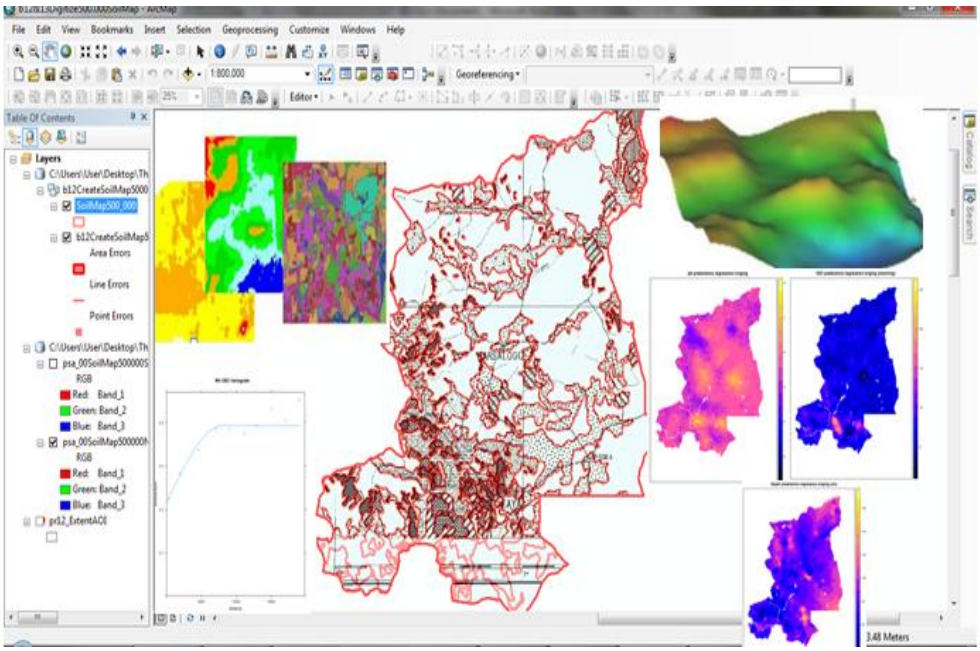
CENTRE FOR GEO-INFORMATION

Thesis Report GIRS-2014-12

DIGITAL SOIL MAPPING USING LEGACY SOIL MAPS FOR PREDICTING SOIL PROPERTIES IN SANMATENGA, BURKINA FASO

Charl Wong



April 2014

WAGENINGEN UNIVERSITY
WAGENINGEN UR

Centre for Geo-Information
Thesis Report GRS-80436

DIGITAL SOIL MAPPING USING LEGACY SOIL MAPS FOR PREDICTING SOIL PROPERTIES IN SANMATENGA, BURKINA FASO

A thesis submitted in partial fulfillment of the degree of Master of Science at Wageningen University and Research Centre, The Netherlands.

Charl Wong

Registration No. 870221972100

Supervision:

Dr Ir Sytze de Bruin	(Laboratory of Geo-information Science and Remote Sensing)
Dr Ir Gerard Heuvelink	(Soil Geography and Landscape / ISRIC – World Soil Information)
Ir Johan Leenaars	(ISRIC – World Soil Information)

April 2014

Acknowledgement

First and foremost I would like to thank the supervisors Gerard Heuvelink, Sytze de Bruin and Johan Leenaars for their valuable guidance and advice, which have guided me through every step of this period of doing my thesis. I really appreciate the time they make to give me suggestions, and constructive criticisms. Therefore, I am deeply indebted to them.

Gerard Heuvelink, thank you for the overall supervision, whenever I am stuck or having difficulties understand something, you were always there to explain and clarify things for me.

Sytze de Bruin, thank you for almost always immediately replying first to my email, whenever I have problems you were the first to reply and give expert advices.

Johan Leenaars, thank you for sharing your expert knowledge of the legacy soil data and the Burkina Faso study area with me.

Furthermore, I would like to thank Bas Kempen from ISRIC for sharing and providing some of the environmental covariates data.

I would like to thank John Stuiver for his help and advice on how to digitize soil maps in ArcGIS10.1.

Last but not least I would like to thank my parents and my sister, who always unconditionally support and encourage me in my life. Without their emotional and financial support, I would even not have had the experience and opportunity to study abroad.

Charl Wong

Wageningen, April 2014

Abstract

Digital soil mapping (DSM) is a method to model the relationship between observed soil profile data with environmental covariates to predict soil properties at unvisited locations. The aim of this study is to focus on the usability of legacy soil maps with different scales as environmental covariates in DSM to predict three target variables (pH, Cation Exchange Capacity (CEC) and soil depth) for the area of Sanmetanga province in Burkina Faso. Four methods were used to include legacy soil map information in DSM, namely: 1) Regression kriging with legacy soil map as categorical variable (CAT); 2) Stratified kriging using the delineations of the legacy soil map as map unit boundaries (STK); 3) Legacy soil map used as observed information obtained from the accompanying report (OBS); and 4) Combining the OBS result with kriged residuals (OBSRES). Regression kriging without use of legacy soil map information is used as a reference method (RK). The accuracy of all methods was assessed with the Correlation (r), Mean Error (ME) and the Root Mean Square Error (RMSE). The best method for predicting the soil properties pH and CEC was the CAT method using a legacy soil map with a scale of 1:100.000, where the pH had an r of 0.53, an ME of -0.004 and an RMSE of 0.76 and the CEC had an r of 0.59, an ME of 0.011 [cmol+/kg] and a RMSE of 4.54. The best method for predicting soil depth was the reference method, which had an r of 0.26, an ME of 0.32 [cm], and an RMSE of 61.42. For the reference method, the predicted pH had an r of 0.48, an ME of -0.009 and RMSE of 0.78, for CEC an r of 0.54, an ME of 0.047 and RMSE of 4.71. The obtained results show that the soil map provided disappointingly little information. However, the methods and data used in this thesis create a basis for a possibly more accurate DSM using legacy soil maps for the province in Burkina Faso, which should be extended to include a much denser soil profile dataset and more adequate environmental covariates including more detailed soil information from the legacy soil maps.

Keywords: Digital Soil Mapping, Legacy Soil Map, Regression Kriging, Stratified Kriging, Variogram

Table of Contents

List of Figures

List of Tables

1. Introduction

1.1 Context and background

Soil is very important part of the natural environment and can be seen as a life support system. Without soil, human life would be impossible. A few of many important soil functions are: to filtrate groundwater, hold and provide nutritious substances and water necessary for plant growth, act like a living source for many organisms and decomposing substances, and to absorb, store and reflect the sun's energy (Mitschang, 2008).

Soil mapping started many decades ago, meaning that the world is rich with soil data (Mayr et al., 2008). Most of the data are unused (Rossiter, 2008) due to the lack of digital availability, technology and inconsistencies in mapping, extent and spatial distribution of the numerous surveys and associated data. These legacy data may consist of soil maps with legends, soil survey reports and/or soil profile descriptions (Minasny and McBratney, 2010). Legacy soil data can be a valuable source of information on the spatial variation of soil properties (Mayr et al., 2010). Practical studies have shown that soil properties derived from soil survey maps have the same quality as maps obtained using spatial interpolation (Bregt et al., 1987).

The rise and the advances in technology such as GPS, remote and proximal sensing, digital elevation models (DEM) and Geographic Information Systems (GIS), have resulted in the production of digital soil data and class maps, needing only limited, expensive, fieldwork and laboratory analysis. This approach is now widely known as *digital soil mapping* (DSM). DSM is formulated as ''the creation and population of spatial soil information systems by numerical models inferring the spatial and temporal variation of soil types and soil properties from soil observation and knowledge and from related environmental variables'' (Lagacherie et al., 2006).

DSM combines soil observations with auxiliary data or covariates (such as correlated environmental variables and remote sensing images), using statistical models to predict soil type and properties at unobserved locations in a landscape (Dobos, 2006). The well-known equation from Jenny (1941) $S = f\,(CL, O, R, P, T)$ represents the soil (S) as a function of the covariates: climate (CL), organisms (O), relief (R), parental material (P), and time (T). The equation can be seen as the foundation for DSM and it offers an idea for understanding the relationships between soil and environmental variables (McBratney et al., 2003). This equation is later adapted by McBratney et al. (2003) to $S = f\,(s, c, o, r, p, a, n)$, where soil is included as a factor because soil can be predicted from other soil properties. The soil (S) is a function of the covariates: soil (s), climate (c), organisms (o), relief (r), parental material (p), age (a) and spatial position (n).The s refers to soil information either from a prior map, from remote or proximal sensing or expert knowledge.

There is an urgent need for accurate, up-to-date and spatially referenced soil information (soil maps) by the modeling community (GlobalSoilMap.Net, 2013), farmers and land users, and policy and decision makers (European_Commission, 2006; UNEP, 2007). As has been demonstrated by many studies, legacy soil data can play an important role in DSM, especially when there is a lack of resources to collect new soil data (Bui and Moran, 2001, 2003; McBratney et al., 2003). However, the use of legacy soil data is challenging due to a number of problems related to the variable nature of the data, such as varying availability of numeric data, varying methods applied and the associated need for harmonization, varying precision of soil descriptions, varying precision or lack of georeferencing, varying adequacy of location, distribution of soil observations and varying consistencies of map extent, legend etc., as well as related to the licensing of the data.

According to McBratney et al. (2003) and Minasny and McBratney (2010) the appropriate method to develop useful digital soil maps depends on the availability, amount and type of data, which preferably includes both legacy soil maps and soil point data. Heuvelink and Bierkens (1992) showed that if legacy soil maps and soil profile data are jointly included in DSM a more accurate soil map is obtained than using either of them separately. Legacy soil maps could be used as soil covariates (Hartemink et al., 2008) or as a source for calibrating DSM procedures that take into account the soil surveyor knowledge (Lagacherie et al., 1995; Bui and Moran, 2001; Bui, 2004).

Stein et al. (1988) and Boucneau (1998) used soil polygons from legacy soil maps to perform kriging within strata. Goovaerts and Journel (1995), Hengl et al. (2004) and Liu et al. (2006) use soil map information to inform the local mean of the random function used to model the soil property statistically, followed by kriging the stationary residuals and hereafter combined together to obtain final estimates. Odgers et al. (2014) disaggregated and harmonized the legacy soil map units to sample the polygons of a legacy soil map to generate a number of realizations of the potential soil class distribution by using classification trees. Grimm et al. (2008) used the legacy soil map as a factor in machine learning techniques, such as random forests, which is an ensemble learning method for classification and/or regression. In another machine learning method, Artificial Neural Networks (ANN), the legacy soil map is used to learn algorithms (Behrens et al., 2005).

1.2 Problem definition

The people of Burkina Faso are highly dependent on pastoral and agricultural resources to support their livelihood. Growing population pressure has led to overgrazing and increased cropping intensity, leading to serious degradation of the natural environment, such as reduced vegetation cover and degraded soil fertility (van Lieshout et al., 1997). As a consequence, there is an increasing risk for food shortage and famine. In order to improve the life quality of the Burkina Faso population and recover the agricultural production, one must properly use and manage the soil. In order to know how to manage which soil to yield optimal agricultural production, information about soil type or soil properties that influence agronomic production must be known. Therefore producing accurate and relevant soil property maps would not only contribute to worldwide DSM activities but would more importantly be very useful for Burkina Faso and its policy and decision makers. In this research three key soil properties, acidity or alkalinity (pH in H_2O), Cation Exchange Capacity (CEC) and the soil depth (Depth), will be studied.

Considering the available methods, the problem considered is which method is best for the case study area to include legacy soil map in DSM. And whether this method or this inclusion will produce a more accurate soil property map than DSM using another method or without using legacy soil maps.

1.3 Research objective

General objective:

The objective of this research is to devise, apply and test methods for including legacy soil maps in digital soil mapping to improve the spatial prediction of the top soil pH, Cation Exchange Capacity (CEC) and soil depth for Sanmatenga province, Burkina Faso.

Research questions:

1. Which methods can be used to include legacy soil maps in DSM to model the relationship between soil properties and environmental covariates?
2. How can the accuracy of the results of each method be assessed?
3. How can these methods be implemented in R software?
4. Which results are obtained when the methods are applied to the area of Sanmatenga province with different legacy soil map scales?
5. Which of the selected methods produces the most accurate soil maps and how accurate are these maps compared to DSM without using legacy soil maps?

1.4 Outline

This report includes five chapters. The first chapter introduced the purpose of this study as well as the background and problem of including legacy soil map information in DSM. The second chapter describes the study area, the input data (soil profile data and environmental covariates), the methods and software (R and ArcGIS) used in this research. The third chapter presents the results of the predicted soil properties maps and summary statistics (correlation, mean error, root mean squared error, mean, median, min, max) of predicted soil properties. The fourth chapter discusses the results presented in Chapter 3. In the last chapter, a conclusion is made on this study.

2. Materials and methods

2.1 Study area description

The study area is the province of Sanmatenga located in the centre-north of Burkina Faso (Figure 2.1) and has an area around 7579 km^2. It stretches north-south from Universe Transverse Mercator (UTM) coordinates 1544 to 1418 (126 km) and east-west from UTM co-ordinates 753 to 657 (96 km).

Two-thirds of the area is situated in the Soudano-Sahelian agroclimatic zone, where the north part stretches into the South-Sahelian zone and the south into the North-Soudanian zone. The average annual rainfall is around 650 mm with an average temperature of 35 °C.

Geologically, two major strata can be distinguished in the area: the Birimian and the Antè-Birimian stratum. The Birimian, dated between 2400-2170 million years ago and consisting of mainly meta-vulcanic rock, is situated in the southern, and a little bit in the north-eastern part of the province. The Antè-Birimian stratum, dated before 2400 million years ago, covers the larger part of the province.

The major kinds of land use in the study area are traditional rain-fed agriculture and extensive grazing.

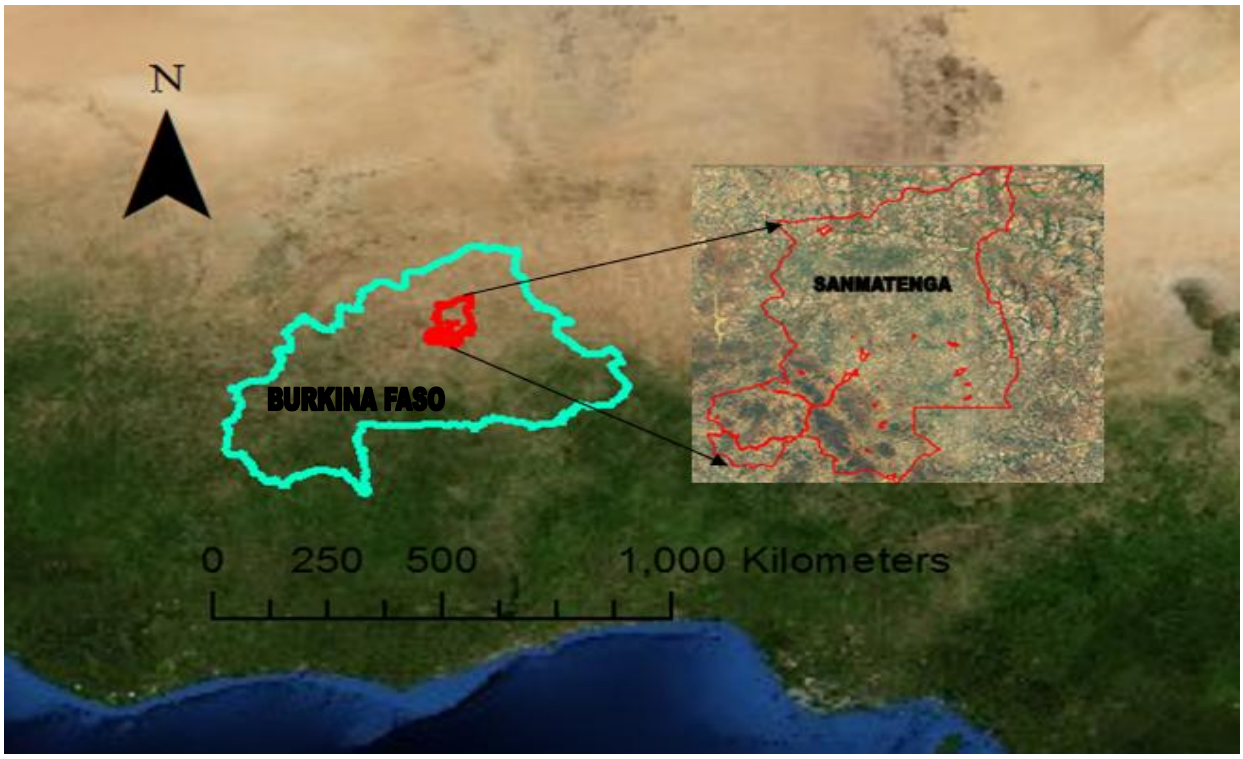


Figure 2. 1 Map of Burkina Faso and study area Sanmatenga province

2.2 Methods for Including Legacy Soil Maps in DSM

Soil properties can be predicted from covariates measured at the same (or neighboring) locations. The predictive approaches may be different in every case study. In this research, the following geostatistical methods to include legacy soil maps in DSM were used.

2.2.1 Common DSM with regression kriging (RK) without legacy soil map

A commonly used DSM method is regression-kriging (Odeh et al., 1994; Goovaerts, 1997; Hengl et al., 2004), which is further referred as RK in this research. The method involves linear regression of the soil properties with the covariates as explanatory variables, followed by kriging of the regression residuals. Next, the regression predictions are summed up with the kriged residuals to obtain a final prediction map. This method will be used as a reference model. Assume the observations of soil properties be specified as $z(s_1)$, $z(s_2)$, . . ., $z(s_n)$, where s_i (i=1,...,n) is a geographic location with x and y coordinates and n is the number of observations. In the case of RK, a soil property at a new, unvisited location (s_0) is predicted by summing up the linear predicted values and the kriged residuals (Eq. 1.):

$$\hat{z}(s_0) = \hat{m}(s_0) + \hat{e}(s_0) \quad \text{(Eq. 1)}$$

where the $\hat{m}(s_0)$ is fitted using linear regression analysis, and the residuals $\hat{e}$ are interpolated using simple kriging (Eq.2.):

$$\hat{z}(s_0) = \sum_{k=0}^{p} \hat{\beta}_k \cdot q_k(s_0) + \left(\sum_{k=1}^{p} \lambda_k \cdot e(s_i)\right) \quad \text{(Eq. 2)}$$

where the $\hat{\beta}_k$ are the estimated regression coefficients of the environmental covariates, $q_k(s_0)$ is the k-th external covariate at location s_0, $q_0 = 1$ so that $\hat{\beta}_{k0}$ is the intercept of the regression, p is the number of covariates, λ_k are kriging weights, which is selected in such a way that minimize the expected squared prediction error, and $e(s_i)$ are the regression residuals. The kriging weights are derived from a semivariogram. The semivariogram is a measure of spatial similarity between observations as a function of distance and consists of an experimental and model variogram. The experimental variogram is determined by calculating half the variance of each pair of points with regard to the other points in the dataset until every point has paired up with each other. Hereafter the half variances versus distance between the points are plotted. Next, a model variogram is computed with simple mathematics to fit a trend in the experimental variogram. The semivariogram consist of a nugget, sill and range. The nugget indicates the presence of short-range spatial variability including measurement error. The sill value means that the residuals values have a large mean deviation and shows high variability. The range is the distance at which there is little or no autocorrelation among points.

2.2.2 DSM including legacy soil map as categorical variable in RK (CAT)

In this method, the legacy soil map is incorporated in the RK method where the legacy soil map is treated as an additional covariate, to be specific as a categorical explanatory variable $q_t(s_0)$. This categorical covariate variable $q_t(s_0)$ is represented by as many dummy variables as there are categories, where the value of each of the $q_t(s_0)$ is either 0 or 1, indicating the absence or presence of the specific category (i.e. soil type). The following equation shows the incorporated categorical variable in reference method (Eq. 3.):

$$\hat{z}(s_0) = \sum_{k=0}^{p} \hat{\beta}_k . q_k(s_0) + \sum_{t=0}^{r} \hat{\alpha}_t . p_t(s_0) + \left(\sum_{k=1}^{p} \lambda_k \cdot e(s_i)\right) \qquad \text{(Eq. 3)}$$

Where r is the number of soil types, $\hat{\alpha}_t$ is the regression coefficient and p_t (s_0) is the presence-absence of the t-th soil type (i.e. has a value of 1 or 0).

2.2.3 DSM including legacy soil map delineations in Stratified kriging (STK)

In this approach, the delineations of the legacy soil map are used as a boundary to perform a within-stratum interpolation. Soil map delineations can be used as an auxiliary variable to improve the spatial interpolation of soil properties (Stein et al., 1988; Boucneau, 1998). The boundaries of the soil map divide the study area into a number of polygons that are represented by a legend that consists of different or associated soil types. Observations located in polygons of the same soil type are kriged using ordinary kriging (Eq. 4.). Observations outside a stratum are not used. For each stratum, at least ten soil observations should be available, otherwise the stratum has to be merged. Some mapping units will have a small number of soil observations to fit a variogram, therefore the observations in each stratum have to be standardized, where a general variogram was computed for kriging per stratum. The standardizing process was done per stratum by dividing each observation in a stratum by the standard deviation of the observations in that particular stratum. Hereafter a general variogram was computed by increasing the coordinates of the soil observations of each stratum with an increment that should be at least twice the width of the study area in order to group the far lying observations and polygon from the same stratum together and to make a distinction between the observations of the other strata.

$$\hat{z}(s_0) = \sum_{k=1}^{p} \lambda_k \, z(s_i) \qquad \text{(Eq. 4)}$$

where λ_k are the kriging weights, $z(s_i)$ are the observations, and $\hat{z}(s_0)$ are the predicted soil properties. The kriging weights are calculated by minimizing the expected squared prediction error, under the condition of unbiasedness. The latter is achieved by the constraint $\sum_{k=1}^{p} \lambda_k = 1$.

2.2.4 DSM including legacy soil map as observed information (OBS)

Based on expert knowledge, lookup tables and/or representative soil profiles attributed to the legend or mapping unit from an accompanying report, the given information about soil properties from the legacy soil map was used as observed information. Soil properties values, measured from the soil profile(s) in the report that are supposed to represent a mapping unit, are assigned to that particular mapping unit. Soil associations in a map unit will be taken under consideration such as when the percentage soil type proportion is given. If more than one soil profile was given to represent a mapping unit and the portion of the soil class association was given in the report of the legacy soil map than the soil property value was determined by a weighed mean (Eq. 5). When no information is given on the proportions within soil associations than the arithmetic mean was taken. When mapping units did not have a representative soil profile or no observed information was available for such units than this mapping unit was merged with other mapping units that had similar soil types.

$$Weighed\ mean = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} \quad \text{(Eq. 5)}$$

where w_i is given weight and x_i is the value.

2.2.5 OBS combined with kriged residual (OBSRES)

In this method, the OBS predicted soil properties are overlaid with the soil profile to calculate and interpolate the residuals with simple kriging, followed by summing up the kriged residuals with the OBS predicted soil property. In other words, the approach applies simple kriging to the residuals of the OBS method and combines the two results.

2.3 Data description and pre-processing

2.3.1 Soil profile data description

As target variables were used, the cation exchange capacity of the fine earth fraction (CEC) expressed in cmol+/kg of the topsoil, the pH measured in H_2O of the topsoil and the depth of the soil expressed in cm. The soil depth, that is further referred as Depth, are censored data obtained by taking the maximum observed depth of the soil profile sample. These soil profile data were obtained from the Africa Soil Profile Database (AfSP) version 1.1 which is maintained by Leenaars (2013). The AfSP is a compilation of standardized legacy soil data of 16,711 soil profiles, of which 15,500 are geo-referenced, for 38 Sub-Saharan African countries. The soil profile data were derived from over 450 data sources, both digital and analogue, from holdings (organizations) such as ISRIC, FAO, WOSSAC and IRD. The data were standardized to the e-SOTER conventions and validated according to routine rules (Leenaars, 2013).

2.3.2 Environmental covariates

The following auxiliary maps were used as environmental covariates: the 100m resolution digital elevation model (DEM) with the derivatives slope (SLOPE), relative elevation (RELEV), curvature (CURVATURE) and aspect (ASPECT). Curvature and aspect was derived from the DEM in ArcGIS 10.1 and the DEM and the other derivates were obtained from International Soil Reference and Information Centre (ISRIC), which was preprocessed by Köthe (2013) from Scilands GmbH. These variables represent the relief in the scorpan model. The European Space Agency (ESA) GlobCover map (LANDCOV), the enhanced vegetation index (EVI) with a resolution of 250m and the soil wetness index (SWI) with 100m resolution was used to represent the organism in the scorpan model. For the parent material in the scorpan model, the geology map (GEOLOGY) of Burkina Faso with scale of 1:1.000.000 was used. For the climate, the average land surface temperature of the day (TEMPDAY) and night (TEMPNIGHT) with a resolution of 1km were used. The following legacy soil maps of Burkina Faso with a scale of 1:100.000 (SM100K), 1:500.000 (SM500K) and 1:1.000.000 (SM1M) was used to represent the soil in the scorpan model. The environmental covariates with their sources and references are given in Table 2.1.

Table 2. 1 Environmental Covariates

Variable	Unit	Source	Reference
EVI	250 m	https://lpdaac.usgs.gov/lpdaac/products/modis_products_Table/vegetation_indi	(LP_DAAC, 2001)
Land cover	250 m	http://due.esrin.esa.int/globcover/	(ESA, 2009)
Soil map BF	Scale 1:100.000	http://library.wur.nl/WebQuery/isric/23904	(Asten, 1995)
Soil map BF (North Center)	Scale 1:500.000	http://library.wur.nl/WebQuery/isric/17498	(Boulet, 1968a)
Soil map BF (South Center)	Scale 1:500.000	http://library.wur.nl/WebQuery/isric/17713	(Kaloga, 1968a)
Soil map BF	Scale 1:1000.000	http://library.wur.nl/WebQuery/isric/17497	(Guillobez, 1985)
TEMPNIGHT	1km	ftp://africagrids.net/1000m/MYD11A2/LST_night/	(AfSIS, 2013a)
TEMPDAY	1km	ftp://africagrids.net/1000m/MYD11A2/LST_day/	(AfSIS, 2013b)
DEM	100m	ISRIC	(Köthe, 2013)
SWI	100m	ISRIC	(Köthe, 2013)
Relative Elevation	100m	ISRIC	(Köthe, 2013)
Slope	100m	ISRIC	(Köthe, 2013)
Aspect	100m	ISRIC	(Köthe, 2013)
Curvature	100m	ISRIC	(Köthe, 2013)
Geology BF	Scale 1:1000.000	ISRIC	(Köthe, 2013)

2.3.3 Legacy soil maps

The SM1M had a physiographic map legend with associated soil classes (Figure 2.2) with no accompanied report. The legend of the SM500K map (Figure 2.3 and 2.4) employs a hierarchical soil class order according to the French Soil Classification System (CPCS, 1967). The main soil classes area are divided into raw mineral soils, weakly developed soils, vertisols, brown soils, ferallitic soils, sodic soils and hydromorphic soils. Raw mineral soils are soils that have not undergone any significant pedological development and that have no or little organic matter in the upper 20 cm of the profile. Weakly developed soils are soils with a humiferous horizon that shows the beginning signs of geochemical alteration. Vertisols are dark-coloured clay soils dominated by swelling and shrinking clay minerals. Brown soils are soils characterized by mull type humus. Ferallitic soils are composed of almost completely weathered primary minerals. Sodic soils are characterized by alkaline salts in solution and the presence of exchangeable sodium. Hydromorphic soils are partly or entirely saturated with water over extended periods of time. A more detailed soil class description can be found in Latham (1982). The SM100K map (Figure 2.5) has a physiographic map legend. The highest level of the legend concerns geology, which is subsequently disaggregated into major landforms, minor landforms, and in some cases a fourth landform level. According to the corresponding report van Lieshout et al. (1997), the soils in the area are strongly correlated with the mapped geomorphological features. The major landforms were divided in: Hills (A), Indurated caps (B), Slopes (C), Bottomlands (D) and Aeolian complexes (E). The major landform classes were presented by capital letters in the map unit code followed by an integer, which represent the minor landforms such as lower, middle and upper slopes. A second integer represents the different land elements such as valleys, bottomland, etc... The ‘soil’ maps will be generalized with other mapping units based on hierarchically higher mapping units or similar soil type properties or soil descriptions. SM100K was already digitized whereas SM500K and SM1M still need to be digitized.

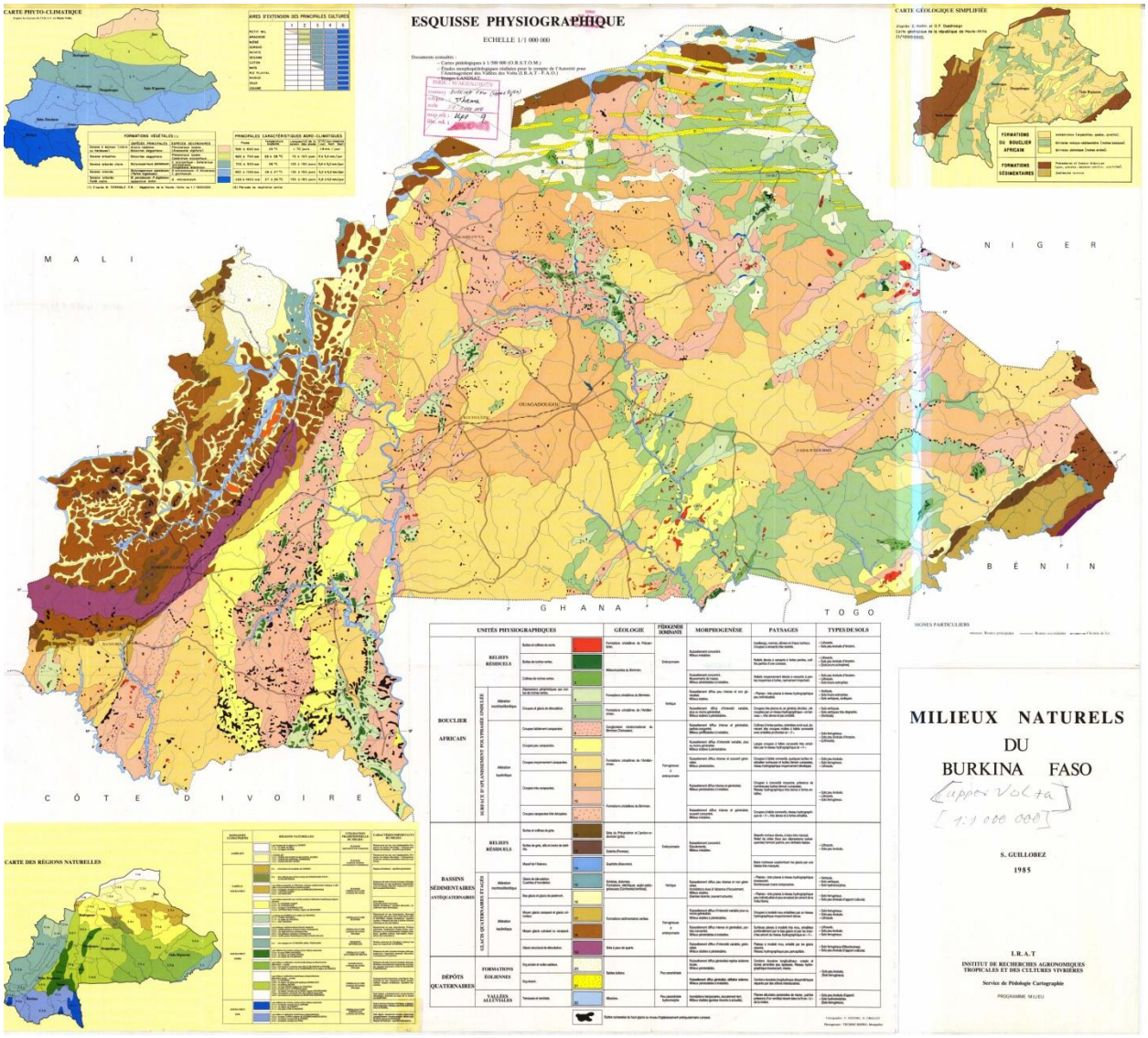


Figure 2. 2 SM1M of Burkina Faso

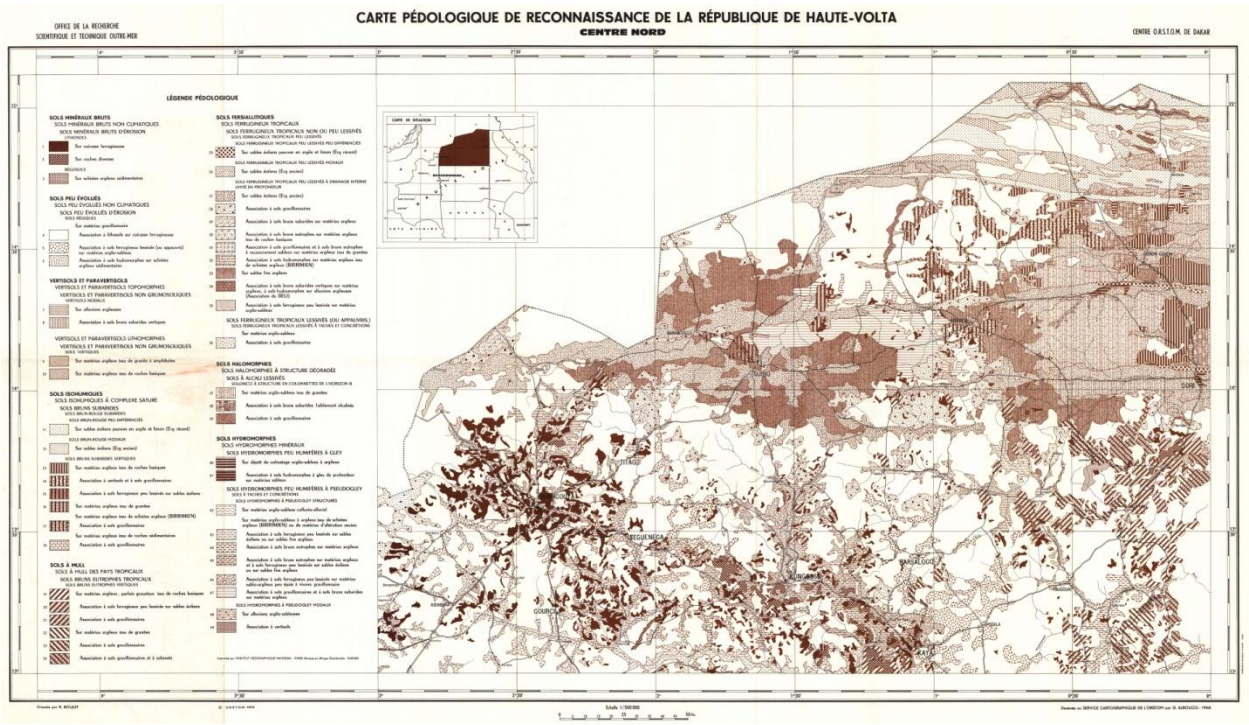


Figure 2. 3 SM500K of north-centre Burkina Faso

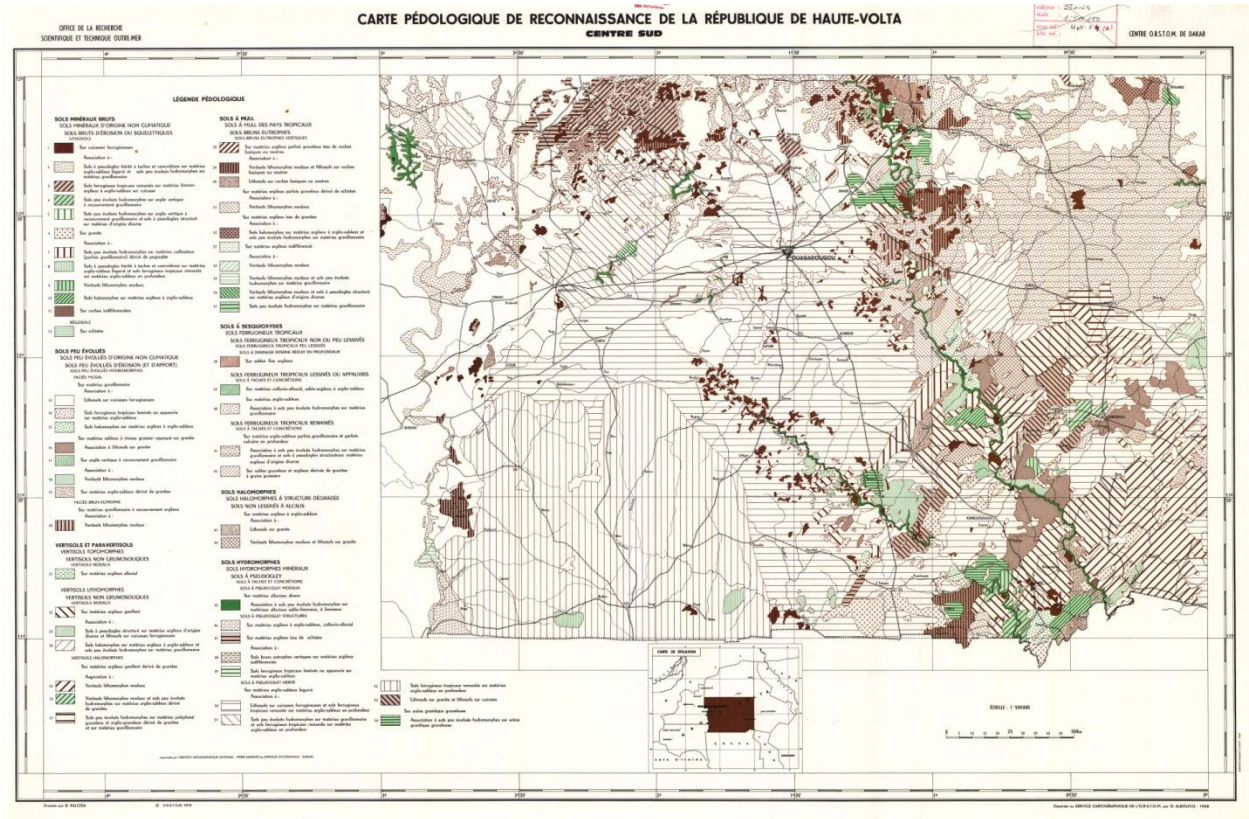


Figure 2. 4 SM500K of south-centre Burkina Faso

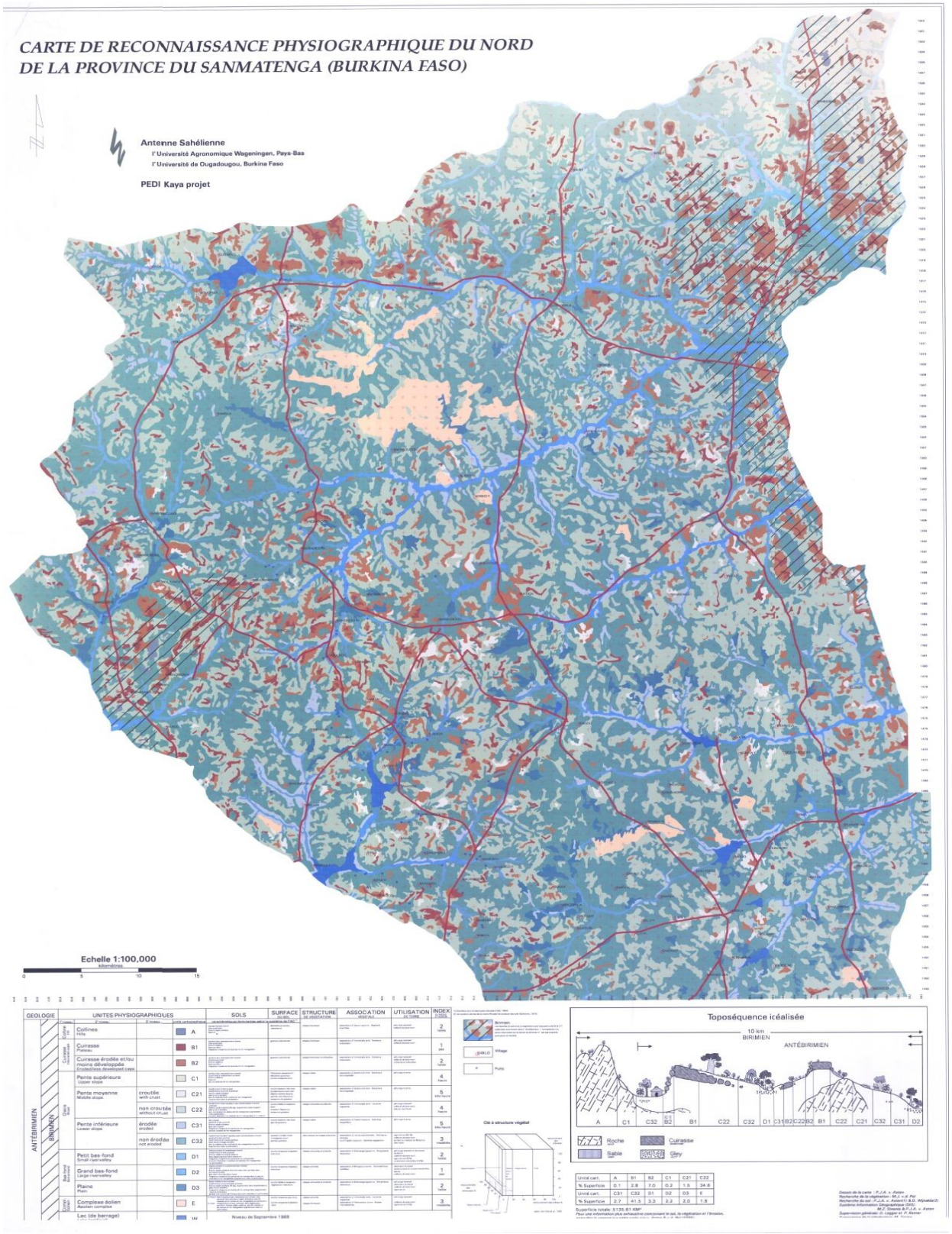


Figure 2. 5 SM100K of part study area, Sanmatenga

2.3.4 Mask, Extent, Resolution and File format

All covariates will be used in the same extent and resolution. The covariates were transformed to 100m resolution using the *resample* tool in ArcGIS. First, a mask was created in ArcGIS, which has the same extent as the SM100K, which is the extent of the province of Sanmatenga. Water bodies were masked out, because digital soil mapping only produces maps for areas that are covered by soil. Hereafter all the variables were masked to the same extent using the *extract by mask* tool in ArcGIS, followed by a conversion to ASCII text file. In Appendix 1 the preprocessed environmental covariates maps are presented.

2.3.5 Exploratory data analysis (EDA)

Exploratory data analysis (EDA) was used prior to the actual geostatistical modeling to examine the properties and quality of the data with the help of graphic techniques and descriptive statistics. The examination consists of checking the marginal distributions, general trends, and checks for outliers and anomalies. In this study, histograms were used to examine the distribution of the data. For geostatistical methods to be optimal, it is required that the data reasonably fit a normal distribution and are stationary. Data were considered stationary when the mean and variance are more or less constant in space (Bohling, 2005). Data transformations were applied to data that are not normally distributed. There are several transformation methods. The most common are natural logarithm, square root type, reciprocal, and Box-Cox. To back transform the predicted values to original scale, the variance needs to be taken under consideration. The following example equations were used to back transform the logarithm type (Eq. 6) and square root type (Eq. 7) to the original scale by considering the variance:

$$E(s_0) = \exp(\hat{z}(s_0) + 0.5*\sigma^2) \quad \text{(Eq. 6)}$$

$$E(s_0) = \hat{z}(s_0)^2 + \sigma \quad \text{(Eq. 7)}$$

where $\hat{z}(s_0)$ is the predicted soil property value, σ is the variance of the predicted soil property and $E(s_0)$ is the back transformed predicted soil property value to original scale.

2.3.6 Digitization of soil maps

As stated before, SM500K and SM1M needed to be digitized. Digitization was done in ArcGIS Desktop10.1. Geo-referencing was necessary to align the maps with existing geographically referenced data. Control points were selected randomly on each map and a polynomial transformation was applied to convert the geometry of the entire map to the reference geometry. The following guidelines were followed to select control points:

- Points should be taken on recognizable places where they represent the same object or geographic location, e.g. street, boundaries, building, etc.

- Points should be taken near each of the corners of the image and spread out over the entire image.

After geo-referencing, the soil maps were clipped to the same extent as the study area with the mask, before the digitization process was started. The first step of digitization was to create a new *feature data set* in a geodatabase and set the coordinate system to *WGS 1984 Web Mercator Auxiliary Sphere*. Hereafter a new *feature class* was created in the *feature data set*. The type of feature stored in the feature class was set on *polygon features* and a new field name of *LegendNR* with the data type *text* was created. Next, the *topology* was created in the *feature data set* to set the some digitization rules for the *feature class*. The *cluster tolerance* was set on 10 m and the rules for the feature class was set to *must not have gaps* and *must not overlap*, where the rest of the settings were left on default. With the *Editor* toolbar in ArcGIS the digitization process can be started by clicking on *start editing* followed by *create features* on the *Editor* toolbar. On the *create features* window the construction tool *polygon* was use to digitize the soil maps. After finishing digitizing, the attribute table with field name *LegendNR* was filled in with the corresponding mapping units' number or code of the legacy soil maps. In SM500K a distinction between legend code was made because the research area occurred both in the north-center and the south-center of the legacy soil maps of Burkina Faso. *LegendNR* that has NC before an integer means that the mapping unit belongs to the legend of the North-Center of the legacy soil map of Burkina Faso where NC stands for North-Center and the integer for the legend number in the legacy soil map of Burkina Faso. The same applies for the *LegendNR* SC that stands for South-Center from the South-Center legacy soil map of Burkina Faso. These legacy soil maps were resampled to 100 m resolution.

2.3.7 Reclassification and generalization of legacy soil maps

Reclassification with specified criteria is conducted to create more generalized layers. Layers with many classes or levels are difficult to handle during the interpolation process, especially with regression kriging. Many levels often cause inaccuracies by not performing accurate predictions. This happens if there is none or too few observations in a soil class to perform stratified kriging and it may also take much calculation time and load. The soil maps were classified to merge classes to avoid inaccuracies in the predictions and save computation time.

2.4 Assessment of models accuracy

Cross-validation was used to evaluate the model fitting and to quantify the accuracy of the methods. In this study Leave-One-Out cross validation was used. Leave-One-Out cross validation uses a single observation from the original data as validation data and the remaining observations were used to build the model. This was repeated until all observations of the original data were used once as validation data. The outcome is an average of the models. As

result, the target variable was predicted. These predictions and prediction errors for all locations were subsequently used to calculate three validation measures: the Correlation (r), the Mean Error (ME) and the Root Mean Squared Error (RMSE). The r measures the degree of linear relationship between the observed and predicted values, where a correlation of 1 or -1 implies a perfect linear correlation. It was calculated with Eq. 8. ME indicates the degree of bias in the predictions (Bello-Pineda and Hernández-Stefanoni, 2007), and should be close to zero. It was calculated with Eq. 9. RMSE measures the difference between the observed value and predicted value; smaller RMSE means that the model has predicted well (Odeh et al., 1994). It is calculated using Eq. 10.

$$r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2 \cdot \sum_{i=1}^{n}(y_i-\bar{y})^2}} \quad \text{(Eq. 8)}$$

$$ME = \frac{1}{n}\sum_{1}^{n}(X_{observed,n} - X_{predicted,n}) \quad \text{(Eq. 9)}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{1}^{n}(X_{observed,n} - X_{predicted,n})^2} \quad \text{(Eq. 10)}$$

where x_i is the observed soil property value, $\bar{x}$ is the mean of he observed soil property, y_i is the predicted soil property value, $\bar{y}$ is the mean predicted soil property, n is the number of observations.

2.5 Software implementation

2.5.1 Modeling Flow Chart

The modeling flow chart in Figure 2.2 represents the methodological framework of this study.

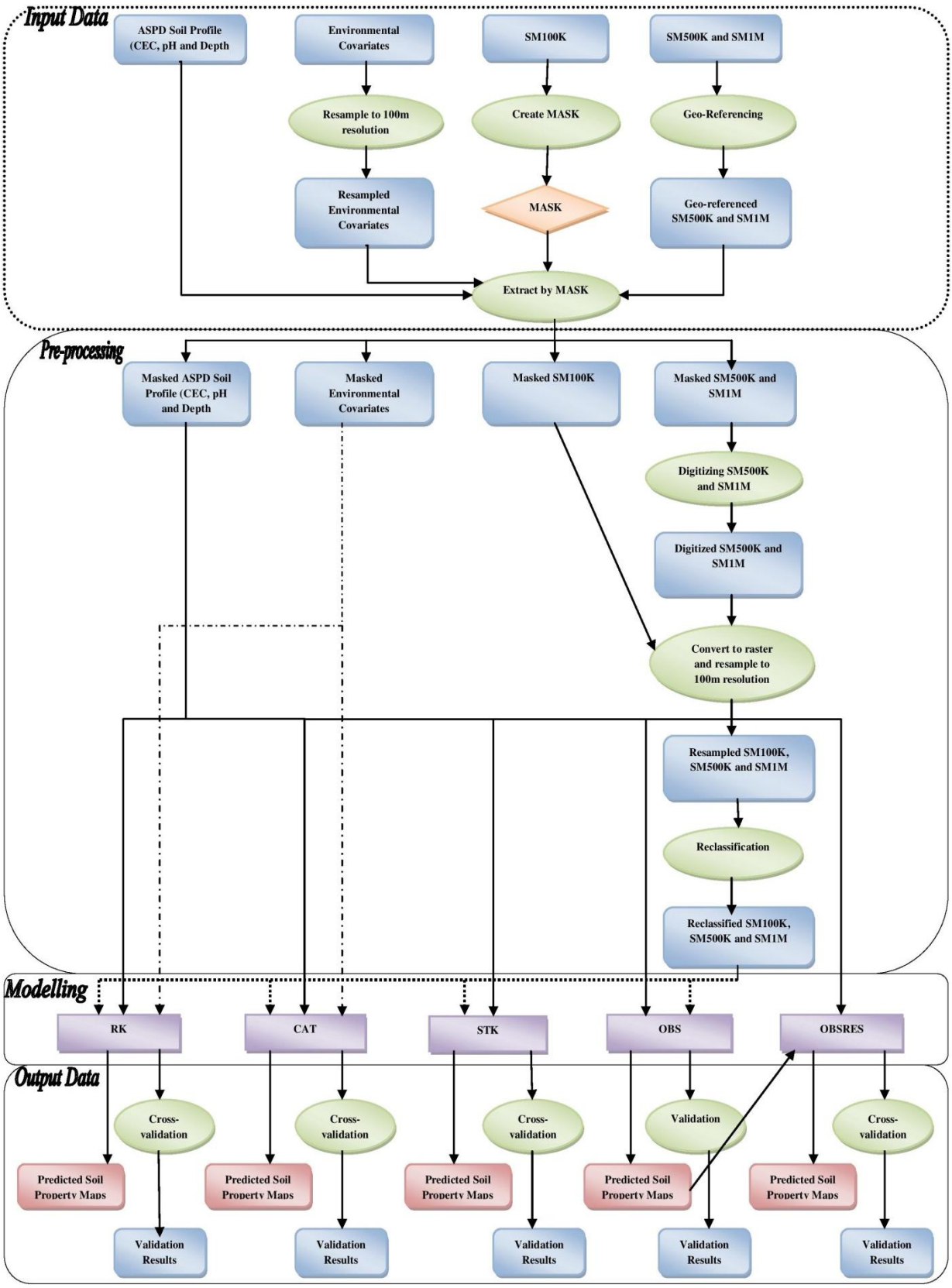


Figure 2. 6 Modeling flow chart

2.5.2 ArcGIS Desktop 10.1

ArcGIS is a geographic information system for mapping, manipulating, visualizing, compiling, analyzing, creating, designing, discovering, sharing and storing maps and geographic information, which is developed by ESRI Company. ArcGIS desktop 10.1 standard version is one of the ArcGIS products, which runs in a Microsoft Windows environment. In this research, ArcGIS was used for digitization, pre-processing the target variables as well the environmental covariates to the same extent and resolution, reclassification and assigning values to the soil maps and conversion of all variables to ASCII files, which were later used in R.

2.5.3 R and contributed packages

In this research R and contributed packages were used. R is open source software for data analysis, statistical computing and graphics. R provides a wide variety of statistical and graphical techniques, and is highly extensible. The packages, which were used in this research, were *gstat*, *raster*, *rgdal*, *sp* and *maptools*. The *gstat* package supports kriging, variogram modelling and visualization and cross-validation (Pebesma et al., 2013). The *sp* package provides classes and methods for spatial data (Pebesma et al., 2012). The *raster* package provides classes and functions to perform reading, writing, manipulating, analyzing and modeling of large geographic (spatial) data sets in 'raster' format (Hijmans et al., 2013). The package *rgdal* provides access to projection/transformation operations as well as read and write access to spatial data in different formats (Keitt et al., 2011). *Maptools* is a set of tools for manipulating and reading geographic data, in particular ESRI shapefiles (Lewin-Koh et al., 2011).

2.5.4 Software implementation for the methods

All methods were performed in R except OBS, which was executed in ArcGIS because this was easier and less time consuming.

2.5.4.1 RK

The following steps for the RK method were executed:

1. *Fit regression model*. Overlay covariates with the locations of observations of the dependent variable, and fit a regression model on the resulting dataset. The tool *over* of the package *sp* was used to perform the overlay. The basic package *stats* provides the tool *lm* to fit linear models and *step* to select covariates that are significant and can improve the model prediction accuracy. The criteria for selecting the covariates are based on the Akaike Information Criterion (AIC).
2. *Compute variogram*. The residuals from the regression models were used to calculate a variogram. First, an experimental semi-variogram was computed with the *variogram* tool followed by *fit.variogram* from the *gstat* package to fit a semi-variogram model.

3. *Apply regression model*. Apply the regression model to all locations in the study area to generate a map of the target variable. This was done with the *predict* tool from base package *stat*.
4. *Krige the residuals*. The residuals were kriged with simple kriging to the same area.
5. *Combine results*. Add the map with kriged residuals to the map of the regression predictions.
6. *Cross-validate results.* Repeat step 4 and use *krige.cv* instead of the *krige* tool. The obtained results were used to calculate the r, ME, and RMSE.

2.5.4.2 CAT

This method was executed exactly the same as the RK method, except that in the first step, the legacy soil map was added to the environmental covariate as a factor with the *as.factor()* tool from the *sp* package. It is possible that the legacy soil map will be left out during the stepwise covariate selection procedure because it is not statistically significant. When this occurs the legacy soil map is manually incorporated in the linear model.

2.5.4.3 STK

The following steps were executed for STK:

1. *Overlay soil property with legacy soil map*. This was done with the *over* tool.
2. *Make subset of data*. To make a subset of each soil type stratum which includes the soil property values, the command *data[which(data$part.a==1),]* was used. The integer in the script indicates the soil type or stratum ID.
3. *Standardize soil property value per stratum*. Calculate per stratum the standard deviation of the soil property and divide the soil property value by the standard deviation for that stratum.
4. *Calculate general variogram*. Stretch the coordinates per stratum with at least twice the width of the study area. Hereafter combine the strata together with the *rbind()* tool to make one dataframe. Calculate the experimental variogram and fit the variogram.
5. *Krige standardize soil property per stratum.* Krige per stratum the standardized values with ordinary kriging using the general variogram.
6. *Combine the strata*. Back-transform the kriged standardized values by multiplying these with the standard deviation of the stratum. Combine the strata with the *rbind()* tool to make one dataframe and plot the prediction maps.
7. *Cross validate model* Step 4 was repeated where *krige.cv* was used instead of *krige*. The obtained results were used to calculate the r, ME and RMSE.

2.5.4.4 OBS

For this method, reports of the legacy soil maps were required to gather information about the soil properties. SM1M does not have an accompanied report thus no research will be done for SMIM in this method. The report of SM100k was van Lieshout et al. (1997) and for SM500K these were Boulet (1968b) and Kaloga (1968b). When the values of the soil properties were determined then these were assigned to the accompanied mapping unit in ArcGIS with the *reclass* tool. In this method, validation was applied by using the soil observations from the AfSP database as observed information and the predicted map as prediction. The differences between these two were used to calculate the validation measurers r, ME and RMSE.

2.5.4.5 OBSRES

The following steps were executed for OBSRES:

1. *Overlay soil property observations with OBS map.* This was done with the *over* tool.
2. *Calculate residuals.* Subtract the overlaid values from the soil property values.
3. *Compute variogram*. The residuals were used to calculate a variogram. First, an experimental semi-variogram was fitted with the *variogram* tool followed by *fit.variogram* from the *gstat* package to fit a semi-variogram model.
4. *Krige the residuals*. The residuals were kriged with simple kriging to the same area with the *krige* function.
5. *Combine results*. Add the map with kriged residuals to the map of the OBS prediction map.
6. *Cross-validate results.* Repeat step 4 and use *krige.cv* instead of the *krige* tool. The obtained results were used to calculate the r, ME, and RMSE.

3. Results

3.1 Model input data

3.1.1 Exploratory data analysis

In total 218 pH observations were used. Figure 3.1 shows the spatial distribution of pH sample points in the study area and the boxplot of the data. The pH range is from 3.9 to 9.1, the mean is 5.6 and median is 5.5. The pH value of the 1st quartile soil profiles is 5 and 3rd quartile is 6.2.

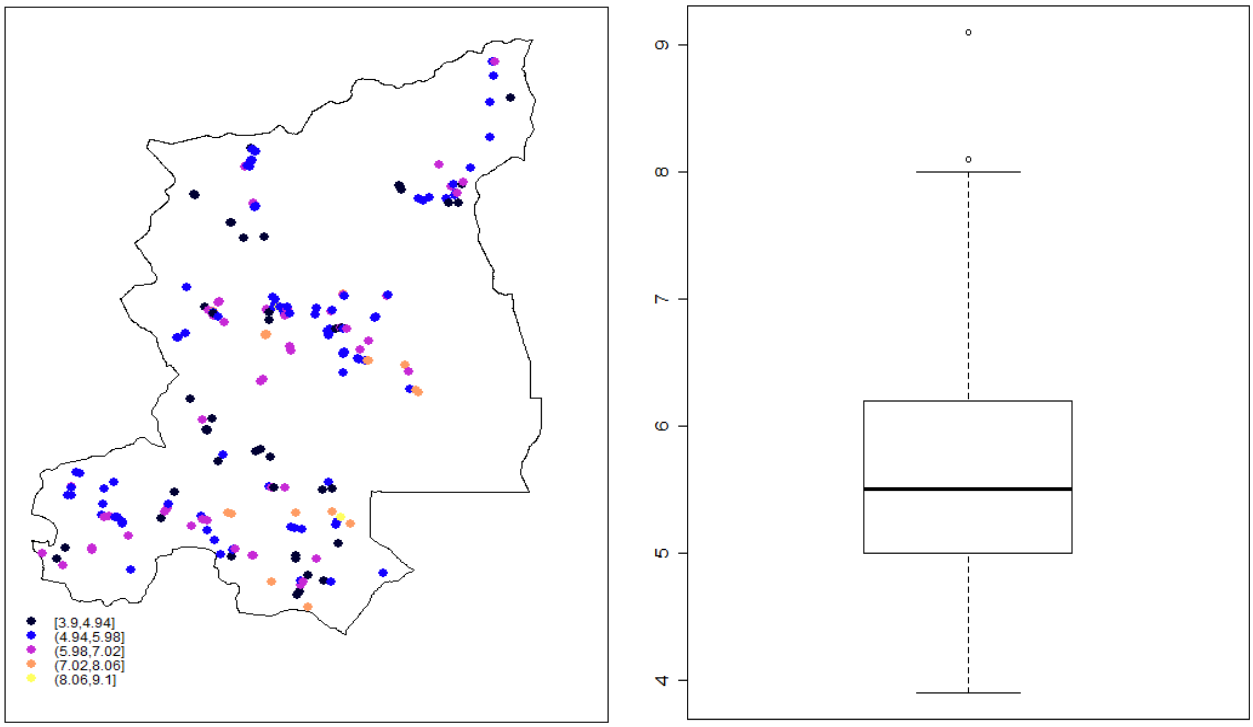


Figure 3. 1 Spatial distribution of pH sample points and box plot of pH values

The spatial distribution of the 206 CEC observations and the corresponding boxplot are shown in Figure 3.2. The CEC value ranges between 1.6 and 29.6 cmol+/kg, the mean is 7.3 and median is 5.5. The CEC value of the 1st quartile soil profiles is 3.8 and 3rd quartile is 8.5.

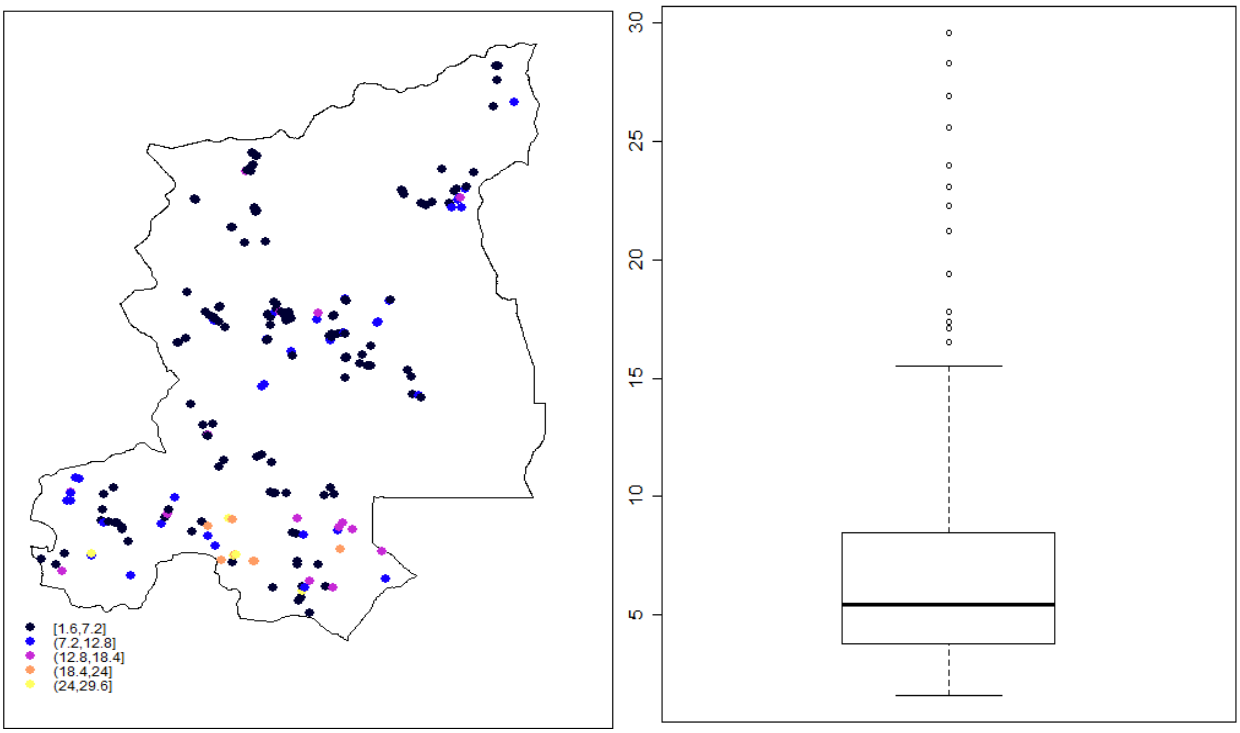


Figure 3. 2 Spatial distribution of CEC sample points and box plot of CEC values

The spatial distribution of the 215 Depth observations and the corresponding boxplot are shown in Figure 3.3. The Depth range is from 3 to 244 cm, the mean is 78.8 and median is 64. The Depth value of the 1st quartile soil profiles is 22.5 and 3rd quartile is 125.

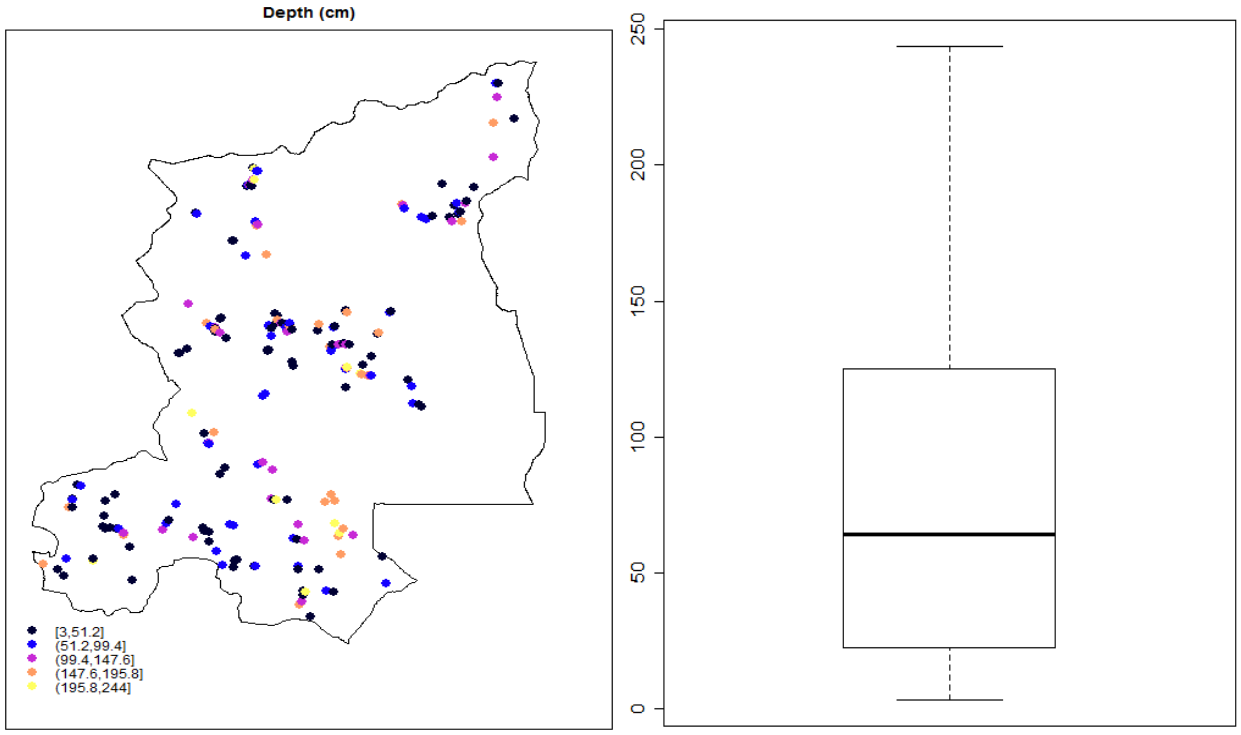


Figure 3. 3 Spatial distribution of Depth sample points and box plot of Depth values

Figure 3.4 shows the histograms of the soil properties and the covariates. It can be observed that CEC, Depth and SLOPE are not normally distributed The CEC data and the SLOPE data were log transformed while for Depth a square root transformation was used. Figure 3.5 shows the histograms of the transformed variables.

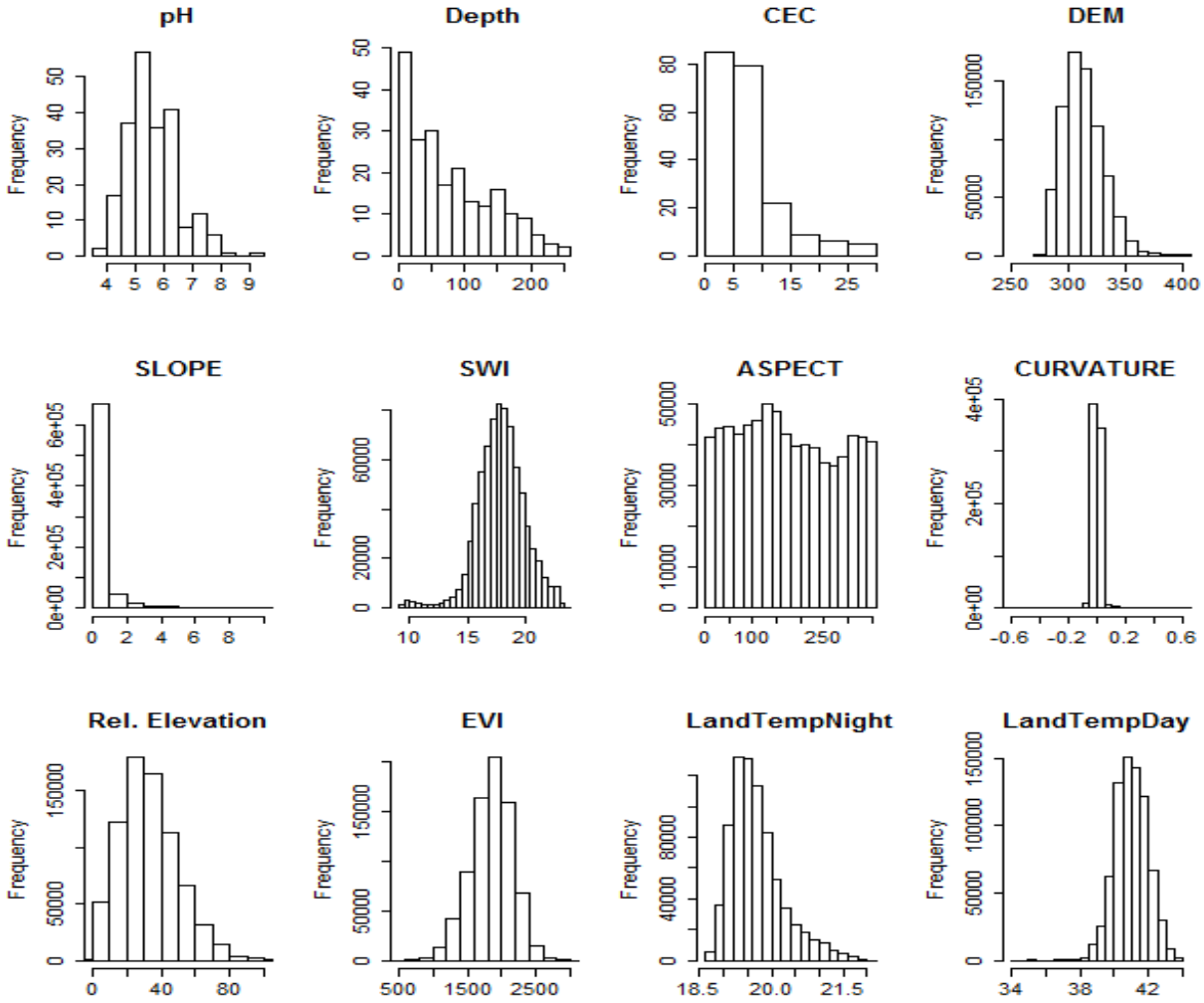


Figure 3. 4 Histograms of the soil properties and the environmental covariates

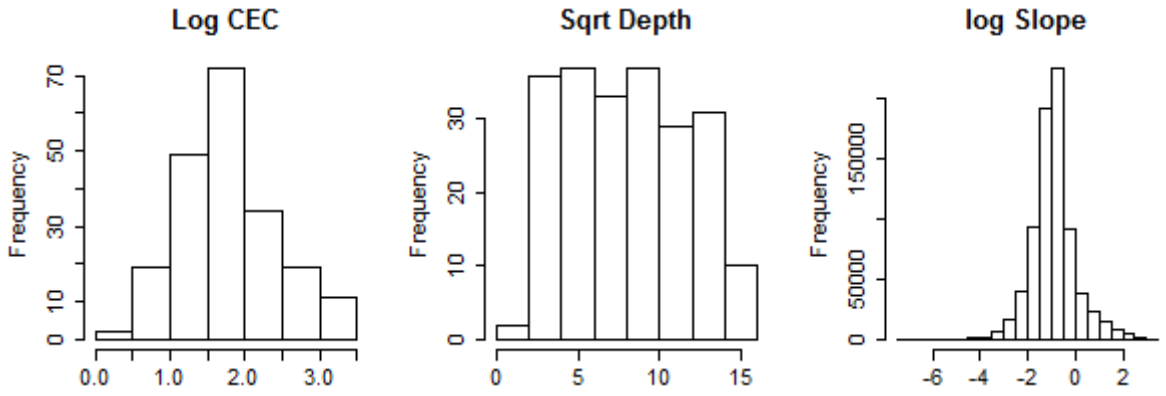


Figure 3. 5 Histogram of the transformed CEC, Depth and Slope data

3.1.2 Results of digitized legacy soil maps

Figure 3.6 shows the SM500K and SM1M soil maps. The SM500K map has 24 mapping units and the SM1M has 10 mapping units. The description of the legend number, which corresponds with the map unit code, for SM1M and SM500K are described in the Appendix 2 in Tables A2. 1 and A2. 3, respectively.

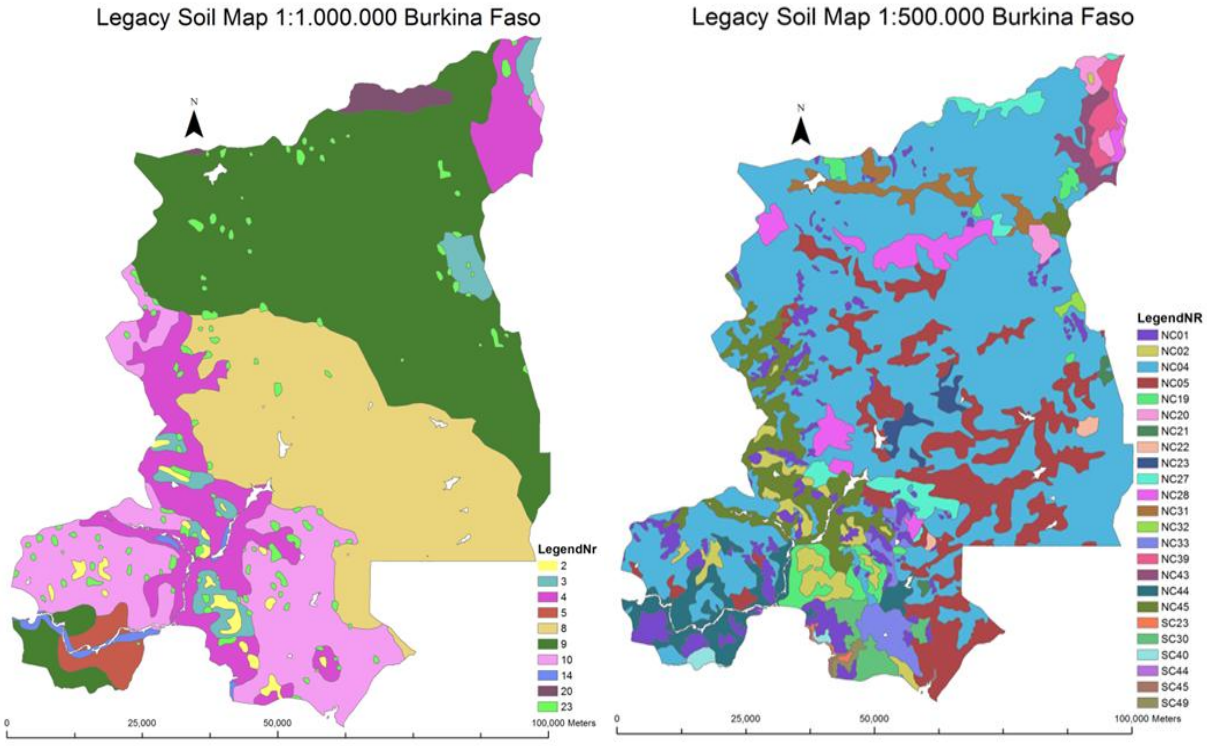


Figure 3. 6 Legacy soil maps SM5000K (left) and SM1M (right)

3.1.3 Results of generalized legacy soil maps

3.1.3.1 Generalized SM1M map

The mapping units of the SM1M map were generalized based on similarity of soil types. The original legend of SM1M is presented in Appendix 2 Table A2.1. Mapping unit with codes 2 and 3 were merged together and the same applies to mapping units 8, 9 and 10. Hereafter a class number was assigned to each of the merged mapping units (see Appendix 2, Table A2.2). Finally, a reclassification was made for the classes that have less than 10 observation points. Since class 3 has only two observations of each soil property it was merged with class 2, which has similar pedogenesis while both units also include vertisols. For similar reasons class 5 was merged with class 1, and class 6 was merged with class 4. Class 7 was merged with class 1, because both comprise raw minerals soils, which are at an embryonal stage of soil development. After reclassification, only 3 classes remained. Table 3.1 shows the generalized classes.

Table 3. 1 Generalized SM1M classes

Original Map Units	Soil Types	Pedogenesis	Nr. Of pH Observations	Nr. Of CEC Observations	Nr. Of Depth Observations	New Class nr.
2, 3, 14 and 23	- Lithosols - Weakly Developed Soils - Brown Soils	- Embryonal	22	20	22	1
4 and 5	- Vertisols - Brown Soils	- Vertic	34	31	34	2
8,9,10 and 20	- Lithosols - Weakly Developed Soils - Ferallitic Soils	- Ferruginous - Embryonal	162	155	159	3

Figure 3.7 shows the generalized SM1M map. Legend with value number 1, which is equal to class 1 of the generalized SM1M . From the map can be seen that class number 3 dominates the study area.

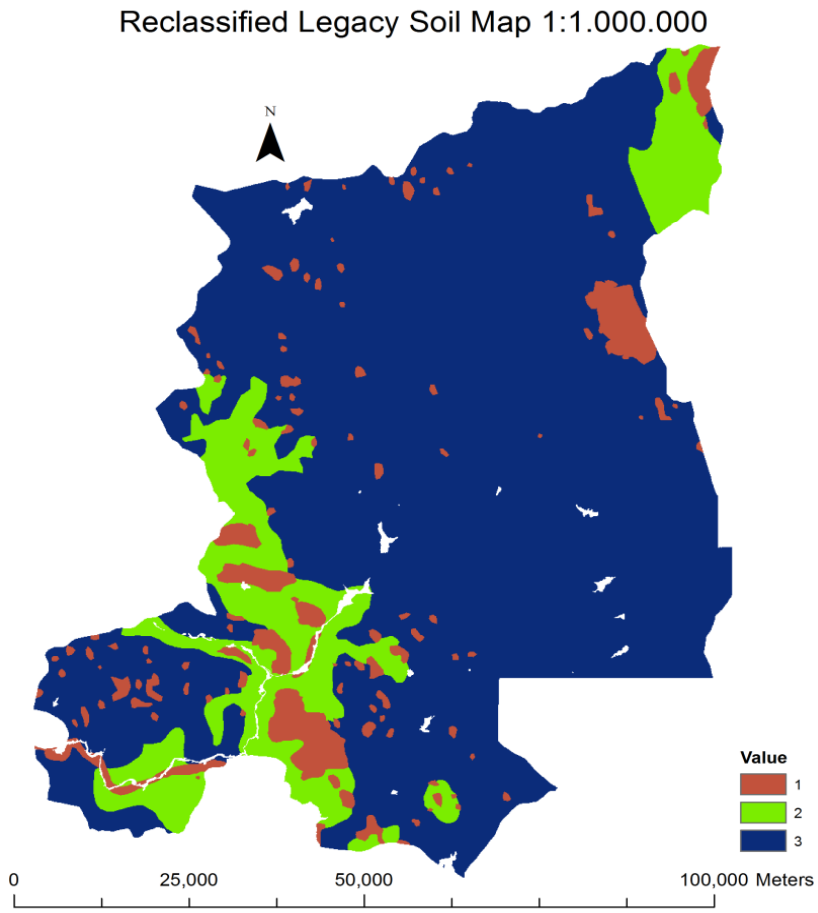


Figure 3. 7 Generalized SM1M map

3.1.3.2 Generalized SM500K map

The original legend of SM500K is presented in Appendix 2 Table A2.3 with the number of soil property observations. The mapping units of the SM500K map were generalized based on hierarchical higher mapping unit followed by the similarity of soil types and/or description, when the former is not applicable. Mapping units belonging to the same main soil type were merged together and a unique class number was assigned to each of the seven generalized soil types (see Appendix 2 Table A2.4). A reclassification was made for the classes that have less than 10 observation points. Class 7 has no observations so this soil class needed to be merged and reclassified into an existing class. Because of the reported similarity of units SC23 and NC01 it was merged with class 1. Class 5 also needs to be reclassified since it has only one observation. The description of class 5 that consist the mapping unit NC39, and SC44 are respectively leached alkali soils associated with gravels and lithomorphic vertisols and lithosols on granite. These description matches most to the description that was provided in class 1 with mapping unit NC01 and NC02 compared to the other class, so class 5 was also reclassified in class 1. To have a numerical order of classes, class 6 was reclassified as class 5. After reclassification, the soil map has 5 classes left as shown in Table 3.2.

Table 3. 2 Generalized SM500K classes

Original Map Units	Soil Class	Nr. Of pH Observations	Nr. Of CEC Observations	Nr. Of Depth Observations	New Class Nr.
NC01, NC02,NC39,SC 23 and SC44	Raw Mineral Soils Vertisols Sodic Soils	26	25	26	1
NC04 and NC05	Weakly Developed Soils	101	95	101	2
NC19, NC20, NC21, NC22, NC23 and SC30	Brown Soils	27	25	27	3
NC27, NC28, NC31, NC32, NC33 and SC40	Ferallitic Soils	28	26	26	4
NC43, NC44, NC45,SC45 and SC49	Hydromorphic Soils	36	35	35	5

Figure 3.8 shows the generalized SM500K map with the legend value number equals to the new class number. From the map can be seen that class number 2 dominates the study area.

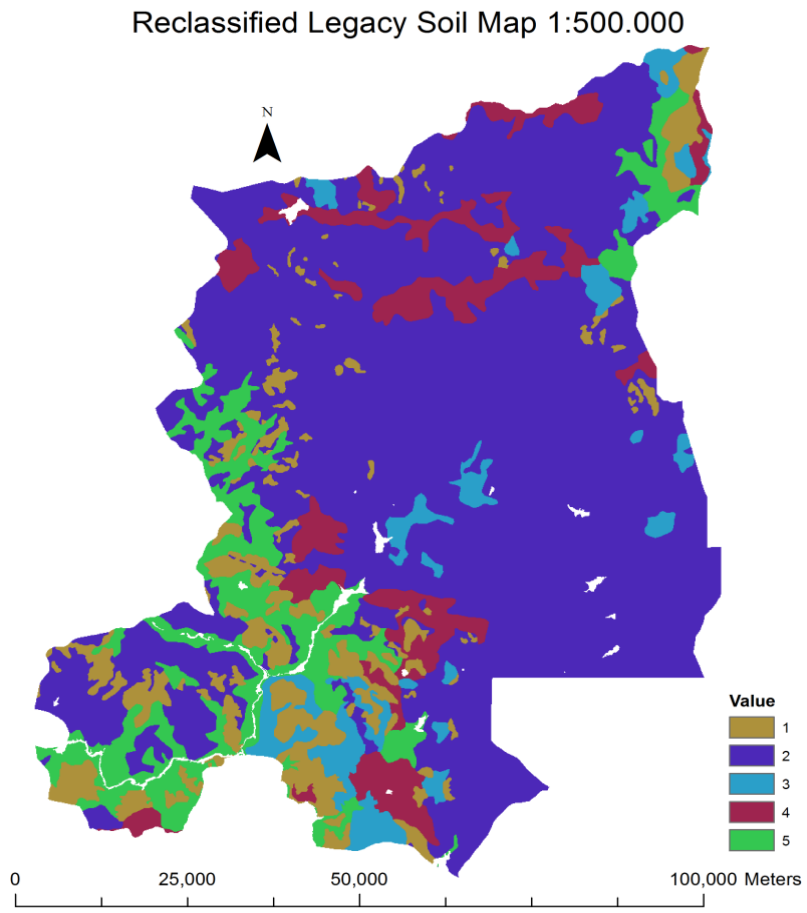


Figure 3. 8 Generalized SM500K map

3.1.3.3 Generalized SM100K map

The map legend of SM100K and the number of accompanying soil properties observations is presented in appendix Table A2.5. As can be seen in the table, mapping units C31, C41, C51, D2, D3 and E has less than 10 observation, so these mapping units needs to be merged. Mapping unit C31 was merged together with C32 because they represent the minor landform, lower slope. The mapping unit C41 was merged with C42 because it has a lower or middle Birimian slope, which was equivalent with mapping unit C42. C51 was merged with C52 due to the similar characteristics of geology and minor landform, the Birimian valleys. D2 were put together with mapping unit D3 because they represent the major landform bottomlands. Mapping unit E is a different individual class and cannot be merged with other mapping units based on the geology and/or landforms. The mapping unit E represents the Aeolian complex which are sand dunes that are formed during the late-pleistocene with a slight relief (van Lieshout et al., 1997). In the same report, 'complex' was described as a wide variation of soil types. The soil type in this mapping unit correspond the most with the mapping unit of D3. Therefore mapping unit E was merged together with D2 and D3. After the merging process, a class number was assigned to the mapping units. Mapping unit W will not have a class number assigned because it is not necessary

in this research and therefore removed. After reclassification, the soil map has 10 classes left as shown in table 3.3.

Table 3. 3 Generalized SM100K classes

Map Unit Code	Physiographic units	Nr. Of pH Observations	Nr. Of CEC Observations	Nr. Of Depth Observations	Class Nr.
AC1	Hills and upper slopes	17	17	17	1
B1	Plateau	13	12	13	2
B2	Eroded or less developed indurated cap	20	19	19	3
C21	Crusted middle slope	12	10	10	4
C22	Non-crusted middle slope	33	31	32	5
C31 and C32	Eroded and non-eroded lower slope	42	42	41	6
C41 and C42	Crusted and non-crusted lower/middle Birimien slope	31	31	32	7
C51 and C52	Crusted and non-crusted Birimien valleys	13	12	12	8
D1	Bottom-land small valley	16	16	16	9
D2, D3 and E	Bottom-land large valley and Aeolian complex	21	16	21	10

Figure 3.9 shows the generalized SM100K map where the legend value number equals to the new class number.

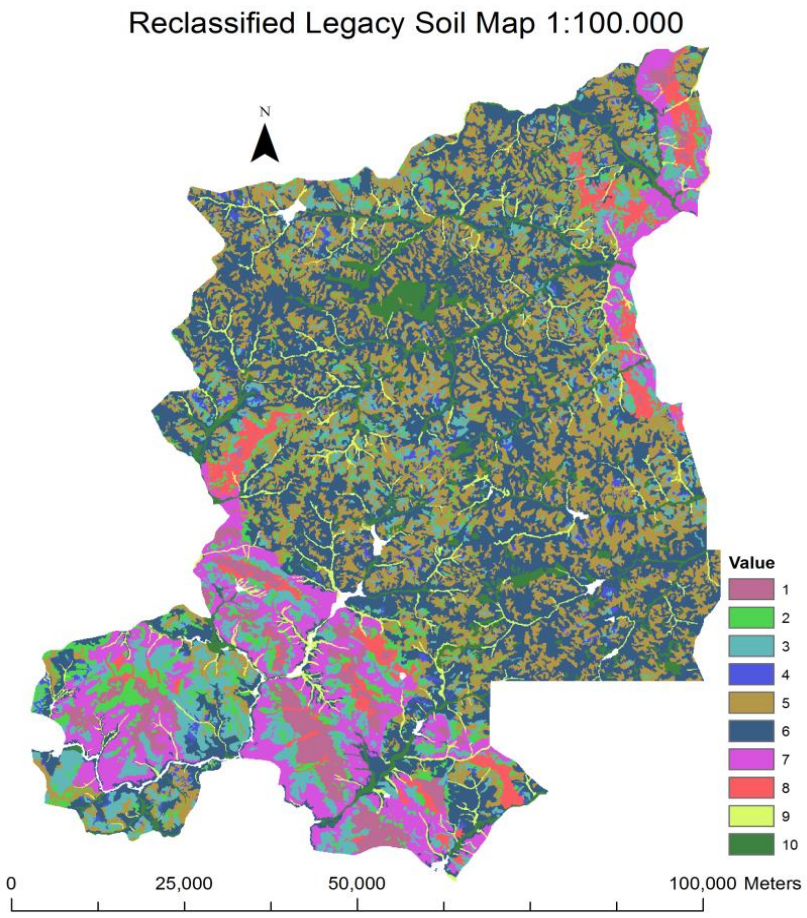


Figure 3. 9 Generalized SM100K map

3.2 Model results

3.2.1 Regression Kriging (RK)

In this section, the intermediate RK results for soil pH are given, while the intermediate RK results for CEC and Depth are provided in Appendix 3. For the latter properties, similar procedures were followed. Table 3.4 lists statistics of the selected covariates and the corresponding regression coefficients obtained by stepwise regression of pH.

Table 3. 4 Statistics of the regression analysis for pH using RK

Variable	Estimate	Std. Error	Significance
(Intercept)	9.008	1.449	2.65e-09 ***
SWI	-0.188	0.063	0.003 **
RELEV	-0.023	0.018	0.005 **
EVI	0.000	0.000	0.063 .

As can be seen in table 3.4, only three covariates were selected by stepwise regression, i.e. only these variables can be used to partially explain the spatial variation in pH in a linear model. The SWI and RELEV variables are relative high significant variables and the EVI less significant. The predictor variables SWI and RELEV have negative regression coefficients while the EVI have a positive regression coefficient. Table 3.5 lists statistics of the regression residuals.

Table 3. 5 pH residuals after regression using RK

Residuals:				
Min	1Q	Median	3Q	Max
-1.98	-0.57	-0.13	0.56	3.56

A variogram was calculated and fitted for the residuals, see Figure 3.10. The variogram has a nugget of 0.40 and a spherical component with a partial sill of 0.43 and a range of 13000 m. The relatively high nugget means that at very small distances there is a large variability in the residuals. The high sill value means that the residuals values have a large mean deviation and shows high variability. The large range means that the points are wider spread compared to each other.

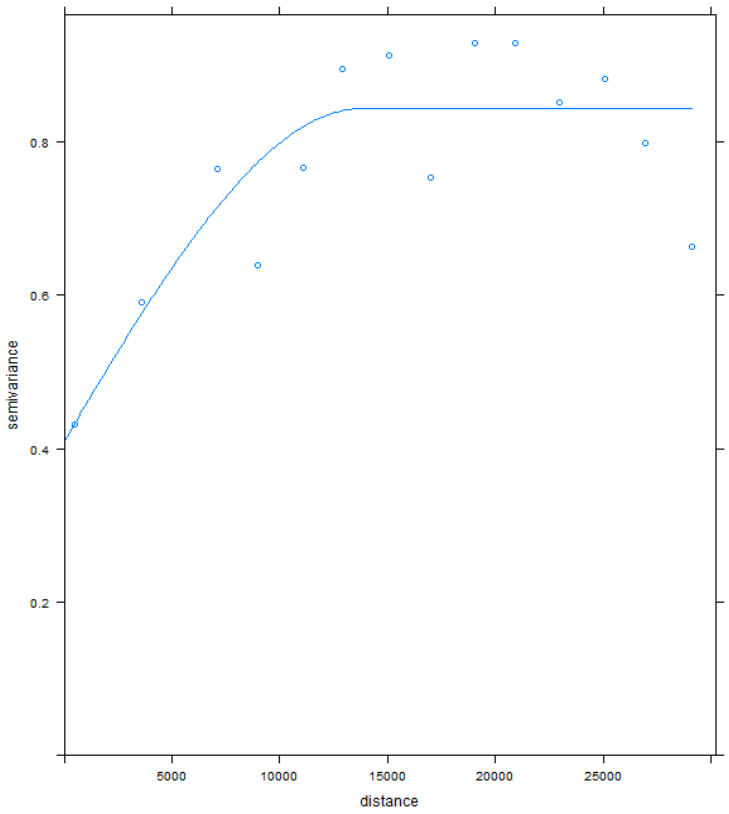


Figure 3. 10 Variogram of pH residuals

The residuals were kriged using the variogram and the obtained results were added to the regressed results to make the final RK prediction. Table 3.6 shows summary statistics of the predicted values for pH, CEC and Depth and Figure 3.11 shows the predicted pH, CEC and Depth map obtained with RK.

Table 3. 6 Summary statistics of pH, CEC and Depth obtained using RK

RK			
	pH	**CEC [cmol+/kg]**	**Depth [cm]**
Min.	2.6	2.4	31.3
1st Quartile	5.4	5.7	71.6
Median	5.6	6.4	80.8
Mean	5.6	6.8	82.4
3rd Quartile	5.8	7.4	92.8
Max.	7.8	24.8	166.9

Table 3.6 shows that the interquartile values of the individual predicted soil properties are close to their mean.

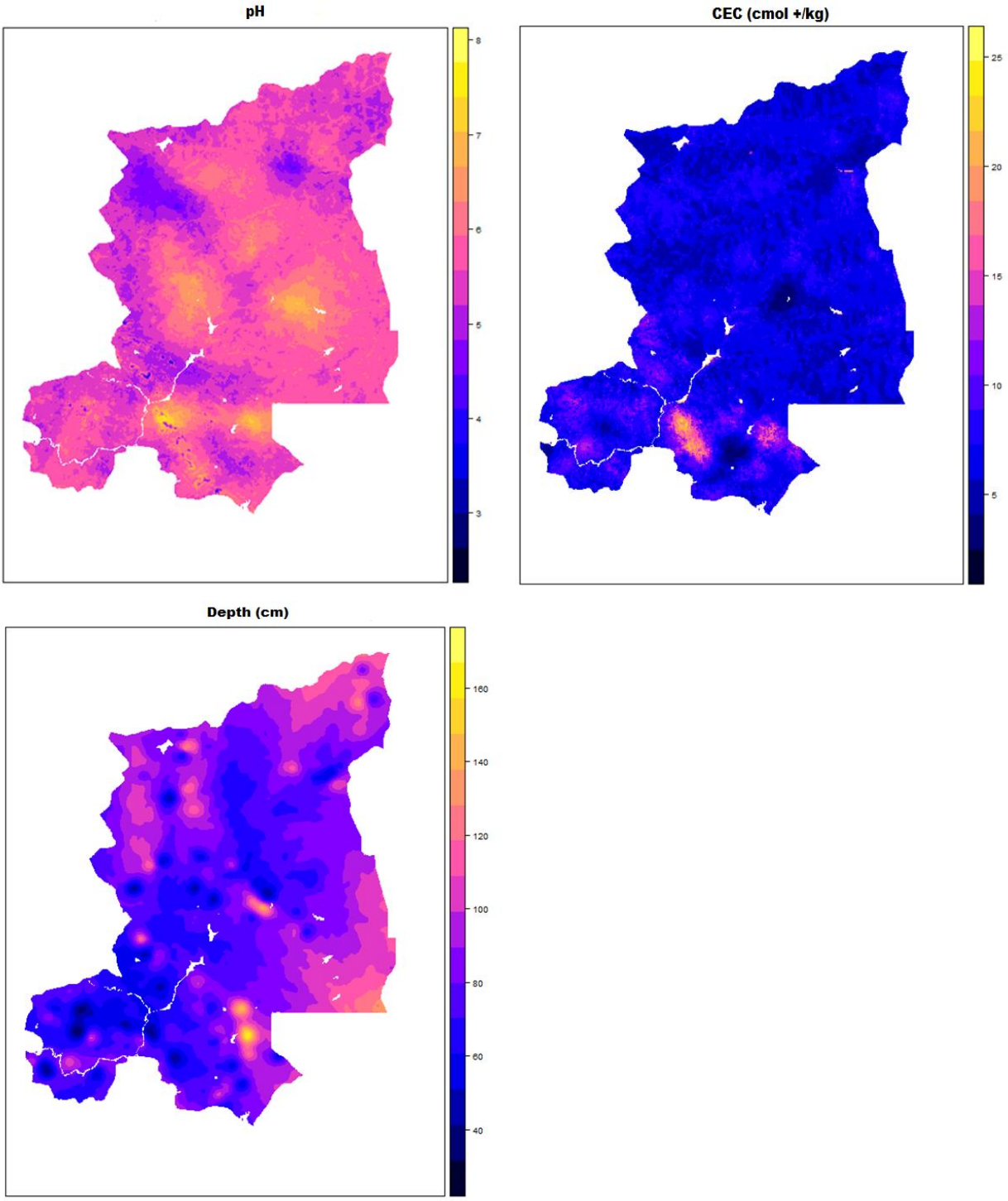


Figure 3. 11 Maps of pH, CEC and Depth predicted using RK

3.2.2 Legacy Soil Map as categorical variable in RK (CAT)

The intermediate results of this method are shown in appendix A3. The method is similar to the RK method, where the only difference lies in the regression part where the legacy soil map is incorporated as categorical variable. Sometimes the legacy soil map was omitted due that they are simply not significant enough to reduce the residual error of the fit for the model. When such case occurs, the legacy soil map is manually incorporated in the regression model even if they are not significant. For the soil property pH the regression model omitted the SM100K and SM500K, for CEC the SM1M and for the Depth SM100K, thus for these model the legacy soil map is manually incorporated.

Table 3. 7 Summary statistics for pH, CEC and Depth predicted using CAT with SM100K, SM500K and SM1M

CAT									
	pH			CEC			Depth		
	SM100K	SM500K	SM1M	SM100K	SM500K	SM1M	SM100K	SM500K	SM1M
Min.	2.2	3.1	2.7	2.1	2.5	2.5	32.6	28.1	27.6
1st Quartile	5.3	5.4	5.4	5.5	5.7	5.7	72.4	69.2	69.2
Median	5.6	5.7	5.7	6.5	6.5	6.4	82.8	79.2	79.4
Mean	5.6	5.7	5.6	6.9	6.8	6.8	83.9	80.1	80
3rd Quartile	5.9	5.9	5.8	7.8	7.5	7.4	95.2	91	91.4
Max.	8	7.6	7.9	35.5	28.9	24.5	157.3	148.8	156.3

Table 3.7 shows summary statistics for pH, CEC and Depth with SM100K, SM500K and SM1M in CAT. The interquartiles of the predicted values for all the soil properties have values that are close to their mean value. The predicted pH, CEC and Depth map using legacy soil maps SM100K, SM500K and SM1M in CAT are shown respectively in Figures 3.12, 3.13 and 3.14.

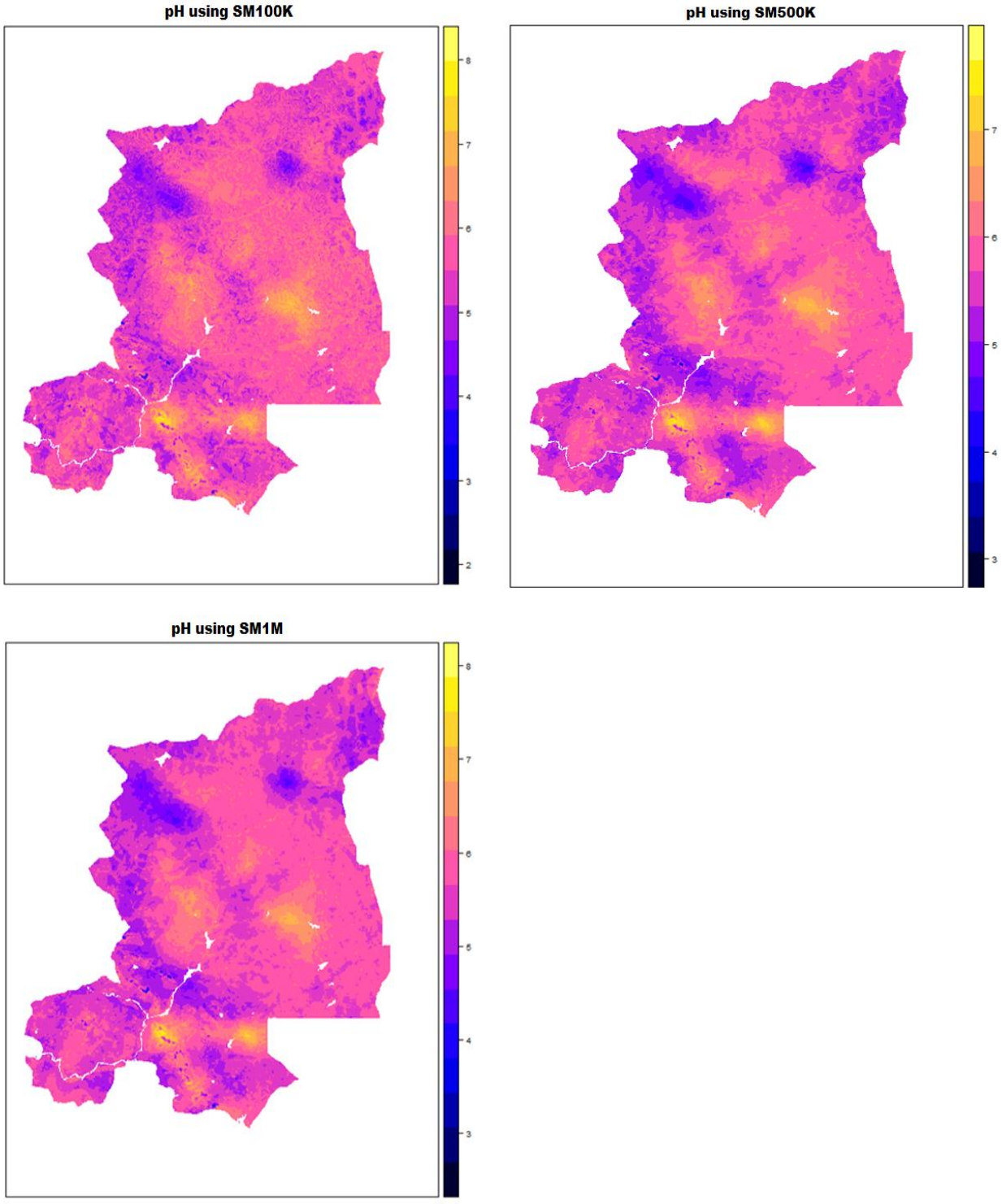


Figure 3. 12 Maps of pH predicted using CAT with SM100K, SM500K and SM1M

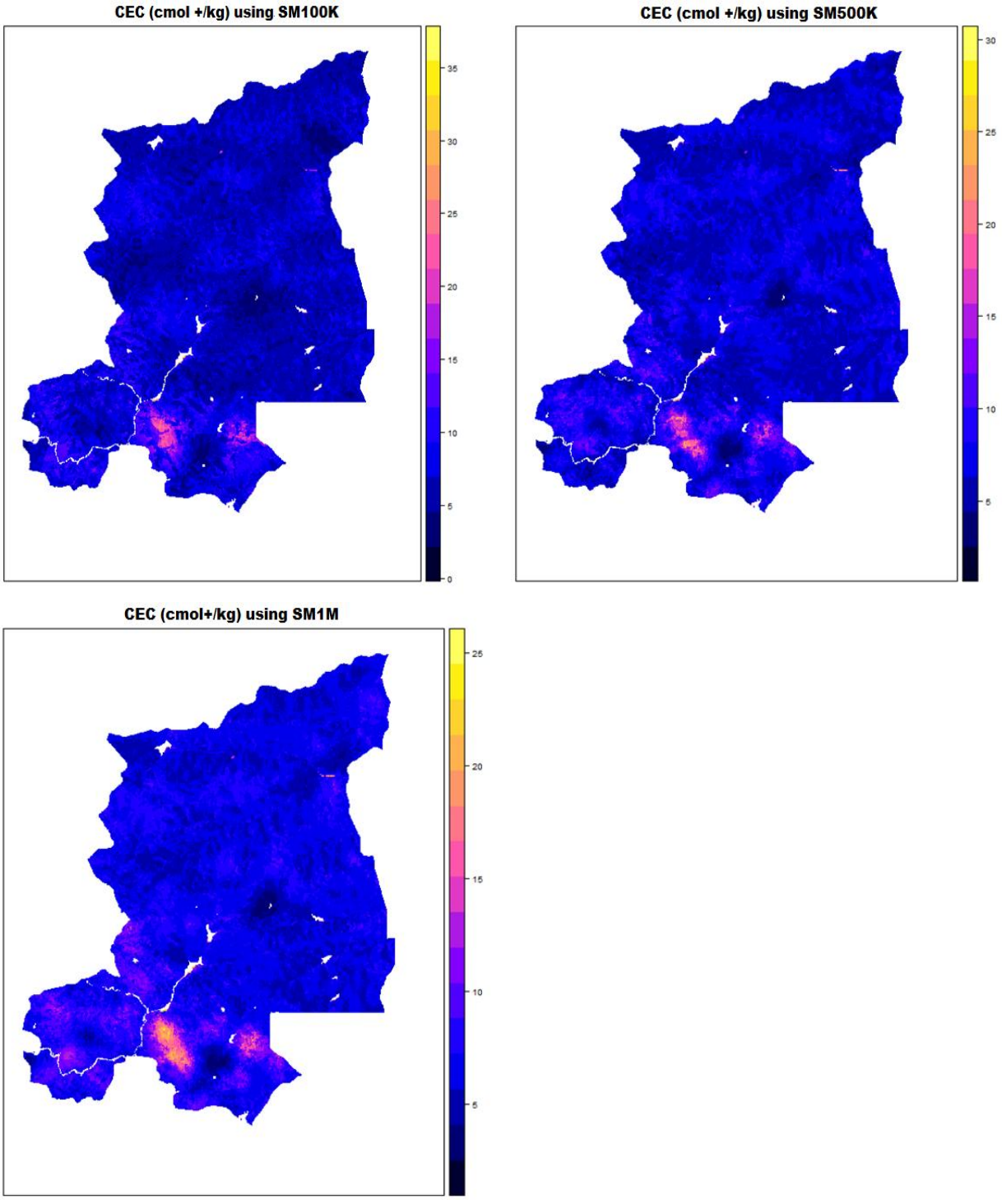


Figure 3. 13 Maps of CEC predicted using CAT with SM100k, SM500K and SM1M

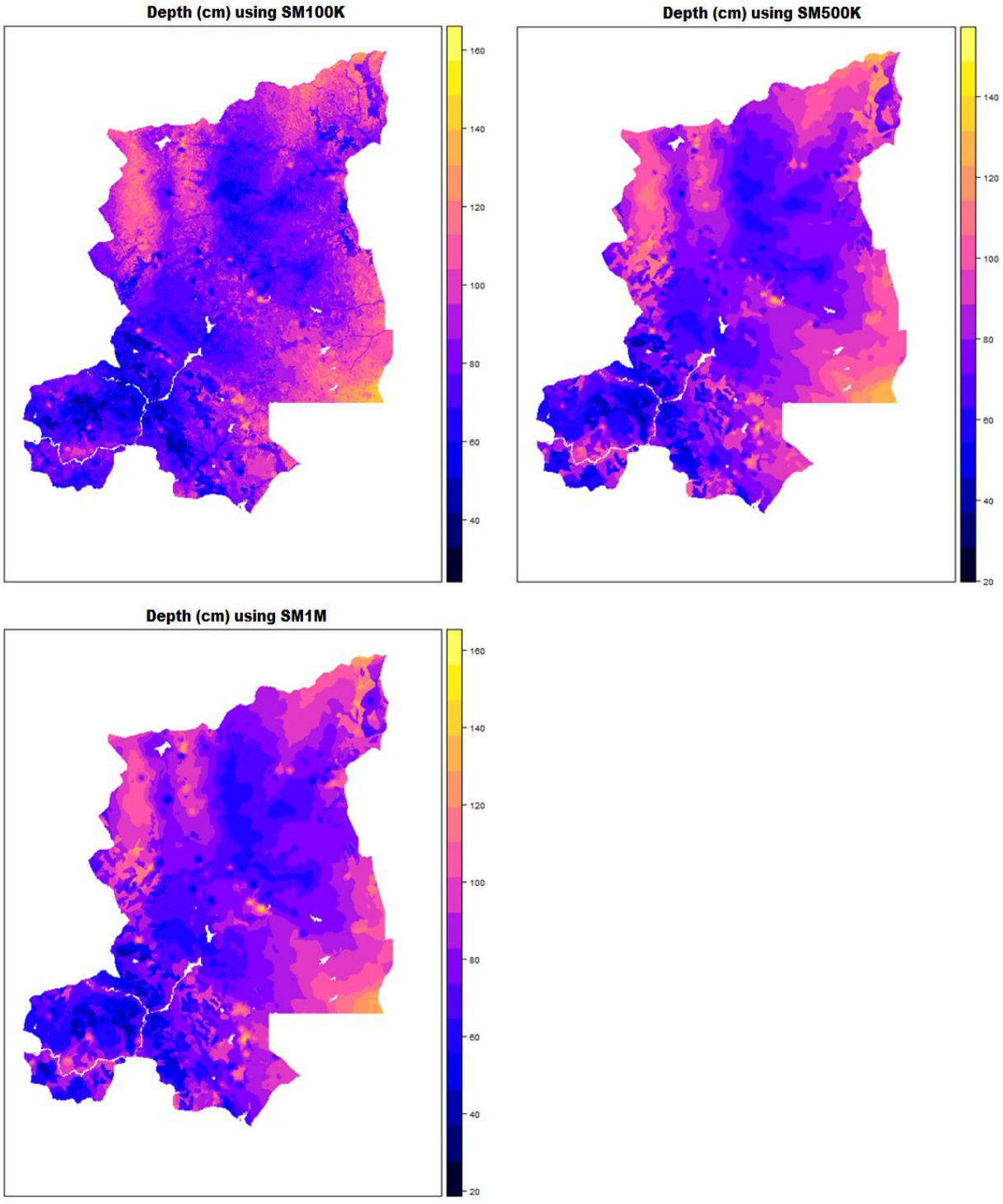


Figure 3. 14 Maps of Depth predicted using CAT with SM100k, SM500K and SM1M

3.2.3 Stratified Kriging (STK) using Legacy Soil Map

The intermediate results of the STK methods (such as variogram) can be found in appendix 3. The summary statistics for pH, CEC and Depth predicted with legacy maps SM100K, SM500K and SM1M using STK are given in table 3.8.

Table 3. 8 Summary statistics for pH, CEC and Depth predicted using STK with SM100K, SM500K and SM1M

STK									
	pH			CEC			Depth		
	SM100K	SM500K	SM1M	SM100K	SM500K	SM1M	SM100K	SM500K	SM1M
Min.	4.4	4.5	4.4	4.2	3.7	3	27.3	38.9	34.9
1st Quartile	5.5	5.6	5.7	6.2	6.9	6.8	81.1	77.5	84.3
Median	5.5	5.8	5.7	6.7	7.2	6.9	84.5	77.5	84.3
Mean	5.6	5.7	5.6	6.8	7.4	7.2	82.9	77.9	80.6
3rd Quartile	5.8	5.8	5.7	6.7	7.3	7.4	89.2	77.7	84.3
Max.	8.2	7.5	7.8	16.8	16.9	20.9	174.6	140.7	125.9

The predicted pH, CEC and Depth maps using legacy soil maps SM100K, SM500K and SM1M STK are shown in figure 3.15, 3.16 and 3.17, respectively. The maps of all the soil properties show smoothing and are less detailed when using coarser resolution legacy data.

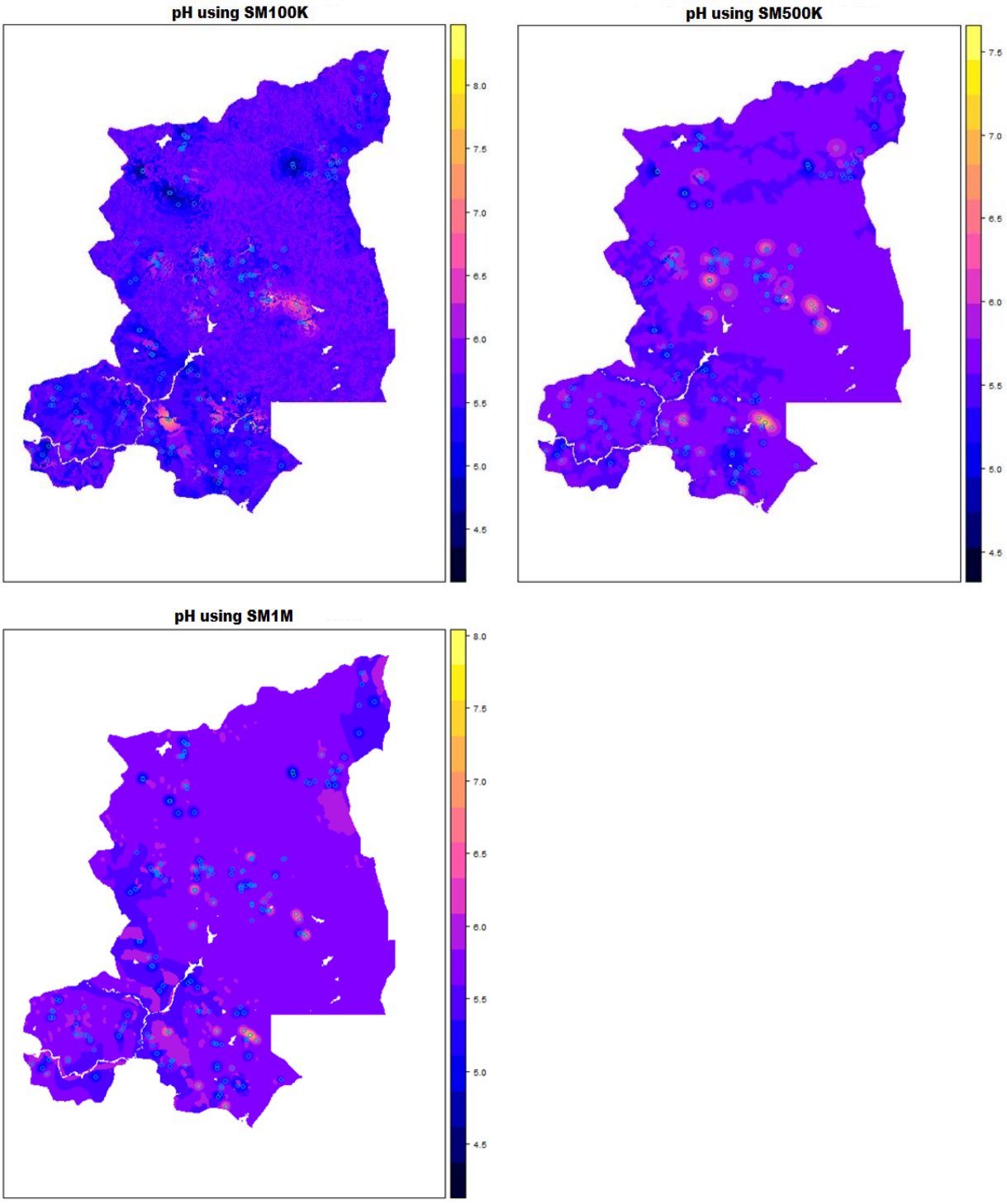


Figure 3. 15 Maps of pH predicted using STK with SM100K, SM500K and SM1M

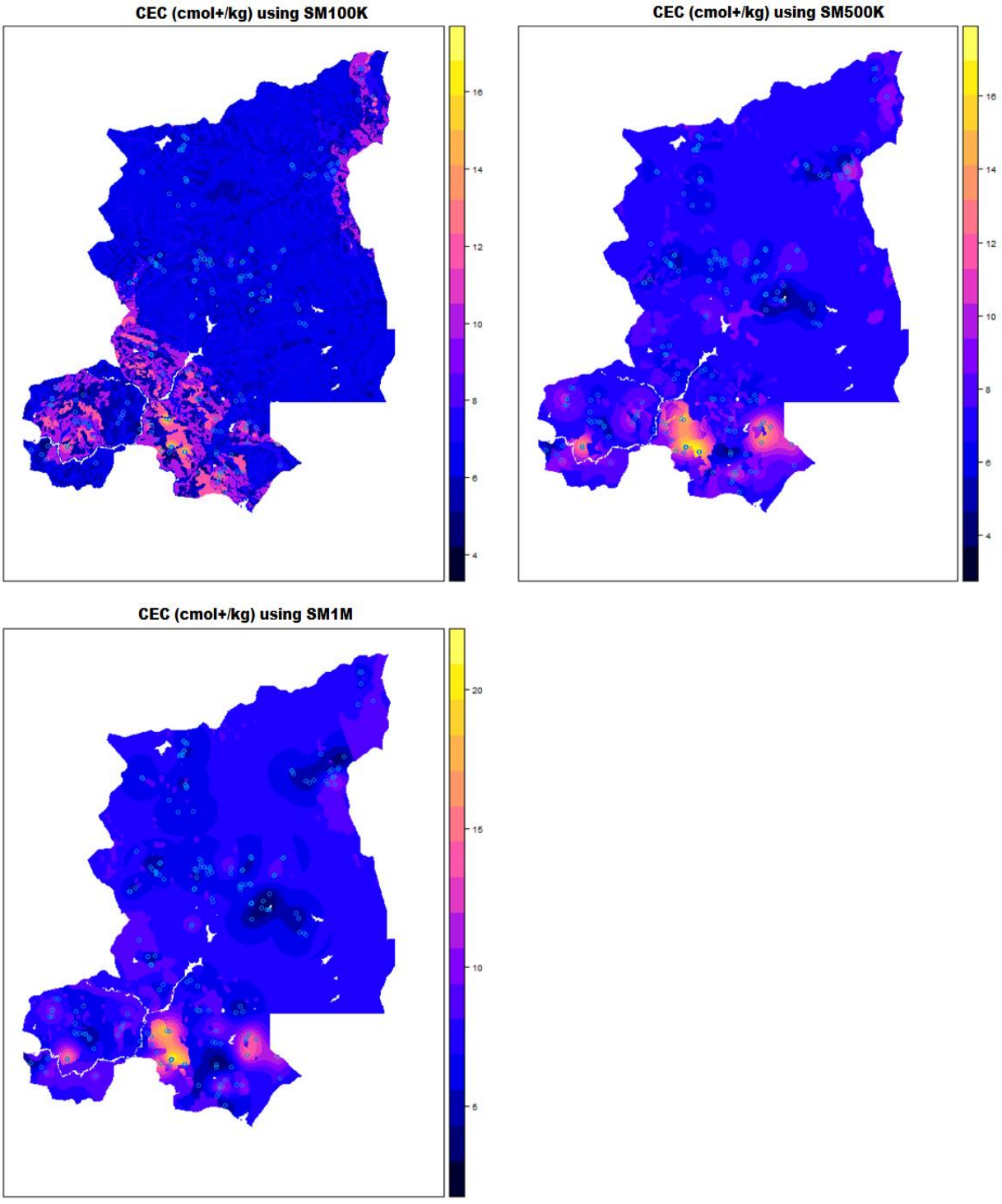


Figure 3. 16 Maps of CEC predicted using STK with SM100K, SM500K and SM1M

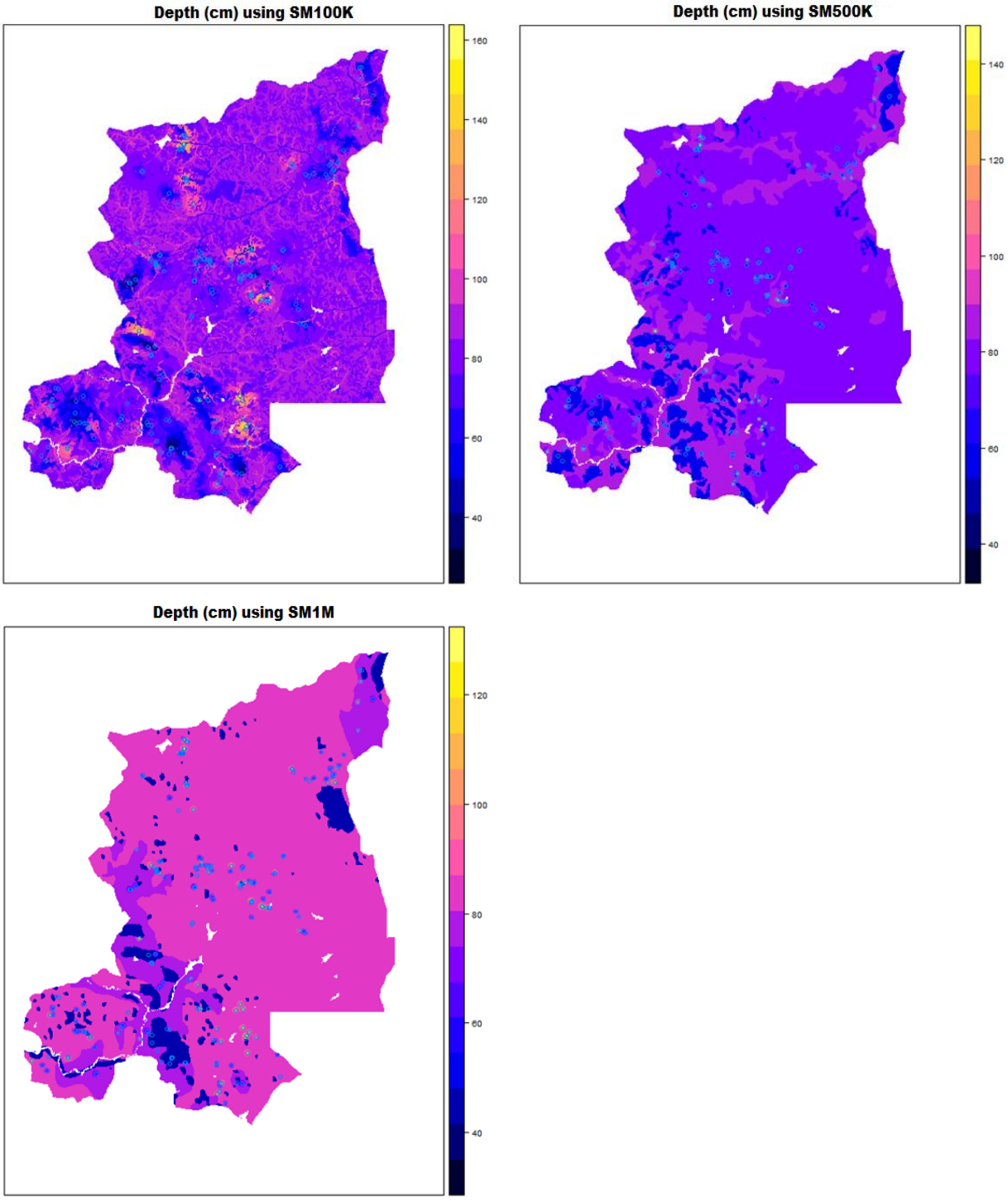


Figure 3. 17 Maps of Depth predicted using STK with SM100K, SM500K and SM1M

3.2.4 Legacy Soil Map as observed trend (OBS)

For this method, the weighted averaged listed in the reports of SM100K and SM500K were mapped. Table A3.23 and A3.24 of appendix 3 give the soil properties values gathered from SM100K and SM500K reports. In some cases mapping units do not have a value because soil profiles (NA) or analytical results (NAR) were lacking. In these cases mapping units were merged and generalized according to Tables 3.2 and 3.3 for SM500K and SM100K, respectively. Tables 3.9 and 3.10 show the soil properties derived from the SM100K and SM500K report, respectively.

Table 3. 9 Weighted mean soil properties values derived from the SM100K report

Map Unit code	pH	CEC (cmol +/kg)	DEPTH (cm)
AC1	7.1	23.3	104.7
B1	8.3	11.0	25
B2	8.3	11.0	19.8
C21	5.5	3.5	102.5
C22	6.4	6.9	86.3
C31	6.5	9.1	125.6
C32	7.1	9.6	121
C41	5.9	8.4	108.3
C42	6.3	11.5	117.7
C51	7.3	25.5	120
C52	7.3	25.5	80
D1	5.9	7.5	100
D2	6.3	10.8	120
D3	6.3	3.7	122
E	6.3	0.8	147.9

Table 3. 10 Weighted mean soil properties values derived from the SM500K report

Map Unit Code	pH	CEC (cmol +/kg)	DEPTH (cm)
NC01	6	6.3	28.3
NC02	6	6.3	28.3
NC04	6	6.7	55.8
NC05	6.7	6.5	123.9
NC19	7.6	11.3	140
NC20 NC21	7.2	9.3	153
NC22	7.6	11.3	120
NC23	7.3	13.8	106.7
NC27 NC28	6.5	4.5	189.4
NC31	6.7	6.3	200
NC32	5.5	5.6	160
NC33	5.5	5.6	160
NC39	6	6.3	28.3
NC43	6	3.5	155
NC44 NC45	7.1	9.5	136
SC23	6	6.3	28.3
SC30	6	6.3	140
SC40	5.9	7.6	164
SC44	6	6.3	28.3
SC45	5.9	7.6	164
SC49	5.9	4.5	171

Table 3.11 shows the summary statistics for the predicted values of pH, CEC and Depth with legacy soil maps SM100K and SM500K using OBS.

Table 3. 11 Summary statistics for pH, CEC and Depth predicted using OBS with SM100K and SM500K

OBS						
	pH		CEC		Depth	
	SM100K	SM500K	SM100K	SM500K	SM100K	SM500K
Min.	5.5	5.5	0.8	3.5	19.8	28.3
1st Quartile	6.4	6	6.9	6.5	86.2	55.8
Median	7.1	6	9.6	6.7	104.7	123.9
Mean	6.9	6.3	9.9	6.9	95	102.2
3rd Quartile	7.1	6.7	11	6.7	121	136
Max.	8.3	7.6	25.5	13.8	147.9	200

The predicted pH, CEC and Depth map using legacy soil maps SM100K and SM500K in OBS are shown in Figures 3.18, 3.19 and 3.20.

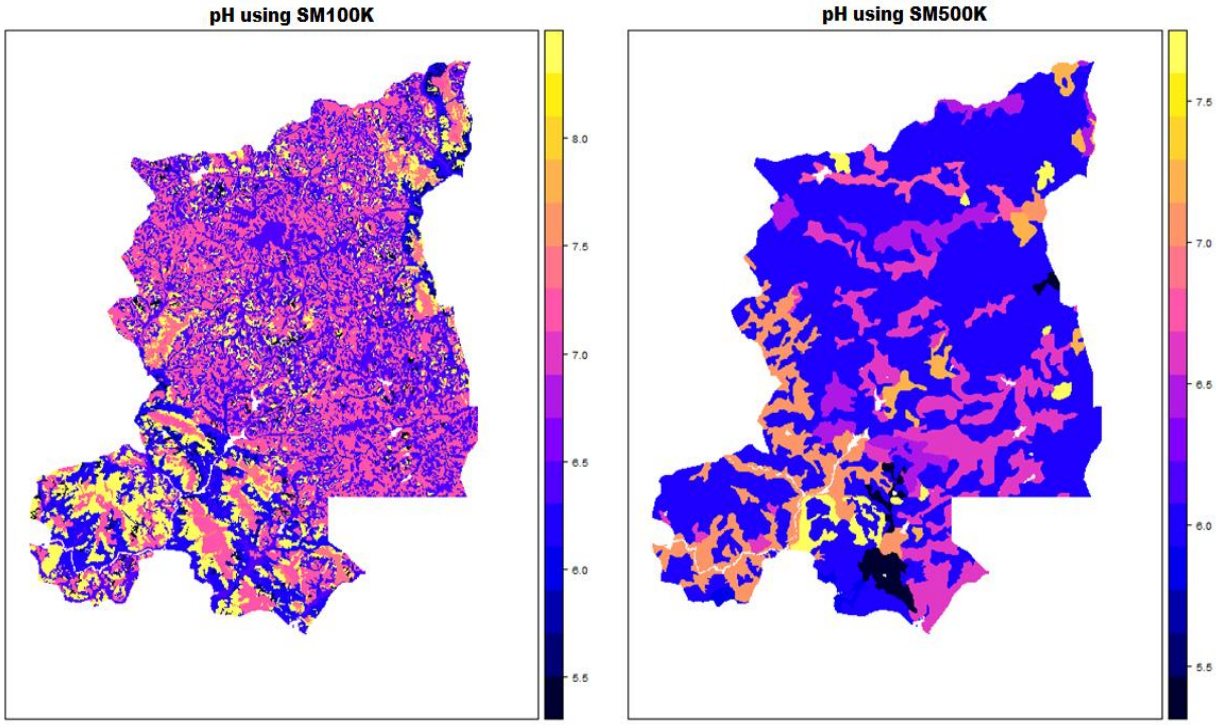


Figure 3. 18 Maps of pH predicted using OBS with SM100K and SM500K

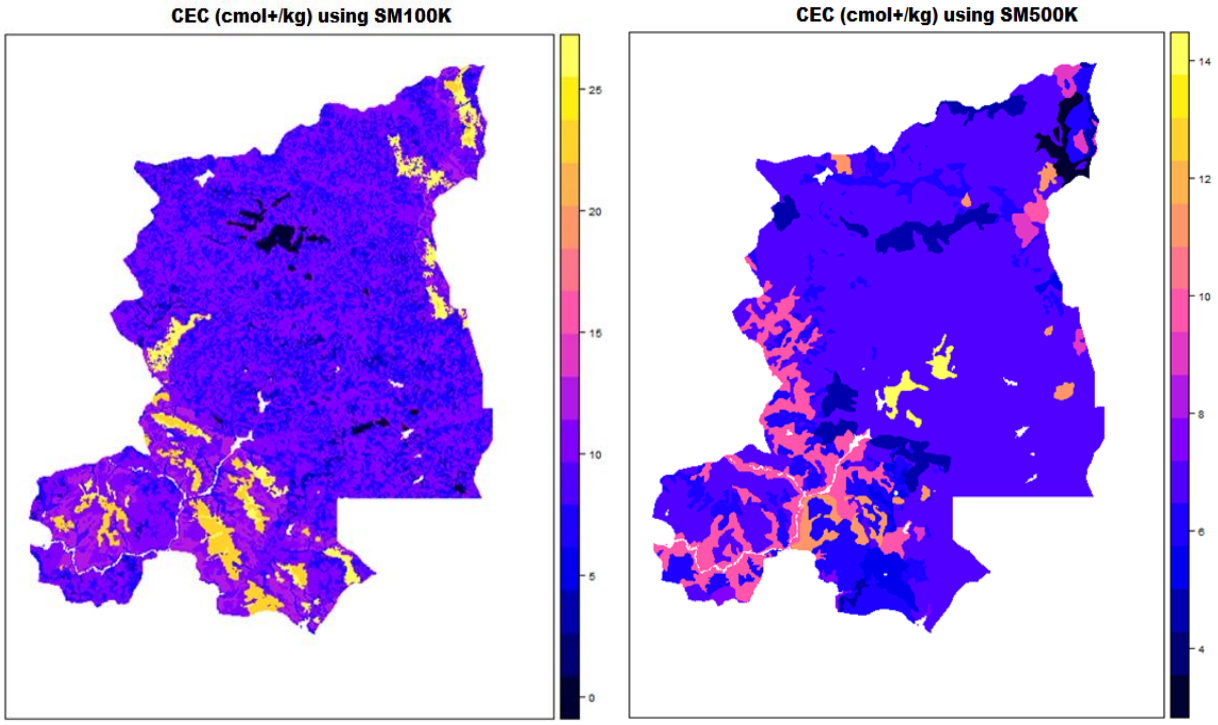


Figure 3. 19 Maps of CEC predicted using OBS with SM100K and SM500K

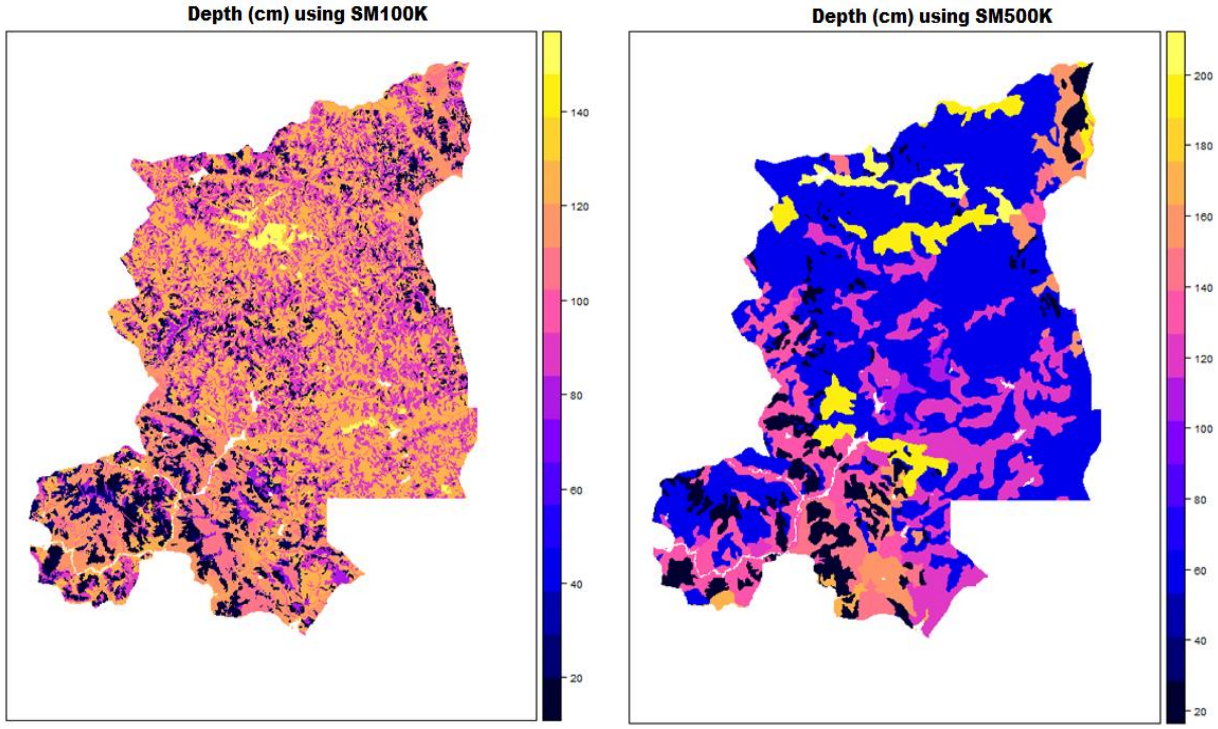


Figure 3. 20 Maps of Depth predicted using OBS with SM100K and SM500K

3.2.5 OBS combined with kriged residuals (OBSRES)

The intermediate results of the OBSRES methods can be found in Appendix 3. Summary statistics of the predicted values of pH, CEC and Depth with legacy maps SM100K and SM500K using OBSRES are given in Table 3.12.

Table 3. 12 Summary statistics for pH, CEC and Depth predicted using OBSRES with SM100K and SM500K

OBSRES						
	pH		CEC		Depth	
	SM100K	SM500K	SM100K	SM500K	SM100K	SM500K
Min.	3.8	4.3	1.3	4.4	33.2	41.1
1st Quartile	6.4	6	8.2	7.8	105.5	74
Median	7.1	6	10.9	8	125.1	74
Mean	7	6.4	11.3	8.3	114.3	104.5
3rd Quartile	7.1	6.7	12.3	8.2	140.3	142
Max.	11	9.8	30.1	17.6	167.5	259.9

Table 3.12 shows the summary statistics of the predicted values of pH, CEC and Depth with legacy soil maps SM100K and SM500K using OBSRES. The predicted pH, CEC and Depth map using legacy soil maps SM100K and SM500K in OBSRES are shown in Figures 3.21, 3.22 and 3.23.

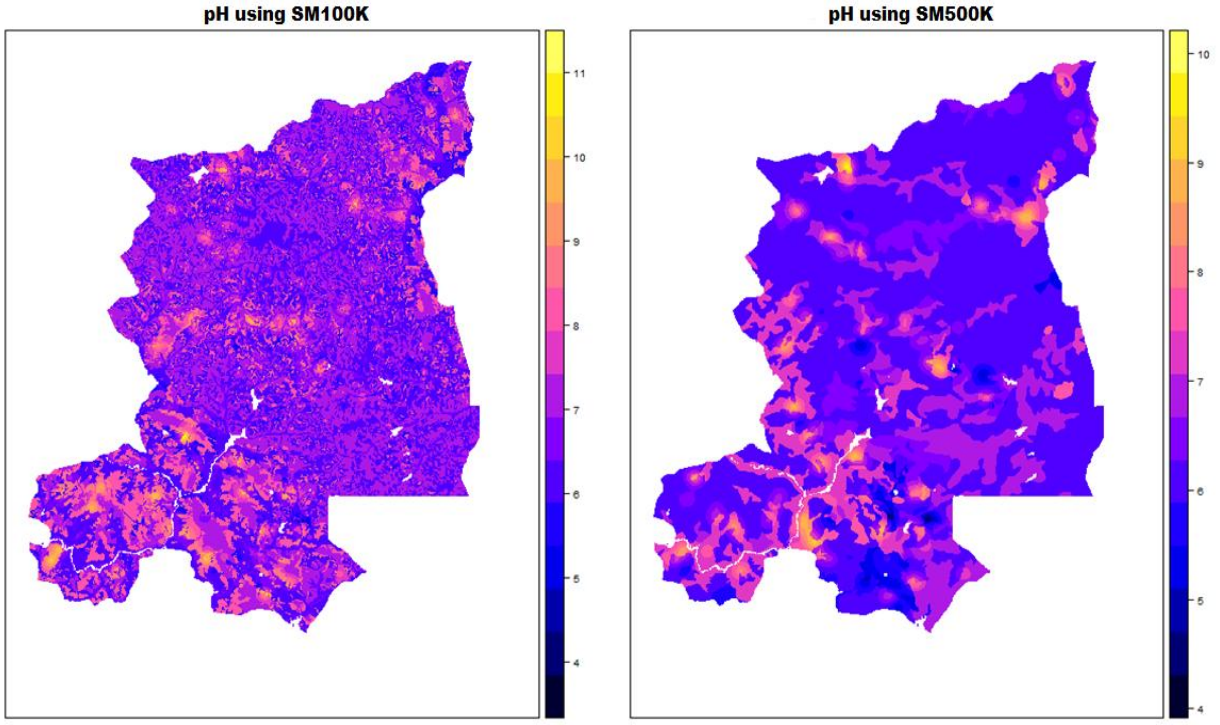


Figure 3. 21 Maps of pH predicted using OBSRES with SM100K and SM500K

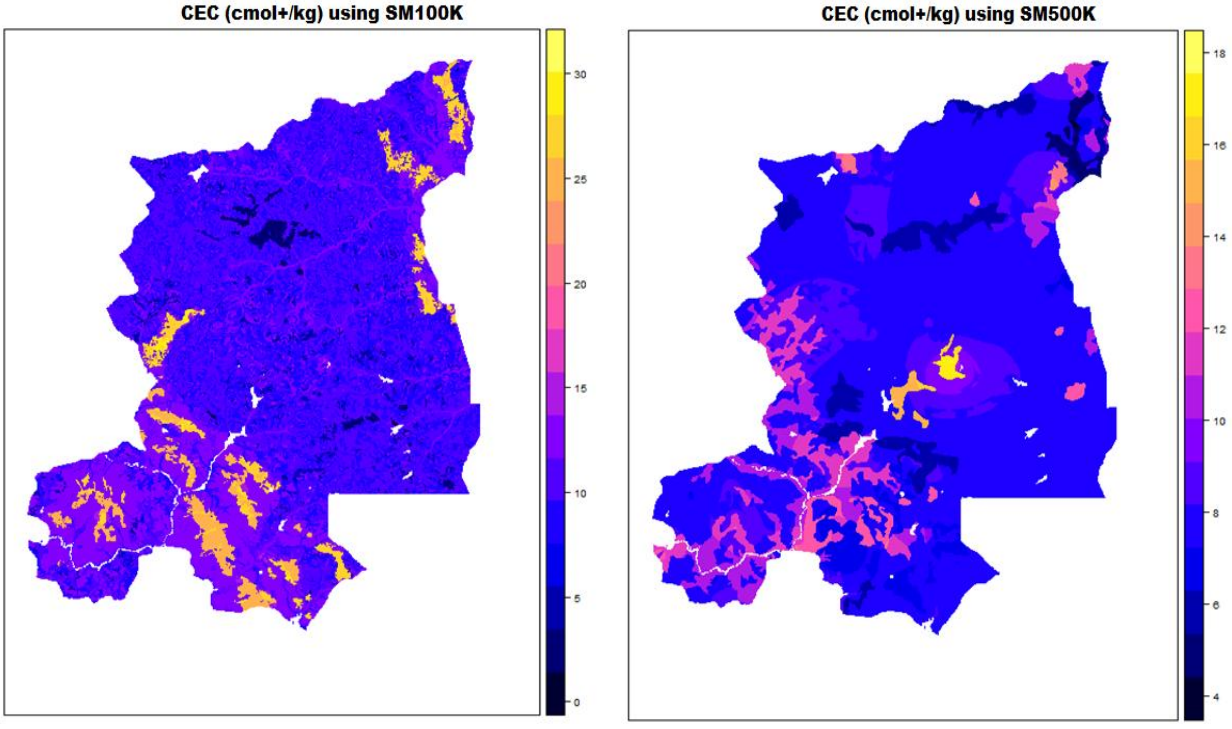


Figure 3. 22 Maps of CEC predicted using OBSRES with SM100K and SM500K

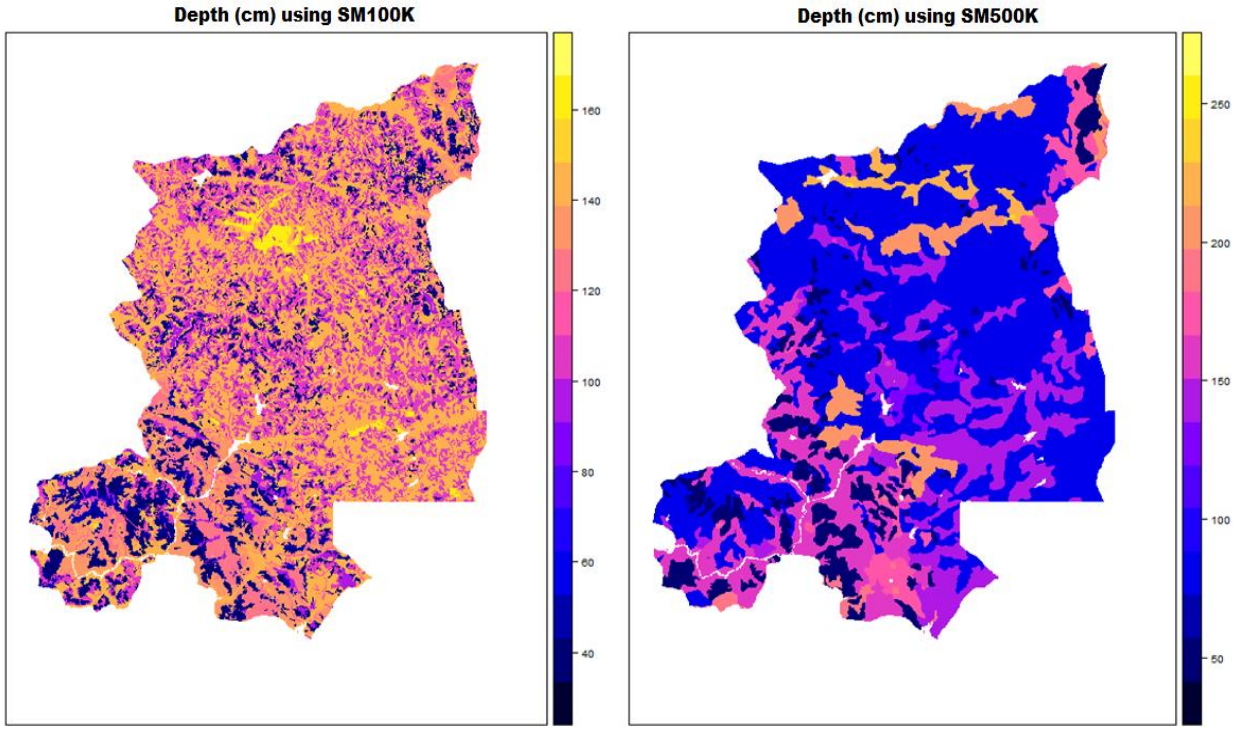


Figure 3. 23 Maps of Depth predicted using OBSRES with SM100K and SM500K

3.3 Assessment of model accuracy

All methods were evaluated by the ME and RMSE. The ME and RMSE of all the methods for pH, CEC and Depth are given in Tables 3.13, 3.14 and 3.15, respectively. For pH, the highest r is found in the CAT method using SM100K. The lowest ME was obtained with the CAT method using the SM1M. The lowest RMSE was obtained with the CAT method using SM100K. The correlation of OBS and OBSRES using both legacy soil maps are negative. The ME of OBS using SM100K and the OBSRES using both legacy soil maps are relatively high. The ME of OBSRES have higher ME compared to the OBS method.

Table 3. 13 Accuracy of pH predictions of all methods

pH											
	RK	CAT			STK			OBS		OBSRES	
		SM100K	SM500K	SM1M	SM100K	SM500K	SM1M	SM100K	SM500K	SM100K	SM500K
r	0.479	0.534	0.492	0.493	0.386	0.349	0.413	-0.166	-0.078	-0.255	-0.256
ME	-0.009	-0.004	-0.007	-0.001	-0.037	-0.002	-0.003	-1.181	-0.823	-1.645	-1.266
RMSE	0.784	0.755	0.783	0.777	0.830	0.837	0.813	1.742	1.365	2.229	1.905

Table 3.14 shows the r, ME and RMSE of the five methods for CEC content. The highest r is found in the CAT method using SM100K. The lowest ME is obtained at the STK method using the SM500K. The lowest RMSE is obtained in the CAT method using SM100K. The correlation of OBS using SM500K and OBSRES using SM500K are negative. The ME of OBS using SM100K and the OBSRES using both legacy soil maps are relatively high. The ME of OBSRES have higher ME compared to the OBS method.

Table 3. 14 Accuracy of CEC predictions of all methods

CEC											
	RK	CAT			STK			OBS		OBSRES	
		SM100K	SM500K	SM1M	SM100K	SM500K	SM1M	SM100K	SM500K	SM100K	SM500K
r	0.541	0.585	0.556	0.541	0.313	0.324	0.392	0.208	-0.002	0.185	-0.089
ME	0.046	0.011	0.048	0.048	-0.012	0.009	-0.035	-3.798	-0.071	-5.434	-1.544
RMSE	4.708	4.537	4.655	4.709	5.335	5.297	5.180	8.089	5.915	9.150	6.345

Table 3.15 shows the r, ME and RMSE of the five methods for Depth. The highest r is found in the RK. The lowest ME is obtained at the STK method using the SM1M. The lowest RMSE is obtained in the CAT method using SM100K. The correlation of STK using SM100K is negative. The ME of OBS and OBSRES using both legacy soil maps are high. The ME of OBSRES have higher ME compared to the OBS method.

Table 3. 15 Accuracy of Depth predictions of all methods

Depth											
	RK	CAT			STK			OBS		OBSRES	
		SM100K	SM500K	SM1M	SM100K	SM500K	SM1M	SM100K	SM500K	SM100K	SM500K
r	0.259	0.256	0.251	0.241	-0.043	0.081	0.174	0.009	0.094	0.010	0.086
ME	0.032	0.093	0.081	0.069	0.066	0.039	0.037	-16.116	-23.396	-35.392	-42.526
RMSE	61.415	61.567	61.466	61.656	66.454	63.761	62.519	73.376	81.012	79.915	90.316

4. Discussion

4.1 Model input: data quality and uncertainty

The model performance is poor for all methods applied, both with and without inclusion of legacy soil maps. One of the main reasons of the poor performance is the low quantity or density of available soil profile data (approximately 1/50 km^2) and the uneven distribution of the data over geographic and covariate space. No data were available for the north-center and west part of the study area. For future improvement, collecting additional samples according to a design that covers geographic and covariate space may well improve the models and the predictions.

Other reasons that contribute to the poor model performance are related to the resolution or scale of the available, and unavailable, environmental covariates. During each preprocessing procedure of the covariates, the uncertainty in the data increases which will affect the accuracy and performance of the models. For example, increasing the resolution of the covariates EVI, LANDCOV, TEMPNIGHT and TEMPDAY. When increasing the resolution, the small-scale information is used to predict information for large-scale image and with every prediction there is an accompanied uncertainty. The fact that the legacy soil maps depict soil associations rather than individual soil types, combined with an even further generalization of the legend units and possible misclassification might have played an important role for the high errors.

4.2 Model results

4.2.1 RK

The RK method uses stepwise regression to select significant covariates based on the AIC criterion and leave out the least significant ones. For soil property pH, only SWI, RELEV and EVI are significant. CEC also has three significant predictors namely ASPECT, LANDCOV and TEMPNIGHT and Depth only two, namely DEM and RELEV. When the SWI value increases the value for pH increases with decreasing SWI and RELEV and increasing EVI. It is remarkable that only few predictors are significant in the regression. This may be due to the combination of the high-resolution predictor variables and the biased or not-randomly spread legacy soil profile locations and the varying accuracy of geo-referencing of the profile data. High-resolution predictors may give much detail, if informed by sufficient well-distributed accurately geo-referenced profile data, otherwise it may lead to insignificant relations with the soil profile data. The predicted interquartile values of each soil property are close to the mean value, this is due to that kriging predict values that are close to the mean based on the relative short range of the fitted variogram.

4.2.2 CAT

The CAT method is the same as the RK whereas the legacy soil map is used as a covariate. For the soil property pH the regression model omitted the SM100K and SM500K, for CEC the SM1M and for the Depth SM100K, thus for these model the legacy soil map is simply not significant enough to reduce the residual error of the fit for the model. For pH, CEC and Depth, the predicted values including legacy soil maps are hardly different from the values predicted when not including the legacy soil maps (RK), which means that none of the soil maps (both considered significant and insignificant) contributes much to the prediction, except for the predicted ranges (min-max values). This may due to the scale of the legacy soil maps with map-units that depict soil associations rather than individual soil types.

4.2.3 STK

The pH values predicted with the three legacy soil maps using STK do not differ much from each other. Similar for the predicted CEC values. The legacy maps had impact on the predicted maximum value for CEC. The Depth predicted values using SM100K was overall higher than the other predicted Depth values with the other soil maps except the minimum values which was the lowest of the other soil maps. This might be cause by soil properties with high values, which tend to contribute more in a large-scale legacy soil map, which consist more polygon and small area than in a small-scale legacy map. The reason for this is when predicting an unknown location of a large-scale map, there is a possibility that there are few observations in a smaller area or polygon and when there are fewer observations than the high value tend to contribute more on the prediction for the unknown areas. In a small-scale map, the area for prediction is relative larger and there are relative more dense observations which minimized the contribution of the high values to the prediction. This also applies to low values. However, there are exceptions when the observation values consist either of only high or low values. An example of this is the CEC using SM100K, which contain only high value. Although these values were standardize before it was use to predict the unknown area, to obtain better results for the CEC and Depth it is recommended to normal distribute these soil property values prior to standardizing.

4.2.4 OBS

The results obtained with the OBS method may be biased or maybe even unrealistic because data of representative soil profiles are averaged and used to represent large areas with heterogeneous soil type assemblages. For the SM100K weighted averages could be assessed, when the percentage proportion of soil types in a mapping unit was provided in the report, which improves the prediction. All together, the soil profiles have been originally identified, after augering, as representative for the soil types occurring in the mapping units, making OBS the reference approximation of the geographic soil property value distribution in the area.

4.2.5 OBSRES

The obtained predicted OBSRES results are mostly higher than the OBS predicted results. This may be due that the residuals obtained from the difference between the observed AfSP soil property value and the OBS soil property value are not normally distributed. Transformation of the residuals might lead to better predictions. The variogram in Figure A3. 15 and A3. 16 show that the pH and Depth, both for the SM100K and SM500K, have high nuggets indicating high short-distance variation, or measurement errors.

4.3 Model comparison

The validation statistics suggest that the CAT method using SM100K is the best method for predicting the soil properties pH and CEC. The Depth was best predicted by the RK and CAT methods using SM100K, giving similar results, but RK resulted in a lower ME, which makes RK the best method for predicting the Depth. Although the SM100K was not significant and the SM1M was significant during the regression for pH (same explanation applies for Depth) in the CAT method, the SM100K performed slightly better than the SM1M. This was due to the manually incorporation of the SM100K which influence the significance of other regression coefficients. For the STK method, the use of SM1M predicted better compared to using the other soil maps. This is due do the fact that the coarse scale of the SM1M map implies a smaller number of mapping units with a larger number of observations in the unit. More observations reduce the prediction error and increase the accuracy. The OBS method using the SM500K predicted pH and CEC well compared to using the SM100K probably due that SM500K is a soil map and the SM100K is a physiographic map with soil association legend. However, the correlation between the observed and predicted values, using the SM500K is lower than when using the SM100K. Similarly, the OBS method using the SM100K predicted Depth better in terms of ME and RMSE than when using the SM500K while the correlation between the observed and predicted values is lower. This is due to SM100K is calculated with a weighted mean that considers the proportion of soil association in a map unit whereas the SM500K is calculated with the arithmetic mean that assumes that a map unit is equally distributed with different soil which might not be the true. The correlation results for the OBSRES method, using the SM100K and SM500K, are better than those for the OBS method. However the ME and RMSE are higher compared to the OBS method this may be due that the residuals, that was kriged and combined, are not normal distributed. Overall, the OBS and OBSRES methods produced the worst results, according to validation statistics. However, the judgment of which method predicts best may be biased because the results of all methods were cross-validated with the same AfSP data. An independent validation data set would be very useful to make an empirical judgment.

5. Conclusions

The aim of this study was to devise methods for including legacy soil maps in digital soil mapping to improve the spatial prediction of soil properties as pH, Cation Exchange Capacity (CEC) and soil depth, which was tested for the area of Sanmatenga province, Burkina Faso. To achieve this goal the following five research questions are answered:

1. **Which methods can be used to include legacy soil maps in DSM to model the relationship between soil properties and environmental covariates?**

   There are several ways to include legacy soil maps in DSM to model the relationship between the selected soil properties and environmental covariates. Five statistical methods were identified and used in this study, namely 1) Regression kriging without use of legacy soil map information is used as a reference method (RK); 2) Regression kriging with legacy soil map as categorical variable (CAT); 3) Stratified kriging using the delineations of the legacy soil map as map unit boundaries (STK); 4) Legacy soil map used as observed information obtained from the accompanying report (OBS); and 5) Combining the OBS result with kriged residuals (OBSRES). These methods are explained in section 2.2.

2. **How can the accuracy of the results of each method be assessed?**

   The accuracy of each method was assessed with the r, ME and RMSE, obtained from the leave one out cross-validation. The method for accuracy assessment is described in section 2.4. Note that the reference for this accuracy assessment is the available AfSP data.

3. **How can these methods be implemented in R software?**

   The methods RK, CAT, STK, OBSRES were successfully implemented in R with the following packages *gstat*, *raster*, *rgdal*, *sp* and *maptools*, whereas OBS was performed in ArcGIS 10.1. The details of how these methods were implemented are briefly described in chapter 2.5 and the used R scripts are given in Appendix 4.

4. **Which results are obtained when the methods are applied to the area of Sanmatenga province with different legacy soil map scales?**

   The results are given in chapter 3.2. Near similar statistic results were obtained when using the different map scales in the CAT method compared to the RK method. The only difference was in the minimum and maximum values. Maps of pH, CEC and Depth using CAT with the legacy soil maps show smoothing and clear indication of the extreme values as the scale of the legacy soil map decreases. The predicted pH and CEC map using SM100K in STK show more detail compared to the other legacy maps. The OBS and OBSRES methods both resulted in different predictions when using different legacy soil maps (SM100K and SM500K) which have different scales and legends.

5. **Which of the selected methods produces the most accurate soil maps and how accurate are these maps compared to DSM without using legacy soil maps?**

   In terms of r, ME and RMSE the CAT method using the SM100K performed best to predict soil properties pH and CEC. Soil Depth is best mapped by the RK method, this without using legacy maps. The performance though is only slightly better than the CAT method using legacy map SM100K. The OBS and the OBSRES methods produced the worst results according to cross validation. Although no methods performed well they are not considered wrong either, because the methods are modeled in different ways with different type of soil maps with different legends and scales. Moreover, the methods were built with only few legacy soil observations with varying accuracy of geo-referencing and different legacy soil maps with inaccuracies and different legends for different degrees of generalization, which possibly is an insufficient basis to construct a qualitative prediction model for this area with high variability at short distances. However, the use of the selected methods create a basis for a possibly more accurate DSM using legacy soil maps for the province in Burkina Faso, which should be extended to include a much denser soil profile dataset and more adequate covariates including more detailed soil information from the legacy soil maps.

References

AfSIS, 2013a, Africa Soil Information Service (AfSIS) MODIS Data Sets: Land Surface Temperature (LST) Night Long-term and Monthly Averages, and Variance. Palisades, NY: Center for International Earth Science Information Network (CIESIN), Columbia University. http://www.africasoils.net/data/datasets. Accessed on 08 11 2013

AfSIS, 2013b, Africa Soil Information Service(AfSIS) MODIS Data Sets: Land Surface Temperature (LST) Day Long-term and MonthlyAverages, and Variance. Palisades, NY: Center for International Earth Science Information Network (CIESIN), Columbia University. http://www.africasoils.net/data/datasets. Accessed 08 11 2013.

Asten, P. J. A., Pol, J.V.D., 1995, Carte de Reconnaissance Physiogeographique du Nord de la Province du Sanmatenga , Burkino Faso. (General Physiogeography Map of Sanmatenga Province, Burkina-Faso).

Behrens, T., Förster, H., Scholten, T., Steinrücken, U., Spies, E. D., and Goldschmitt, M., 2005, Digital soil mapping using artificial neural networks: Journal of plant nutrition and soil science, v. 168, no. 1, p. 21-33.

Bello-Pineda, J., and Hernández-Stefanoni, J. L., 2007, Comparing the performance of two spatial interpolation methods for creating a digital bathymetric model of the Yucatan submerged platform: Pan-American Journal of Aquatic Sciences, v. 2, no. 3, p. 247-254.

Bohling, G., 2005, Introduction to geostatistics and variogram analysis: Kansas geological survey, 20p.

Boucneau, G., Meirvenne, van, M., Thas, O., Hofman, G., 1998, Integrating properties of soil map delineations into ordinary kriging: European Journal of Soil Science, v. 49, no. 2, p. 213-229.

Boulet, R., 1968a, Carte Pédolgoqique de Reconnaissance de la République de Haute-Volta. Centre Nord [Soil Map Center North].

Boulet, R., 1968b, Etude pédologique de la Haute-Volta : région Centre Nord. Site: http://library.wur.nl/WebQuery/isric/30233.

Bregt, A., Bouma, J., and Jellinek, M., 1987, Comparison of thematic maps derived from a soil map and from kriging of point data: Geoderma, v. 39, no. 4, p. 281-291.

Bui, E. N., 2004, Soil survey as a knowledge system: Geoderma, v. 120, no. 1, p. 17-26.

Bui, E. N., and Moran, C. J., 2001, Disaggregation of polygons of surficial geology and soil maps using spatial modelling and legacy data: Geoderma, v. 103, no. 1, p. 79-94.

CPCS, 1967, Classification des sols: édition 1967 : d'après les travaux de G. Aubert, R. Betremieux, P. Bonfils et autres.

Dobos, E., F. Carré, T. Hengl, H. I. Reuter, and G. Tóth, 2006, Digital soil mapping : as a support to production of functional maps, Luxembourg, Office for Official Publication of the European Communities, EUR;22123 EN.

ESA, 2009, GlobCover Land Cover Maps.

European_Commission, 2006, Soil protection – the long story behind the strategy. Office for Official Publications of the European Communities, Luxembourg.

GlobalSoilMap.Net, 2013, http://www.globalsoilmap.net/biblio.

Goovaerts, P., 1997, Geostatistics for natural resources evaluation, Oxford university press.

Goovaerts, P., and Journel, A., 1995, Integrating soil map information in modelling the spatial variation of continuous soil properties: European Journal of Soil Science, v. 46, no. 3, p. 397-414.

Grimm, R., Behrens, T., Märker, M., and Elsenbeer, H., 2008, Soil organic carbon concentrations and stocks on Barro Colorado Island—digital soil mapping using Random Forests analysis: Geoderma, v. 146, no. 1, p. 102-113.

Guillobez, S., 1985, Milieux naturels du Burkina Faso. Esquisse Physiographique. Carte Géologique simplifiée, Phyto-climatique, des Régions Naturelles. [Provisional physiographic map].

Hartemink, A. E., McBratney, A. B., and de Lourdes Mendonça-Santos, M., 2008, Digital soil mapping with limited data, Springer.

Hengl, T., Heuvelink, G., and Stein, A., 2004, A generic framework for spatial prediction of soil variables based on regression-kriging: Geoderma, v. 120, no. 1, p. 75-93.

Heuvelink, G., and Bierkens, M., 1992, Combining soil maps with interpolations from point observations to predict quantitative soil properties: Geoderma, v. 55, no. 1, p. 1-15.

Hijmans, R. J., van Etten, J., Hijmans, M. R. J., and ByteCompile, T., 2013, Package 'raster'.

Jenny, H., 1941, Factors of soil formation, McGraw-Hill Book Company New York, NY, USA.

Kaloga, B., 1968a, Carte Pédologique de Reconnaissance de la République de Haute-Volta. Centre Sud. [Soil Map Central South].

Kaloga, B., 1968b, Etude pédologique de la Haute-Volta : région Centre Sud. Site: http://library.wur.nl/WebQuery/isric/30234.

Keitt, T. H., Bivand, R., Pebesma, E., and Rowlingson, B., 2011, rgdal: bindings for the Geospatial Data Abstraction Library: R package version 0.7-1, URL http://CRAN.R-project.org/package=rgdal.

Köthe, R., 2013, Terrain Analysis Burkina Faso.

Lagacherie, P., Legros, J., and Burfough, P., 1995, A soil survey procedure using the knowledge of soil pattern established on a previously mapped reference area: Geoderma, v. 65, no. 3, p. 283-301.

Lagacherie, P., McBratney, A., and Voltz, M., 2006, Digital soil mapping: An introductory perspective, Access Online via Elsevier.

Latham, M., 1982, French soil classifications and their application in the South Pacific islands.

Leenaars, J. G. B., 2013, Africa Soil Profiles Database (version 1.1). A compilation of geo-referenced and standardized legacy soil profile data for Sub Saharan Africa (with dataset). ISRIC report 2013/03. Africa Soil Information Service (AfSIS) project. ISRIC – World Soil Information, Wageningen, the Netherlands. URL http://www.isric.org/data/africa-soil-profiles-database-version-01-1.

Lewin-Koh, N. J., Bivand, R., Pebesma, E., Archer, E., Baddeley, A., Bibiko, H., Dray, S., Forrest, D., Friendly, M., and Giraudoux, P., 2011, maptools: Tools for reading and handling spatial objects: R package version 0.8-10, URL http://CRAN.R-project.org/package=maptools.

Liu, T.-L., Juang, K.-W., and Lee, D.-Y., 2006, Interpolating soil properties using kriging combined with categorical information of soil maps: Soil Science Society of America Journal, v. 70, no. 4, p. 1200-1209.

LP_DAAC, 2001, NASA Land Processes Distributed Active Archive Center (LP DAAC). ASTER L1B. USGS/Earth Resources Observation and Science (EROS) Center, Sioux Falls, South Dakota.

Mayr, T., Palmer, R., and Cooke, H., 2008, Digital soil mapping using legacy data in the Eden valley, UK, Digital Soil Mapping with Limited Data, Springer, p. 291-301.

Mayr, T., Rivas-Casado, M., Bellamy, P., Palmer, R., Zawadzka, J., and Corstanje, R., 2010, Two Methods for Using Legacy Data in Digital Soil Mapping, Digital Soil Mapping, Springer, p. 191-202.

McBratney, A., Mendonça Santos, M. d. L., and Minasny, B., 2003, On digital soil mapping: Geoderma, v. 117, no. 1, p. 3-52.

Minasny, B., and McBratney, A., 2010, Methodologies for global soil mapping, Digital soil mapping, Springer, p. 429-436.

Mitschang, S., 2008, Bodenschutzrecht in Der EU, P. Lang.

Odeh, I., McBratney, A., and Chittleborough, D., 1994, Spatial prediction of soil properties from landform attributes derived from a digital elevation model: Geoderma, v. 63, no. 3, p. 197-214.

Odgers, N. P., Sun, W., McBratney, A. B., Minasny, B., and Clifford, D., 2014, Disaggregating and harmonising soil map units through resampled classification trees: Geoderma, v. 214–215, no. 0, p. 91-100.

Pebesma, E., Bivand, R., Rowlingson, B., and Gomez-Rubio, V., 2012, sp: classes and methods for spatial data: R package version, p. 1.0-5.

Pebesma, E., Graeler, B., and Pebesma, M. E., 2013, Package 'gstat'.

Rossiter, D., 2008, Digital Soil Mapping as a component of data renewal for areas with sparse soil data infrastructures, Digital Soil Mapping with Limited Data, Springer, p. 69-80.

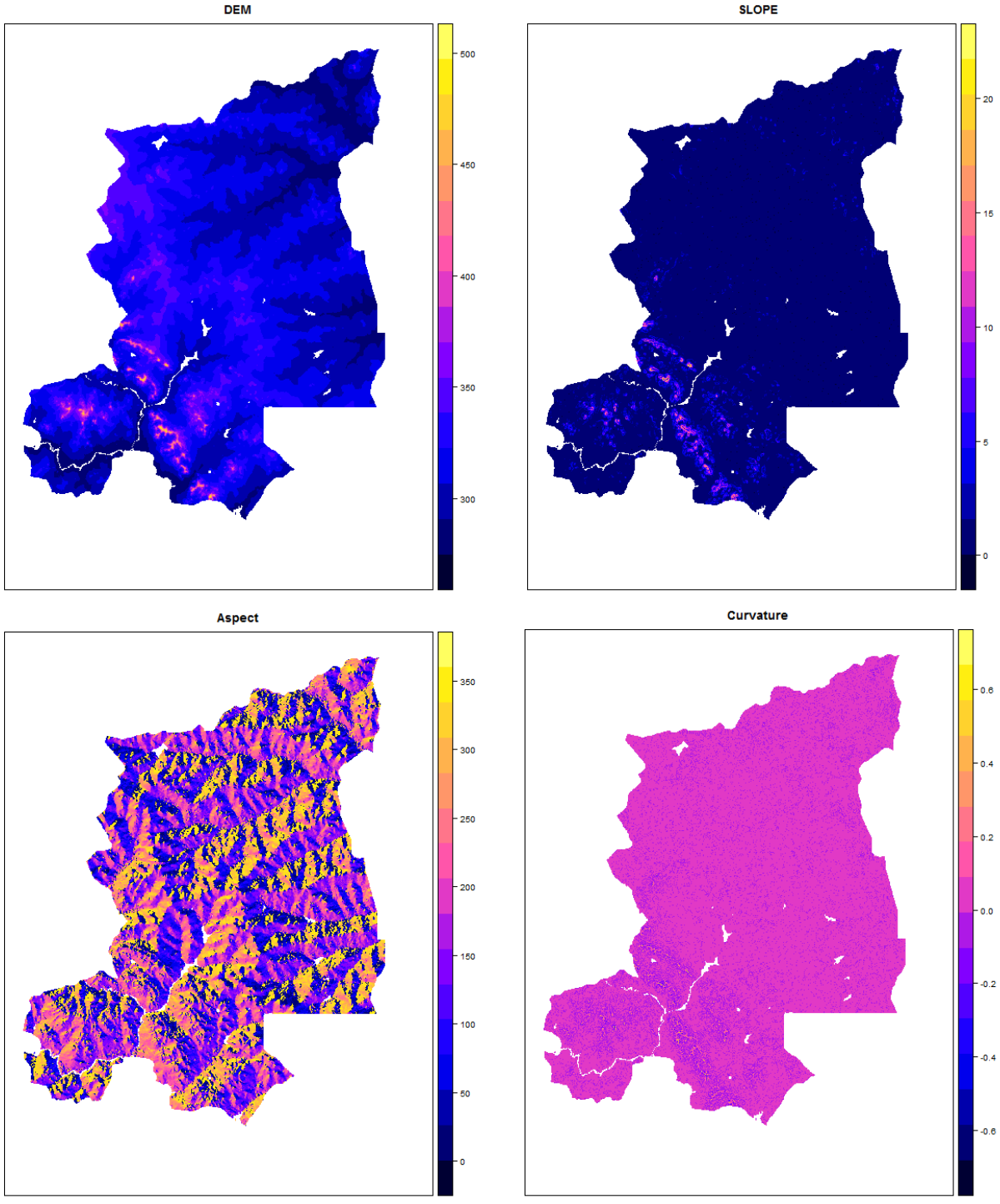
Stein, A., Hoogerwerf, M., and Bouma, J., 1988, Use of soil-map delineations to improve (co-) kriging of point data on moisture deficits: Geoderma, v. 43, no. 2, p. 163-177.

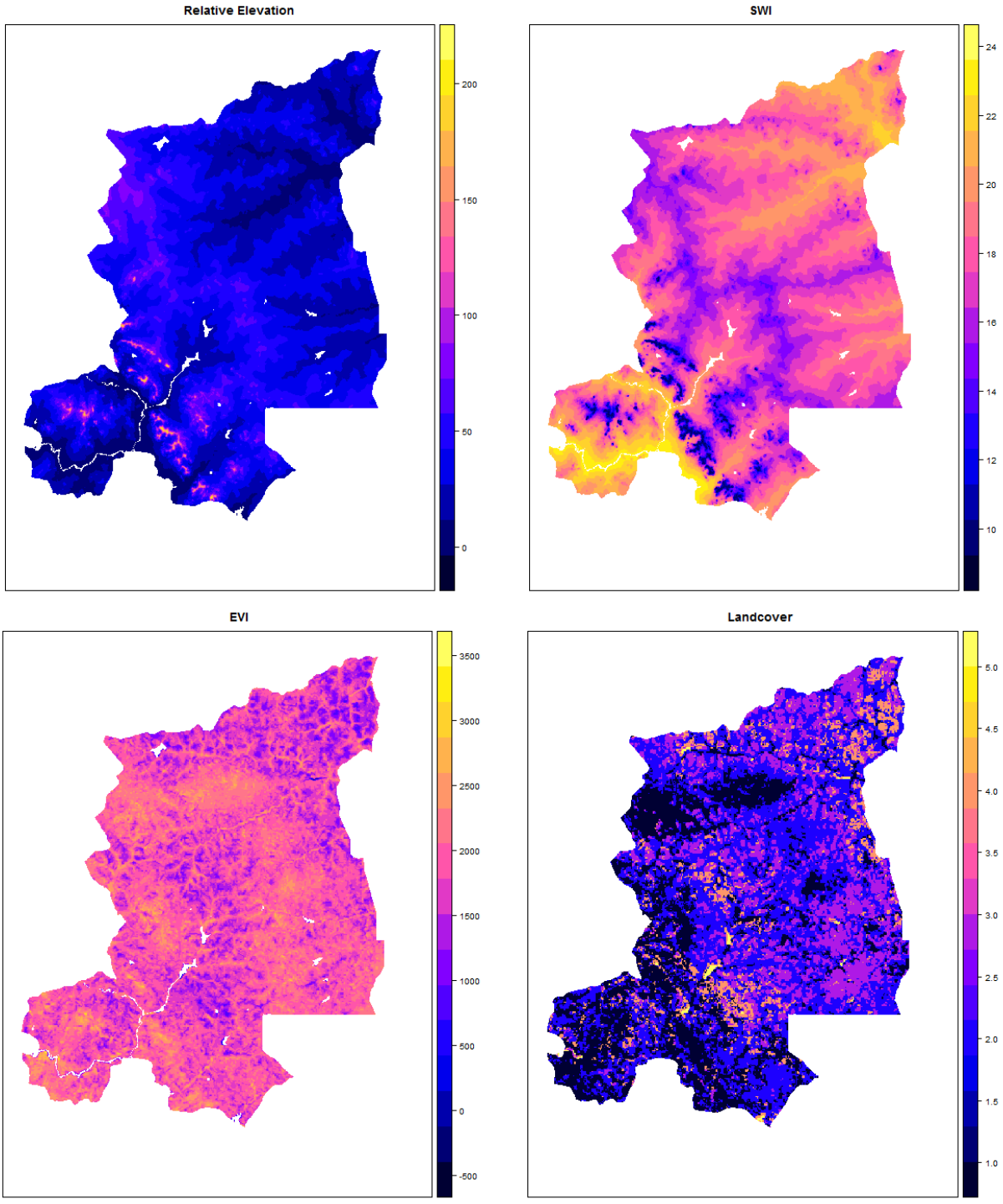
UNEP, 2007, Global environmental outlook GEO4 – environment for development. United Nations Environment Programme, Nairobi.

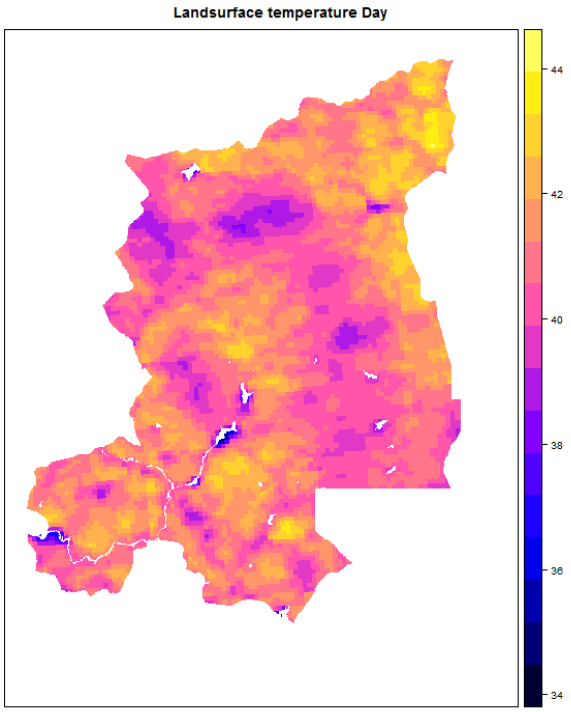
van Lieshout, S. M. J., Dongen, v., and van Winkel, J. H. M., 1997, Soil, Vegetation, Land Use, Erosion Risk and Geology Mapping in Sanmatenga, Burkina Faso: A Guide to the Physiographic, Erosion Risk and Geology Maps, Van Lieshout.

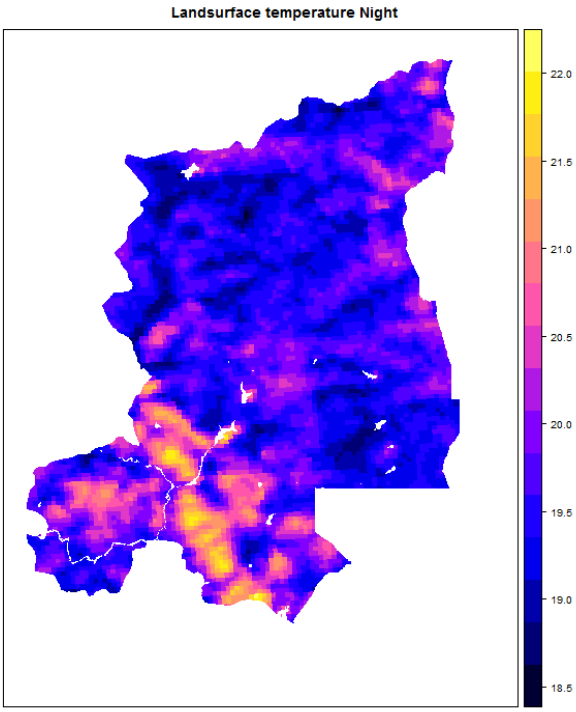
Appendices

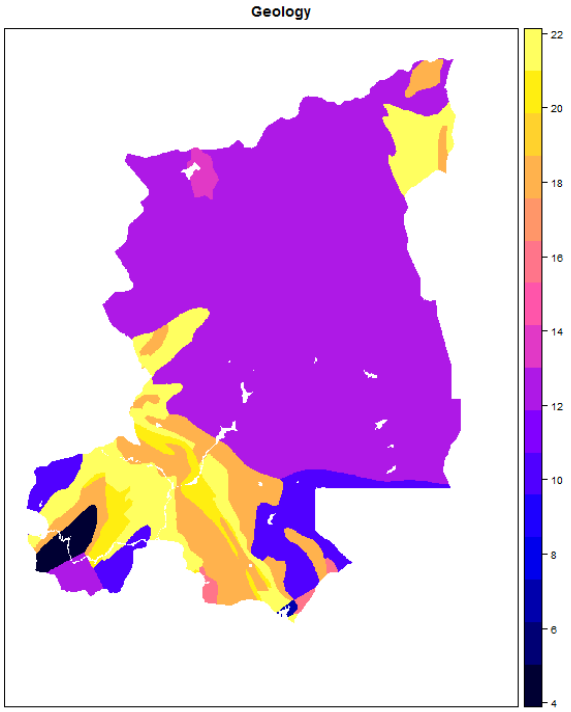
Appendix 1: Covariates




Relative Elevation
200
150
100
50
0
SWI
24
22
20
18
16
14
12
10
EVI
3500
3000
2500
2000
1500
1000
500
0
-500
Landcover
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0


Landsurface temperature Day
44
42
40
38
36
34


Landsurface temperature Night
22.0
21.5
21.0
20.5
20.0
19.5
19.0
18.5


Geology
22
20
18
16
14
12
10
8
6
4

Appendix 2: Generalization and reclassification of legacy soil maps

Table A2. 1 Mapping units with soil types and soil description of SM1M

Map Unit Code	Soil Types	Pedogenesis	Nr. Of pH Observations	Nr. Of CEC Observations	Nr. Of Depth Observations
2	- Lithosols - Weakly Developed Soils - Brown Soils	- Embryonal	1	1	1
3	- Lithosols - Weakly Developed Soils - Brown Soils	- Embryonal	12	12	12
4	- Vertisols - Brown Soils	- Vertic	32	29	32
5	- Vertisols	- Vertic	2	2	2
8	- Lithosols - Weakly Developed Soils - Ferallitic Soils	- Ferruginous - Embryonal	62	58	63
9	- Lithosols - Weakly Developed Soils - Ferallitic Soils	- Ferruginous - Embryonal	57	55	56
10	- Lithosols - Weakly Developed Soils - Ferallitic Soils	- Ferruginous - Embryonal	43	42	40
14	- Lithosols - Weakly Developed Soils	- Embryonal	0	0	0
20	- Weakly Developed Soils - Ferallitic Soils	- Little Characterized	0	0	0
23	- Raw Mineral Soils	--------	9	7	9

Table A2. 2 Generalized SM1M classes

Map Unit Code	Soil Types	Pedogenesis	Nr. Of pH Observations	Nr. Of CEC Observations	Nr. Of Depth Observations	Class nr.
2 and 3	- Lithosols - Weakly Developed Soils - Brown Soils	- Embryonal	13	13	13	1
4	- Vertisols - Brown Soils	- Vertic	32	29	32	2
5	- Vertisols	- Vertic	2	2	2	3
8,9 and 10	- Lithosols - Weakly Developed Soils - Ferallitic Soils	- Ferruginous - Embryonal	162	155	159	4
14	- Lithosols - Weakly Developed Soils	- Embryonal	0	0	0	5
20	- Weakly Developed Soils - Ferallitic Soils	- Little Characterized	0	0	0	6
23	- Raw Mineral Soils	--------	9	7	9	7

Table A2. 3 Mapping units with soil types and soil description of SM500K

Map Unit Code	Soil Description	Soil Class
NC01	Lithosols and iron caps	Raw Mineral Soils
NC02	Lithosols on various rocks	Raw Mineral Soils
NC04	Association of lithosols on iron cap	Weakly Developed Soils
NC05	Association of leached ferruginous soils with sandy clay material	Weakly Developed Soils
NC19	Clay materials with gravels from basic rock	Brown Soils
NC20	Association of leached ferruginous soils with eolian sands	Brown Soils
NC21	Association with gravel soils	Brown Soils
NC22	Clay material from granites	Brown Soils
NC23	Association with gravel soils	Brown Soils
NC27	Leached ferruginous tropical material with low internal drainage and eolian sands	Ferallitic Soils
NC28	Leached ferruginous tropical material with low internal drainage and associated with gravel	Ferallitic Soils
NC31	Leached ferruginous tropical material with low internal drainage associated with gravel and brown soils covered by sandy clay material derived from granites	Ferallitic Soils
NC32	Leached ferruginous tropical material with low internal drainage associated with hydromorphic soils on clayey material derived from slate	Ferallitic Soils
NC33	Leached ferruginous tropical material with low internal drainage and clayey fine sand	Ferallitic Soils
NC39	Leached alkali soils associated with gravels	Sodic Soils
NC43	Hydromorphic soils with pseudogley structures associated with leached ferruginous soils on aeolian or clayey sands	Hydromorphic Soils
NC44	Hydromorphic soils with pseudogley structures associated with brown soils on clay materials	Hydromorphic Soils
NC45	Hydromorphic soils with pseudogley structures associated with brown soils on clay materials and leached ferruginous soils on eolian or clayey sands	Hydromorphic Soils
SC23	Clay and lithosols material on iron cap	Vertisols
SC30	Lithosols on basic or neutral rocks	Brown Soils
SC40	Association of weakly developed hydromorphic soils with gravel materials	Ferallitic Soils
SC44	Lithomorphic Vertisols and Lithosols on granite	Sodic Soils
SC45	Association of weakly developed hydromorphic with sandy loam alluvial materials	Hydromorphic Soils
SC49	Leached ferruginous tropical or depleted sandy clay materials	Hydromorphic Soils

Table A2. 4 Generalized SM500K classes

Map Unit Code	Soil Class	Class Nr.	Nr. Of pH Observations	Nr. Of CEC Observations	Nr. Of Depth Observations
NC01 and NC02	Raw Mineral Soils	1	25	24	25
NC04 and NC05	Weakly Developed Soils	2	101	95	101
NC19, NC20, NC21, NC22, NC23 and SC30	Brown Soils	3	27	25	27
NC27, NC28, NC31, NC32, NC33 and SC40	Ferallitic Soils	4	28	26	26
NC39 and SC44	Sodic Soils	5	1	1	1
NC43, NC44, NC45,SC45 and SC49	Hydromorphic Soils	6	36	35	35
SC23	Vertisols	7	0	0	0

Table A2. 5 Mapping units with their soil types and soil description of SM100K

Map Unit Code	Physio-grapic units	Soil Description	Nr. Of pH Observ.	Nr. Of CEC Observ.	Nr. Of Depth Observ.
AC1	Hills and upper slopes	Soils are shallow to moderately deep, usually dark (reddish) brown and sandy clay loams. At the foot-slopes of the hills, there are colluviums deposition consisting many accumulated stones, forming shal-low soils. Soils on the hill are mostly moderately deep. The A-horizon is dark brown and contains up to abundant stones. The B-horizon has dark red color and contains few stones. The B/C - horizon consist of olive yellow in situ weathered greenstone with texture varying from sandy loam to silty clay loam.	17	17	17
B1	Plateau	Nearly flat to gently sloping; medium to many do-minant surface gravel; moderately often exposure of hard hardened plinthite; slight sheet erosion; no to medium hard surface sealing; moderately to ex-cessively well drained; very shallow (depth limited by hardened plinthite and/or dominant ironstone gravel);mottles absent; yellowish brown, light brown grey or dark brown sandy loams; abundant ironstones nodules present.	13	12	13
B2	Eroded or less developed indurated cap	Nearly flat to gently sloping; very few to abundant surface gravel; slight to moderate sheet erosion; moderately well drained; shallow to moderately deep; none to few mottles; yellowish brown to strong brown sandy loams or silt loams; few to many abundant ironstones gravel.	20	19	19
C21	Crusted middle slope	Nearly flat; surface crust with none to common gravel; no rock outcrops; moderate sheer erosion, sometimes also some gully erosion; medium to thick hard crust, sometimes extremely hard, alternated by areas with a small (<20cm) sand cover; (moderately) well drained.	12	10	10
C22	Non-crusted middle slope	Nearly flat; surface often covered with, up to many, gravel; sometimes rock outcrops; slight to severe sheet erosion, sometime gully erosion ; wind deposit on can occur; thin and sometimes thick slightly hard to very hard surface sealing; moderately well to somewhat excessively drained.	33	31	32
C31	Eroded lower slope	Flat to nearly flat; medium thick and very hard surface crust; moderately well to well drained; texture is sandy clay loam.	4	4	4
C32	Non-eroded lower slope	Flat to nearly flat: soils have none to many gravel; some have deposition; other soils have slight to moderate sheer or rill erosion or water/wind deposition; none to medium thick, slightly hard to sometimes very hard surface sealing; somewhat excessively to moderately well drained.	38	38	37
C41	Crusted lower /middle Birimien	Flat/level to gently sloping; none to common medium gravel; moderate to severe sheet erosion; slightly hard to hard medium crust; poorly to moderately well drained.	4	4	4

	slope				
C42	Non-crusted lower/ middle Birimien slope	Mainly water deposition, sometimes water erosion; none to thin slightly hard to hard surface sealing; sometimes fine surface cracks; moderately well to well drained.	27	27	28
C51	Crusted Birimien valleys	Nearly flat to gently sloping; common coarse gravel; severe sheet erosion also some gullies; hard medium surface sealing; somewhat excessively drained; deep yellowish red sands to loamy sands; no rocks; no mottles; no nodules.	1	1	1
C52	Non-crusted Birimien valleys	Gently sloping; few to dominant medium gravel to boulders; both water erosion as water deposition; none to slightly hard thin surface sealing; sometimes-fine cracks; moderately well to well drained.	12	11	11
D1	Bottom-land small valley	Flat /level; surface, sometimes very few gravel; no rock outcrops; water deposition; thin hard surface sealing to medium hard surface sealing; moderately well to poorly drained.	16	16	16
D2	Bottom-land large valley	Flat/level; surface gravel is rare; no rock outcrops; water deposition; strongly variable surface sealing ranging from thin to very thick and from slightly hard to very hard surface sealing; moderately well to poorly drained.	10	8	10
D3	Bottom-land plain	No surface gravels; no rock outcrops; mostly water deposition; none to medium (slightly) hard surface sealing; somewhat poorly (imperfectly) drained.	9	6	9
E	Aeolian complex	(Nearly) flat; slight sheet erosion features on top of the dunes; thin to medium (slightly) hard surface sealing on top of the dunes; in depressions thick and up to extremely hard; somewhat excessively drained on top of the old dunes to poorly in depressions.	2	2	2
W		Water	0	0	0

Appendix 3: Model interim results

1. RK

Table A3. 1 Statistics of the regression analysis for CEC using RK

Variable	Estimate	Std. Error	Pr(>|t|)
(Intercept)	-2.2002818	1.4272943	0.12477
aspect	0.0007594	0.0004100	0.06549 .
landcov2	-0.2094790	0.0987654	0.03516 *
landcov3	-0.2235865	0.1534659	0.14672
landcov4	-0.3088596	0.1614157	0.05713 .
landcov5	0.6503222	0.4410013	0.14189
tempnight	0.1997272	0.0703478	0.00499 **

Table A3. 2 CEC residuals after regression using RK

Residuals:				
Min	1Q	Median	3Q	Max
-1.5480	-0.3831	-0.0511	0.34083	1.5340

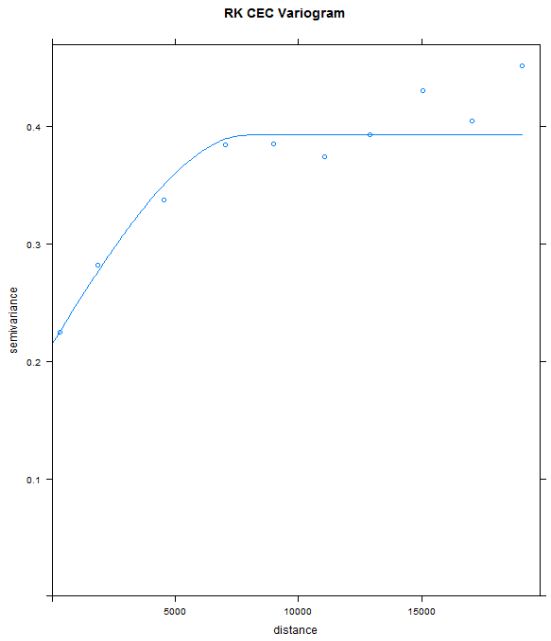


Figure A3. 1 Variogram of CEC residuals using RK

Table A3. 3 Statistics of the regression analysis for Depth using RK

Variable	Estimate	Std. Error	Significance
(Intercept)	55.20284	17.30838	0.00164 **
dem	-0.16760	0.06149	0.00696 **
relelev	0.16721	0.06313	0.00869 **

Table A3. 4 Depth residuals after regression using RK

Residuals:				
Min	1Q	Median	3Q	Max
-7.0703	-3.1727	0.0807	3.0524	7.9470

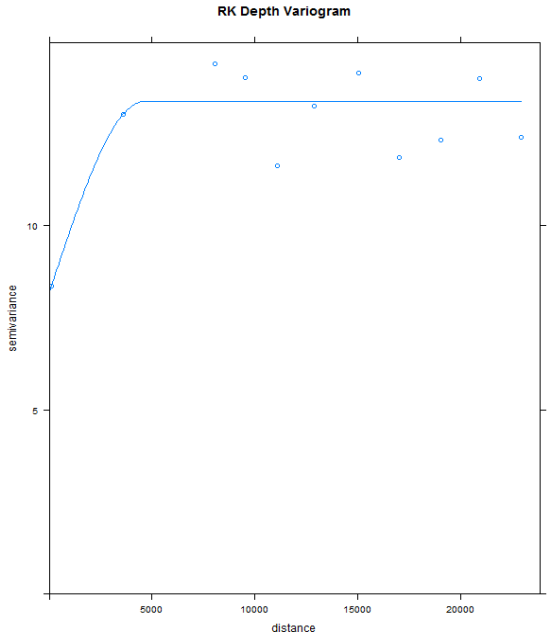


Figure A3. 2 Variogram of Depth residuals using RK

2. CAT

Table A3. 5 Statistics of the regression analysis for pH using CAT with SM100K

Variable	Estimate	Std. Error	Significance
(Intercept)	9.5125135	1.5691790	6.36e-09 ***
swi	-0.2053337	0.0714213	0.00447 **
relelev	-0.0251636	0.0091951	0.00675 **
evi	0.0003950	0.0002369	0.09696 .
sm100k2	-0.4496914	0.3358269	0.18204
sm100k3	-0.3757355	0.3021964	0.21516
sm100k4	0.2505545	0.3745730	0.50431
sm100k5	0.0814647	0.2745407	0.76697
sm100k6	-0.2360588	0.2628704	0.37024
sm100k7	0.0088190	0.3009095	0.97665
sm100k8	0.0524462	0.3390502	0.87722
sm100k9	-0.0988936	0.3271627	0.76275
sm100k10	0.0830636	0.3072840	0.78719

Table A3. 6 pH residuals after regression in CAT using SM100K

Residuals:				
Min	1Q	Median	3Q	Max
-1.96	-0.63	-0.10	0.58	3.22

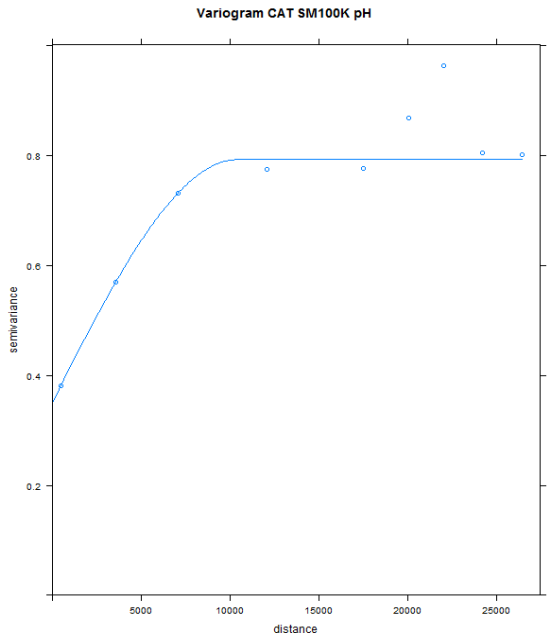


Figure A3. 3 Variogram of pH residuals using CAT with SM100K

Table A3. 7 Statistics of the regression analysis for pH using CAT with SM500K

Variable	Estimate	Std. Error	Significance
(Intercept)	8.2851491	1.6102107	6.12e-07 ***
swi	-0.1507861	0.0699935	0.0324 *
relelev	-0.0196503	0.0086706	0.0245 *
evi	0.0003941	0.0002264	0.0832 .
sm500k2	0.0884845	0.2052198	0.6668
sm500k3	-0.0485744	0.2464691	0.8440
sm500k4	-0.1876843	0.2521435	0.4575
sm500k5	-0.1311167	0.2490568	0.5991

Table A3. 8 pH residuals after regression in CAT using SM500K

Residuals:				
Min	1Q	Median	3Q	Max
-2.070	-0.57	-0.06	0.56	3.45

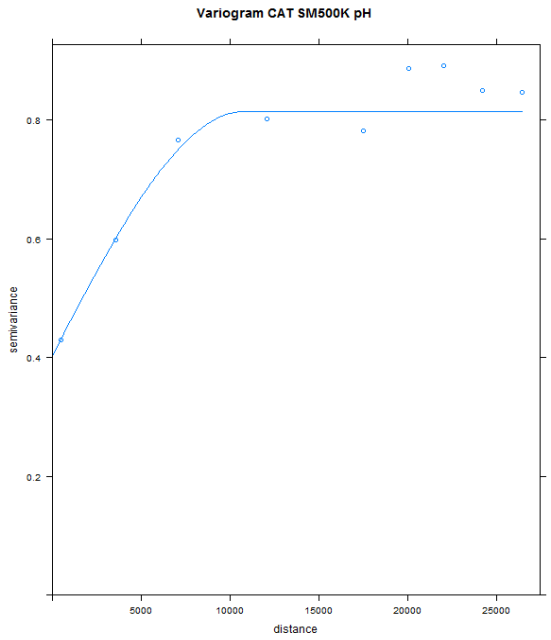


Figure A3. 4 Variogram of pH residuals using CAT with SM500K

Table A3. 9 Statistics of the regression analysis for pH using CAT with SM1M

Variable	Estimate	Std. Error	Significance
(Intercept)	8.6748729	1.4664014	1.31e-08 ***
swi	-0.1604806	0.0648571	0.0141 *
relelev	-0.0211424	0.0081437	0.0101 *
evi	0.0004096	0.0002242	0.0691 .
sm1m2	-0.4295761	0.2547062	0.0932 .
sm1m3	-0.1896094	0.2074514	0.3618

Table A3. 10 pH residuals after regression in CAT using SM1M

Residuals:				
Min	1Q	Median	3Q	Max
-1.99	-0.55	-0.12	0.57	3.53

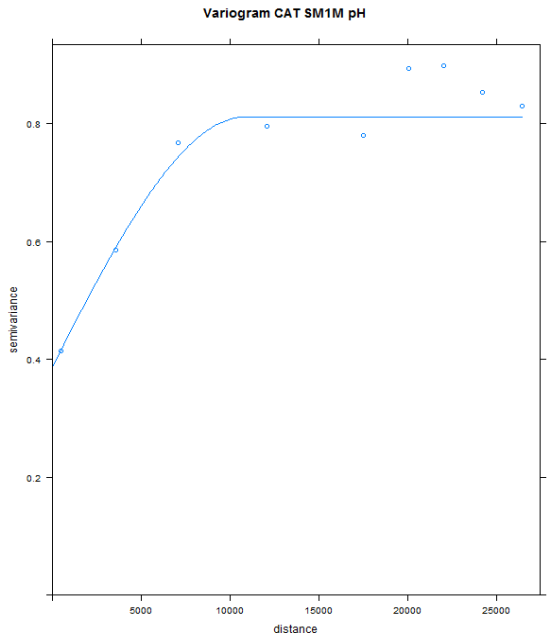


Figure A3. 5 Variogram of pH residuals using CAT with SM1M

Table A3. 11 Statistics of the regression analysis for CEC using CAT with SM100K

Variable	Estimate	Std. Error	Pr(>|t|)
(Intercept)	-2.0065438	2.0622466	0.3318
aspect	0.0006368	0.0004268	0.1373
landcov2	-0.2226903	0.1025242	0.0311 *
landcov3	-0.2293779	0.1611281	0.1562
landcov4	-0.2286556	0.1684736	0.1763
landcov5	0.9061246	0.4601081	0.0504 .
tempnight	0.1971768	0.0969156	0.0433 *
sm100k2	-0.4219083	0.2406668	0.0812 .
sm100k3	-0.3845320	0.2177471	0.0790 .
sm100k4	-0.2948728	0.2905577	0.3115
sm100k5	-0.1307320	0.2189196	0.5511
sm100k6	0.0372434	0.2224775	0.8672
sm100k7	0.0957079	0.1996669	0.6322
sm100k8	-0.3273343	0.2296264	0.1557
sm100k9	-0.0822245	0.2499809	0.7426
sm100k10	-0.3855908	0.2739394	0.1609

Table A3. 12 CEC residuals after regression in CAT using SM100K

Residuals:				
Min	1Q	Median	3Q	Max
-1.26	-0.39	-0.04	0.37	1.30

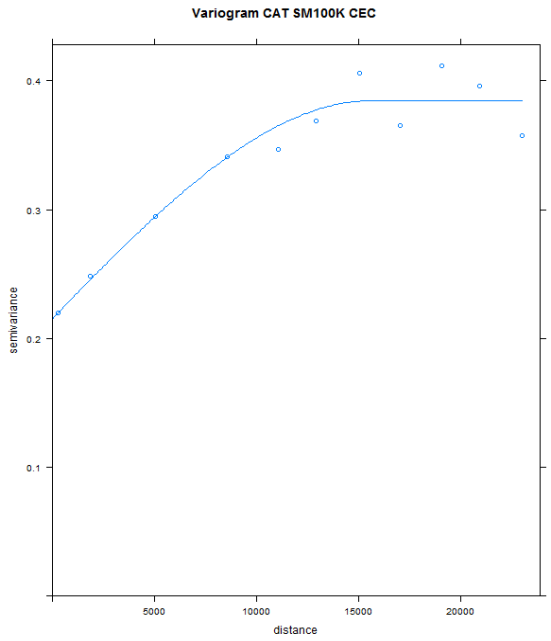


Figure A3. 6 Variogram of CEC residuals using CAT with SM100K

Table A3. 13 Statistics of the regression analysis for CEC using CAT with SM500K

Variable	**Estimate**	**Std. Error**	**Pr(>|t|)**
(Intercept)	-4.0609927	1.7427110	0.02081 *
aspect	0.0008099	0.0004159	0.05295 .
landcov2	-0.2317720	0.0994935	0.02086 *
landcov3	-0.2591814	0.1567155	0.09977 .
landcov4	-0.3608365	0.1625236	0.02756 *
landcov5	0.7377421	0.4479245	0.10116
tempnight	0.2819231	0.0836850	0.00091 ***
sm500k2	0.3344040	0.1612199	0.03937 *
sm500k3	0.2633161	0.1753517	0.13481
sm500k4	0.2348063	0.1938898	0.22735
sm500k5	0.1112017	0.1669902	0.50625

Table A3. 14 CEC residuals after regression in CAT using SM500K

Residuals:				
Min	1Q	Median	3Q	Max
-1.49	-0.37	-0.07	0.39	1.59

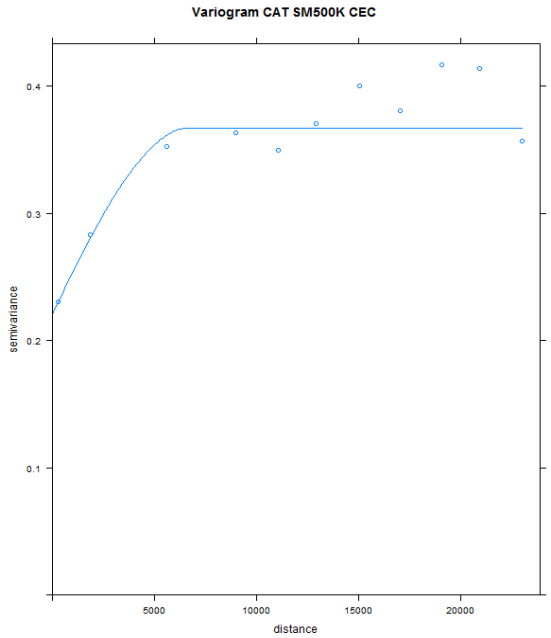


Figure A3. 7 Variogram of CEC residuals using CAT with SM500K

Table A3. 15 Statistics of the regression analysis for CEC using CAT with SM1M

Variable	Estimate	Std. Error	Pr(>|t|)
(Intercept)	-1.7686495	1.7018435	0.3000
aspect	0.0007372	0.0004142	0.0766 .
landcov2	-0.2089368	0.0994307	0.0369 *
landcov3	-0.2222728	0.1542473	0.1512
landcov4	-0.3113140	0.1622158	0.0564 .
landcov5	0.6680590	0.4443648	0.1343
tempnight	0.1808507	0.0807821	0.0263 *
sm1m2	-0.0158124	0.1824278	0.9310
sm1m3	-0.0665846	0.1669652	0.6905

Table A3. 16 CEC residuals after regression in CAT using SM1M

Residuals:				
Min	1Q	Median	3Q	Max
-1.52	-0.36	-0.04	0.35	1.53

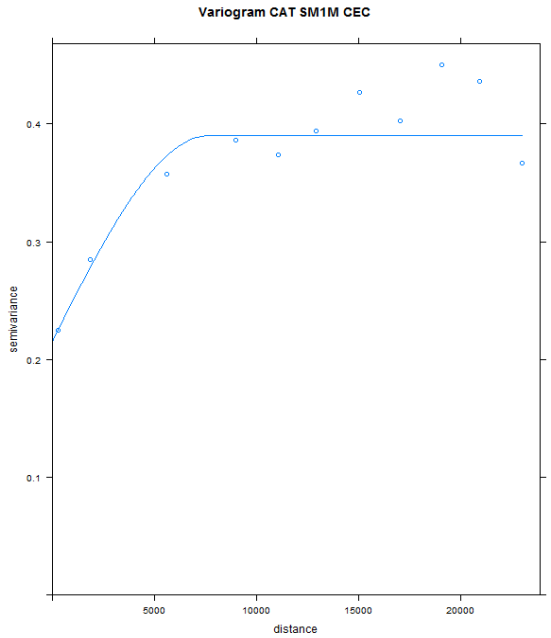


Figure A3. 8 Variogram of CEC residuals using CAT with SM1M

Table A3. 17 Statistics of the regression analysis for Depth using CAT with SM100K

Variable	Estimate	Std. Error	Pr(>|t|)
(Intercept)	55.57704	19.64379	0.00513 **
dem	-0.17461	0.06912	0.01229 *
relelev	0.18552	0.07053	0.00918 **
sm100k2	0.75728	1.40728	0.59109
sm100k3	2.02174	1.25948	0.11000
sm100k4	2.04101	1.55491	0.19079
sm100k5	1.15929	1.17903	0.32665
sm100k6	1.47552	1.13056	0.19333
sm100k7	1.65737	1.20417	0.17023
sm100k8	0.06361	1.37825	0.96323
sm100k9	1.44662	1.39936	0.30247
sm100k10	0.51575	1.29899	0.69176

Table A3. 18 Depth residuals after regression in CAT using SM100K

Residuals:				
Min	1Q	Median	3Q	Max
-7.42	-3.11	-0.16	2.98	7.51

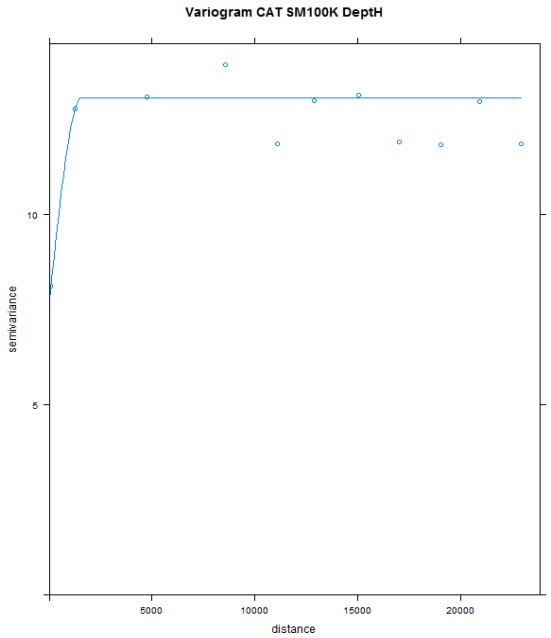


Figure A3. 9 Variogram of Depth residuals using CAT with SM100K

Table A3. 19 Statistics of the regression analysis for Depth using CAT with SM500K

Variable	Estimate	Std. Error	Pr(>|t|)
(Intercept)	54.90419	18.87808	0.00403 **
dem	-0.17166	0.06626	0.01026 *
relelev	0.18213	0.06707	0.00717 **
sm500k2	0.88946	0.87194	0.30887
sm500k3	1.58896	1.04673	0.13053
sm500k4	0.77195	1.13874	0.49859
sm500k5	2.01295	1.02934	0.05185 .

Table A3. 20 Depth residuals after regression in CAT using SM500K

Residuals:				
Min	1Q	Median	3Q	Max
-7.66	-3.15	-0.08	3.07	7.69

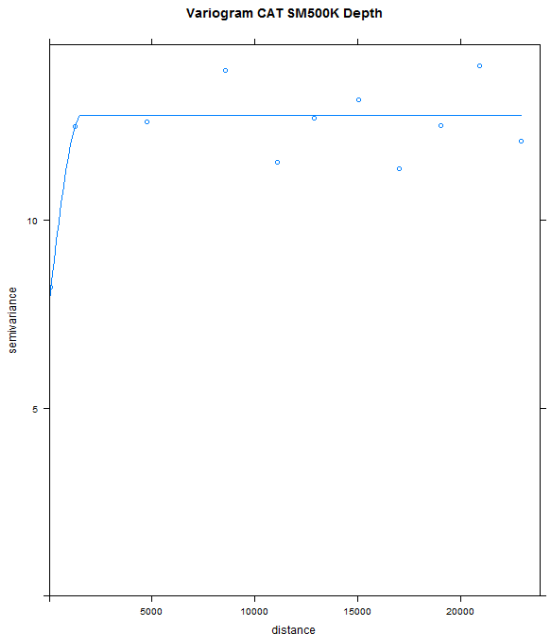


Figure A3. 10 Variogram of Depth residuals using CAT with SM500K

Table A3. 21 Statistics of the regression analysis for Depth using CAT with SM1M

Variable	Estimate	Std. Error	Pr(>|t|)
(Intercept)	43.07347	18.65218	0.0219 *
dem	-0.13120	0.06521	0.0455 *
relelev	0.13929	0.06578	0.0354 *
sm1m2	1.75600	1.05577	0.0978 .
sm1m3	1.76623	0.91294	0.0544 .

Table A3. 22 Depth residuals after regression in CAT using SM1M

Residuals:				
Min	1Q	Median	3Q	Max
-7.05	-3.15	0.06	2.91	7.56

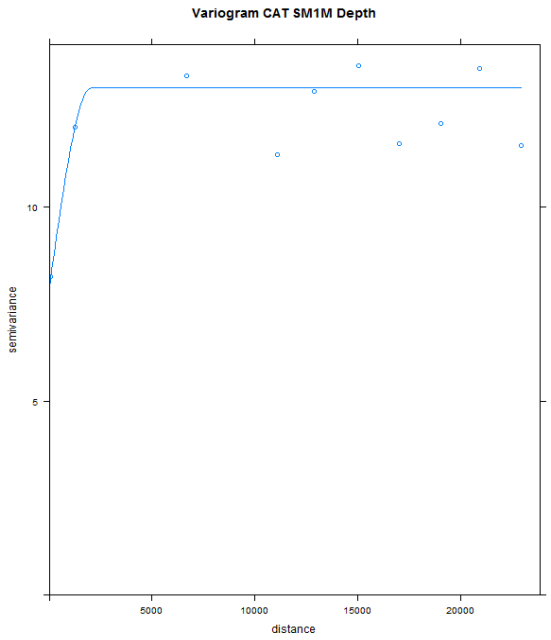


Figure A3. 11 Variogram of Depth residuals using CAT with SM1M

3. STK

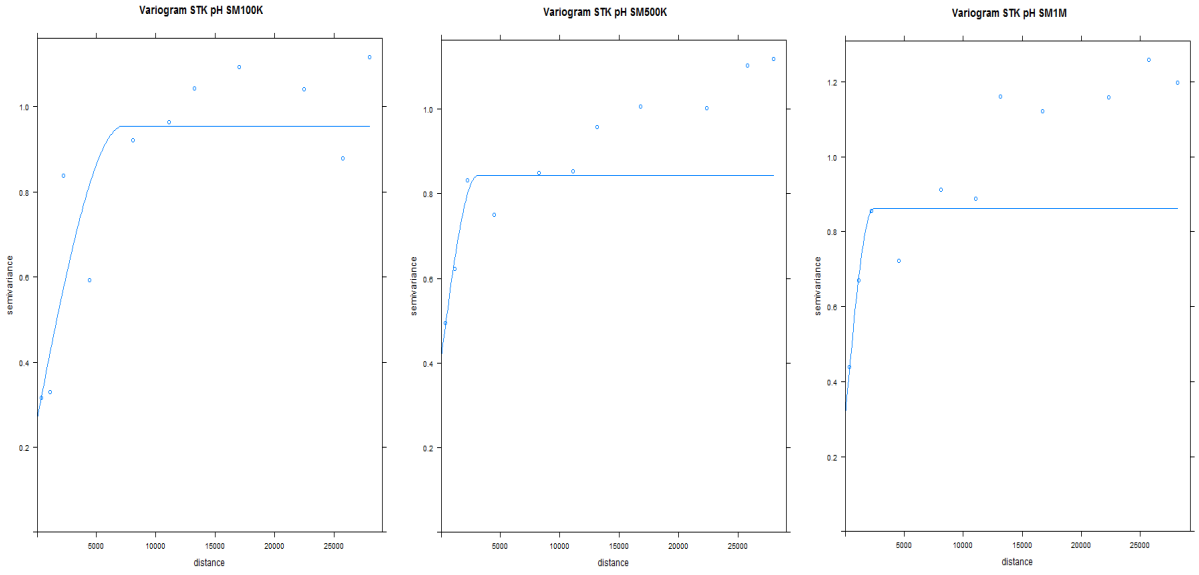


Figure A3. 12 Variogram of pH using STK with SM100K, SM500K and SM1M

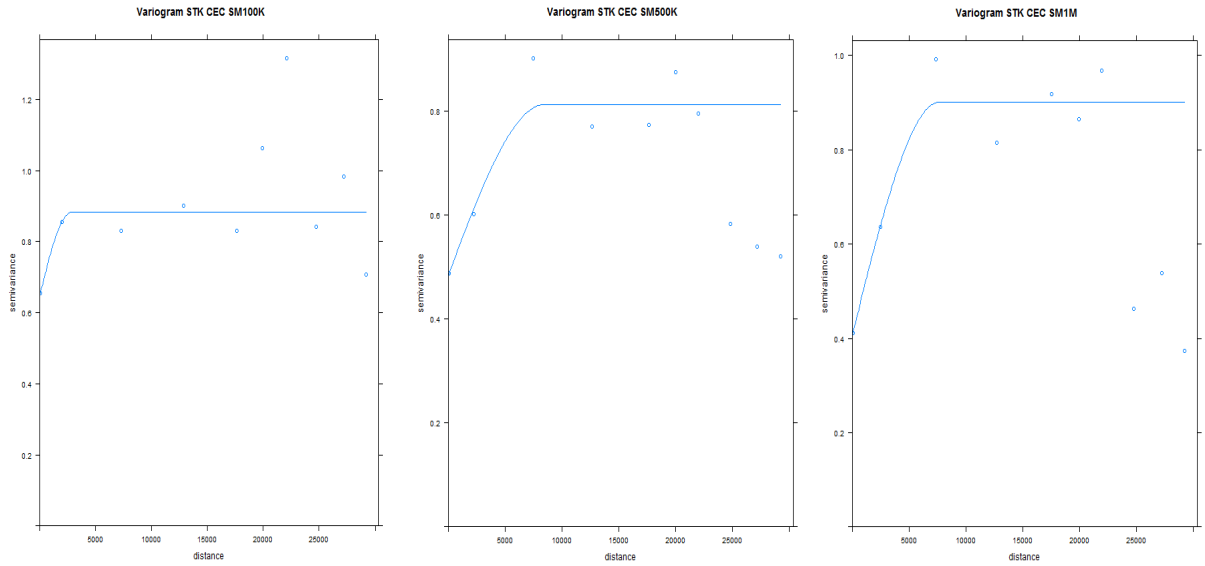


Figure A3. 13 Variogram of CEC using STK with SM100K, SM500K and SM1M

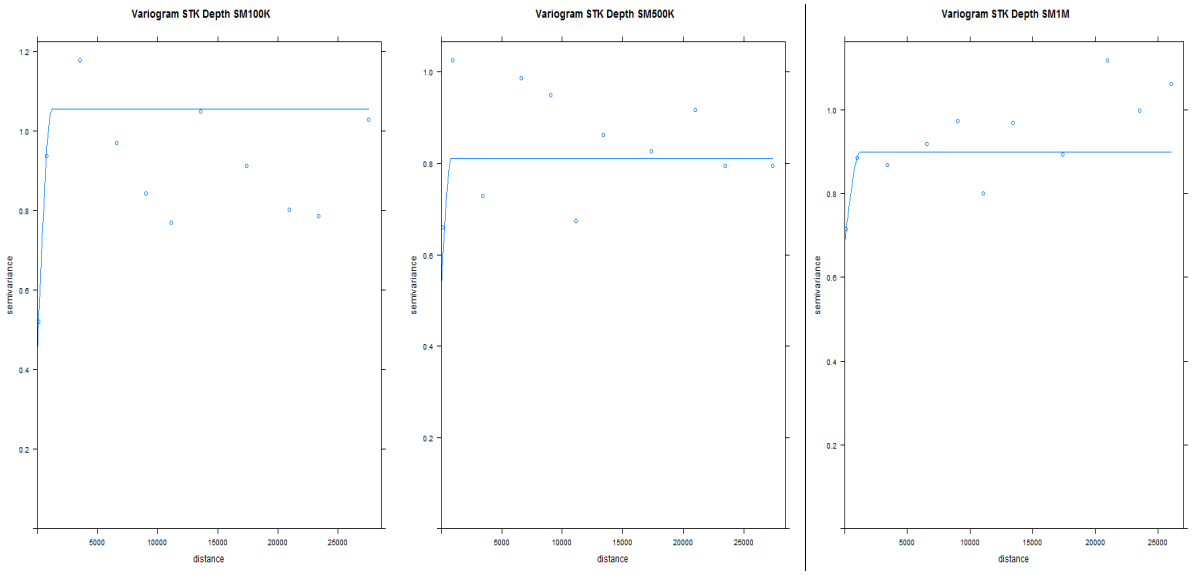


Figure A3. 14 Variogram of Depth using STK with SM100K, SM500K and SM1M

4. OBS

Table A3. 23 OBS SM100K results from report

Map Unit code	Soil Type	Soil Type Pro-portion (%)	Repre-sentive Soil Profiles	Avg. Soil Profile pH	Avg. Soil Profile CEC (cmol +/kg)	Avg. Soil Profile DEPTH (cm)	Weighted Mean Map Unit pH	Weighted Mean Map Unit CEC (cmol +/kg)	Weighted Mean Map Unit DEPTH (cm)
AC1	AC1a	10	26.2	NAR	NAR	40	7.1	23.3	104.7
			33.1	7	14.8	25			
	AC1b	35	740	6.6	20.9	120			
			768	7.5	27.9	120			
			769	7.3	28.8	40			
			807	6.9	23.5	110			
	AC1c	55	741	7	20.2	120			
			805	7.2	24.5	120			
			918	7.2	25.9	127			
B1	B1a	NA	2.4	NAR	NAR	40	NA	NA	25
	B1b	NA	715	NAR	NAR	10			
B2	B2a	50	32.3	NAR	NAR	5	8.3	11	19.8
	B2b	10	801	8.3	11	100			
	B2c	5	712	NAR	NAR	70			
			760	NA	NA	NA			
			908	NA	NA	NA			
	Bare rock	35	NA	NA	NA	NA			
C21	C21a	35	22.1	NAR	NAR	70	5.5	3.5	102.5
	C21b	65	12.1	NAR	4.1	120			
			12.2	NAR	1.3	120			
			12.7	5.5	5.2	120			
C22	C22a	40	12.8	5.5	6.2	120	6.4	6.89	86.3
			33.2	6	4.9	120			
	C22b	30	27.1	NAR	NAR	25			
			13.5	7.2	8.7	50			
	C22c	30	30.4	NAR	NAR	90			
C31	C31a	65	13.4	6.3	9	130	6.5	9.1	125.8

	C31b	10	719	NA	NA	NA			
			913	NAR	NAR	110			
	D1c	10	13.2	5.9	7.5	120			
	C32a	15	12.3	NAR	9.7	130			
			13.3	7.1	7.7	115			
			34.2	7.8	14.1	120			
C32	C32a	50	12.3	NAR	9.7	130	7.1	9.6	121
			13.3	7.1	7.7	115			
			34.2	7.8	14.1	120			
	C32b	10	12.4	NAR	6.1	130			
			33.5	6.5	4.8	120			
	C32c	10	31.6	NAR	NAR	90			
	C32d	10	33.3	6.7	12.1	120			
	C32e	10	2.6	NAR	NAR	120			
	C32f	10	13.1	6.2	4.8	130			
			34.4	6.4	9.2	120			
C41	C41a	NA	917	6.5	12	100	6	8.4	108.3
	C41b	NA	743	5.6	6.4	120			
			803	5.7	6.7	105			
C42	C42a	20	738	6.9	14.9	120	6.3	11.5	117.7
			739	6.1	8.7	100			
			742	5.9	7	120			
	C42b	30	772	6.5	18.8	120			
			920	5.9	8.5	120			
	C42c	50	776	6.6	18.8	120			
			778	5.6	12.1	120			
			780	5.7	7.8	110			
			790	6.7	7.4	120			
			792	6.8	7.3	120			
C51	C51a	100	799	NAR	NAR	120	NA	NA	120
C52	C52a	25	770	7.3	28.8	40	7.3	25.5	80
			919	7.3	22.3	120			
	C52b	15	911	NAR	NAR	120			
	C52c	25	746	NAR	NAR	120			
			905	NA	NA	NA			
	C52d	20	728	NAR	NAR	30			
	C52e	5	706	NAR	NAR	60			

	C52f	10	703	NAR	NAR	30			
D1	D1a	40	17.2	NAR	NAR	70	5.9	7.5	100
	D1b	20	28.4	NAR	NAR	120			
	D1c	40	13.2	5.9	7.5	120			
D2	D2a	40	33.4	6.4	11.5	120	6.3	10.8	120
	D2b	35	13.7	6	8.6	120			
			34.3	6.5	9.2	120			
	D2c	25	13.6	6.3	10.6	120			
			34.5	6.5	13.8	120			
D3	D3a	40	12.5	NAR	0.8	100	NA	3.7	122
			16.2	NAR	NAR	120			
	D3b	60	12.6	NAR	5.7	130			
E	Ea	45	16.3	NAR	NAR	220	NA	0.8	147.9
	Eb	45	16.4	NAR	NAR	80			
	D3a	5	12.5	NAR	0.8	100			
			16.2	NAR	NAR	120			
	NA	5	NA	NA	NA	NA			

Table A3. 24 OBS SM500K results from report

Map Unit Code	Representive Soil Profiles	Avg. Soil Profile pH	Avg. Soil Profile CEC (cmol +/kg)	Avg. Soil Profile DEPTH (cm)	Average Map Unit pH	AverageMap Unit CEC (cmol +/kg)	Average Map Unit DEPTH (cm)
NC01	HVF99	6.2	6.7	20	6	6.3	28.3
	HVC58	5.9	3.4	30			
	HVG50	5.9	8.8	35			
NC02	NA	NA	NA	NA	NA	NA	NA
NC04	HVC19	NAR	NAR	30	6	6.7	55.8
	HVG75	NAR	NAR	170			
	HVF99	6.2	6.7	20			
	HVE57	NAR	NAR	50			
	HVC58	5.9	3.4	30			
	HVG50	5.9	8.8	35			
NC05	HVC79	6.6	4.3	55	6.7	6.5	123.9

	HVD39	NAR	NAR	180			
	HVC86	NAR	NAR	140			
	HVD70	5.8	7.2	200			
	HVD49	NAR	NAR	160			
	HVC60	8.3	6.1	160			
	HVC86	6.4	5.2	140			
	HVD71	7.2	9.4	60			
	HVC85	5.8	4.6	20			
NC19	HVE99	NAR	NAR	180	NAR	NAR	140
	HVE50	NAR	NAR	100			
NC20 NC21	HVC53	NAR	NAR	115	7.2	9.3	153
	HVC41	6.5	5.4	200			
	HVC52	7.9	13.2	110			
	HVC99	NAR	NAR	180			
	HVD1	NAR	NAR	160			
NC22	HVE83	7.6	11.3	120	7.6	11.3	120
NC23	HVB47	NAR	NAR	120	7.3	13.8	106.7
	HVB50	7.3	13.8	140			
	HVD74	NAR	NAR	60			
NC27 NC28	HVB7	NAR	NAR	250	6.5	4.5	189.4
	HVA3	NAR	NAR	230			
	HAV5	6.5	6.6	140			
	HVA16	6.9	4.5	170			
	HVB7	6.5	3.1	175			
	HVA5	5.8	1.9	200			
	HVG23	5.5	3	180			
	HVD30	7.5	5.5	180			
	HVD31	6.8	7.2	180			
NC31	HVD40	7	6.3	200	6.7	6.3	200
NC32	NA	NA	NA	NA	NA	NA	NA
NC33	HVD75	5.5	5.6	160	5.5	5.6	160
NC39	NA	NA	NA	NA	NA	NA	NA
NC43	HVE2	6	3.5	150	6	3.5	155
	HVE43	NAR	NAR	160			
NC44 NC45	HVE51	NAR	NAR	140	7.1	9.5	136
	HVD84	6.9	11.4	170			
	HVE92	6.4	9.9	160			

	HVE94	7.7	9.1	160			
	HVE93	7.5	7.8	50			
SC23	NA	NA	NA	NA	NA	NA	NA
SC30	VOB41	NAR	NAR	140	NAR	NAR	140
SC40	VOH26	5.9	7.6	164	5.9	7.6	164
SC44	NA	NA	NA	NA	NA	NA	NA
SC45	NA	NA	NA	NA	NA	NA	NA
SC49	VOA10	5.9	4.5	171	5.9	4.5	171

5. OBSRES

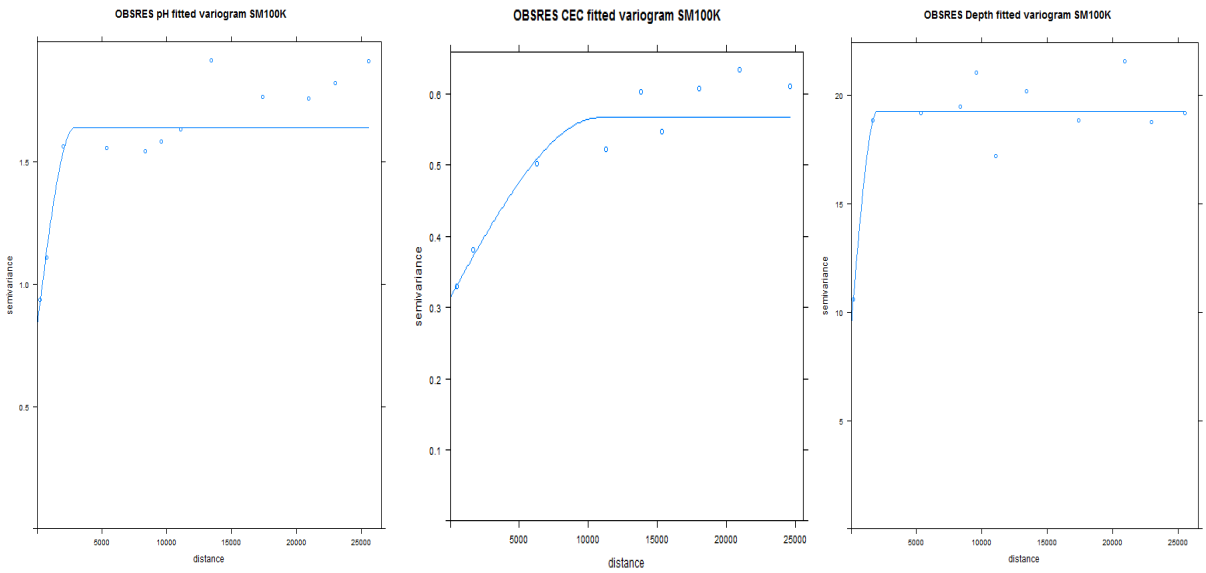


Figure A3. 15 Variogram of pH, CEC and Depth using OBSRES with SM100K

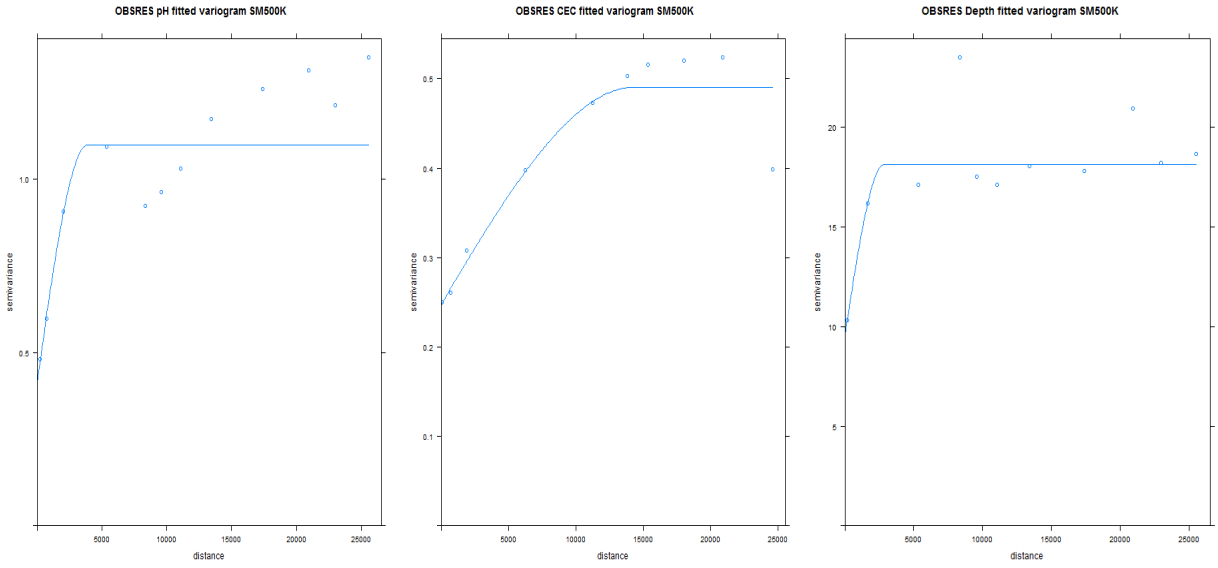


Figure A3. 16 Variogram of pH, CEC and Depth using OBSRES with SM500K

Appendix 4: R scripts

1. EDA

```
setwd('C:/Users/User/Desktop/Thesis/c002CharlWong/Workspace/a1ExternalFormattedData/Exploratory data analysis')
# Load libraries
library(sp)
library(rgdal)
library(maptools)
library(gstat)
library(raster)
#Read Soil Profiles
SpH <- read.table("SoilProfilepH.txt", header = TRUE)
SDepth <- read.table("SoilProfileDepth.txt", header = TRUE)
SCEC <- read.table("SoilProfileCEC.txt", header = TRUE)
# make spatial
coordinates(SpH) <- ~x+y
coordinates(SDepth) <- ~x+y
coordinates(SCEC) <- ~x+y
# read boundary study area
StudyArea <- readShapePoly("MaskAOI.shp")
# plot observations on study area
spplot(SpH, zcol = "pH", xlim = c(-178000,-60000),
      ylim = c(1420000,1575000), cex = 1.4, main = "pH",
      key.space = list(x = 0.02, y = 0.25, corner = c(0,2)),
      sp.layout = list("sp.polygons", StudyArea),col.regions = bpy.colors(5))
dev.print(png, file="pHProfilesonStudyArea.png", width=600, height=700)
spplot(SDepth, zcol = "Depth", xlim = c(-178000,-60000),
      ylim = c(1420000,1575000), cex = 1.4, main = "Depth (cm)",
      key.space = list(x = 0.02, y = 0.25, corner = c(0,2)),
      sp.layout = list("sp.polygons", StudyArea),col.regions = bpy.colors(5))
dev.print(png, file="DepthProfilesonStudyArea.png", width=600, height=700)
spplot(SCEC, zcol = "cec", xlim = c(-178000,-60000),
      ylim = c(1420000,1575000), cex = 1.4, main = "CEC (cmol+/kg)",
      key.space = list(x = 0.02, y = 0.25, corner = c(0,2)),
      sp.layout = list("sp.polygons", StudyArea),col.regions = bpy.colors(5))
dev.print(png, file="CECProfilesonStudyArea.png", width=600, height=700)
#Boxplot for outliers
par(mfrow=c(1,1))
boxplot(SpH$pH, main='ph')
dev.print(png, file="BoxPlotPH.png", width=600, height=700)
boxplot(SDepth$Depth, main='Depth')
dev.print(png, file="BoxPlotDepth.png", width=600, height=700)
boxplot(SCEC$cec, main='CEC')
dev.print(png, file="BoxPlotCEC.png", width=600, height=700)
# Read covariates
DEM <- readGDAL("prb2_demofaoi.txt")
SLOPE <- readGDAL("prb3_slopeaoi.txt")
SWI <- readGDAL("psb4_swiofaoi.txt")
ASPECT <- readGDAL("prb5_aspectaoi.txt")
CURVATURE <- readGDAL("prb6_curvatureaoi.txt")
RELATIVEELEVATION <- readGDAL("prb17_relativeElevationaoi.txt")
EVI <- readGDAL("psb07_eviofaoi.txt")
LANDCOVER <- readGDAL("psb08_landcoveraoi.txt")
```

```
LANDSURFACTEMPNIGHT <- readGDAL("psb09_landsurfacetempnightaoi.txt")
LANDSURFACTEMPDAY <- readGDAL("psb10_landsurfacetempdayaoi.txt")
GEOLOGY <- readGDAL("prb18_geologyaoi.txt")
SM100K <- readGDAL("soilmap100k.txt")
SM500K <- readGDAL("soilmap500k.txt")
SM1M <- readGDAL("soilmap1m.txt")
#Plot covariates
spplot(DEM, col.regions = bpy.colors(), xlim = c(-178000,-60000),
      ylim = c(1420000,1575000), main="DEM")
dev.print(png, file="DEM.png", width=600, height=700)
spplot(SLOPE, col.regions = bpy.colors(), xlim = c(-178000,-60000),
      ylim = c(1420000,1575000), main="SLOPE")
dev.print(png, file="SLOPE.png", width=600, height=700)
spplot(SWI, col.regions = bpy.colors(), xlim = c(-178000,-60000),
      ylim = c(1420000,1575000), main="SWI")
dev.print(png, file="SWI.png", width=600, height=700)
spplot(ASPECT, col.regions = bpy.colors(), xlim = c(-178000,-60000),
      ylim = c(1420000,1575000), main="Aspect")
dev.print(png, file="Aspect.png", width=600, height=700)
spplot(CURVATURE, col.regions = bpy.colors(), xlim = c(-178000,-60000),
      ylim = c(1420000,1575000), main="Curvature")
dev.print(png, file="Curvature.png", width=600, height=700)
spplot(EVI, col.regions = bpy.colors(), xlim = c(-178000,-60000),
      ylim = c(1420000,1575000), main="EVI")
dev.print(png, file="EVI.png", width=600, height=700)
spplot(LANDCOVER, col.regions = bpy.colors(), xlim = c(-178000,-60000),
      ylim = c(1420000,1575000), main="Landcover")
dev.print(png, file="Landcover.png", width=600, height=700)
spplot(GEOLOGY, col.regions = bpy.colors(), xlim = c(-178000,-60000),
      ylim = c(1420000,1575000), main="Geology")
dev.print(png, file="Geology.png", width=600, height=700)
spplot(LANDSURFACTEMPNIGHT, col.regions = bpy.colors(), xlim = c(-178000,-60000),
      ylim = c(1420000,1575000), main="Landsurface temperature Night")
dev.print(png, file="LANDSURFACTEMPNIGHT.png", width=600, height=700)
spplot(LANDSURFACTEMPDAY, col.regions = bpy.colors(), xlim = c(-178000,-60000),
      ylim = c(1420000,1575000), main="Landsurface temperature Day")
dev.print(png, file="LANDSURFACTEMPDAY.png", width=600, height=700)
spplot(RELATIVEELEVATION, col.regions = bpy.colors(), xlim = c(-178000,-60000),
      ylim = c(1420000,1575000), main="Relative Elevation")
dev.print(png, file="RELATIVEELEVATION.png", width=600, height=700)
spplot(SM100K, col.regions = bpy.colors(), xlim = c(-178000,-60000),
      ylim = c(1420000,1575000), main="Soil map 1:100.000")
dev.print(png, file="SM100K.png", width=600, height=700)
spplot(SM500K, col.regions = bpy.colors(), xlim = c(-178000,-60000),
      ylim = c(1420000,1575000), main="Soil map 1:500.000")
dev.print(png, file="SM500K.png", width=600, height=700)
spplot(SM1M, col.regions = bpy.colors(), xlim = c(-178000,-60000),
      ylim = c(1420000,1575000), main="Soil map 1:1.000.000")
dev.print(png, file="SM1M.png", width=600, height=700)
#Plot histograms of Soil profiles and environmental covariates
par.ori <- par(no.readonly = TRUE)  # save original setting graphics
par(mfrow=c(3,4))
hist(SpH$pH, main = "pH", xlab=NA)
hist(SDepth$Depth, main = "Depth", xlab=NA)
hist(SCEC$cec, main = "CEC", xlab=NA)
hist(DEM$band1, main = "DEM", xlab = NA, xlim=c(250,400))
```

```
hist(SLOPE$band1, main = "SLOPE", xlab = NA, xlim=c(0,10))
hist(SWI$band1, main = "SWI", xlab = NA)
hist(ASPECT$band1, main = "ASPECT", xlab = NA)
hist(CURVATURE$band1, main = "CURVATURE", xlab = NA, xlim=c(-0.6,0.6))
hist(RELATIVEELEVATION$band1, main = "Rel. Elevation", xlab = NA, xlim=c(0,100))
hist(EVI$band1, main = "EVI", xlab = NA, xlim=c(500,3000))
hist(LANDSURFACTEMPNIGHT$band1, main = "LandTempNight", xlab = NA)
hist(LANDSURFACTEMPDAY$band1, main = "LandTempDay", xlab = NA)
dev.print(png, file="Histograms.png", width=600, height=700)
#Transform data
SCEC$logcec <- log(SCEC$cec)
SDepth$sqrtdepth <- sqrt(SDepth$Depth)
SLOPE$logslope <- log(SLOPE$band1)
#Histogram of log transformed data
par.ori <- par(no.readonly = TRUE)
par(mfrow=c(3,3))
hist(SCEC$logcec, main= 'Log CEC', xlab=NA)
hist(SDepth$sqrtdepth, main= 'Sqrt Depth', xlab=NA)
hist(SLOPE$logslope, main= 'log Slope', xlab=NA)
dev.print(png, file="TransformedHistograms.png", width=600, height=700)
#Look at correlation of data
cor(SpH@data)
```

2. RK_pH

```
setwd('C:/Users/User/Desktop/Thesis/c002CharlWong/Workspace/a1ExternalFormattedData/RegressionKriging')
# Load libraries
library(sp)
library(rgdal)
library(maptools)
library(gstat)
library(raster)
#Read Soil Profiles
SpH <- read.table("SoilProfilepH.txt", header = TRUE)
# Make spatial
coordinates(SpH) <- ~x+y
# Read covariates
DEM <- readGDAL("prb2_demofaoi.txt")
SLOPE <- readGDAL("prb3_slopeaoi.txt")
SWI <- readGDAL("psb4_swiofaoi.txt")
ASPECT <- readGDAL("prb5_aspectaoi.txt")
CURVATURE <- readGDAL("prb6_curvatureaoi.txt")
RELATIVEELEVATION <- readGDAL("prb17_relativeElevationaoi.txt")
EVI <- readGDAL("psb07_eviofaoi.txt")
LANDCOVER <- readGDAL("psb08_landcoveraoi.txt")
LANDSURFACTEMPNIGHT <- readGDAL("psb09_landsurfacetempnightaoi.txt")
LANDSURFACTEMPDAY <- readGDAL("psb10_landsurfacetempdayaoi.txt")
GEOLOGY <- readGDAL("prb18_geologyaoi.txt")
# Add explanatory data to SpH object
SpH$dem <- over(SpH, DEM)$band1
SpH$slope <- over(SpH, SLOPE)$band1
SpH$swi <- over(SpH, SWI)$band1
SpH$aspect <- over(SpH, ASPECT)$band1
SpH$curv <- over(SpH, CURVATURE)$band1
```

```
SpH$relelev <- over(SpH, RELATIVEELEVATION)$band1
SpH$evi <- over(SpH, EVI)$band1
SpH$landcov <- over(SpH, LANDCOVER)$band1
SpH$tempnight <- over(SpH, LANDSURFACTEMPNIGHT)$band1
SpH$tempday <- over(SpH, LANDSURFACTEMPDAY)$band1
SpH$geology <- over(SpH, GEOLOGY)$band1
#Log transform data
SpH$logslope <- log(SpH$slope)
#Hist new log transformed data
hist(SpH$logslope, main = "Log SLOPE", xlab = NA)
#Look at correlation of data
cor(SpH@data)
#Convert to factors
SpH$geology <- as.factor(SpH$geology)
SpH$landcov <- as.factor(SpH$landcov)
#Regression of ph data
regph <- step(lm(pH~dem+logslope+swi+aspect+curv+relelev+evi+landcov+tempnight+tempday+geology, data =
SpH))
summary(regph)
# Read MAsk
mask <- readGDAL("maskaoi.txt")
# Add residual to SpH
SpH$res <- regph$residuals
# Calculate variogram of residual
SpHres <- gstat(id="resids", formula = res~1, data = SpH)
vSpHres <- variogram(SpHres, boundaries=c(1000, 3:15*2000))
plot(vSpHres, plot.nu=T)
vgmph1 <- vgm(nugget = .35, psill = 0.45, range = 10000, model = "Sph")
plot(vSpHres,vgmph1, plot.nu=T)
vgmres <- fit.variogram(vSpHres, vgmph1, fit.method=7)
plot(vSpHres, vgmres, main="RK pH Variogram")
dev.print(png, file="RKVariogramFittedResidualPH.png", width=600, height=700)
vgmres
# Interpret results
summary(regph)
summary(regph)$r.squared
# Kriging of data
SpH.rk <- krige(res~1, SpH, newdata = mask, vgmres, beta=0, debug.level=-1)
#Create data frame
regdata <- data.frame(swi=SWI$band1, relelev=RELATIVEELEVATION$band1, evi=EVI$band1)
#Predict NA areas
lm_pred <- predict(regph, regdata)
#Add Residuals
SpH.rk$pH.pred <- SpH.rk$var1.pred + lm_pred
#Plot Predictions of pH
spplot(SpH.rk, zcol = "pH.pred", col.regions = bpy.colors(), xlim = c(-178000,-60000),
      ylim = c(1420000,1575000), main="pH ")
dev.print(png, file="RK_PHPred.png", width=600, height=700)
#Calculate evaluation measure
crossval <- function(linmodel, vgm1, alldata){
  output <- numeric()
  for (i in 1:nrow(alldata)){
    pred <- krige(res~1, alldata[-i,], newdata = alldata[i,], vgm1, beta=0)
    lin <- predict(linmodel, alldata[i,])
    pred <- pred$var1.pred+lin
    output <- c(output, pred - alldata[i,]$pH)
```

```
  }
 return(output)
}
# call function crossval
cv <- crossval(regph, vgmres, SpH)
obs <- SpH$pH
pred <- cv+SpH$pH
summary(obs)
summary(pred)
# correlation observed and predicted, ideally 1
cor(obs, pred)
#Mean Absolute Error
me <- function(obs, pred)(mean(obs-pred))
me(obs, pred)
#RMSE
rmse <- function(obs, pred) sqrt(mean((obs-pred)^2))
rmse(obs, pred)
```

3. CAT_pH with SM1M

```
setwd('C:/Users/User/Desktop/Thesis/c002CharlWong/Workspace/a1ExternalFormattedData/CATRegressionKrigin
g')
# Load libraries
library(sp)
library(rgdal)
library(maptools)
library(gstat)
library(rgeos)
library(raster)
#Read Soil Profiles
SpH <- read.table("SoilProfilepH.txt", header = TRUE)
# make spatial
coordinates(SpH) <- ~x+y
# Read covariates
DEM <- readGDAL("prb2_demofaoi.txt")
SLOPE <- readGDAL("prb3_slopeaoi.txt")
SWI <- readGDAL("psb4_swiofaoi.txt")
ASPECT <- readGDAL("prb5_aspectaoi.txt")
CURVATURE <- readGDAL("prb6_curvatureaoi.txt")
RELATIVEELEVATION <- readGDAL("prb17_relativeElevationaoi.txt")
EVI <- readGDAL("psb07_eviofaoi.txt")
LANDCOVER <- readGDAL("psb08_landcoveraoi.txt")
LANDSURFACTEMPNIGHT <- readGDAL("psb09_landsurfacetempnightaoi.txt")
LANDSURFACTEMPDAY <- readGDAL("psb10_landsurfacetempdayaoi.txt")
GEOLOGY <- readGDAL("prb18_geologyaoi.txt")
SM1M <- readGDAL("soilmap1m.txt")
# Add explanatory data to SpH object
SpH$dem <- over(SpH, DEM)$band1
SpH$slope <- over(SpH, SLOPE)$band1
SpH$swi <- over(SpH, SWI)$band1
SpH$aspect <- over(SpH, ASPECT)$band1
SpH$curv <- over(SpH, CURVATURE)$band1
SpH$relelev <- over(SpH, RELATIVEELEVATION)$band1
SpH$evi <- over(SpH, EVI)$band1
```

```
SpH$landcov <- over(SpH, LANDCOVER)$band1
SpH$tempnight <- over(SpH, LANDSURFACTEMPNIGHT)$band1
SpH$tempday <- over(SpH, LANDSURFACTEMPDAY)$band1
SpH$geology <- over(SpH, GEOLOGY)$band1
SpH$sm1m <- over(SpH, SM1M)$band1
#Log transform data
SpH$logslope <- log(SpH$slope)
#Convert to factors
SpH$geology <- as.factor(SpH$geology)
SpH$landcov <- as.factor(SpH$landcov)
SpH$sm1m <- as.factor(SpH$sm1m)
#Regression of ph data
regph <- step(lm(pH~dem+logslope+swi+aspect+curv+relelev+evi+landcov+tempnight+tempday+geology+sm1m,
data = SpH))
summary(regph)
regph <-lm(pH~swi+relelev+evi+sm1m, data = SpH)
summary(regph)
# Read MAsk
mask <- readGDAL("maskaoi.txt")
# Add residual to SpH
SpH$res <- regph$residuals
# Calculate variogram of residual
SpHres <- gstat(id="resids", formula = res~1, data = SpH)
vSpHres <- variogram(SpHres, boundaries=c(1000,6000,8000,16000,19000,21000,10:12*2300))
plot(vSpHres, plot.nu=T)
vgmph1 <- vgm(nugget = .36, psill = 0.45, range = 9500, model = "Sph")
plot(vSpHres, vgmph1, main="Variogram CAT SM1M pH")
vgmres <- fit.variogram(vSpHres, vgmph1, fit.method=7)
plot(vSpHres, vgmres, main="Variogram CAT SM1M pH")
dev.print(png, file="CATVariogramFittedResidualPHSM1M.png", width=600, height=700)
vgmres
# Interpret results
summary(regph)
summary(regph)$r.squared
# Kriging of residual data
SpH.CAT <- krige(res~1, SpH, newdata = mask, vgmres, beta=0, debug.level=-1)
#Create data frame
regdata <- data.frame(swi=SWI$band1, relelev=RELATIVEELEVATION$band1, evi=EVI$band1,
sm1m=as.factor(SM1M$band1))
#Predict NA areas
lm_pred <- predict(regph, regdata)
#Add Residuals
SpH.CAT$pH.pred <- SpH.CAT$var1.pred + lm_pred
#Plot Predictions of pH
spplot(SpH.CAT, zcol = "pH.pred", col.regions = bpy.colors(), xlim = c(-178000,-60000),
      ylim = c(1420000,1575000), main="pH using SM1M")
dev.print(png, file="CAT_PHPredSM1M.png", width=600, height=700)
#Calculate evaluation measure
crossval <- function(linmodel, vgm1, alldata){
  output <- numeric()
  for (i in 1:nrow(alldata)){
    pred <- krige(res~1, alldata[-i,], newdata = alldata[i,], vgm1, beta=0, debug.level=-1)
    lin <- predict(linmodel, alldata[i,])
    pred <- pred$var1.pred+lin
    output <- c(output, pred - alldata[i,]$pH)
  }
```

```
  return(output)
}
# call function crossval
cv <- crossval(regph, vgmres, SpH)
obs <- SpH$pH
pred <- cv+SpH$pH
summary(obs)
summary(pred)
# correlation observed and predicted, ideally 1
cor(obs, pred)
#Mean Error
me <- function(obs, pred)(mean(obs-pred))
me(obs, pred)
#RMSE
rmse <- function(obs, pred) sqrt(mean((obs-pred)^2))
rmse(obs, pred)
```

4. STK_pH with SM1M

```
setwd('C:/Users/User/Desktop/Thesis/c002CharlWong/Workspace/a1ExternalFormattedData/StratifiedKriging')
# Load libraries
library(sp)
library(rgdal)
library(maptools)
library(gstat)
library(rgeos)
library(raster)
#Read Soil Profiles
SpH <- read.table("SoilProfilepH.txt", header = TRUE)
# Make spatial
coordinates(SpH) <- ~x+y
# Read Soil Map 1:1.000.000
soilmap1m <- readGDAL("soilmap1m.txt")
#Overlay Soil Map Soil Profiles
SpH$soilmap1m <- over(SpH, soilmap1m)$band1
#Convert to factors
SpH$soilmap1m <- as.factor(SpH$soilmap1m)
#Create subset
SpHpart.1 <- SpH[ which(SpH$soilmap1m==1), ]
SpHpart.2 <- SpH[ which(SpH$soilmap1m==2), ]
SpHpart.3 <- SpH[ which(SpH$soilmap1m==3), ]
#Calculate Standaard deviation pH per stratum
Stdev1 <- sd(SpHpart.1$pH)
Stdev2 <- sd(SpHpart.2$pH)
Stdev3 <- sd(SpHpart.3$pH)
#Standardized pH values
SpHpart.1$StdpH <- (SpHpart.1$pH)/Stdev1
SpHpart.2$StdpH <- (SpHpart.2$pH)/Stdev2
SpHpart.3$StdpH <- (SpHpart.3$pH)/Stdev3
#Create dummy coordinates
SpHpart.1$dummycoordinates <- coordinates(SpHpart.1) + 400000
SpHpart.2$dummycoordinates <- coordinates(SpHpart.2)
SpHpart.3$dummycoordinates <- coordinates(SpHpart.3) - 700000
#Make data frame of the part to add the dummy coordinates
```

```
SpHpart.1 <- as(SpHpart.1, "data.frame")
coordinates(SpHpart.1) = ~ dummycoordinates.x + dummycoordinates.y
SpHpart.2 <- as(SpHpart.2, "data.frame")
coordinates(SpHpart.2) = ~ dummycoordinates.x + dummycoordinates.y
SpHpart.3 <- as(SpHpart.3, "data.frame")
coordinates(SpHpart.3) = ~ dummycoordinates.x + dummycoordinates.y
#Combine the data frame and add dummy coordinates
SpH.std <- rbind((SpHpart.1), (SpHpart.2),(SpHpart.3))
#Create one part variogram to use as general variogram in prediction
gph1 <- gstat(formula = StdpH~1, data = SpH.std)
vgph1 <- variogram(gph1,width= 1e5, boundaries =
c(800,1500,3000,6000,10000,12000,14500,19500,25500,26000,30000))
plot(vgph1, plot.nu=T)
vgmph1 <- vgm(nugget = .46, psill = 0.44, range = 8000, model = "Sph")
plot(vgph1, vgmph1, plot.nu=T)
vgmph1 <- fit.variogram(vgph1, vgmph1, fit.method=7)
plot(vgph1, model = vgmph1, main="Variogram STK pH SM1M")
vgmph1
dev.print(png, file="STKVariogramFittedResidualpH1M.png", width=600, height=700)
#Make data frame of the part
SpHpart.1 <- as(SpHpart.1, "data.frame")
coordinates(SpHpart.1) = ~ x + y
SpHpart.2 <- as(SpHpart.2, "data.frame")
coordinates(SpHpart.2) = ~ x + y
SpHpart.3 <- as(SpHpart.3, "data.frame")
coordinates(SpHpart.3) = ~ x + y
#Predict
x1 <- krige(StdpH ~ 1, SpHpart.1, newdata= subset(soilmap1m,soilmap1m$band1 ==1), vgmph1, debug.level=-1)
x2 <- krige(StdpH ~ 1, SpHpart.2, newdata= subset(soilmap1m,soilmap1m$band1 ==2), vgmph1,debug.level=-1)
x3 <- krige(StdpH ~ 1, SpHpart.3, newdata= subset(soilmap1m,soilmap1m$band1 ==3), vgmph1, debug.level=-1)
#Transform predicted value back by multiplying with the st.dev.
x1$pred.val <- x1$var1.pred * Stdev1
x2$pred.val <- x2$var1.pred  * Stdev2
x3$pred.val <- x3$var1.pred  * Stdev3
#Combine and plot predictions
SpH.stk <- rbind(as.data.frame(x1), as.data.frame(x2),as.data.frame(x3))
coordinates(SpH.stk) <- c("x", "y")
SpH.stk <- as(SpH.stk, "SpatialPixelsDataFrame")
#Plot prediction map
spplot(SpH.stk["pred.val"], col.regions = bpy.colors(), main="pH using SM1M",
      xlim = c(-178000,-60000), ylim = c(1420000,1575000), sp.layout=list("sp.points",SpH,pch=1))
dev.print(png, file="STK_pHPredSM1M.png", width=600, height=700)
#Calculate evaluation measure
SpH.stk.cv1 <- krige.cv(StdpH ~ 1, SpHpart.1, vgmph1, debug.level=-1)
SpH.stk.cv2 <- krige.cv(StdpH ~ 1, SpHpart.2, vgmph1,debug.level=-1)
SpH.stk.cv3 <- krige.cv(StdpH ~ 1, SpHpart.3,  vgmph1, debug.level=-1)
#Transform predicted value back by multiplying with the st.dev.
y1 <- SpH.stk.cv1$observed  * Stdev1
y2 <- SpH.stk.cv2$observed * Stdev2
y3 <- SpH.stk.cv3$observed  * Stdev3
z1 <- SpH.stk.cv1$var1.pred  * Stdev1
z2 <- SpH.stk.cv2$var1.pred  * Stdev2
z3 <- SpH.stk.cv3$var1.pred  * Stdev3
obs <- c(y1, y2,y3)
pred <- c(z1,z2,z3)
summary(obs)
```

```
summary(pred)
# correlation observed and predicted, ideally 1
cor(obs, pred)
#Mean Error
me <- function(obs, pred)(mean(obs-pred))
me(obs, pred)
#RMSE
rmse <- function(obs, pred) sqrt(mean((obs-pred)^2))
rmse(obs, pred)
```

5. OBS_pH with SM500K

```
setwd('C:/Users/User/Desktop/Thesis/c002CharlWong/Workspace/a1ExternalFormattedData/OBS')
# Load libraries
library(sp)
library(rgdal)
library(maptools)
library(gstat)
library(raster)
# Read pH Soil Map 1:500.000
soilmap500kph <- readGDAL("obsph500k.txt")
#Read Soil Profiles
SpH <- read.table("SoilProfilepH.txt", header = TRUE)
# make spatial
coordinates(SpH) <- ~x+y
#Overlay Observed soil map 1:500.000 with soil profiles
SpH$soilmap500kph <- over(SpH, soilmap500kph)$band1
obs <- SpH$pH
pred <-SpH$soilmap500kph
summary(obs)
summary(pred)
# correlation observed and predicted, ideally 1
cor(obs, pred)
#Mean Absolute Error
me <- function(obs, pred)(mean(obs-pred))
me(obs, pred)
#RMSE
rmse <- function(obs, pred) sqrt(mean((obs-pred)^2))
rmse(obs, pred)
 #Plot soil property pH sm500K
spplot(soilmap500kph, col.regions = bpy.colors(), main="pH using SM500K",
     xlim = c(-178000,-60000), ylim = c(1420000,1575000))
dev.print(png, file="OBS_pHPredSM500K.png", width=600, height=700)
```

6. OBSRES_pH with SM500K

```
setwd('C:/Users/User/Desktop/Thesis/c002CharlWong/Workspace/a1ExternalFormattedData/OBSRES')
# Load libraries
library(sp)
library(rgdal)
library(maptools)
library(gstat)
```

```
library(raster)
#Read Soil Profiles
SpH<- read.table("SoilProfilepH.txt", header = TRUE)
# make spatial
coordinates(SpH) <- ~x+y
# Read covariates
SM500KpH<- readGDAL("obsph500k.txt")
#Add SpH$pH to obs
obs <- SpH$pH
# Make new variable pred1
pred1 <- over(SpH, SM500KpH)$band1
# Calculate and residual and add to SpH
SpH$res <- pred1 - obs
# Calculate variogram of residual
SpHres <- gstat(formula = res~1, data = SpH)
vSpHres <- variogram(SpHres, boundaries=c(500, 1000,
3000,7500,9100,10000,12000,15000,20000,22000,24000,27000))
plot(vSpHres, plot.nu=T)
vgmDepth1 <- vgm(nugget = 2.4, psill = 11, range = 4500, model = "Sph")
plot(vSpHres,vgmDepth1)
vgmres <- fit.variogram(vSpHres, vgmDepth1, fit.method=7)
plot(vSpHres, vgmres, main="OBSRES pH fitted variogram SM500K")
dev.print(png, file="OBSRESVariogramFittedResidualpHSM500K.png", width=600, height=700)
vgmres
# Read MAsk
mask <- readGDAL("maskaoi.txt")
# Simple kriging of residual
SpH.obsres <- krige(res~1, SpH, newdata = mask, vgmres, beta=0, debug.level=-1)
#Add Residuals to OBS result
SpH.obsres$pH.pred <- SpH.obsres$var1.pred+ SM500KpH$band1
#Plot Predictions of pH
spplot(SpH.obsres, zcol = "pH.pred", col.regions = bpy.colors(), xlim = c(-178000,-60000),
      ylim = c(1420000,1575000), main=" pH using SM500K")
dev.print(png, file="OBSRES_pHPredSM500K.png", width=600, height=700)
#Calculate evaluation measure
crossval <- function(vgm1, alldata){
  output <- numeric()
  for (i in 1:nrow(alldata)){
    pred <- krige(res~1, alldata[-i,], newdata = alldata[i,], vgm1, beta=0)
    pred <- pred$var1.pred
    output <- c(output, pred )
  }
 return(output)
}
# call function crossval
cv <- crossval(vgmres, SpH)
obs <- SpH$pH
pred <- cv+pred1
summary(obs)
summary(pred)
# correlation observed and predicted, ideally 1
cor(obs, pred)
#Mean Error
me <- function(obs, pred)(mean(obs-pred))
me(obs, pred)
#RMSE
```

```
rmse <- function(obs, pred) sqrt(mean((obs-pred)^2))
rmse(obs, pred)
```