the value of the soil properties were compared between the original soil dataset top horizon and new dataset (Figure 13). The figure shows that the fitted values agree quite well with the original values. The correlation between fitted and original value are all high, 0.9957 for pH, 0.9961 for SOC and 0.9961 for Clay content.

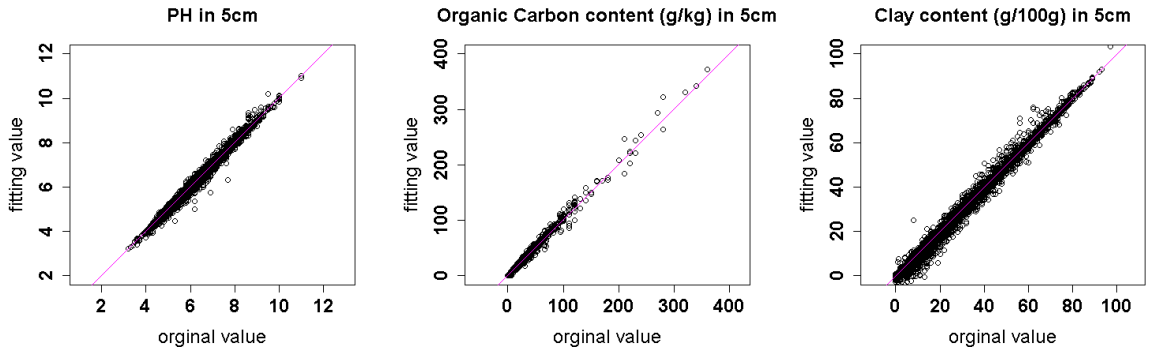


Figure 13 Fitted values (0-5 cm) against original values (top horizon) for the three soil properties

3.1.2 Predictor covariates

The predictor covariates are derived from the world grids repository (Table 1), and masked by the SMKMOD0a mask file. The predictor covariates maps of SSA are shown in Appendix 1.

The example map is shown in Figure 14. Some maps of predictor variables of land cover (i.e. G01ESA0a- G21ESA0a) are shown almost all dark blue, that means only a little area is the meaning of land cover in SSA, such as figure 15. The number of legend bar for G01ESA0a - G21ESA0a means how much percentage land cover of maps title in one pixel. For example, figure 9 shows that there was almost no post-flooding or irrigated croplands area in SSA, except the red point in the yellow circle.

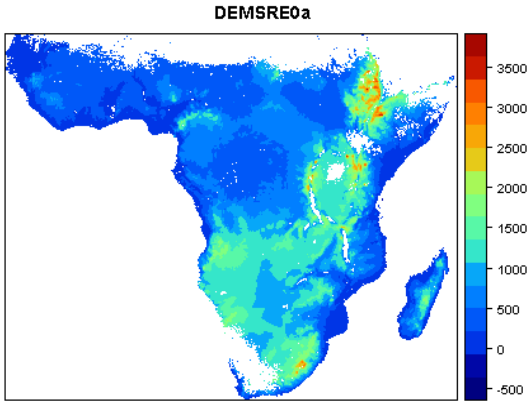


Figure 14 Predictor variable of DEM in Sub- Saharan Africa

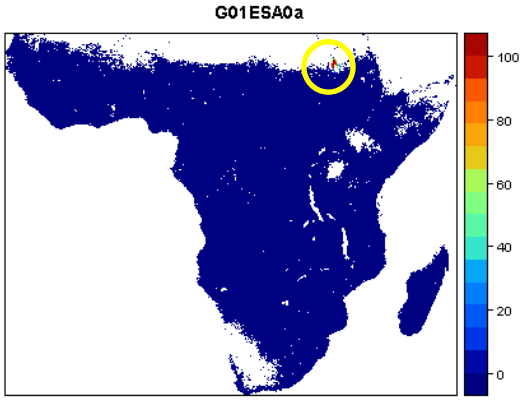


Figure 15 Predictor variable of G01ESA0a (post-flooding or irrigated croplands)

3.2 Model Interim results (pH)

This section interprets the interim results of each statistical model for soil pH, while the other two soil properties (clay content and organic carbon content) interim results are shown in Appendix 2. The procedures used to select predictive variables and setting model parameters for SOC and Clay are similar to that used for pH.

3.2.1 Linear regression without interaction

After LRW model fitting and selection of predictor variables by stepwise regression, the model can be summarized by Table 2. It shows the names of the predictor variables, the regression coefficients and their standard errors, and the statistical significance. The significance is displayed based on the p-value. The significance code shown depends on the p-value, as follows: '0,‘***’; 0.001, ‘**’; 0.01, ‘*’; 0.05, ‘.’, 0.1; ‘ ’ 1’.

The model using formula and input variables are:

MPH ~ DEMSRE0a + SLPSRT0a + TDMMOD0a + TDSMOD0a +
IFLGRE0a + TDHMOD0a + TNSMOD0a + EVMMOD0a + EVSMOD0a + PREGSM0a +
G01ESA0a + G02ESA0a + G03ESA0a + G04ESA0a + G05ESA0a + G06ESA0a +
G07ESA0a + G09ESA0a + G11ESA0a + G12ESA0a + G13ESA0a + G14ESA0a +
G15ESA0a + G17ESA0a + G19ESA0a + G20ESA0a + G21ESA0a

Table 2 Regression coefficients, standard errors and significance of predictor variables in pH linear regression without interaction

Variables	Estimate	Std.Err	Signif.	Variables	Estimate	Std.Err	Signif.
(Intercept)	4.20E+00	2.38E-01	***	G04ESA0a	4.87E-03	9.68E-04	***
DEMSRE0a	-4.22E-05	2.55E-05	.	G05ESA0a	6.90E-03	1.22E-03	***
SLPSRT0a	1.05E-02	7.40E-03		G06ESA0a	2.37E-03	1.09E-03	*
TDMMOD0a	9.28E-02	7.93E-03	***	G07ESA0a	5.18E-03	1.09E-03	***
TDSMOD0a	1.63E-01	1.37E-02	***	G09ESA0a	7.48E-03	5.11E-03	
IFLGRE0a	-4.00E-01	1.49E-01	**	G11ESA0a	8.32E-03	1.13E-03	***
TDHMOD0a	-4.63E-02	5.47E-03	***	G12ESA0a	3.79E-03	1.36E-03	**
TNSMOD0a	-5.60E-02	1.15E-02	***	G13ESA0a	5.81E-03	1.01E-03	***
EVMMOD0a	1.65E-04	2.72E-05	***	G14ESA0a	8.30E-03	1.00E-03	***
EVSMOD0a	1.32E-04	5.09E-05	**	G15ESA0a	1.44E-02	1.64E-03	***
PREGSM0a	-7.96E-04	3.38E-05	***	G17ESA0a	5.38E-03	2.25E-03	*
G01ESA0a	2.44E-02	3.13E-03	***	G19ESA0a	8.06E-03	2.04E-03	***
G02ESA0a	5.03E-03	1.08E-03	***	G20ESA0a	1.50E-02	1.58E-03	***
G03ESA0a	5.20E-03	1.05E-03	***	G21ESA0a	8.03E-03	2.76E-03	**

Comparison with the input covariates (Table 1) shows that TNMMOD0, G10ESA0, G16ESA0, G18ESA0 are deleted during the stepwise regression, which means that these four parameters do not provide useful additional information to explain spatial variation in pH. The Relief (DEMSRE0a and SLPSRT0a) and G09ESA0a variables are the least significant variables. The Climate variables (TDMMOD0a, TDSMOD0a, TDHMOD0a, TNSMOD0a and PREGSM0a) are all highly significant. The regression coefficients of DEMSRE0a, IFLGRE0a, TDHMOD0a, TNSMOD0a and PREGSM0a have negative signs while the regression coefficients of all other variables have a positive sign.

3.2.2 Linear regression with interaction

The LRI model results are summarized in Table 3. The difference with Table 2 is that interaction terms are included as well. When two covariates show negative coefficient with pH value, their interaction will give a positive coefficient, although the coefficient s is low, such as DEMSRE0a and SLPSRT0a. When two variables have positive coefficient with pH, these two variables interaction has positive coefficient with pH as well, such as EVSMOD0a and TDHMOD0a. When one variable has positive coefficient with pH and another one not, their interaction has the negative coefficient, such as DEMSRE0a and EVSMOD0a.

Table 3 Regression coefficients, standard errors and significance of predictor variables in pH linear regression with interaction model

Variables	Estimate	Std.Error	signif.	Variables	Estimate	Std.Error	signif.
(Intercept)	1.22E+01	2.27E+00	***	DEMSRE0a:TNMMOD0a	-5.56E-05	8.64E-06	***
DEMSRE0a	-2.36E-03	3.70E-04	***	DEMSRE0a:TNSMOD0a	-1.00E-04	3.15E-05	**
SLPSRT0a	-2.48E-02	1.37E-01		DEMSRE0a:PREGSM0a	8.32E-07	1.17E-07	***
EVSMOD0a	1.22E-03	3.13E-04	***	DEMSRE0a:IFLGRE0a	-1.53E-03	7.63E-04	*
EVMMOD0a	-3.21E-04	2.55E-04		SLPSRT0a:EVSMOD0a	-7.34E-05	3.11E-05	*
TDHMOD0a	-5.77E-01	6.44E-02	***	SLPSRT0a:EVMMOD0a	-2.81E-05	1.46E-05	.
TDMMOD0a	4.48E-01	8.52E-02	***	SLPSRT0a:TDMMOD0a	-6.86E-03	3.77E-03	.
TDSMOD0a	8.75E-01	1.64E-01	***	SLPSRT0a:TDSMOD0a	-1.94E-02	7.29E-03	**
TNMMOD0a	-2.87E-01	8.49E-02	***	SLPSRT0a:TNMMOD0a	2.72E-02	5.11E-03	***
TNSMOD0a	-1.62E-01	1.99E-01		SLPSRT0a:PREGSM0a	-6.66E-05	2.55E-05	**
PREGSM0a	-2.97E-03	5.87E-04	***	EVSMOD0a:EVMMOD0a	-1.07E-07	6.65E-08	
IFLGRE0a	6.50E+00	3.04E+00	*	EVSMOD0a:TDSMOD0a	-1.51E-04	3.34E-05	***
G01ESA0a	1.75E-02	3.09E-03	***	EVSMOD0a:PREGSM0a	4.50E-07	1.29E-07	***
G02ESA0a	4.83E-03	1.10E-03	***	EVMMOD0a:TDHMOD0a	3.90E-05	8.46E-06	***
G03ESA0a	4.96E-03	1.08E-03	***	EVMMOD0a:TDMMOD0a	-2.50E-05	1.06E-05	*
G04ESA0a	3.51E-03	9.69E-04	***	EVMMOD0a:TDSMOD0a	-1.24E-04	2.42E-05	***
G05ESA0a	3.81E-03	1.23E-03	**	EVMMOD0a:TNSMOD0a	8.02E-05	2.25E-05	***
G06ESA0a	3.82E-03	1.11E-03	***	EVMMOD0a:PREGSM0a	1.16E-07	5.58E-08	*
G07ESA0a	5.40E-03	1.10E-03	***	TDHMOD0a:TDMMOD0a	4.19E-03	1.15E-03	***
G09ESA0a	7.72E-03	5.10E-03		TDHMOD0a:TDSMOD0a	8.72E-03	3.24E-03	**
G11ESA0a	6.92E-03	1.14E-03	***	TDHMOD0a:TNMMOD0a	8.44E-03	2.25E-03	***
G12ESA0a	3.07E-03	1.34E-03	*	TDHMOD0a:TNSMOD0a	2.20E-02	3.84E-03	***
G13ESA0a	5.02E-03	1.03E-03	***	TDMMOD0a:TDSMOD0a	-3.76E-02	6.93E-03	***
G14ESA0a	4.67E-03	1.02E-03	***	TDMMOD0a:TNMMOD0a	-7.74E-03	2.30E-03	***
G15ESA0a	7.87E-03	1.73E-03	***	TDMMOD0a:TNSMOD0a	-1.88E-02	7.24E-03	**
G17ESA0a	6.28E-03	2.39E-03	**	TDMMOD0a:PREGSM0a	-6.95E-05	1.33E-05	***
G19ESA0a	6.91E-03	2.00E-03	***	TDSMOD0a:TNMMOD0a	2.52E-02	5.57E-03	***
G20ESA0a	1.04E-02	1.67E-03	***	TNMMOD0a:TNSMOD0a	-2.06E-02	6.35E-03	**
G21ESA0a	1.01E-02	2.73E-03	***	TNMMOD0a:PREGSM0a	1.60E-04	2.26E-05	***
DEMSRE0a:SLPSRT0a	9.51E-05	2.23E-05	***	TNMMOD0a:IFLGRE0a	-3.70E-01	1.41E-01	**
DEMSRE0a:EVSMOD0a	-2.85E-07	8.83E-08	**	PREGSM0a:IFLGRE0a	5.11E-04	2.86E-04	.
DEMSRE0a:TDHMOD0a	7.66E-05	7.61E-06	***				

3.2.3 Regression tree

Table 4 shows the default grown regression tree attribute, which is an initial big tree. The tree was pruned based on 5- fold cross-validation, that is the data were randomly split into 5 sections, each time using 4 sections to build the tree and one section reserved for validation, so the tree building and pruning process were performed 5 times. In the table, CP is the Complex Parameter. “nsplit” is the numbers of splits. “rel error” means the error of prediction estimates. “xerror” column contains the value of the 5 – fold cross-validated prediction error. “xstd” column indicates the variance of “xerror” among the 5-fold cross-validated prediction . The smallest “xerror” is when cp=0.01, so the tree does not need to prune.

Table 4 The attributes of grown tree based on 5- fold cross-validation

	CP	nsplit	rel error	xerror	xstd
1	0.113876	0	1	1.00015	0.016622
2	0.044383	1	0.886124	0.888901	0.014384
3	0.03932	2	0.841741	0.846636	0.014171
4	0.019138	3	0.802421	0.813846	0.013902
5	0.015204	4	0.783283	0.80497	0.013847
6	0.013763	5	0.768079	0.790051	0.013656
7	0.01	6	0.754316	0.770759	0.013312

Figure 16 represents the information from Table 4 as a partition structure tree. There are 9 terminal root nodes. The terminal nodes represent the final groups' numeric prediction for the response variable pH. Under each terminal node, the first value means the predicted value. Below that, the "n =" means the number of observed value meeting this node requirements. Four variables, PREGSM0a, TDMMOD0a, TDHMOD0a and DEMSRE0a are presented in the tree that determines the tree splits.

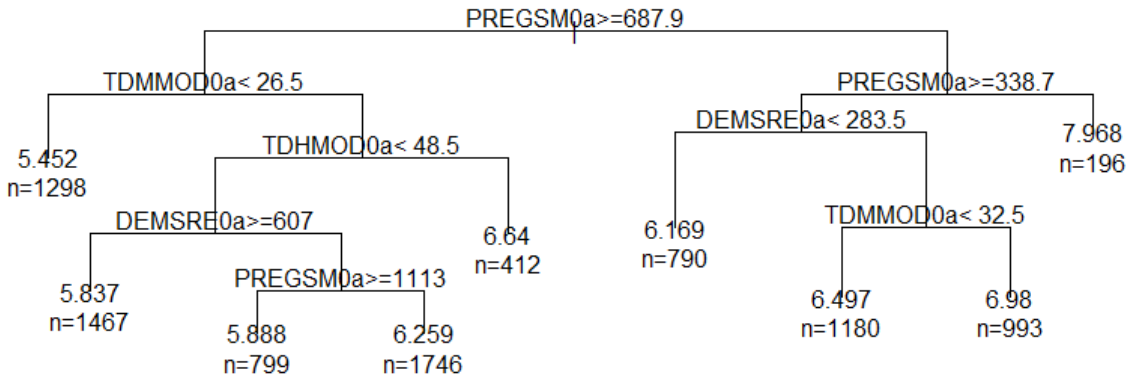


Figure 16 Result regression tree for pH

3.2.4 Random forest

As discussed in Section 2.3.4, the accuracy of the random forest model is controlled by the node size (nz), the number of splits (mtry) and the number of trees (ntree). In this study, nz is taken as 10, 15, 20, 25, 30, splits number(mtry) as 3, 4, 5, and tree number (ntree) as 500, 750, 1000, 1250.The model is tested for all combinations of nz, mtry and ntree, the result table are shown in Table 5 and Table 6.

The highlight number in Table 5 is the highest value R^2: 0.439 and the highlight number in Table 6 is the lowest RMSE: 0.759. Both tables indicate the best combination of parameters in the random forest model is nz as 15, mtry as 5 and ntree as 1250. However, the differences between different combinations are quite small.

Figure 17 shows the importance of predictor variables in the random forest model. The x-axis 'increasing mean squared error %' indicates when remove one variable, how much error will be increased. When removing upper variables, the error of the model increases more than the lower variables. This means that upper variables are more important than lower variables. The variables elevation (DEMSRE0a), enhanced vegetation index (EVSMOD0a and EVMMOD0a), land use of mosaic grassland and crop land (G04ESA0a), precipitation (PREGSM0a) data and temperature variables (T*******) are the most important predictors.

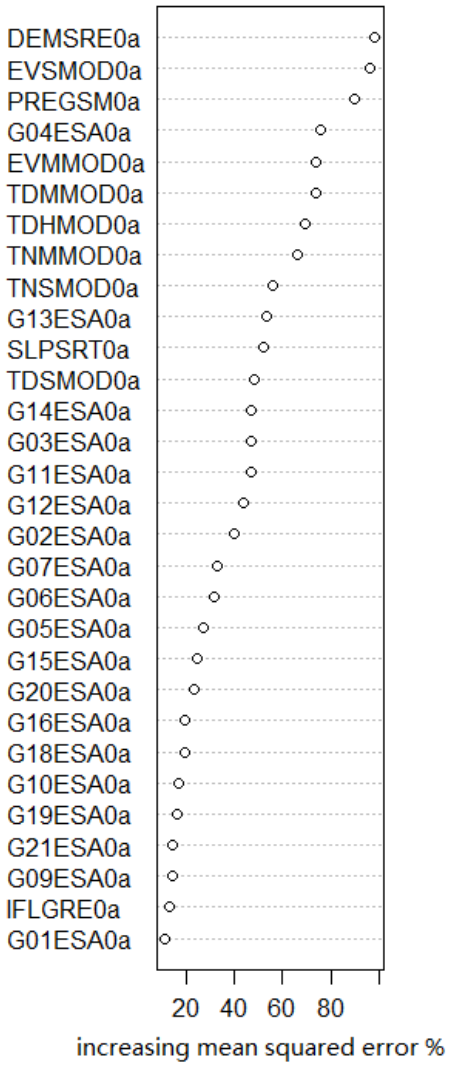


Figure 17 Importance of predictor variables in random forest model for pH, x-axis means the percentage of increased mean squared error when remove one variable.

Table 5 R^2 of pH in random forest

ntree	500			750			1000			1250		
nz						mtry						
	3	4	5	3	4	5	3	4	5	3	4	5
10	0.4280	0.4378	0.4362	0.4286	0.4385	0.4367	0.4284	0.4383	0.4370	0.4281	0.4382	0.4373
15	0.4247	0.4367	0.4388	0.4254	0.4371	0.4387	0.4255	0.4372	0.4389	0.4257	0.4372	0.4390
20	0.4218	0.4333	0.4354	0.4225	0.4336	0.4354	0.4228	0.4339	0.4353	0.4228	0.4339	0.4351
25	0.4179	0.4265	0.4285	0.4179	0.4273	0.4294	0.4178	0.4269	0.4296	0.4178	0.4266	0.4297
30	0.4102	0.4193	0.4215	0.4107	0.4198	0.4221	0.4112	0.4200	0.4225	0.4114	0.4201	0.4230

Table 6 RMSE of pH in random forest

ntree	500			750			1000			1250		
						mtry						
nz	3	4	5	3	4	5	3	4	5	3	4	5
10	0.7666	0.7600	0.7611	0.7662	0.7596	0.7607	0.7663	0.7597	0.7606	0.7665	0.7598	0.7604
15	0.7688	0.7608	0.7593	0.7684	0.7605	0.7594	0.7683	0.7604	0.7593	0.7681	0.7604	0.7592
20	0.7707	0.7630	0.7617	0.7703	0.7628	0.7617	0.7701	0.7627	0.7617	0.7701	0.7627	0.7619
25	0.7734	0.7676	0.7663	0.7734	0.7671	0.7657	0.7734	0.7673	0.7656	0.7734	0.7675	0.7655
30	0.7784	0.7725	0.7709	0.7781	0.7721	0.7706	0.7778	0.7719	0.7703	0.7777	0.7719	0.7700

3.2.5 Artificial neural network

The structure of the ANN model was increased from only one layer in the beginning to four layers at the end. The number of hidden neurons for each layer was chosen according to the setting strategy that was introduced in Section 2.3.5.

The results of SSE with different number hidden neurons in different number hidden layers are given in Table 7. Firstly, only one layer with neuron numbers of 17, 23, 50 or 60 were tested. The smallest SSE (yellow highlight number) is 3472.95 in one layer with 50 neurons. Next the second layer was set with 26, 35, 50, 70 or 90 neurons. Now the smallest SSE is 3193.44 in two layers that first layer owning 50 neurons and second layer owning 26 neurons. Based on this result, the third layer neurons numbers were set as 14, 17, 25, 35 or 52. The best result in three layers is obtained when the first layer has 50 neurons, the second layer 26, and the third layer 14 neurons. The SSE is 2549.94. Finally, the fourth layer neurons numbers were set as 9, 12, 16, 20 and 36. The smallest SSE is 2819.15 in four layers, but it is bigger than the best result with three layers, so the best structure in ANN is three layers, where the first layer has 50 neurons, the second 26, and the third layer 14 neurons.

Table 7 Sum of squared error in one layer artificial neural network model

one layer	SSE	Two layers	SSE	Three layers	SSE	Four layers	SSE
17	4028.9	50, 26	3193.44	50, 26, 14	3185.51	50, 26, 17, 9	3030.44
23	3788.7	50, 35	3196.29	50, 26, 17	2549.94	50, 26, 17, 12	2819.15
40	3684.13	50, 50	3358.76	50, 26, 25	3216.72	50, 26, 17, 16	3205.04
50	3472.95	50, 70	3526.47	50, 26, 35	2996.91	50, 26, 17, 20	3079.28
64	3507.27	50, 90	3562.76	50, 26, 52	2905.41	50, 26, 17, 32	3150.81

ANN models stated extremely bad results for soil properties, where RMSE in pH is 1.81 and R^2 is -2.1. That means the ANN model was totally under fitting and failed to produce acceptable predictions of soil properties.

3.3 Assessment of model fitting

All models except ANN are evaluated by the value of R^2 and RMSE, because ANN presents unacceptable result, which will be excluded from the comparison with other prediction models. The histogram (Figure 18) illustrates that the R^2 of RF models in three soil properties are all the highest, above 40%. The LRW models have the lowest R^2 for all three soil properties, with an exceptionally small value in the case of clay, below 0.05. For LRI the R^2 for clay is below0.15.

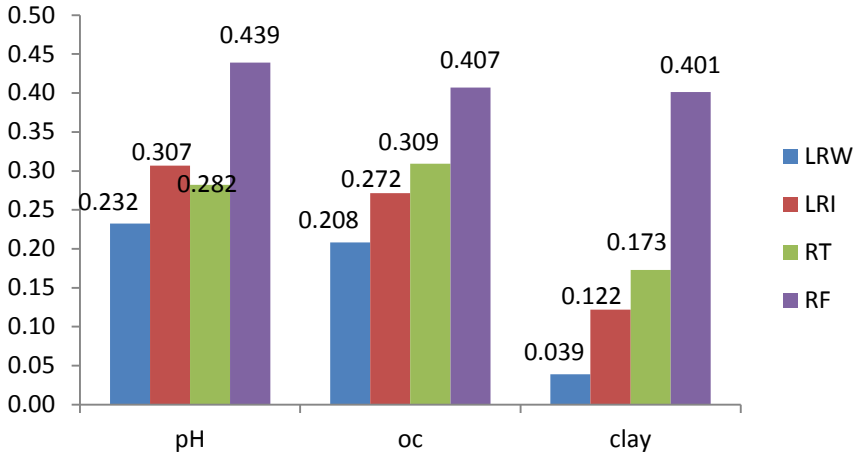


Figure 18 R^2 of the four prediction models for pH, organic carbon content and clay content

Histograms (Figure 19 to 21) show the RMSE of the four models for pH, SOC and clay content. The corresponding values are given in Table 8. Conversely with R^2， the RF models always had the lowest RMSE value, while LRW models always had the highest RMSE value. In fact, the highest R^2 and the lowest RMSE agree to each other, both of them indicate the best model.

Table 8 RMSE of the four prediction models for pH, organic carbon content and clay content

soil property	LRW	LRI	RT	RF
pH	0.890	0.847	0.882	0.759
SOC (g/kg)	17.038	16.430	15.910	14.746
Clay(g/100g)	18.533	17.714	17.192	14.629

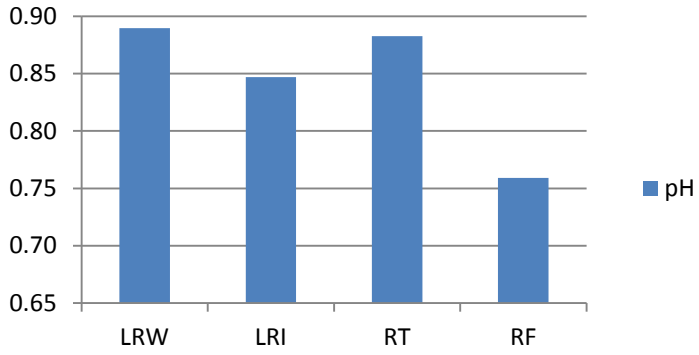


Figure 19 RMSE of pH for the four predict models.

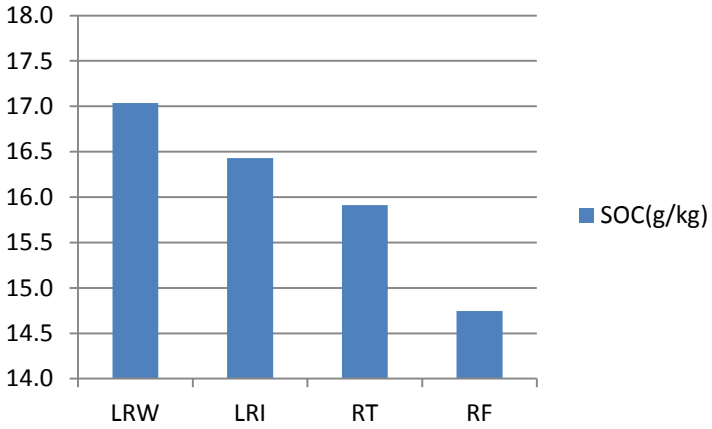


Figure 20 RMSE of soil organic carbon content (g/kg) for the four predict models.

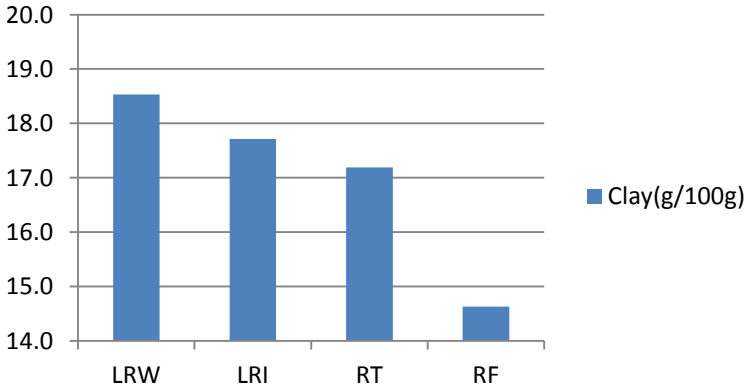


Figure 21 RMSE of clay content (g/100g) for the four predict models.

Table 9 shows the correlation between observed value and predicted values at observation locations. The predicted values from the RF model always have the highest correlation with the observed values, while the LRW always has the lowest correlation in three soil properties. The correlation between the two linear models is high.

Table 9 Correlation between observed value and predicted value from four model for pH, organic carbon content and clay content

	Observation			LRW			LRI			RT			RF		
	pH	SOC	Clay	pH	SOC	Clay	pH	SOC	Clay	pH	SOC	Clay	pH	SOC	Clay
Observation	1.00	1.00	1.00												
LRW	0.48	0.51	0.36	1.00	1.00	1.00									
LRI	0.55	0.55	0.45	0.87	0.89	0.79	1.00	1.00	1.00						
RT	0.51	0.61	0.42	0.76	0.71	0.56	0.81	0.73	0.64	1.00	1.00	1.00			
RF	0.79	0.78	0.78	0.69	0.73	0.54	0.78	0.78	0.65	0.73	0.83	0.62	1.00	1.00	1.00

3.4 Independent validation

Section 3.3 used the whole soil dataset to fit models and use these models to predict at observation locations, which may lead to overoptimistic results. Thus this section shows the independent validation results of soil properties prediction using all models, excluding the ANN model. The evaluation parameters are again RMSE and R^2. Tables 13 and 14 show results for soil pH, table 15 and 16 were the results for SOC, and Tables 16 and 17 for clay content.

3.4.1 Soil pH

Table 10 shows that the best performance model is the RF model, where the RMSE are the lowest when the dataset split to 50-50%, 60-40% and 70-30% (training- testing dataset). The LRW and RT model have the highest error. Correspondingly, Table 11 shows that the R^2 in RF is the highest, where both training and testing dataset above 40%. The LRW model has a poor result, where the R^2 is only about 22%. Though the RMSE results of LRW and RT are similar, the regression tree performs better than LRW according to the R^2. The difference between training set and testing set are small in all prediction models. The training data always do a bit better than the testing data, because all models were firstly developed by training data, when using the predictive model to predict the same dataset, it must have better results than the other dataset.

Table 10 RMSE of linear regression without interaction, linear regression with interaction, regression tree and random forest for different percentage of training and testing dataset for pH

Model	LRW		LRI		RT		RF	
train-test%	Training	Testing	Training	Testing	Training	Testing	Training	Testing
50-50%	0.90	0.88	0.85	0.85	0.90	0.88	0.79	0.78
60-40%	0.90	0.88	0.85	0.85	0.89	0.87	0.78	0.77
70-30%	0.89	0.88	0.85	0.85	0.88	0.87	0.77	0.78

Table 11 R^2 of linear regression without interaction, linear regression with interaction, regression tree and random forest for different percentage of training and testing dataset for pH

Model	LRW		LRI		RT		RF	
train-test%	Training	Testing	Training	Testing	Training	Testing	Training	Testing
50-50%	0.2238	0.2234	0.3086	0.2702	0.2805	0.2519	0.4088	0.4012
60-40%	0.2248	0.2308	0.3062	0.2812	0.2776	0.2539	0.4209	0.4035
70-30%	0.2299	0.2277	0.3096	0.2798	0.2767	0.2532	0.4315	0.4001

3.4.2 Soil organic carbon content

Table 12 shows that the RF model has the best performance with the lowest RMSE, whereas the LRW model has the opposite result. The differences between results for the training and testing data sets are small.

Table 13 shows that the RF model also has the best performance for SOC. The noticeable thing is that when the data split is 50-50, the differences between the R^2 of the tree models are high. The RT model has 34.58% against 23.34% in training and testing dataset and the RF model has 40.92% against 33.15%. However, when the training dataset randomly selected by 60% and 70% of the data, the difference of R^2 is small, but the performance of training data is worse than the testing data. The performance in LRW, LRI and RT models show better results in training data than testing data, but the difference between training data and testing data of RT models are significantly higher. The R^2 of the training data is around 9-11% higher than that of the testing data, which indicates that the RT model is unstable and may over-fit. The R^2 in linear models are small, around 22% for the LRW model and about 29% for the LRI model.

Table 12 RMSE of linear regression without interaction, linear regression with interaction, regression tree and random forest for different percentage of training and testing dataset for organic carbon content

Model	LRW		LRI		RT		RF	
train-test%	Training	Testing	Training	Testing	Training	Testing	Training	Testing
50-50%	16.97	17.10	16.28	16.58	15.46	16.79	14.91	15.34
60-40%	17.06	16.95	16.26	16.43	15.1	16.97	15.46	13.91
70-30%	17.4	15.98	16.63	15.63	15.66	15.18	15.11	14.02

Table 13 R^2 of linear regression without interaction, linear regression with interaction, regression tree and random forest for different percentage of training and testing dataset for organic carbon content

Model	LRW		LRI		RT		RF	
train-test%	Training	Testing	Training	Testing	Training	Testing	Training	Testing
50-50%	0.2122	0.2048	0.28	0.25	0.3458	0.2334	0.4092	0.3315
60-40%	0.2265	0.1840	0.30	0.23	0.3941	0.1822	0.3847	0.3881
70-30%	0.2203	0.1919	0.29	0.23	0.3687	0.2710	0.3880	0.4226

3.4.3 Soil clay content

Table 14 illustrates that the RF model had the smallest RMSE and LRW has the largest. As before, the training dataset yields better performance indices compared to the testing data.

Table 15 clearly shows that the linear model had the worst and poor result, where the R^2 is only around 4.5% for the LRW model and around 10% for the LRI model. The RT model also has a very poor result, with an R^2 below 20%. The RF model has a much better result than the other three models, and in this case the training dataset yields a similar result as the testing dataset.

Table 14 RMSE of linear regression without interaction, linear regression with interaction, regression tree and random forest for different percentage of training and testing dataset for clay content

Model	LRW		LRI		RT		RF	
train-test%	Training	Testing	Training	Testing	Training	Testing	Training	Testing
50-50%	18.14	18.99	17.39	18.51	16.96	17.46	14.96	15.21
60-40%	18.46	18.56	17.63	18.04	17.12	17.34	14.81	15.18
70-30%	18.44	18.61	17.65	18.00	17.06	17.54	14.79	14.91

Table 15 R^2 of linear regression without interaction, linear regression with interaction, regression tree and random forest for different percentage of training and testing dataset for clay content

Model	LRW		LRI		RT		RF	
train-test%	Training	Testing	Training	Testing	Training	Testing	Training	Testing
50-50%	0.0508	0.0199	0.1276	0.0690	0.1736	0.1720	0.3722	0.3590
60-40%	0.0401	0.0452	0.1248	0.0982	0.1771	0.1666	0.3845	0.3630
70-30%	0.0442	0.0416	0.1245	0.1038	0.1840	0.1494	0.3902	0.3752

3.5 Predicted maps for soil properties

This section presents the resulting soil property maps of SSA that are derived with the four predictive models. Because the ANN model could not be fitted properly, the associated soil property maps were not produced.

3.5.1 Soil pH (H_2O)

Soil pH maps (Figure 22) predicted by four models show the similar patterns, the low pH values occur in central Africa and high pH values present in West Africa near the coast and the North Africa near the Saharan desert. The maps derived from RT model is the most smooth one, which has less legends than the other pH maps, because its predicted value do not have the extremely values.

The box plot (Figure 23) describes the pH predictions in the whole study area. The median of the pH predictions for the regression tree method is small and equal to the first quartile (bottom of box) , while for the other statistical models the medians are in the box centre and very similar. The whiskers are also similar for the linear models and random forest model. The lower whisker values are a bit larger than 4 and the higher whiskers are around 8. The locations of four boxes are all in the centre of whiskers, which means the pH predictions are normality distribution. The regression tree box size is the narrowest, indicating that predicted pH value concentrate a small range. The other three box sizes are similar. The linear models have more extreme values, while the predicted value from regression tree model only has one extreme value.

The specific values of the box plots (Figure 23) can be read from Table 16. In all four models, the median values and mean values are around 6.1. The predicted pH value range in regression tree,

from 5.45 to 7.97, is lower than the other three predictor models. The predicted values of the 1st quartiles (25%) are around 5.7 and the 3rd quartiles (75%) are about 6.5 in all four models.

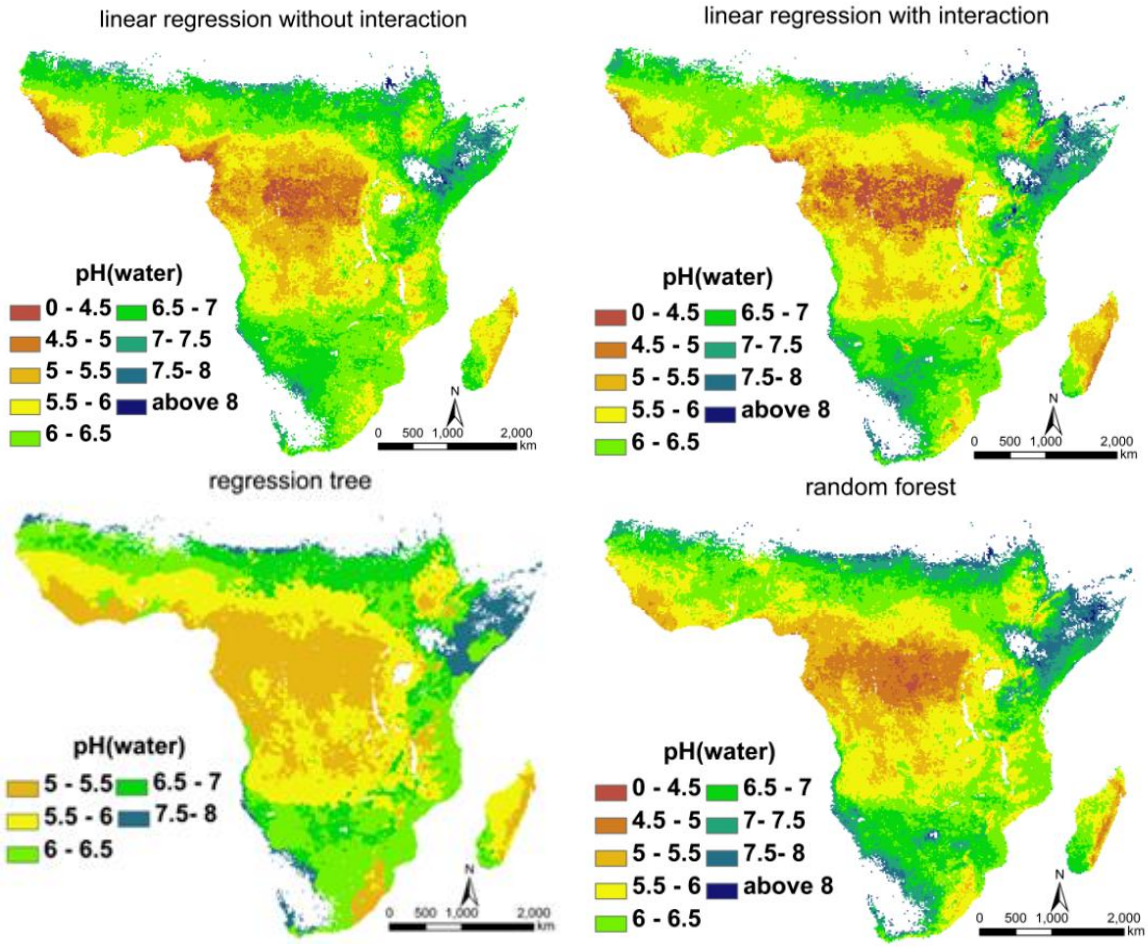


Figure 22 pH in top 5cm maps with different statistics approach

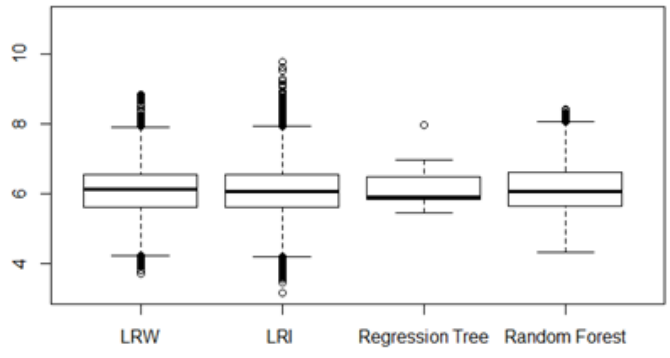


Figure 23 Box plots of soil pH value in different prediction maps

Table 16 The value of soil pH in different prediction maps

pH	LRW	LRI	RT	RF
Min.	3.72	3.17	5.45	4.32
1st	5.62	5.61	5.84	5.64
Median	6.13	6.07	5.89	6.06
Mean	6.08	6.08	6.14	6.13
3rd	6.54	6.54	6.50	6.61
Max.	8.83	9.76	7.97	8.43

3.5.2 Soil organic carbon content

SOC maps (Figure 24) predicted by four models show the similar patterns. The low SOC values occur in South and North Africa, the closer to the desert the lower SOC value. The high SOC value presents in the central Africa, where has forest and grassland. The maps derived from RT model is the most smooth one, which has less legends than the other SOC maps, because its predicted value do not have the lowest values and lack range 35-40 values.

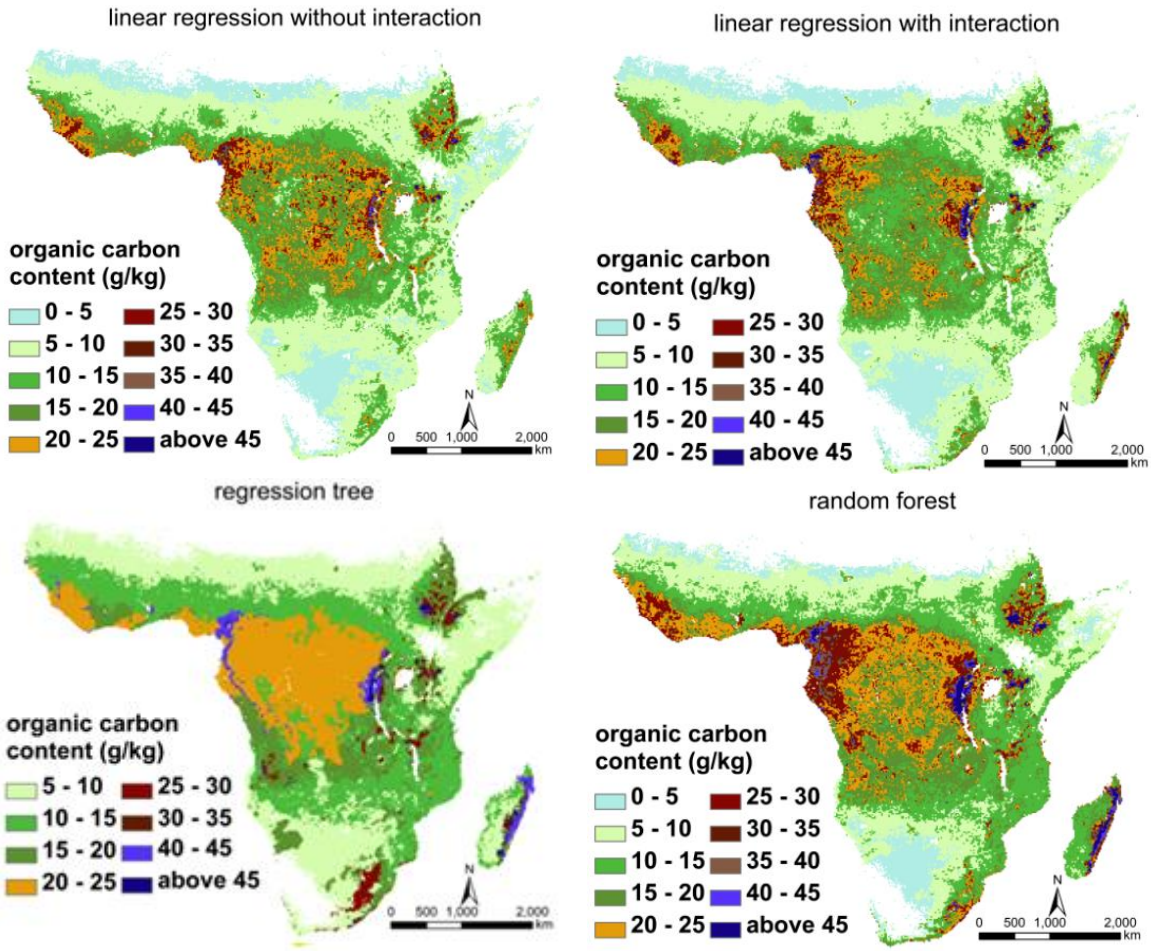


Figure 24 Organic carbon content in top 5cm maps with different statistics approach

The box plots (Figure 24) describe the soil organic carbon content (SOC) value information in the whole study area. The whiskers location in linear models and random forest are similar. The location of all box within the whiskers are low, which means the SOC value are skewed. The extreme values in all models are much higher than the median. The RT model has less extreme values.

The specific value of the box plots (Figure25) can be read from Table 17. The predictions of the linear models (LRW and LRI) show similar result in all statistical measures. The ranges of SOC are from 0 to around 420, the 1st quartiles are around 6.6 and the 3rd quartiles are around 16.5. The median is around 10.3 and the mean is around 12.1. While the tree models (RT and RF) illustrate a range of SOC values that are much smaller than those of the linear models, from 7.3 to 192.7 in RT and from 2.5 to 112.8 in RF compare to from 0 to around 420 in linear models. The other statistical items in tree models are all a bit higher than for the linear models.

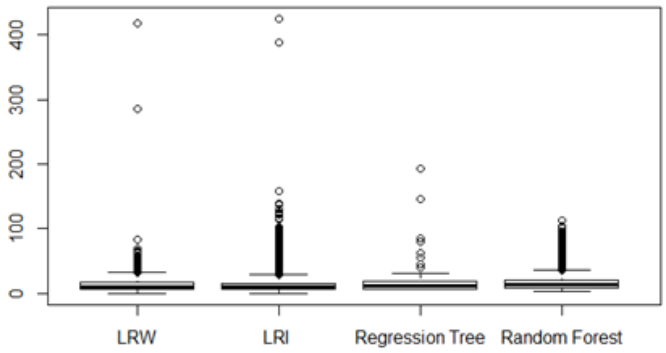


Figure 25 Box plots of soil organic carbon content in different prediction map

Table 17 The value of soil organic carbon content in different prediction maps

Organic Carbon	LRW	LRI	RT	RF
Min.	0	0	7.3	2.5
1st	6.5	6.9	7.3	9.1
Median	10.21	10.3	12.7	13.5
Mean	12.1	12.2	15.1	15.3
3rd	17.0	16.0	18.0	20.24
Max.	417.1	424.8	192.7	112.8

3.5.3 Soil clay content

Soil clay content maps (Figure 26) predicted by four models show the similar patterns. The low clay content values occur in South and North Africa, the closer to the desert the lower clay content value. The high SOC values occur in the north-east part of Africa. The maps derived from RT model is the most smooth one, which has less legends than the other SOC maps, because its predicted value are discrete value, which lack some range values.

The box plot (Figure 27) describes the soil clay content value information in the whole study area. The specific value of the box plot (Figure 27) can be read from Table 18. The clay content value range, from 0 to around 97, in linear models are much wider than in RT and RF models, from 0 to about 58. The statistical values in other items of tree models are higher than in linear models. The other items of Table 18 can be read as same way as Table 16 and Table 17.

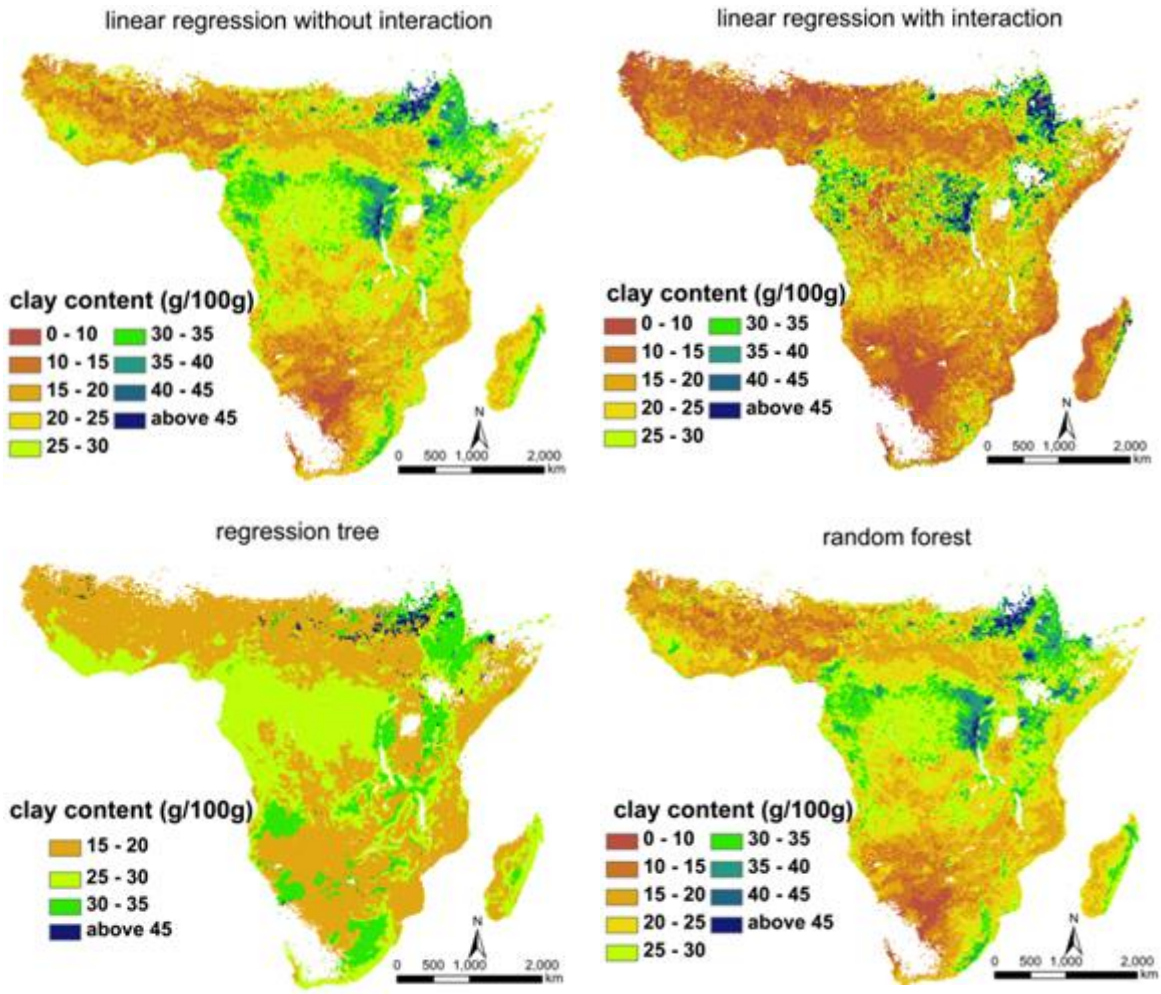


Figure 26 Clay content map in top 5cm map with different statistics approach

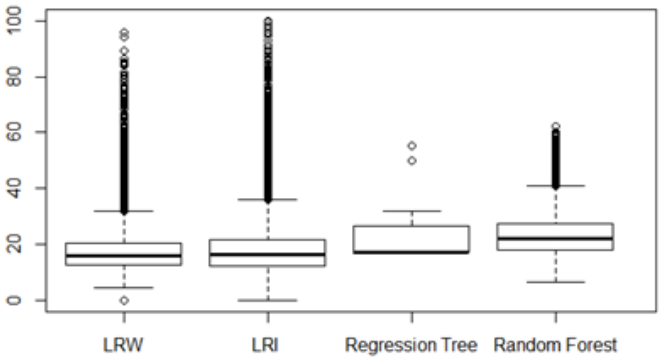


Figure 27 Box plots of soil clay content in different prediction map

Table 18 The value of soil clay content in different prediction maps

Clay	LRW	LRI	RT	RF
Min.	0	0	17.25	6.32
1st	12.59	12.37	17.25	18.03
Median	15.85	16.43	17.25	22.19
Mean	17.2	18.38	22.28	23.02
3rd	20.25	21.76	26.71	27.22
Max.	95.8	99.91	55.26	61.87

4 Discussion

4.1 Data pre-processing

The predictive qualities of the soil property values at fixed depth are very good. The correlations between the fitted and original values are very high for all three soil properties, all above 99.5%. Therefore, the processed soil data in the fixed depth, 5 cm, can be seen as the observed value, In addition, because the depths of top horizon in different location are different, which are not all equal to 5 cm, so the fitting values is not necessary exactly equal to the original value.

The good-fitting is partly due to the spline functions, which has much flexibility. The spline using the quadratic polynomials fit each observed location profile, that means the variances of soil properties values in different depth of the same location have been considered, therefore the variances do not affect the fitting value anymore (Bishop et al., 1999). In addition, the spline function uses the equal-area spline curve to fit the soil profiles to decrease the effect of horizon samples (Ponce-Hernandez et al., 1986).

The predictor variables are masked by MODIS LAI file from the world grids and clipped by the study area, which actually contain the same data in the study area, so they do not give any negative effect for models.

4.2 Model results

This section discusses how the four statistical models select the variables and their interim results as presented in Section 3.2 and compares these results presented in Section 3.3.

4.2.1 Linear regression without interaction

The LRW model uses stepwise regression to select the predictors, which is based on the AIC criterion. The criterion seeks the maximum likelihood estimators from all random generated linear regression models, and does not delete less significant predictors, such as relief variables DEMSRE0a and SLPSRT0a (Table 2). Except the predictor of land use in needle forest (G09ESA0a), the other land use covariates are all shown to be highly significant.

The resulting model holds multiple predictive variables, so the coefficient shows how much the response variable would increase or decrease with one unit increase of the associated predictive variable (and holding all the other predictors constant). The positive or negative coefficient indicates the direction of change. For example, all organism variables have positive regression coefficients for pH, which means that the pH value increases when these variables increase. PREGSM0a has high negative regression coefficient for pH, which means that higher precipitation leads to lower pH. TDMMOD0a has high positive regression coefficient for pH, which means that higher mean temperature in day time the higher pH. However, TDHMOD0a shows a negative regression coefficient, which means the higher maximum temperature in day

time contributes to the lower pH. The two temperature variables gave different direction effects to pH, but because the size of TDMMOD0a is larger than TDHMOD0a, the former one has more effects than the latter one for pH.

4.2.2 Linear regression with interaction

The method of fitting models and selecting predictor variables in LRI model is the same as LRW. The only difference is that the response variable has interaction terms, which leads to the effect of one variable on the dependent may depend on the value of another variable (Aiken and West, 1991). For example, PREGSM0a (the mean monthly precipitation) and DEMSRE0a (digital elevation model) in LRW are negative value, the former one is significant and the latter one not. This means that more precipitation or higher elevation lead to lower pH in varying degrees of influence. However, in LRI the interaction term DEMSRE0a * PREGSM0a with positive coefficient, when DEMSRE0a is lower, higher PREGSM0a decreases pH and lower PREGSM0a increases pH, which means that the effect of precipitation on pH depends on elevation. One more example, SLPSRT0a (slope) has not significant positive effect on pH and TNMMOD0a (earth surface night time temperature) is also not selected as a predictor variable in LRI. However, the interaction term with these two variables SLPSRT0a * TNMMOD0a shows a significant positive effect on pH. Apparently, the influence of mean monthly night temperature on PH depends on the degree of inclination.

4.2.3 Regression tree

The predict results in regression tree is discrete, which look like the classification result, but the actually data are continuous. The resulting model used only a few important predictor variables, which are PREGSM0a, TDMMOD0a, TDHMOD0a and DEMSRE0a. From the finial tree (Figure 16), we can find the higher PREGRM0a (precipitation), lower TDHMMOD0a (maximum temperature in day time) or TDMMOD0a (mean temperature in day time) all can lead to the lower pH individually. However, the influence of DEMSRE0a (DEM) was depending on the PREGRM0a. When the PREGSM0a was larger than 687.9, the higher DEMSRE0a lead to lower pH, while when the PREGSM0a was lower than 687.9 and higher than 338.7, the higher DEM contributed to higher pH. These four predictor variables are also important for the other models, except DEMSRE0a in LRW model. In addition, PREGSM0a, TDMMOD0a had the same direction effects for pH in other models, but TDHMOD0a had different direction influence compared to the other models. Moreover, PREGSM0a and DEMSRE0a in RT model were explained as similar as LRI model. In LRI model, we only know PREGSM0a and DEMSRE0a have interaction influence to pH, but in Regression tree model, we find the effect of DEMSRE0a to pH was rely on the level of PREGSM0a.

4.2.4 Random forest

The calculation time of RF model was much longer than for the other models. This is because firstly, the optimal model is found by numerical search, so the model has to be built again and again. Here we built 75 models to get the most appropriate one. Secondly, the RF model does

not need to consider the over fitting problem, the tree could be grown as deep as possible, which gives more terminal nodes (set the "nz" bigger) compared to RT model, and a deeper tree leads to more calculation time. However, more terminal nodes would not always give the best prediction, see the "nz" column of Tables 5 and 6. Thirdly, data are randomly split to many sections to grow many trees for one RF model.

In the RF model, the variables of importance are evaluated by cross-validation. The results indicate that most variables were useful for the RF model. The most important variables are DEMSRE0a, EVSMOD0a, EVMMOD0a, G04ESA0a, PREGSM0a, SLPSRE0a and temperature variables (T*******)('*' means anyone number or letter), remove any of them may increase the error more than 50%.

4.2.5 Comparison of model interim results

The four developed models used different variables. The linear regression models select the most useful variables by using stepwise regression based on AIC criteria. RT and RF model select predictors using 5-fold cross-validation. Compared with RT models, combining the information of Figure 17, RF models use much more variables than the RT model. However, the most important variables in four models are similar, that are PREGSM0a, EVMMOD0a, DEMSRE0a and temperature variables (T*******). Whereas, most Organism variables (G**ESA0a) are less important, that may cause by these variables obtain 0 value in wide area (check in Appendix 1), which means if the predictive variables in many locations have 0 value, they may would be less useful to develop and predict models.

All figures and tables in Section 3.3 show that the RF models have the best performance for all three soil properties, such as having the highest correlation between the observed value and predicted value at observation locations, the lowest prediction error and highest explained variance. Whereas the LRW have the worst results, especially in Figure 18 and Figure 21, the extremely low R^2 and RMSE of clay content indicates that the response variable has no linear relationship with predictor variables. In soil pH and clay content, the LRI model has somewhat better results than the RT model and in SOC, it is the opposite.

4.3 Predicted maps

Form the results of Section 3.5, the box plots, tables and maps show that the predictions produced by linear regression models always have a larger data range than the RT and RF models, which show that linear regression models produce more extreme values than the RT and RF models. That may because in linear regression models the response variable is affected by the sum of each used predictive variables, but in RT and RF models the response variable is decided by the mean of each splits data. That is why the linear regression models get wider range predicted values and the predicted values in RT and RF models are closer to the mean or median value.

Predicted maps from RT models produce the least extreme value. In addition, the maps produced by RT produce smoother maps that have less small-scale spatial variation, with fewer details than the maps derived using the other models. This is because RT uses the fewest predictor variables. In the pH RT model, the important predictor variables are PREGSM0a, TDMMOD0a, TDHMOD0a and DEMSRE0a, which produce 9 leaves. In the linear regression models and RF model there are many more predictors. In all of the pH models, the variables PREGSM0a, TDMMOD0a, TDHMOD0a and DEMSRE0a that stand for Relief and Climate in Jenny's model are important, although different models apply them in different ways. Therefore the main differences of each predict models predictors are Organism, so the difference of maps is more contributed by Organism variables.

5 Conclusions

The aim was to determine which statistical models can be used to characterise the relationship between soil properties and environmental covariates and to assess the performance of these statistical models for SSA by using (1) multiple statistical methods, (2) the African Soil Profile database and (3) generally available gridded covariate layers, and to compare the results of different statistical methods. To do so, the 5 research questions are answered as following:

1 Which statistical methods can be used to model the relationship between soil properties and environmental covariates and how do these statistical methods work?

In this study we applied five statistical models (LRI, LRW, RT, RF and ANN) to explore the potential of digital soil mapping for three soil properties (pH, organic carbon content and clay content) for SSA. Except ANN model, the other statistical models were developed to model the relationship between soil properties and environmental covariates. The ways of statistical methods working were explained in the Section 2.3.

2 Are software implementations in R available for these methods and how can these be used?

The model implementation in R was successful for LRI, LRW, RT and RF models, but failed in the case of ANN models. This is probably because the ins and outs of the ANN package in R were not fully understood, and future research could analyse this more closely and extend the comparison with ANN. We also used R to pre-processes data and to help building and validating models. The details of how the R packages working were introduced in Section 2.5.

3 How can the results of each method be validated?

The results of the four successful statistical models were validated with an independent validation method, which randomly separated the dataset in training and testing datasets. Most models have similar RMSE and R^2 for different splits in training and testing dataset, which indicates that the models are stable. The main exceptions to this were the RT and linear models for clay content.

4 What do the result maps look like and which results are obtained when the methods are applied to soil property prediction in SSA?

The predicted maps were presented in Section 3.5 and discussed in Section 4.3. From these maps, we can find that the low pH and high SOC value occur in the central of SSA, where mainly is tropical forest and grassland. In addition, the maps show the closer to the desert the lower clay content and SOC value. These characters of maps indicate the soil properties were affected by the land cover and climate. The digital soil maps using the RT models show a block-structured area, which look more like classification maps than continuous maps. This is because the RT model uses discrete thresholds to branch at the nodes of the tree. LRW and LRI models obtain more accurate predictions for extreme value whereas RT and RF models obtain the extreme value more concentrate to the mean value. That may because in linear regression models the response variable is affected by the sum of each used predictive variables, but in RT and RF models the response variable is decided by the mean of each splits data, which may be analysed in more detail in future studies.

5 What can we learn from a case study applying the different statistical methods on data from SSA?

In different models, the relationship between response variables and predicted variables are modelled in a different way. The diversity of variable selection and the variables' importance are different as well. In this study, the results shown that soil pH and SOC content have a stronger relationship with environmental covariates than soil clay content, which means that soil pH and SOC content are more sensitive to environmental change. On the other hand, because the clay content largely depends on parent material (geology) and we may not include that factor very well in the predictive variables, the developed models could not show a clear relationship between clay content and predictive variables.

It has to be noticed that RF models clearly outperformed than the other models for all three soil variables in terms of assessment criteria RMSE and R^2, but the performance is still not very high, with 43.9% of R^2 for pH, 40.69% of R^2 for SOC and 40.13% for clay content . The LRW model has the worst fit between response and predictive variables, the R^2 are 23.2% for pH, 20.8% for SOC and only 3.9% for Clay content. The very low R^2 in clay is remarked that the predictor variables do not find linear relationship with clay content. The reason may cause by that there were not enough relative predictive variables to predict clay.

This study showed that digital soil mapping for SSA using the available covariates and the tested statistical models is able to explain a substantial part of the spatial variation in three selected soil properties, although much of the variation remains unexplained and the accuracy of the resulting maps is limited. The quality of the digital soil maps is limited because of the relatively poor density of observations and coarse environmental covariates data. Using geo-statistical interpolation which considers the spatial correlation in the residuals of the models might help to improve the quality of digital soil mapping for SSA.

References

1996. Distribution of soil orders. Office of Agriculture, Global Programs, U.S. Agency for International Development, USA.

A.S.Kauzeni, I.S.K., S.A.Mohamed, J.G.Lyimo, 1993. LAND USE PLANNING AND RESOURCE ASSESSMENT IN

TANZANIA: A CASE STUDY, The Environmental Planning Group

The International Institute for Environment and Development, London.

Aiken, L.S., West, S.G., 1991. Multiple Regression: Testing and Interpreting Interactions. Sage Publications, USA.

Akaike, H., 1974. A new look at the statistical model identification. IEEE Trans, pp. 716-723.

Alvarez, R., Lavado, R.S., 1998. Climate, organic matter and clay content relationships in the Pampa and Chaco soils, Argentina. Geoderma 83(1–2), 127-141.

Barrios, S., Bertinelli, L., Strobl, E., 2006. Climatic change and rural–urban migration: The case of sub-Saharan Africa. Journal of Urban Economics 60(3), 357-371.

Batjes, N.H., 2008. ISRIC-WISE - Global Soil Profile Data, Wageningen.

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

Berberoglu, S., Lloyd, C.D., Atkinson, P.M., Curran, P.J., 2000. The integration of spectral and textural information using neural networks for land cover mapping in the Mediterranean. Computers & Geosciences 26(4), 385-396.

Bishop, T.F.A., McBratney, A.B., Laslett, G.M., 1999. Modelling soil attribute depth functions with equal-area quadratic smoothing splines. Geoderma 91(1–2), 27-45.

Bourennane, H., King, D., Couturier, A., 2000. Comparison of kriging with external drift and simple linear regression for predicting soil horizon thickness with different sample densities. Geoderma 97(3–4), 255-271.

Breiman, L., 2001. Random Forests. Machine Learning 45(1), 5-32.

Breiman., L., Cutler., A., Liaw., A., MatthewWiener., 2012. Package 'randomForest'.

Dictionary, C., 2011. Collins English Dictionary. HarperCollins, Glasgow.

Efroymson， MA, 1960. Multiple regression analysis. In: H. A. and Wilf (Ed.), Mathematical Methods for Digital Computers. Wiley, Ralston.

Endre Dobos, F.C., Tomislav Hengl, Hannes I. Reuter, Gergely Tóth, 2006. Digital Soil Mapping as a support to production of functional maps, Luxemburg.

Eswaran, H., Almaraz, R., Evert, v.d.B., Reich, P., 1996. An Assessment of the Soil Resources of Africa in Relation to Productivity. World Soil Resources, Soil Survey Division, USDA Natural Resources Conservation Service.

FAO, 2012. Soil Organic Carbon Accumulation and Greenhouse Gas Emission Reductions from Conservation Agriculture: A literature review. Integrated Crop Management 16.

Foley, N.K., 1999. Environmental Characteristics of Clays and Clay Mineral Deposits. U.S. Geological Survey

Gazey, C., 2009. Soil acidity needs your attention.

Gershenson, C., 2013. Artificial Neural Networks for Beginners.

Graham Dy, A.N.M.C.D.R., et al., 2002. Ulcer prevention in long-term users of nonsteroidal anti-inflammatory drugs: Results of a double-blind, randomized, multicenter, active- and placebo-controlled study of misoprostol vs lansoprazole. Archives of Internal Medicine 162(2), 169-175.

Grimm, R., Behrens, T., Märker, M., Elsenbeer, H., 2008. Soil organic carbon concentrations and stocks on Barro Colorado Island — Digital soil mapping using Random Forests analysis. Geoderma 146(1–2), 102-113.

Hastie, T., Tibshirani, R., Friedman, J., 2001. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition (Springer Series in Statistics). springer series in statistics, 9. Springer.

Hengl, T., H.I. Reuter, 2013. Worldgrids — a public repository and a WPS for global environmental layers. ISRIC.

Ivry, T., Michal, S., 2013. License Plate Number Recognition Using Artificial Neural Network.

Jenny, H., 1941. FACTORS OF SOIL FORMATION, A System of Quantitative Pedology. General Publishing Company, Canada.

Johnston, A.E., 2004. Soil acidity - resilience and thresholds. CABI Publishing, Wallingford, pp. 35-46.

Jones, M.J., 1973. THE ORGANIC MATTER CONTENT OF THE SAVANNA SOILS OF WEST AFRICA. Journal of Soil Science 24(1), 42-53.

LADA, 2008. Land use systems of the world - Sub-Saharan Africa.

Lake, B., 2000. Understanding Soil pH. New South Wales Acid Soil Action Program.

Lal, R., 2001. Soil degradation by erosion. Land Degradation & Development 12(6), 519-539.

Lal, R., 2004. Soil Carbon Sequestration Impacts on Global Climate Change and Food Security. Science 304(5677), 1623-1627.

Lawrence, S., Giles, L., Tsoi, A.C., 1996. What Size Neural Network Gives Optimal Generalization? Convergence Properties of Backpropagation.

Leenaars, J.G.B., 2012. Africa Soil Profiles Database Version 1.0, ISRIC.

Leo Breiman, J.F., Charles J. Stone, R.A. Olshen, 1984. Classification and Regression Trees. Chapman and Hall/CRC.

Lugo, A., Sanchez, M., Brown, S., 1986. Land use and organic carbon content of some subtropical soils. Plant Soil 96(2), 185-196.

Malone, B., Hengl, T., 2012a. Fits a mass preserving spline.

Malone, B., Hengl, T., 2012b. Global Soil Information Facilities version 0.3-1.

Marion Mertens, Inga Nestler, Huwe, B., 2001. GIS-based regionalization of soil profiles with Classification and Regression Trees(CAER).

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

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

Minasny, B., McBratney, A.B., Bristow, K.L., 1999. Comparison of different approaches to the development of pedotransfer functions for water-retention curves. Geoderma 93(3–4), 225-253.

Oades, J., 1995. Krasnozems - organic-matter. Soil Research 33(1), 43-57.

Odgers, N.P., Libohova, Z., Thompson, J.A., 2012. Equal-area spline functions applied to a legacy soil database to create weighted-means maps of soil organic carbon at a continental scale. Geoderma 189–190(0), 153-163.

Owuor, P.O.O., C. O.; Kamau, D. M.; Wanyoko, J. K., 2012. Effects of long-term fertilizer use on a high-yielding tea clone AHPS15/10: soil pH, mature leaf nitrogen, mature leaf and soil phosphorus and potassium. International Journal of Tea Science (IJTS) 2011/2012 Vol. 8 No. 1 pp. 15-51.

Parikh, S.J.J., B. R., 2012. Soil: The Foundation of Agriculture. Nature Education Knowledge 3(10):2.

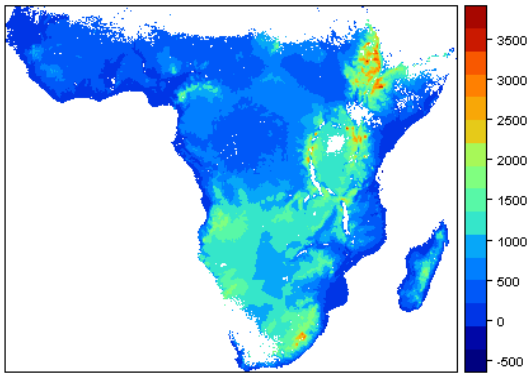
Ponce-Hernandez, R., Marriott, F.H.C., Beckett, P.H.T., 1986. An improved method for reconstructing a soil profile from analyses of a small number of samples. Journal of Soil Science 37(3), 455-467.
Ripley, B., 2013. Package 'rpart'.
Rumelhart, D.E., McClelland, J.L., 1986. Parallel Distributed Processing. The MIT press, Cambridge.
Russell, J.M., AW, 1968. Comparison of different depth weigshting in the numerical analysis of anisotropic soil profile data, Transactions of the 9th International Congress on Soil Science Int.Soil Sci.Soc. and Angus And Robertson Sydney.
Sanchez, P.A., 2002. Soil Fertility and Hunger in Africa. Science 295(5562), 2019-2020.
Sanchez, P.A., Ahamed, S., Carré, F., Hartemink, A.E., Hempel, J., Huising, J., Lagacherie, P., McBratney, A.B., McKenzie, N.J., Mendonça-Santos, M.d.L., Minasny, B., Montanarella, L., Okoth, P., Palm, C.A., Sachs, J.D., Shepherd, K.D., Vågen, T.-G., Vanlauwe, B., Walsh, M.G., Winowiecki, L.A., Zhang, G.-L., 2009. Digital Soil Map of the World. Science 325(5941), 680-681.
Sheets, K.R., Hendrickx, J.M.H., 1995. Noninvasive Soil Water Content Measurement Using Electromagnetic Induction. Water Resour. Res. 31(10), 2401-2409.
Stefan Fritsch, F.G., Marc Suling, 2012. Package ‘neuralnet’.
Terry Therneau , B.A., Brian Ripley 2013. Package ‘rpart’.
Torgo, L.F.R.A., 1999. Inductive Learning of Tree-based Regression Models : Chapter 4 Overfitting Avoidance in Regression Trees, University of Porto.
W. N. Venables, D.M.S., 2013. An Introduction to R， Notes on R: A Programming Environment for Data Analysis and Graphics
Version 3.0.0
Webster, R., Burrough, P.A., 1972. COMPUTER-BASED SOIL MAPPING OF SMALL AREAS FROM SAMPLE DATA. Journal of Soil Science 23(2), 210-221.
Wiesmeier, M., Barthold, F., Blank, B., Kögel-Knabner, I., 2011. Digital mapping of soil organic matter stocks using Random Forest modeling in a semi-arid steppe ecosystem. Plant Soil 340(1), 7-24.
Xu, M., Qi, Y., 2001. Budgets of soil erosion and deposition for sediments and sedimentary organic carbon across the conterminous United States. Global Biogeochem. Cycles 15(3), 687-696.

Appendices

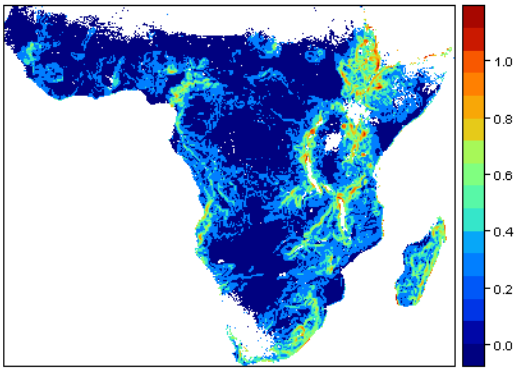
Appendix 1 Predictor variable maps

This appendix contains a list of predictor variables available via the Worldgrids repository (Hengl, 2013 #90). Each layer comes with a separate profile page and following links (Table A 2) were ways to access full metadata for each layer.

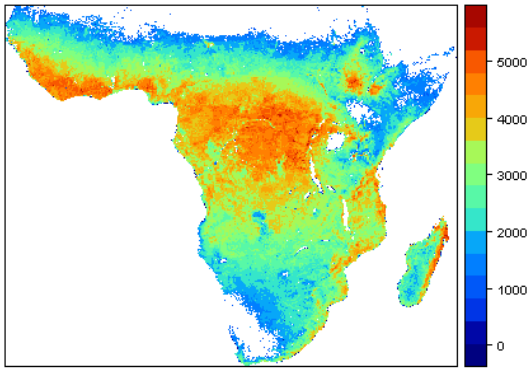
DEMSRE0a



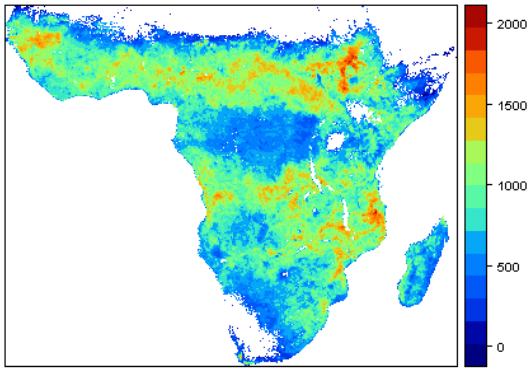
SLPSRT0a



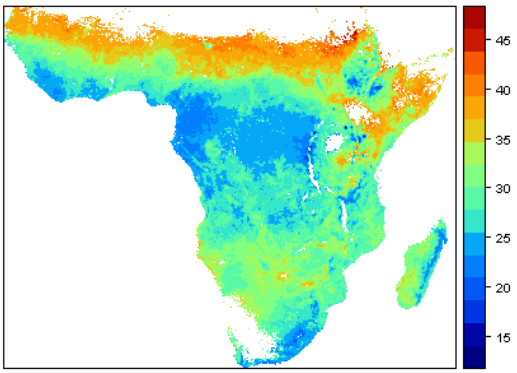
EVMMOD0a



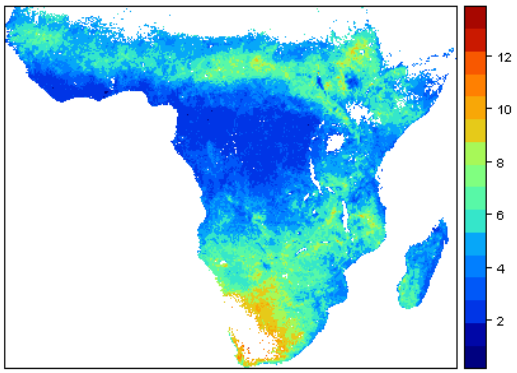
EVSMOD0a



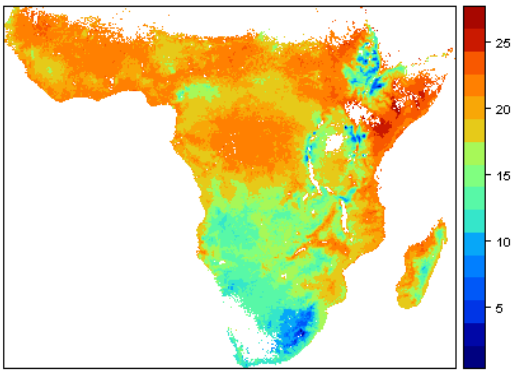
TDMMOD0a



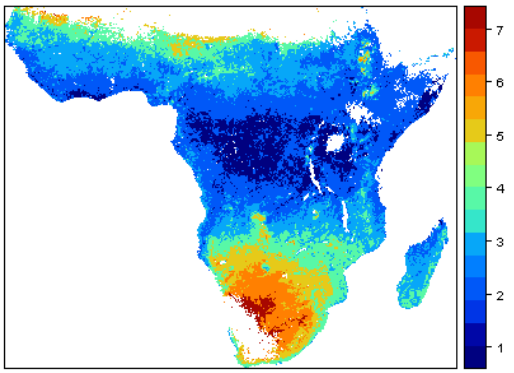
TDSMOD0a



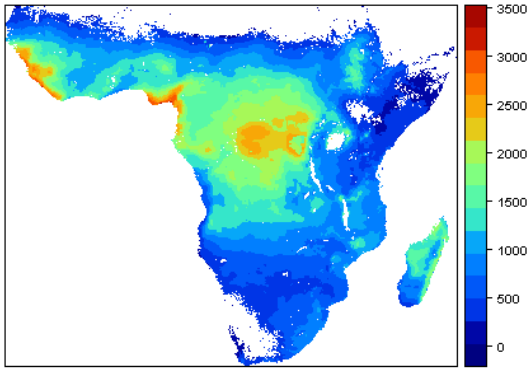
TNMMOD0a



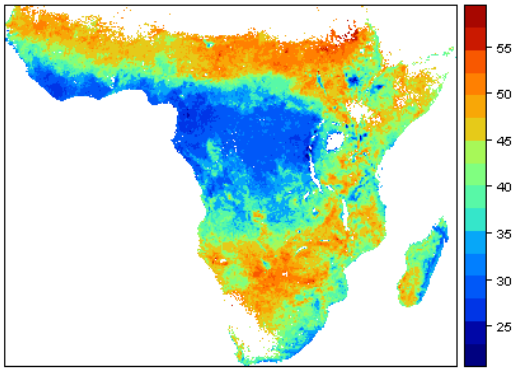
TNSMOD0a



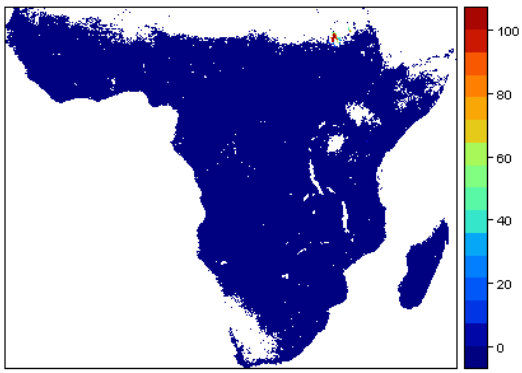
PREGSM0a



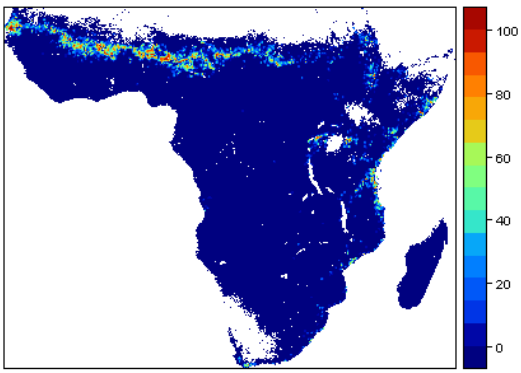
TDHMOD0a



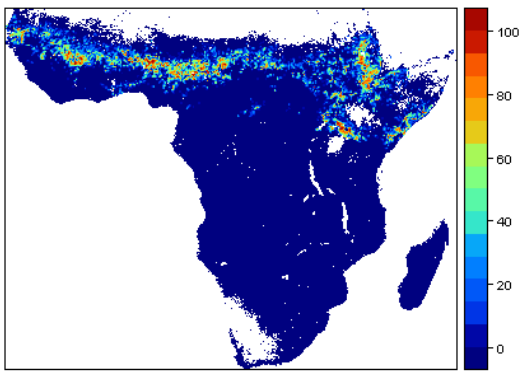
G01ESA0a



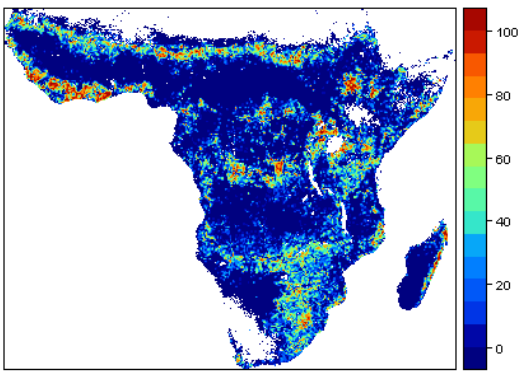
G02ESA0a

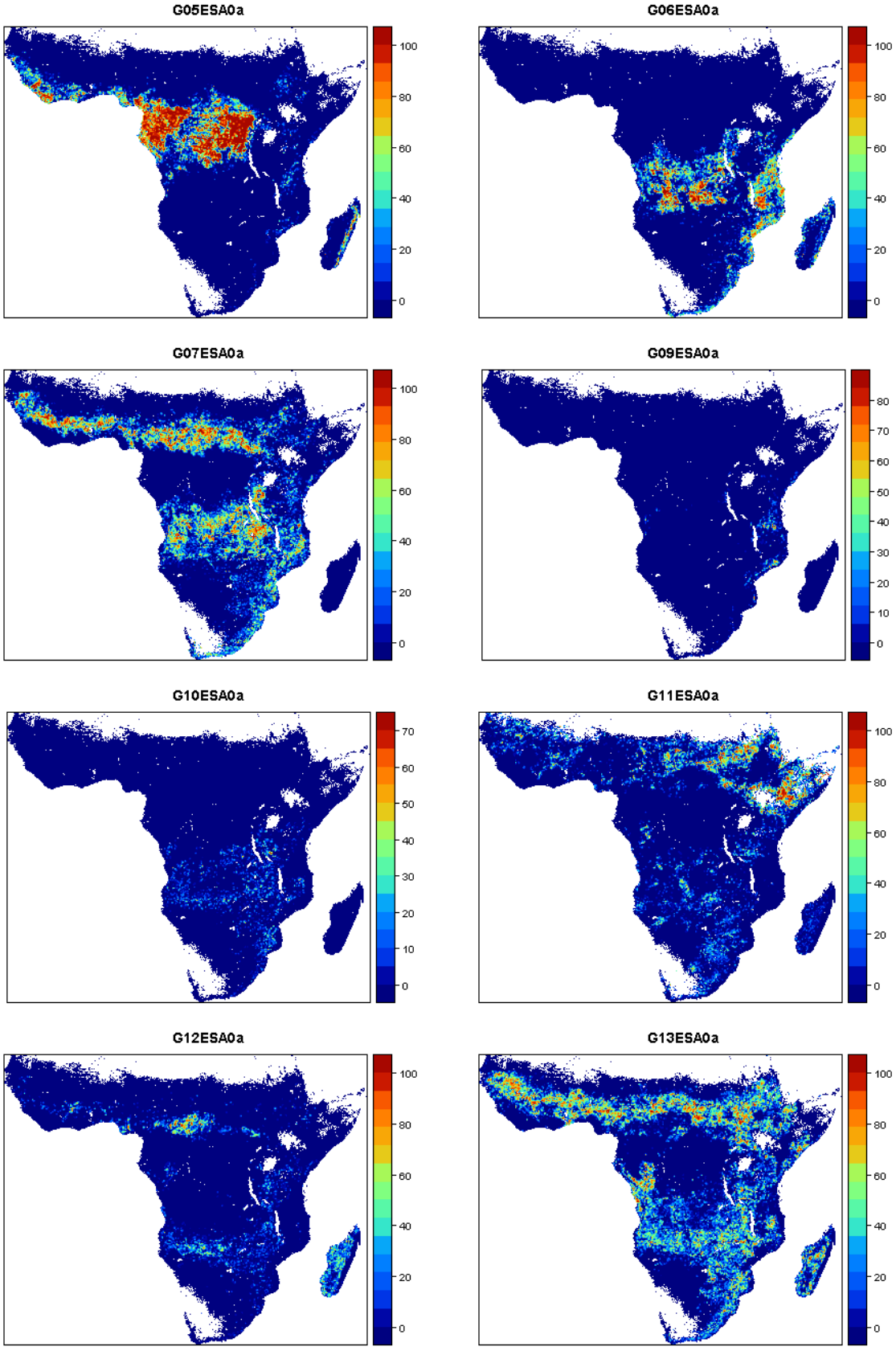


G03ESA0a



G04ESA0a




G05ESA0a
G06ESA0a
G07ESA0a
G09ESA0a
G10ESA0a
G11ESA0a
G12ESA0a
G13ESA0a

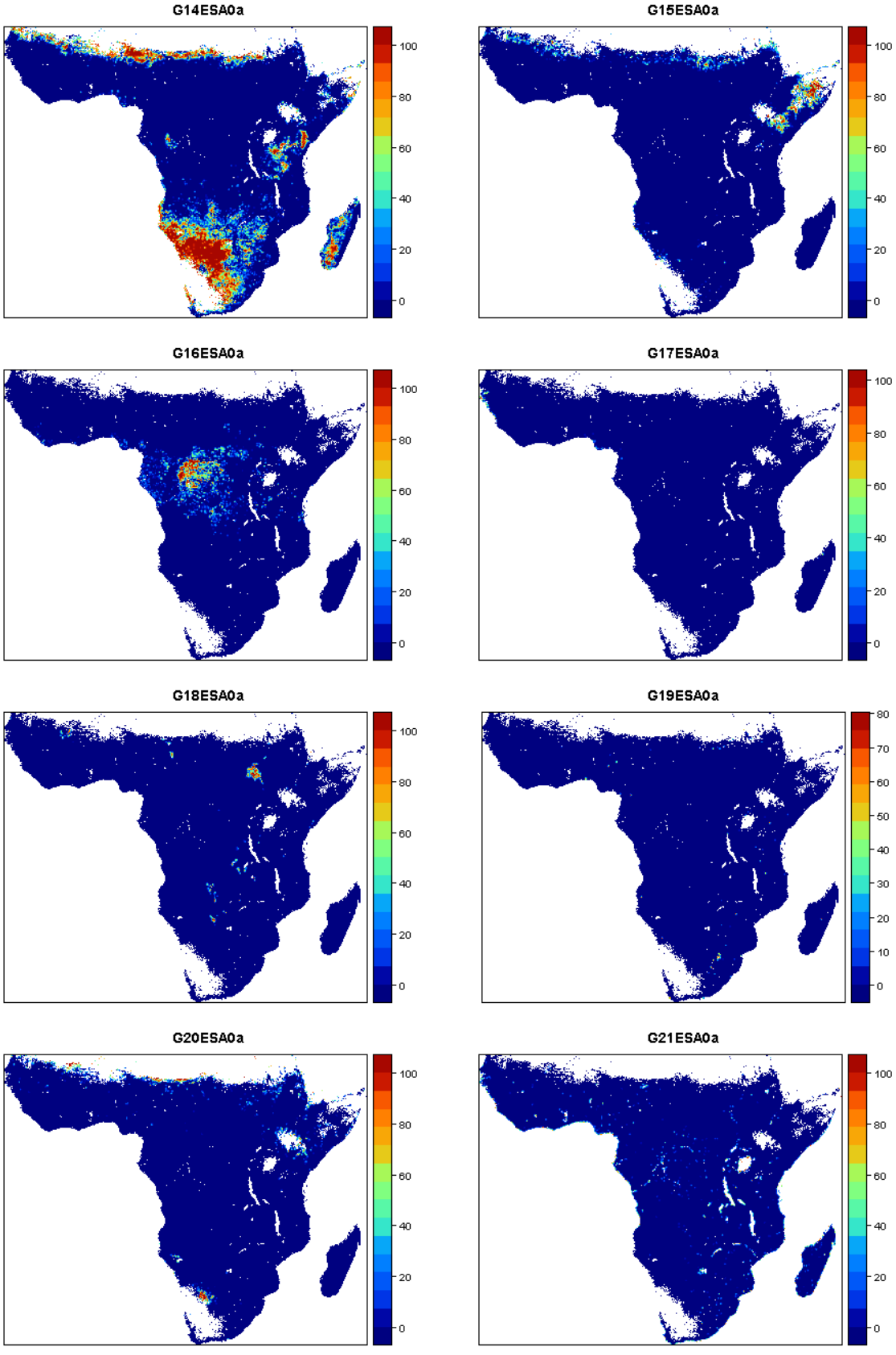

G14ESA0a
G15ESA0a
G16ESA0a
G17ESA0a
G18ESA0a
G19ESA0a
G20ESA0a
G21ESA0a

Table A 1 Links of metadata of predictor variables (Last Accessed on 20th March 2012)

	type	Abbreviated name	Description
1	Climate	PREGSM0a	http://www.worldgrids.org/doku.php?id=wiki:pregsm1
2		TDMMOD0a	http://www.worldgrids.org/doku.php?id=wiki:tdmmod3
3		TDSMOD0a	http://www.worldgrids.org/doku.php?id=wiki:tdsmod3
4		TDHMOD0a	http://www.worldgrids.org/doku.php?id=wiki:tdhmod3
5		TNMMOD0a	http://www.worldgrids.org/doku.php?id=wiki:tnmmod3
6		TNSMOD0a	http://www.worldgrids.org/doku.php?id=wiki:tnsmod3
7	Relief	DEMSRE0a	http://www.worldgrids.org/doku.php?id=wiki:demsre3
8		SLPSRT0a	http://www.worldgrids.org/doku.php?id=wiki:slpsrt3
9	Organisms, vegetation or human activity	EVMMOD0a	http://www.worldgrids.org/doku.php?id=wiki:evmmod3
10		EVSMOD0a	http://www.worldgrids.org/doku.php?id=wiki:evmmod3
11		IFLGRE0a	http://www.worldgrids.org/doku.php?id=wiki:iflgre
12		G01ESA0a	http://www.worldgrids.org/doku.php?id=wiki:g01esa3
13		G02ESA0a	http://www.worldgrids.org/doku.php?id=wiki:g02esa3
14		G00ESA0a	http://www.worldgrids.org/doku.php?id=wiki:g03esa3
15		G04ESA0a	http://www.worldgrids.org/doku.php?id=wiki:g04esa3
16		G05ESA0a	http://www.worldgrids.org/doku.php?id=wiki:g05esa3
17		G06ESA0a	http://www.worldgrids.org/doku.php?id=wiki:g06esa3
18		G07ESA0a	http://www.worldgrids.org/doku.php?id=wiki:g07esa3
19		G09ESA0a	http://www.worldgrids.org/doku.php?id=wiki:g09esa3
20		G10ESA0a	http://www.worldgrids.org/doku.php?id=wiki:g10esa3
21		G11ESA0a	http://www.worldgrids.org/doku.php?id=wiki:g11esa3
22		G12ESA0a	http://www.worldgrids.org/doku.php?id=wiki:g12esa3
23		G13ESA0a	http://www.worldgrids.org/doku.php?id=wiki:g13esa3
24		G14ESA0a	http://www.worldgrids.org/doku.php?id=wiki:g14esa3
25		G15ESA0a	http://www.worldgrids.org/doku.php?id=wiki:g15esa3
26		G16ESA0a	http://www.worldgrids.org/doku.php?id=wiki:g16esa3
27		G17ESA0a	http://www.worldgrids.org/doku.php?id=wiki:g17sa3
28		G18ESA0a	http://www.worldgrids.org/doku.php?id=wiki:g18esa3
29		G19ESA0a	http://www.worldgrids.org/doku.php?id=wiki:g19esa3
30		G20ESA0a	http://www.worldgrids.org/doku.php?id=wiki:g20esa3
31		G21ESA0a	http://www.worldgrids.org/doku.php?id=wiki:g21esa3

Appendix 2 Interim results of soil organic carbon content and clay content in four statistical models

1 Linear regression without interaction

Organic carbon content

The formula is: *log(1 + MO) ~ DEMSRE0a + SLPSRT0a + TDSMOD0a + IFLGRE0a + TNMMOD0a + TDHMOD0a + TNSMOD0a + EVMMOD0a + EVSMOD0a + PREGSM0a + G01ESA0a + G02ESA0a + G03ESA0a + G04ESA0a + G05ESA0a + G06ESA0a + G07ESA0a + G09ESA0a + G10ESA0a + G11ESA0a + G12ESA0a + G13ESA0a + G14ESA0a + G15ESA0a + G16ESA0a + G17ESA0a + G18ESA0a + G19ESA0a + G20ESA0a + G21ESA0a*

Table A2. 1The coefficients, stand error and significant of predictor variables in organic carbon content linear regression without interaction model

	Estimate	Std.	Sign.
(Intercept)	1.49E+01	1.96E+00	***
DEMSRE0a	2.27E-04	3.01E-05	***
SLPSRT0a	5.34E-02	5.55E-03	***
TDSMOD0a	-5.38E-02	1.02E-02	***
IFLGRE0a	-3.18E-01	1.10E-01	**
TNMMOD0a	-3.44E-02	6.10E-03	***
TDHMOD0a	-1.83E-02	2.86E-03	***
TNSMOD0a	-8.88E-02	9.85E-03	***
EVMMOD0a	7.25E-05	1.88E-05	***
EVSMOD0a	3.27E-04	3.87E-05	***
PREGSM0a	3.06E-04	2.58E-05	***
G01ESA0a	-1.14E-01	1.95E-02	***
G02ESA0a	-1.15E-01	1.94E-02	***
G03ESA0a	-1.15E-01	1.94E-02	***
G04ESA0a	-1.17E-01	1.94E-02	***
G05ESA0a	-1.14E-01	1.94E-02	***
G06ESA0a	-1.16E-01	1.94E-02	***
G07ESA0a	-1.17E-01	1.94E-02	***
G09ESA0a	-1.30E-01	1.98E-02	***
G10ESA0a	-1.10E-01	1.96E-02	***
G11ESA0a	-1.13E-01	1.94E-02	***
G12ESA0a	-1.14E-01	1.95E-02	***
G13ESA0a	-1.17E-01	1.94E-02	***
G14ESA0a	-1.17E-01	1.94E-02	***
G15ESA0a	-1.20E-01	1.95E-02	***
G16ESA0a	-1.19E-01	1.95E-02	***
G17ESA0a	-1.11E-01	1.95E-02	***
G18ESA0a	-1.17E-01	1.94E-02	***
G19ESA0a	-1.14E-01	1.95E-02	***
G20ESA0a	-1.13E-01	1.94E-02	***
G21ESA0a	-1.09E-01	1.95E-02	***

Clay content

The formula is : *Log(1+MClay) ~ DEMSRE0a + SLPSRT0a + TDMMOD0a + TDSMOD0a + IFLGRE0a + TNMMOD0a + TDHMOD0a + TNSMOD0a + EVMMOD0a + EVSMOD0a + PREGSM0a + G01ESA0a + G02ESA0a + G03ESA0a + G04ESA0a + G05ESA0a + G06ESA0a + G07ESA0a + G08ESA0a + G09ESA0a + G10ESA0a + G11ESA0a + G12ESA0a + G13ESA0a + G14ESA0a + G15ESA0a + G16ESA0a + G17ESA0a + G18ESA0a + G19ESA0a + G20ESA0a + G21ESA0a + G22ESA0a*

Table A2. 2 The coefficients, stand error and significant of predictor variables in organic carbon content linear regression with interaction model

	Estimate	Std. Error	significant
(Intercept)	1.51E+01	2.25E+00	***
DEMSRE0a	4.96E-04	2.82E-05	***
SLPSRT0a	7.80E-02	6.65E-03	***
TDMMOD0a	-2.14E-02	6.76E-03	**
TDSMOD0a	-7.29E-02	1.19E-02	***
TNMMOD0a	3.09E-02	6.44E-03	***
TDHMOD0a	1.50E-02	4.76E-03	**
EVSMOD0a	1.16E-04	3.87E-05	**
G01ESA0a	-1.17E-01	2.26E-02	***
G02ESA0a	-1.31E-01	2.24E-02	***
G03ESA0a	-1.27E-01	2.24E-02	***
G04ESA0a	-1.30E-01	2.24E-02	***
G05ESA0a	-1.27E-01	2.24E-02	***
G06ESA0a	-1.29E-01	2.24E-02	***
G07ESA0a	-1.30E-01	2.24E-02	***
G09ESA0a	-1.34E-01	2.29E-02	***
G10ESA0a	-1.26E-01	2.26E-02	***
G11ESA0a	-1.23E-01	2.24E-02	***
G12ESA0a	-1.31E-01	2.25E-02	***
G13ESA0a	-1.32E-01	2.24E-02	***
G14ESA0a	-1.32E-01	2.24E-02	***
G15ESA0a	-1.32E-01	2.25E-02	***
G16ESA0a	-1.29E-01	2.25E-02	***
G17ESA0a	-1.23E-01	2.25E-02	***
G18ESA0a	-1.27E-01	2.25E-02	***
G19ESA0a	-1.30E-01	2.25E-02	***
G20ESA0a	-1.19E-01	2.24E-02	***
G21ESA0a	-1.24E-01	2.25E-02	***

2 Linear regression with interaction

Organic carbon content

The formula is:

log(1+MO) ~ DEMSRE0a(SLPSRT0a+EVSMOD0a+EVMMOD0a+TDHMOD0a+TDMMOD0a+TDSMOD0a+TNMMOD0a+TNSMOD0a+PREGSM0a+IFLGRE0a)+SLPSRT0a*(EVSMOD0a+EVMMOD0a+TDHMOD0a+TDMMOD0a+TDSMOD0a+TNMMOD0a+TNSMOD0a+PREGSM0a+IFLGRE0a)+EVSMOD0a*(EVMMOD0a+TDHMOD0a+TDMMOD0a+TDSMOD0a+TNMMOD0a+TNSMOD0a+PREGSM0a+IFLGRE0a)+EVMMOD0a*(TDHMOD0a+TDMMOD0a+TDSMOD0a+TNMMOD0a+TNSMOD0a+PREGSM0a+IFLGRE0a)+TDHMOD0a*(TDMMOD0a+TDSMOD0a+TNMMOD0a+TNSMOD0a+PREGSM0a+IFLGRE0a)+TDMMOD0a*(TDSMOD0a+TNMMOD0a+TNSMOD0a+PREGSM0a+IFLGRE0a)+TDSMOD0a*(TNMMOD0a+TNSMOD0a+PREGSM0a+IFLGRE0a)+TNMMOD0a*(TNSMOD0a+PREGSM0a+IFLGRE0a)+TNSMOD0a*(IFLGRE0a+PREGSM0a)+IFLGRE0a*PREGSM0a +G01ESA0a + G02ESA0a + G03ESA0a + G04ESA0a + G05ESA0a + G06ESA0a + G07ESA0a + G09ESA0a + G10ESA0a + G11ESA0a + G12ESA0a + G13ESA0a + G14ESA0a + G15ESA0a + G16ESA0a + G17ESA0a + G18ESA0a + G19ESA0a + G20ESA0a + G21ESA0a*

Table A2. 3 The coefficients, stand error and significant of predictor variables in organic carbon content linear regression with interaction model

	Estimate	Std.	Sign.	Estimate	Std.	Sign.	
(Intercept)	2.18E+01	2.51E+00	***	G19ESA0a	-1.18E-01	1.91E-02	***
DEMSRE0a	-1.24E-03	3.94E-04	**	G20ESA0a	-1.18E-01	1.90E-02	***
SLPSRT0a	-2.98E-01	7.80E-02	***	G21ESA0a	-1.16E-01	1.91E-02	***
EVSMOD0a	6.65E-04	4.35E-04		DEMSRE0a:EVSMOD0a	-2.52E-07	1.09E-07	*
EVMMOD0a	-4.24E-04	1.97E-04	*	DEMSRE0a:EVMMOD0a	1.61E-07	4.98E-08	**
TDHMOD0a	-3.20E-02	2.34E-02		DEMSRE0a:TDMMOD0a	7.36E-05	1.02E-05	***
TDMMOD0a	-2.61E-01	3.91E-02	***	DEMSRE0a:TDSMOD0a	-4.66E-05	1.90E-05	*
TDSMOD0a	-3.62E-01	1.07E-01	***	DEMSRE0a:TNMMOD0a	-5.03E-05	6.54E-06	***
TNMMOD0a	-9.78E-02	7.78E-02		DEMSRE0a:IFLGRE0a	1.02E-03	5.54E-04	.
TNSMOD0a	6.22E-02	1.00E-01		SLPSRT0a:EVMMOD0a	4.95E-05	1.03E-05	***
PREGSM0a	1.03E-03	1.90E-04	***	SLPSRT0a:TDHMOD0a	-1.35E-02	2.76E-03	***
IFLGRE0a	-5.04E+00	2.25E+00	*	SLPSRT0a:TDMMOD0a	1.86E-02	3.44E-03	***
G01ESA0a	-1.20E-01	1.91E-02	***	SLPSRT0a:TDSMOD0a	2.89E-02	7.62E-03	***
G02ESA0a	-1.19E-01	1.90E-02	***	SLPSRT0a:TNSMOD0a	9.92E-03	5.66E-03	.
G03ESA0a	-1.18E-01	1.90E-02	***	EVSMOD0a:TDHMOD0a	2.16E-05	5.81E-06	***
G04ESA0a	-1.20E-01	1.90E-02	***	EVSMOD0a:TNMMOD0a	-4.03E-05	1.95E-05	*
G05ESA0a	-1.19E-01	1.90E-02	***	EVSMOD0a:TNSMOD0a	-1.17E-04	3.67E-05	**
G06ESA0a	-1.20E-01	1.90E-02	***	EVMMOD0a:TNMMOD0a	1.49E-05	9.17E-06	
G07ESA0a	-1.20E-01	1.90E-02	***	TDHMOD0a:TDMMOD0a	1.23E-03	6.22E-04	*
G09ESA0a	-1.29E-01	1.94E-02	***	TDHMOD0a:TNSMOD0a	-4.68E-03	3.28E-03	
G10ESA0a	-1.14E-01	1.92E-02	***	TDMMOD0a:TDSMOD0a	1.27E-02	2.86E-03	***
G11ESA0a	-1.17E-01	1.90E-02	***	TDMMOD0a:TNMMOD0a	4.11E-03	1.75E-03	*
G12ESA0a	-1.18E-01	1.91E-02	***	TDMMOD0a:TNSMOD0a	-1.22E-02	4.69E-03	**
G13ESA0a	-1.20E-01	1.90E-02	***	TDSMOD0a:TNMMOD0a	-6.71E-03	3.76E-03	.
G14ESA0a	-1.20E-01	1.90E-02	***	TDSMOD0a:TNSMOD0a	1.41E-02	8.16E-03	.
G15ESA0a	-1.23E-01	1.91E-02	***	TNMMOD0a:TNSMOD0a	1.92E-02	3.17E-03	***
G16ESA0a	-1.20E-01	1.91E-02	***	TNMMOD0a:PREGSM0a	-5.88E-05	8.74E-06	***
G17ESA0a	-1.17E-01	1.91E-02	***	TNMMOD0a:IFLGRE0a	2.23E-01	9.96E-02	*
G18ESA0a	-1.21E-01	1.90E-02	***	TNSMOD0a:PREGSM0a	1.33E-04	2.65E-05	***

Clay content

The formula is：

log(1+MClay) ~ DEMSRE0a(SLPSRT0a+EVSMOD0a+EVMMOD0a+TDHMOD0a+TDMMOD0a+TDSMOD0a+TNMMOD0a+TNSMOD0a+PREGSM0a+IFLGRE0a)+SLPSRT0a*(EVSMOD0a+EVMMOD0a+TDHMOD0a+TDMMOD0a+TDSMOD0a+TNMMOD0a+TNSMOD0a+PREGSM0a+IFLGRE0a)+EVSMOD0a*(EVMMOD0a+TDHMOD0a+TDMMOD0a+TDSMOD0a+TNMMOD0a+TNSMOD0a+PREGSM0a+IFLGRE0a)+EVMMOD0a*(TDHMOD0a+TDMMOD0a+TDSMOD0a+TNMMOD0a+TNSMOD0a+PREGSM0a+IFLGRE0a)+TDHMOD0a*(TDMMOD0a+TDSMOD0a+TNMMOD0a+TNSMOD0a+PREGSM0a+IFLGRE0a)+TDMMOD0a*(TDSMOD0a+TNMMOD0a+TNSMOD0a+PREGSM0a+IFLGRE0a)+TDSMOD0a*(TNMMOD0a+TNSMOD0a+PREGSM0a+IFLGRE0a)+TNMMOD0a*(TNSMOD0a+PREGSM0a+IFLGRE0a)+TNSMOD0a*(IFLGRE0a+PREGSM0a)+IFLGRE0a*PREGSM0a +G01ESA0a + G02ESA0a + G03ESA0a + G04ESA0a + G05ESA0a + G06ESA0a + G07ESA0a + G09ESA0a + G10ESA0a + G11ESA0a + G12ESA0a + G13ESA0a + G14ESA0a + G15ESA0a + G16ESA0a + G17ESA0a + G18ESA0a + G19ESA0a + G20ESA0a + G21ESA0a*

Table A2. 4 The coefficients, stand error and significant of predictor variables in clay content linear regression with interaction model

Covariates	Estimate	Std. Error	Sig.	interaction covariates	Estimate	Std. rror	Sig.
(Intercept)	2.33E+01	2.69E+00	***	DEMSRE0a:EVMMOD0a	1.23E-07	4.91E-08	*
DEMSRE0a	-1.90E-03	3.14E-04	***	DEMSRE0a:TDHMOD0a	-2.69E-05	8.09E-06	***
SLPSRT0a	-3.78E-01	9.78E-02	***	DEMSRE0a:TDMMOD0a	1.29E-04	1.33E-05	***
EVSMOD0a	1.31E-03	6.48E-04	*	DEMSRE0a:TNMMOD0a	-5.01E-05	7.13E-06	***
EVMMOD0a	1.95E-04	2.81E-04		DEMSRE0a:TNSMOD0a	-7.13E-05	1.81E-05	***
TDHMOD0a	-1.43E-01	4.05E-02	***	DEMSRE0a:PREGSM0a	2.97E-07	5.87E-08	***
TDMMOD0a	-3.23E-01	5.37E-02	***	DEMSRE0a:IFLGRE0a	3.13E-03	1.15E-03	**
TDSMOD0a	9.37E-02	1.38E-01		SLPSRT0a:EVSMOD0a	6.43E-05	2.83E-05	*
TNMMOD0a	6.67E-02	2.59E-02	*	SLPSRT0a:EVMMOD0a	3.51E-05	1.34E-05	**
TNSMOD0a	9.29E-01	1.50E-01	***	SLPSRT0a:TDHMOD0a	-8.51E-03	3.58E-03	*
PREGSM0a	-1.30E-03	2.32E-04	***	SLPSRT0a:TDMMOD0a	1.55E-02	4.33E-03	***
IFLGRE0a	-1.29E+01	5.06E+00	*	SLPSRT0a:TDSMOD0a	1.89E-02	8.87E-03	*
G01ESA0a	-1.32E-01	2.21E-02	***	SLPSRT0a:TNSMOD0a	2.60E-02	7.12E-03	***
G02ESA0a	-1.42E-01	2.19E-02	***	EVSMOD0a:EVMMOD0a	-3.25E-07	7.21E-08	***
G03ESA0a	-1.38E-01	2.19E-02	***	EVSMOD0a:TDHMOD0a	-9.19E-05	1.89E-05	***
G04ESA0a	-1.41E-01	2.19E-02	***	EVSMOD0a:TDMMOD0a	1.06E-04	2.83E-05	***
G05ESA0a	-1.41E-01	2.19E-02	***	EVSMOD0a:TDSMOD0a	1.05E-04	5.16E-05	*
G06ESA0a	-1.40E-01	2.19E-02	***	EVSMOD0a:TNMMOD0a	-2.81E-05	1.85E-05	
G07ESA0a	-1.40E-01	2.19E-02	***	EVSMOD0a:TNSMOD0a	1.92E-04	4.66E-05	***
G09ESA0a	-1.47E-01	2.25E-02	***	EVSMOD0a:PREGSM0a	-1.82E-07	1.28E-07	
G10ESA0a	-1.37E-01	2.21E-02	***	EVSMOD0a:IFLGRE0a	2.54E-03	9.67E-04	**
G11ESA0a	-1.35E-01	2.19E-02	***	EVMMOD0a:TDHMOD0a	4.40E-05	8.33E-06	***
G12ESA0a	-1.42E-01	2.20E-02	***	EVMMOD0a:TDMMOD0a	-5.60E-05	1.14E-05	***
G13ESA0a	-1.42E-01	2.19E-02	***	EVMMOD0a:TDSMOD0a	-7.62E-05	2.39E-05	**
G14ESA0a	-1.42E-01	2.19E-02	***	EVMMOD0a:TNMMOD0a	1.43E-05	9.25E-06	
G15ESA0a	-1.42E-01	2.20E-02	***	EVMMOD0a:TNSMOD0a	-1.31E-04	2.34E-05	***
G16ESA0a	-1.43E-01	2.20E-02	***	EVMMOD0a:PREGSM0a	2.05E-07	4.86E-08	***
G17ESA0a	-1.33E-01	2.20E-02	***	TDHMOD0a:TDMMOD0a	6.23E-03	8.33E-04	***
G18ESA0a	-1.39E-01	2.20E-02	***	TDHMOD0a:TDSMOD0a	4.91E-03	2.89E-03	.
G19ESA0a	-1.41E-01	2.20E-02	***	TDHMOD0a:TNSMOD0a	-1.82E-02	2.63E-03	***
G20ESA0a	-1.32E-01	2.20E-02	***	TDMMOD0a:TDSMOD0a	-1.13E-02	3.86E-03	**
G21ESA0a	-1.37E-01	2.20E-02	***	TDSMOD0a:PREGSM0a	7.18E-05	2.93E-05	*
				TNMMOD0a:IFLGRE0a	4.70E-01	1.99E-01	*
				TNSMOD0a:PREGSM0a	9.60E-05	3.53E-05	**

3 regression tree

Soil organic carbon content

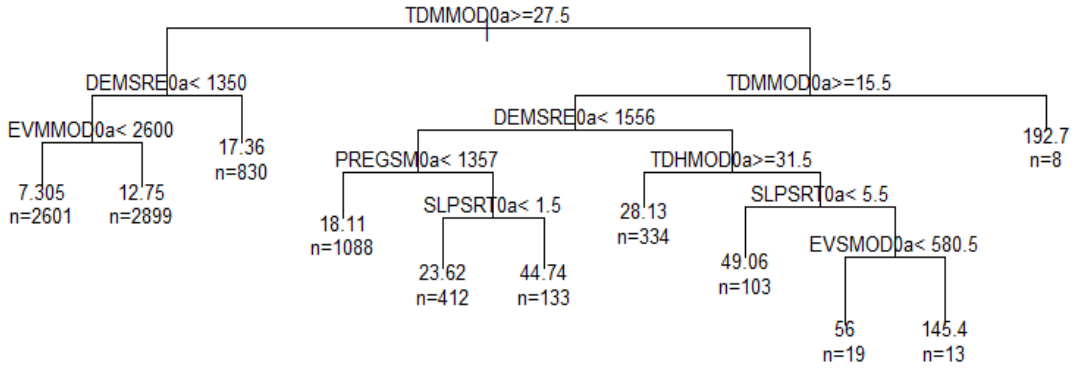


Figure A2. 5 Soil organic carbon content tree pruned tree by cp = 0.01

There are 7 important variables, and the deepest tree is 6.

Clay content

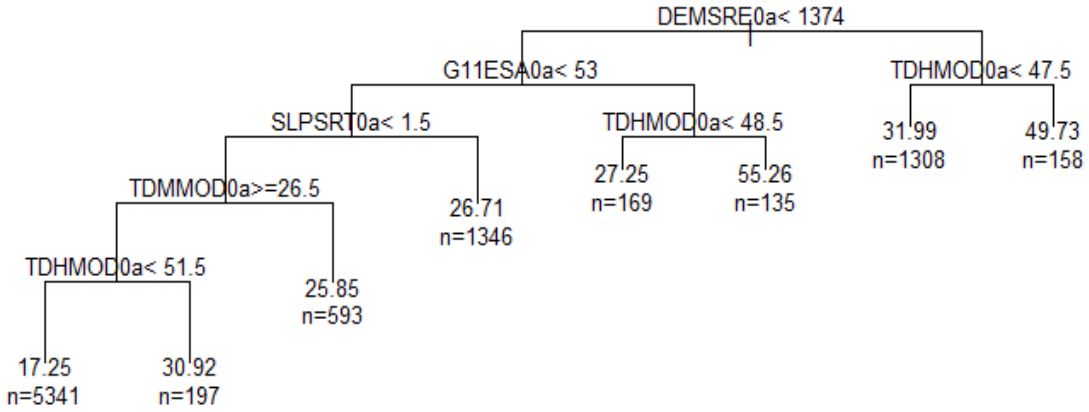


Figure A2. 6 Clay content tree pruned tree by cp = 0.01

There are 5 important variables, and the deepest tree is 5.

4 random forest

Organic carbon content

Table A2. 5 R2 of organic carbon content in random forest

Ntree / Nz	500			750			1000			1250		
	mtry											
	3	4	5	3	4	5	3	4	5	3	4	5
10	0.4001	0.3998	0.3980	0.4019	0.4006	0.3983	0.4014	0.4001	0.3985	0.4016	0.4004	0.3991
15	0.4002	0.4050	0.4059	0.4018	0.4041	0.4069	0.4037	0.4051	0.4065	0.4034	0.4052	0.4048
20	0.4025	0.4039	0.4054	0.4024	0.4052	0.4051	0.4023	0.4056	0.4063	0.4013	0.4055	0.4059
25	0.3978	0.4042	0.4068	0.3965	0.4051	0.4067	0.3964	0.4056	0.4064	0.3957	0.4048	0.4059
30	0.3945	0.4023	0.4004	0.3944	0.4024	0.4011	0.3937	0.4011	0.4023	0.3937	04009	0.4024

Table A2. 6 RMSE of organic carbon content in random forest

Ntree / Nz	500			750			1000			1250		
	mtry											
	3	4	5	3	4	5	3	4	5	3	4	5
10	14.830	14.835	14.856	14.808	14.825	14.853	14.815	14.831	14.851	14.812	14.828	14.843
15	14.829	14.770	14.759	14.810	14.781	14.746	14.786	14.769	14.752	14.790	14.768	14.772
20	14.801	14.784	14.765	14.802	14.768	14.768	14.804	14.763	14.754	14.815	14.763	14.759
25	14.859	14.780	14.748	14.845	14.769	14.749	14.876	14.762	14.752	14.885	14.773	14.759
30	14.899	14.804	14.827	14.901	14.802	14.819	14.910	14.818	14.803	14.909	14.820	14.802

According to the Table A2. 9 and Table A2. 10, the highlight number are the highest $R^{2:}$ 40.69%, and lowest RMSE: 14.746, which indicates the optimal model with combination of mtry=5, nodesize=15 and ntree =750.

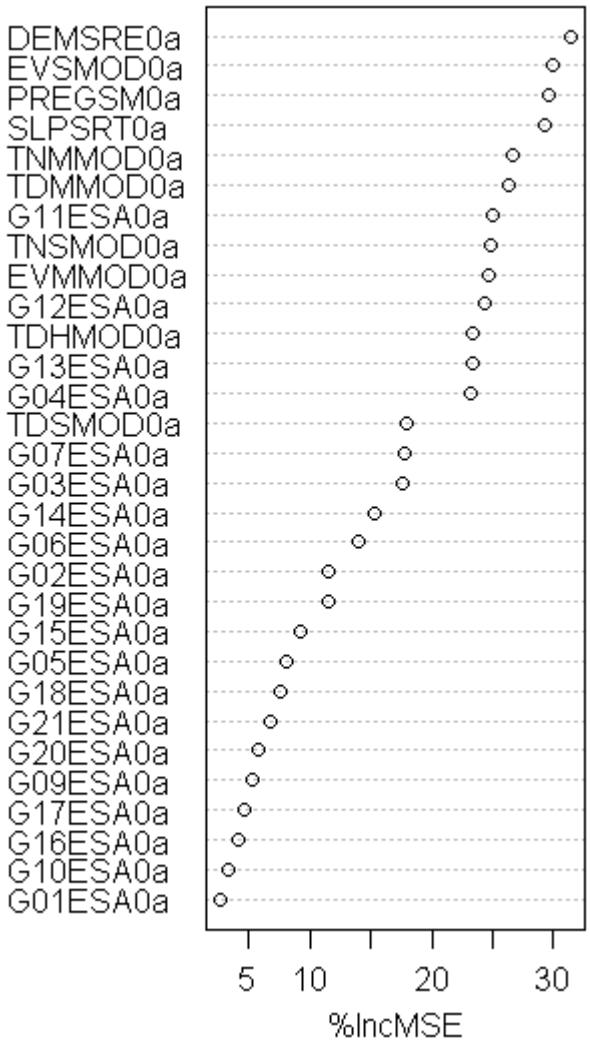


Figure A2. 7 Importance of predictor variables in random forest model for soil organic carbon content

Clay Content

Table A2. 7 R2 of Clay content in random forest

Ntree / Nz	500			750			1000			1250		
	mtry											
	3	4	5	3	4	5	3	4	5	3	4	5
10	0.3890	0.3999	0.4004	0.3892	0.4004	0.4004	0.3896	0.4004	0.4006	0.3898	0.4007	0.4006
15	0.3843	0.3991	0.3997	0.3843	0.3988	0.4002	0.3846	0.3992	0.4003	0.3848	0.3989	0.4002
20	0.3813	0.3929	0.3949	0.3814	0.3931	0.3956	0.3821	0.3932	0.3956	0.3823	0.3934	0.3955
25	0.3761	0.3853	0.3886	0.3763	0.3856	0.3891	0.3765	0.3856	0.3891	0.3765	0.3859	0.3891
30	0.3790	0.3771	0.3811	0.3691	0.3772	0.3809	0.3697	0.3773	0.3812	0.3699	0.3773	0.3812

Table A2. 8 RMSE of clay content in random forest

Ntree / Nz	500			750			1000			1250		
	mtry											
	3	4	5	3	4	5	3	4	5	3	4	5
10	14.791	14.658	14.652	14.789	14.653	14.653	14.783	14.652	14.650	14.781	14.649	14.650
15	14.848	14.668	14.661	14.847	14.671	14.655	14.844	14.667	14.654	14.842	14.671	14.655
20	14.884	14.744	14.720	14.882	14.741	14.711	14.875	14.740	14.710	14.782	14.738	14.713
25	14.946	14.836	14.796	14.944	14.832	14.790	14.941	14.832	14.790	14.941	14.829	14.790
30	15.031	14.934	14.887	15.030	14.934	14.889	15.023	14.932	14.885	15.020	14.932	14.885

According to Table A2. 11 and Table A2. 12, the highlight number are the highest $R^{2:}$ 40.07%, and lowest RMSE: 14.649, which indicates the optimal model with combination of mtry=4, nodesize=10 and ntree =1250.

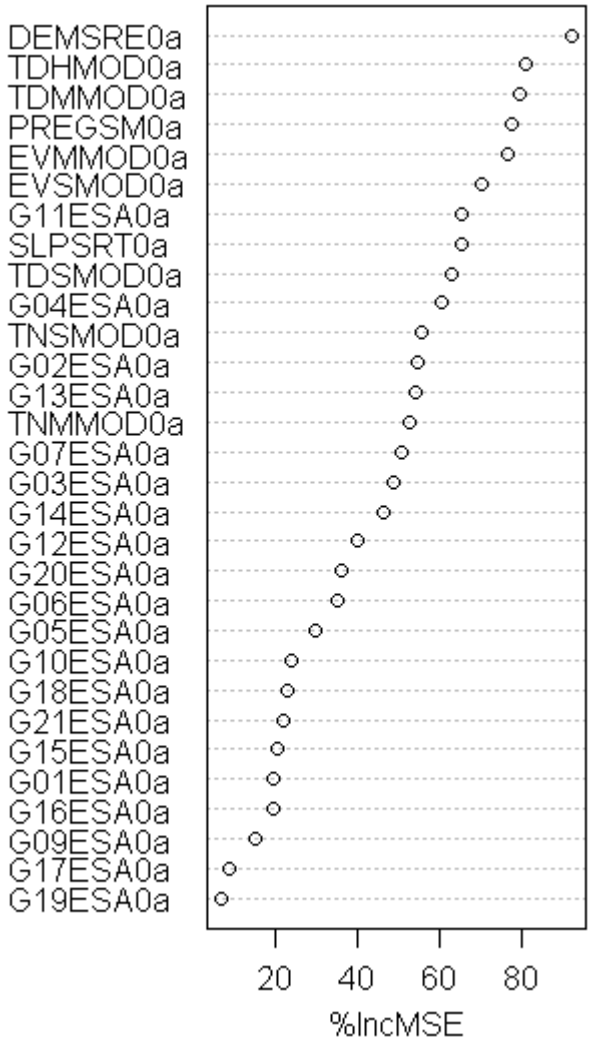


Figure A2. 8 Importance of predictor variables in random forest model for soil clay content

Appendix 3 R scripts

1 data pre-processing (mass-preserved-spline)

```
library(GSIF)
library(aqp)
data(afsp)
Pdata <- subset(afsp$horizons, select=c("SOURCEID", "PHIHO5", "PHIKCL", "ORCDRC", "CLYPPT", "LSQINT", "UHDICM", "LHDICM", "MCOMNS"))
TD<-join(Pdata,afsp$sites)
#subset database TD suit to MPS, when the layerID "LSQINT" =0, it should be removed
TD.p<-subset(TD,!(TD$SOURCEID %in% "TZ 13575W3_0137")&TD$LSQINT>0)
#SOURCEID=TZ #13575W3_0137 is missing layer1, Td.p will become to soilProfileCollection later
TD.O<-TD.p ##orginal data
depths(TD.p) <- SOURCEID ~ UHDICM + LHDICM
site(TD.p) <- ~ LONWGS84 + LATWGS84 + TAXGWRB
coordinates(TD.p) <- ~ LONWGS84 + LATWGS84
proj4string(TD.p) <- CRS("+proj=longlat +datum=WGS84")
ORCDRC.w<- mpspline(TD.p, var.name="ORCDRC",d = t(c(0,5,15,30,60,100,200)))
str(ORCDRC.w)
PHIHO5.w<- mpspline(TD.p, var.name="PHIHO5",d = t(c(0,5,15,30,60,100,200)))
CLYPPT.w<- mpspline(TD.p, var.name="CLYPPT",d = t(c(0,5,15,30,60,100,200)))
##############
MO<- colMeans(ORCDRC.w$var.1cm[1:6,]) #mean value in top soil(0-5cm) of organic
MPH<-colMeans(PHIHO5.w$var.1cm[1:6,])#mean value in top soil(0-5cm) of PH
MClay<-colMeans(CLYPPT.w$var.1cm[1:6,])#mean value in top soil(0-5cm) of clay%
A<-data.frame(ORCDRC.w$idcol,MO,MPH,MClay)
D1<-subset(TD.O,TD.O$LSQINT==1)##subset original value of top soil
afsp.o<-merge(D1,A,by.x="SOURCEID",by.y="ORCDRC.w.idcol")#prepared dataset
afsp.w <- afsp.o[,c("SOURCEID", "LONWGS84", "LATWGS84", "MO", "MClay", "MPH", "TAXGWRB")]
afsp.ph <- subset(afsp.w,MPH>=0) ##PH
coordinates(afsp.ph) <- ~ LONWGS84 + LATWGS84
afsp.Clay <- subset(afsp.w,MClay>=0) ##Clay
coordinates(afsp.Clay) <- ~ LONWGS84 + LATWGS84
afsp.oc<- subset(afsp.w,MO>=0) ##SOC
coordinates(afsp.o) <- ~ LONWGS84 + LATWGS84
```

2 linear regression without interaction model （pH）

```
f.PH <- as.formula(paste('MPH ~ DEMSRE0a + SLPSRT0a + TDMMOD0a + TDSMOD0a + IFLGRE0a +
                         TNMMOD0a + TDHMOD0a +TNSMOD0a + EVMMOD0a + EVSMOD0a + PREGSM0a +', paste("G0", 1:9, "ESA0a", sep="", collapse="+"), '+',paste("G", 10:22, "ESA0a", sep="", collapse="+")))
PH.lm <- step(lm(f.PH, data=afsp.ph,na.action=na.exclude))
PHlr.pr<-predict(PH.lm,afsp.ph) #predict only for the observation location
PHlr.pr<-as.data.frame(PHlr.pr)
PHLR.pr<-predict(PH.lm,grids0, type="response", se.fit=TRUE)#predict for the whole area
PHLRW.pr <- as.data.frame(PHLR.pr)
boxplot(PHLR.pr[,1])
hist(PHLR.pr[,1])
##plot the map in R
grids0$PHLR.pr <- PHLR.pr[,1]
spplot(grids0, zcol="PHLR.pr", col.regions=colorRampPalette(c('yellow','orange','darkorange4','goldenrod4','darkgreen')),main="PH")
```

```
writeOGR(grids0, ".", "PH.LRWIA", driver="ESRI Shapefile")
```

3 linear regression with interaction model (pH)

```
f.PH<-as.formula(paste('  MPH  ~  DEMSRE0a*(SLPSRT0a+EVSMOD0a+EVMMOD0a+TDHMOD0a+TDMMOD0a+
TDSMOD0a+TNMMOD0a+TNSMOD0a+PREGSM0a+IFLGRE0a)+
SLPSRT0a*(EVSMOD0a+EVMMOD0a+TDHMOD0a+TDMMOD0a+TDSMOD0a+TNMMOD0a+TNSMOD0a+PREGS
M0a+IFLGRE0a)+
EVSMOD0a*(EVMMOD0a+TDHMOD0a+TDMMOD0a+TDSMOD0a+TNMMOD0a+TNSMOD0a+PREGSM0a+IFLGR
E0a)+
EVMMOD0a*(TDHMOD0a+TDMMOD0a+TDSMOD0a+TNMMOD0a+TNSMOD0a+PREGSM0a+IFLGRE0a)+
                TDHMOD0a*(TDMMOD0a+ TDSMOD0a+TNMMOD0a+TNSMOD0a+PREGSM0a+IFLGRE0a)+
                TDMMOD0a*(TDSMOD0a+TNMMOD0a+TNSMOD0a+PREGSM0a+IFLGRE0a)+
                TDSMOD0a*(TNMMOD0a+TNSMOD0a+PREGSM0a+IFLGRE0a)+
                TNMMOD0a*(TNSMOD0a+PREGSM0a+IFLGRE0a)+
                TNSMOD0a*(IFLGRE0a+PREGSM0a)+IFLGRE0a*PREGSM0a +
                ', paste("G0", 1:9, "ESA0a", sep="", collapse="+"), '+', paste("G", 10:22, "ESA0a", sep="",
collapse="+")))
PH.LRia<-step(lm(f.PH,data=afsp.ph))
PHlr.prI<- predict(PH.LRia,afsp.ph)
```

4 regression tree model (pH)

```
library(rpart)
#####Default setting####
rt.phD<-rpart(f.PH, data=afsp.ph)  # regression tree
pr.phD<-predict(rt.phD)
ph.ob<-sum((afsp.ph$MPH-mean(afsp.ph$MPH))^2)
ph.pr<-sum((afsp.ph$MPH-pr.phD)^2)
phD.R2=(ph.ob-ph.pr)/ph.ob  ## R2=0.246
phD.RMSE=sqrt(ph.pr/nrow(afsp.ph))  ##RMSE= 0.8827
#
set.seed(323)
ph.rte<-rpart(f.PH, data=afsp.ph,cp=0.01,method="anova",usesurrogate=2,xval=10,surrogatestyle=1)#
cp <- ph.rte$cptable[which.min(ph.rte$cptable[,"xerror"]),"CP"]
ph.rt<-prune(ph.rte,cp=cp)
plotcp(ph.rte)
pr.phrt<-predict(ph.rte)#predict for the observation area
pr.phrtM<-predict(ph.rte,grids0)# predict whole area
phpr.R2<-(ph.ob-sum((afsp.ph$MPH-pr.phrt)^2))/ph.ob# 0.2456841
phpr.RMSE<-sqrt(sum((afsp.ph$MPH-pr.phrt)^2)/nrow(afsp.ph))# 0.8803541
```

5 random forest model (pH)

```
library(randomForest)
f.PH <- as.formula(paste('MPH ~ DEMSRE0a + SLPSRT0a + TDMMOD0a + TDSMOD0a + IFLGRE0a + TNMMOD0a
+ TDHMOD0a +
                TNSMOD0a + EVMMOD0a + EVSMOD0a + PREGSM0a +',
                paste("G0", 1:9, "ESA0a", sep="", collapse="+"), '+',paste("G", 10:22, "ESA0a", sep="",
collapse="+")))
########optimal model #####
##get the optimal combination of parameters nodesize, mtry and ntree combination
rsq.PH=c(1:75)# for saving R2
rmse.PH=c(1:75)# for saving RMSE
nz=c(10,15,20,25,30) ##node size testing list:b
nt=c(100,500,750,1000,1250) #tree number(ntree) testing list:c
```

```
a=3# the a=1 or a= 2 have been checked, which contribute to bad results and memory problems
#a>5,also get worse result.
i=1# the order of the calculation loop.
##the below loop run around 10 hours
for(a in 3:5) {#mtry, split number: a
 b=1  #nodesize
 for(b in 1:5){
  c=1 # ntree
  for(c in 1:5){
   set.seed(4476)
   rf.PH<-randomForest(f.PH, data=afsp.ph,mtry=a,
nodesize=nz[b],ntree=nt[c],importance=T,na.action=na.omit)
   n=nt[c]
   rsq.PH[i]=rf.PH$rsq[n]
   rmse.PH[i]=sqrt(rf.PH$mse[n])
   print(paste(i,"mtry =",a,"nodesize =",nz[b],"ntree =",nt[c]," R2 =",rsq.PH[i],"rmse =",rmse.PH[i]))
    i=i+1}
 }
}
set.seed(4476)
rf.PH<-randomForest(f.PH, data=afsp.ph, mtry=5, nodesize=15,ntree=1250,importance=T,na.action=na.omit)
round(importance(rf.PH), 2)
pr.ph<-predict(rf.PH,afsp.ph,type="response", norm.votes=TRUE, predict.all=FALSE,proximity=FALSE,
nodes=TRUE)
```

6 independent validation （linear regression without interaction for pH）

```
part<-c(0.4,0.5,0.6,0.7,0.8)# training percent
r2.PHTrain<-c(1:5)
rmse.PHTrain<-c(1:5)
r2.PHTest<-c(1:5)
rmse.PHTest<-c(1:5)
PH.it<-as.data.frame(afsp.ph)
####
for(p in 1:5){
 set.seed(4476)
 PH.T1<- sample(1:nrow(PH.it), round(part[p]*nrow(PH.it)))#training part of data
 PH.tt <- step(lm(f.PH, PH.it[PH.T1,],na.action=na.exclude))
 pr.PHTr<-predict(PH.tt,PH.it[PH.T1,])
 pr.PHTr<-as.data.frame(pr.PHTr)
  A=sum((PH.it[PH.T1,]$MPH-mean(PH.it[PH.T1,]$MPH))^2) # training set
 B=sum((PH.it[PH.T1,]$MPH-pr.PHTr)^2)  #training set
 r2.PHTrain[p]<-(A-B)/A
 rmse.PHTrain[p]<-sqrt(B/length(pr.PHTr[,1]))
  pr.PH<-predict(PH.tt,PH.it[-PH.T1,])
 pr.PH<-as.data.frame(pr.PH)
 A=sum((PH.it[-PH.T1,]$MPH-mean(PH.it[-PH.T1,]$MPH))^2) # test set
 B=sum((PH.it[-PH.T1,]$MPH-pr.PH)^2)  #test set
 r2.PHTest[p]<-(A-B)/A
 rmse.PHTest[p]<-sqrt(B/length(pr.PH[,1]))
 print (c(part[p],rmse.PHTrain[p],r2.PHTrain[p],rmse.PHTest[p],r2.PHTest[p]))
}
```