PLOS ONE

RESEARCH ARTICLE

High Resolution Mapping of Soil Properties Using Remote Sensing Variables in South-Western Burkina Faso: A Comparison of Machine Learning and Multiple Linear Regression Models

Gerald Forkuor[1☯], Ozias K. L. Hounkpatin[2☯*], Gerhard Welp[2], Michael Thiel[3]

1 West African Science Service Centre on Climate Change and Adapted Land Use—WASCAL, Burkina Faso, **2** University of Bonn, Institute of Crop Science and Resource Conservation (INRES), Soil Science and Soil Ecology, Nussallee 13, Bonn, Germany, **3** University of Wuerzburg, Remote Sensing Unit, Oswald-Kuelpe-Weg 86, Wuerzburg, Germany

☯ These authors contributed equally to this work.
* hozias@uni-bonn.de

Citation: Forkuor G, Hounkpatin OKL, Welp G, Thiel M (2017) High Resolution Mapping of Soil Properties Using Remote Sensing Variables in South-Western Burkina Faso: A Comparison of Machine Learning and Multiple Linear Regression Models. PLoS ONE 12(1): e0170478. doi:10.1371/journal.pone.0170478

Editor: Dafeng Hui, Tennessee State University, UNITED STATES

Received: June 20, 2016

Accepted: January 5, 2017

Published: January 23, 2017

Data Availability Statement: The data underlying the findings can be accessed from the Figshare public repository (https://figshare.com/s/abe41649a5b3c5950372) except for the RapidEye data. Some access restrictions applied to the RapidEye data due to the agreement between WASCAL and RapidEye Science Archive (RESA) that these data will be used only within the WASCAL project. However, researchers who are interested in the RapidEye data can write to Michael Thiel (email: michael.thiel@uni-wuerzburg.

Abstract

Accurate and detailed spatial soil information is essential for environmental modelling, risk assessment and decision making. The use of Remote Sensing data as secondary sources of information in digital soil mapping has been found to be cost effective and less time consuming compared to traditional soil mapping approaches. But the potentials of Remote Sensing data in improving knowledge of local scale soil information in West Africa have not been fully explored. This study investigated the use of high spatial resolution satellite data (RapidEye and Landsat), terrain/climatic data and laboratory analysed soil samples to map the spatial distribution of six soil properties–sand, silt, clay, cation exchange capacity (CEC), soil organic carbon (SOC) and nitrogen–in a 580 km^2 agricultural watershed in south-western Burkina Faso. Four statistical prediction models–multiple linear regression (MLR), random forest regression (RFR), support vector machine (SVM), stochastic gradient boosting (SGB)–were tested and compared. Internal validation was conducted by cross validation while the predictions were validated against an independent set of soil samples considering the modelling area and an extrapolation area. Model performance statistics revealed that the machine learning techniques performed marginally better than the MLR, with the RFR providing in most cases the highest accuracy. The inability of MLR to handle non-linear relationships between dependent and independent variables was found to be a limitation in accurately predicting soil properties at unsampled locations. Satellite data acquired during ploughing or early crop development stages (e.g. May, June) were found to be the most important spectral predictors while elevation, temperature and precipitation came up as prominent terrain/climatic variables in predicting soil properties. The results further showed that shortwave infrared and near infrared channels of Landsat8 as well as soil specific indices of redness, coloration and saturation were prominent predictors in digital soil mapping. Considering the increased availability of freely available Remote Sensing data (e.g. Landsat, SRTM, Sentinels), soil information at local and

de) at the Department of Geography and Geology, Würzburg University, Germany for data access arrangements.

Funding: The authors would like to acknowledge the German Federal Ministry of Education and Research (BMBF) for providing financial support for conducting this research. The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

regional scales in data poor regions such as West Africa can be improved with relatively little financial and human resources.

Introduction

Accurate and detailed spatial soil information is essential for sustainable land use and management as well as environmental modelling and risk assessment. In West Africa, where land degradation and loss in soil fertility has been reported by numerous studies [1–3], such information is increasingly required by governments and development partners to aid in improving land management [4]. High resolution spatial information on soils can assist decision makers to better target areas for soil fertility interventions and implement knowledge-based policies that aim at increasing agricultural production and improving livelihoods of small scale farmers in the subregion. This is even crucial for the sustainable use of the soil resources particularly in the context of climate change [5].

Traditional soil mapping approaches have mostly relied on ground-based surveys. Classical field surveys including soil sampling and laboratory analyses are reported to be time consuming and expensive, especially when mapping is being done at national, regional or global scales [6–8]. In view of this bottleneck, new techniques for obtaining high-resolution soil information are being developed and still have to be optimized. Digital soil mapping, which incorporates secondary (non-soil) data sources into the mapping process, has been identified as a potential means of overcoming the limitations of traditional approaches and improving the detail and spatial coverage of soil databases [8–10]. Apart from its cost effectiveness, digital approaches allow a determination of an objective quantitative measure of prediction uncertainty, which is often not provided when traditional approaches are employed [11]. Remote Sensing (RS) data have come up in the last few decades as promising secondary data sources for improving digital soil mapping at all scales. Remotely sensed data sources: (1) contain extractable soil information, e.g. spectral reflectance, (2) have large spatial coverage and therefore permit mapping of inaccessible areas, (3) produce consistent and comprehensive data both in time and space and (4) offer possibilities of supplementing or at least reducing traditional soil sampling in soil surveys [12]. Based on these advantages, numerous studies have explored the use of RS data with varying spatial, temporal and spectral characteristics in digital soil mapping [8,13,14].

Saadat et al. [15] combined imagery from the Advanced Spaceborne Thermal Emission and Reflectance Radiometer (ASTER) sensor and a digital elevation model (DEM) for landform classification in Iran. They found that the spectral information in the RS data increased the possibility of distinguishing topographically similar landforms and subsequently improved the classification. Ehsani and Quiel [16] arrived at a similar conclusion when they used Landsat and Shuttle Radar Topographic Mission (SRTM) DEM for analysing landscape elements in Eastern Europe. Dobos et al. [6] found the combination of coarse resolution AVHRR (Advanced Very High Resolution Radiometer) data and DEM derived terrain derivatives to be promising data for characterizing soil forming environments and delineating soil patterns at national and continental scales. Compared to the terrain derivatives, the spectral information in the AVHRR bands were noted to have contributed more to the accurate delineation of soil types. Hahn and Gloaguen [17] underscored the importance of remotely sensed terrain variables (e.g. altitude, aspect, slope) as input to soil type classification in Germany. In a regional scale analysis, Scudiero et al. [18] found, among various variables, that surface reflectance of multi-year Landsat data was a useful indicator for characterizing the spatial variability of soil salinity in the western San Joaquin valley of California. Other studies also demonstrated the contribution of RS data in

mapping soil properties such as sand, silt, clay and soil organic carbon (SOC) based on reasonable correlations between soil properties and reflectance spectra [14,19,20].

Despite many advances, further exploration of the application of RS data to soil mapping is required, especially in data poor regions such as West Africa. This is in light of the increasing availability of RS data, some of which are provided free of charge (e.g. Landsat, SRTM, Sentinel-1, -2) [8,21]. Research on the potential of RS data to improve digital soil mapping in West Africa is sparse [22]. Recent digital mapping initiatives on the continent (e.g. African Soil Information Service - http://africasoils.net/) [11] and at national scales (e.g. [23]) have used RS and other environmental variables in mapping soil units and properties. However, the spatial resolution of these studies is still coarse (ca. 250–1000 m), and may be of limited use for local scale (e.g. watershed) analysis. The derivation of digital soil data at local scales is important for assessing landscape scale resource needs and subsequently aid in regional, national and global soil and agricultural monitoring efforts [4,6]. Moreover, the success of digital soil mapping is to a large extent dependent on the availability, quality and timing of RS data acquisitions [24]. Land surface characteristics, especially on agricultural lands, are subject to temporal changes and it is not always clear which periods of the year are suitable for acquiring RS data for accurate soil property prediction. The use of multi-temporal images permits an investigation on the impact of the temporal window of RS data acquisition on prediction accuracies.

This paper reports findings of a digital soil mapping effort that integrated RS data and conventionally analysed soil samples to map the spatial distribution of soil properties (sand, silt, clay, cation exchange capacity, SOC and nitrogen) in a 580 km^2 agricultural watershed in south-western Burkina Faso. High spatial resolution multi-temporal RapidEye and Landsat imagery together with ASTER Global DEM terrain derivatives were tested to determine their suitability for improving the availability and accuracy of spatial soil information in rural African landscapes. Since typical farm sizes in West Africa are small (i.e. less than one hectare [25]), the use of such high spatial resolution RS data for digital soil mapping at local scales is important and beneficial for optimizing farm management. However, such studies, to the best of our knowledge, are rare.

Four statistical methods which have proved their suitability for digital soil mapping in previous studies—multiple linear (MLR), random forest regression (RFR), support vector machine (SVM) and stochastic gradient boosting (SGB) [26–29] were explored to ascertain the most suitable method for high resolution RS data in the study region. Comparison of the traditional regression method (MLR) and different machine learning methods to spatially predict soil properties in West Africa are scarce. The research questions that the study addresses are: (1) which regression method offers the best accuracy for predicting soil properties? (2) What is the optimal time of RS data acquisition for predicting soil properties?

Materials and Methods

Study area

The study was conducted in a rural watershed that falls in the Ioba province in south-western Burkina Faso (South-west Region). It has an area of about 580 km^2 and lies between latitudes 11˚ 21' 50" and 11˚ 04' 27"N and longitudes 003˚ 08' 37" and 002˚ 50' 15"W (Fig 1). Detailed soil sampling was carried out in a sub-watershed which is about one-quarter of the watershed (Fig 1). The watershed has a uni-modal rainfall distribution (May-October), with an annual rainfall average of about 900 mm [30] while daily temperature ranges between 20.1 (minimum) and 34^0 C (maximum). The lithology is composed of partly volcanic formations from the middle precambrian period and is made up mainly of andesic rocks with massive texture, basalt, diabase, gabbro and quartz-rich andesites. The study area is dominated by Plinthosols

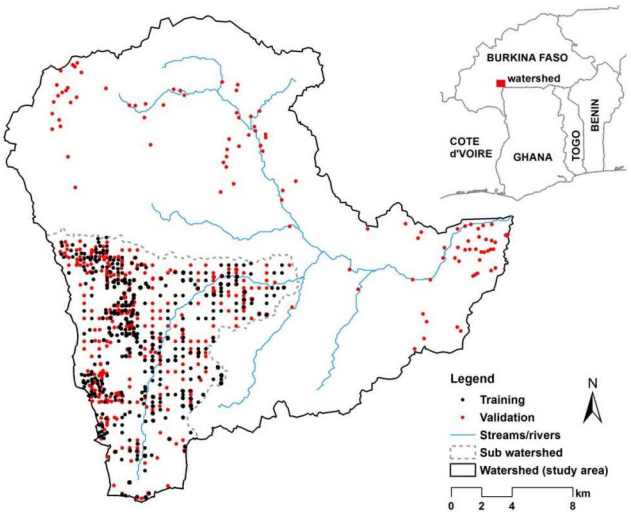

Fig 1. Map of the study watershed and locations of soil sampling.

doi:10.1371/journal.pone.0170478.g001

(70%) with inclusions of haplic Gleysols (15%), vertic Cambisols (14%), haplic Leptosols (1.7%) and stagnic Lixisols (0.2%). The terrain is flat, with elevation below 600 m above mean sea level and average slope of less than 5˚ [31].

Population density of the study area ranges between 40–60 persons/km^2, most of whom engage in small scale rainfed farming as their main source of occupation [32]. Agriculture is the major land use and covers about 45% of the watershed, while different forms of vegetation (e.g. forest, shrub, grass), artificial surfaces (e.g. settlements, bare areas) and water bodies cover 52%, 2% and 1%, respectively [25].

Input data

Soil sampling and analysis. Representative soil units were chosen for sampling based on existing soil [33], land use [34] and DEM [31] data of the watershed. The focus of the sampling was a sub-watershed (see Fig 1). A total of 1104 soil samples (1002 in sub-watershed and 102 outside) coming mainly from the topsoil (0–30 cm), were considered in this study. They were taken from the topsoil of 35 profiles along with intensive auger sampling carried out from July to October 2012 and from July to October 2013. At each auger sampling point, composite samples were taken from the topsoil (0–30 cm). These samples were dried at 40˚C and sieved to 2 mm.

Because of the high number of soil samples, we analysed only 100 samples conventionally for the soil properties under study (i.e. texture–sand, silt, clay; nitrogen (N), SOC, cation exchange capacity (CEC)) following the procedures described in [35]. For the rest of the sample set, we used mid infrared spectroscopy (MIRS) to predict the above mentioned soil properties. The estimation of soil properties is generated by calibrating spectral information against conventionally obtained data using multivariate statistical procedures such as partial least squares regression (PLSR [36]).

For spectra measurement, about 20 mg of the soil samples were set into microplates and compacted with a plunger to get a level and plain surface in five replicates. The spectra were recorded using a Bruker Tensor 27 with an automated high throughput device (Bruker HTS-XT; Bruker Optik, Ettlingen, Germany). For each spectrum, 120 scans were recorded from 8000 to 580 cm^{-1} at a resolution of 4 cm^{-1} [37]. The software package OPUS QUANT (© 2006 Bruker Optik GmbH) was used for spectra analyses and for the prediction of soil properties with PLSR. OPUS QUANT

Table 1. Statistical parameters of the mid infrared spectroscopy-partial least squares regression prediction models (n = 100 samples) and of the predicted dataset (n = 1104 samples).

Parameters	Full cross-validation				Test-set validation (V = 10%)				Predicted dataset			
	R^2 (%)	RMSECV	RPD	Slope	R^2 (%)	RMSEP	RPD	Slope	Min	Max	Mean	SD
Sand (%)	70.5	6.8	1.8	0.7	80.9	5.7	2.5	0.7	3.5	66.2	37.6	9.6
Silt (%)	75.8	4.9	2	0.8	88.2	3.9	3	0.8	22.4	67.1	45.1	8.6
Clay (%)	77.6	6.2	2.1	0.8	80.6	5.5	2.4	0.8	0	54.5	20.8	8
CEC (cmolc kg^{-1})	75.6	3.6	2	0.8	90.5	3.2	3.6	0.8	0	36.3	12.5	5.9
SOC (%)	95.3	0.1	4.6	0.9	92.2	0.2	3.6	0.9	0	4.4	1.5	0.6
Nitrogen (%)	85.5	0	2.6	0.9	85.7	0	3	0.8	0	0.3	0.1	0.1

RMSECV: root mean square error of cross validation, RMSEP: root mean square error of prediction, RPD: ratio of performance to deviation, V: validation set, SD: standard deviation.

doi:10.1371/journal.pone.0170478.t001

uses a routine that automatically tests combinations of varying spectral ranges and data treatments for the optimum prediction power of the model. For each soil parameter, we conducted calibration procedures employing a leave–one–out, full–cross validation as well as a test-set calibration for checking model robustness as described by Bornemann et al. [37] (Table 1). The quality of the different models for each soil property was assessed based on their predictive ability with the R^2, ratio of performance to deviation (RPD) and the standard error of prediction (SEP). For more technical information, readers are referred to [38]. Only models exhibiting good predictive ability (RPD>2) or close to that (RPD 1.7–2.0) [39] were used to make predictions for the remaining samples (Table 1). As seen in the table, the MIRS cross validation showed that SOC, followed by N presented the best prediction accuracy based on the R^2 and the RPD. Additionally, the error metrics from the MIRS test-set validation confirmed the robustness of the different calibration models for all soil properties with $R^2 \geq 80\%$ and with RPD>2.

Satellite spectral data. Multi-temporal data from two optical sensors, RapidEye and Landsat, were used in this study. The images were acquired on 1st March, 1st April, 3rd May 2013 (RapidEye) and 13th June 2013 (Landsat). This period was selected to coincide with the peak of the dry season and the ploughing/planting period during which there's little or no vegetation especially on croplands. RapidEye was obtained from the RapidEye Science Archive team of the German Aerospace Center (DLR) (https://resa.blackbridge.com/), while Landsat 8 was downloaded from the United States Geological Survey's GLOVIS website (http://glovis.usgs.gov). The RapidEye data has five spectral channels (blue, green, red, rededge and near infrared (NIR)) and a spatial resolution of 5 m (i.e. orthorectified, level 3A) [40], while Landsat has eleven spectral channels [41] and a spatial resolution of 30 m, which was later resampled to 5 m to ensure integration with the RapidEye data. Six out of the eleven spectral channels of Landsat (see Table 2) were used in the analysis. Images from both sensors were atmospherically corrected using the ENVI ATCOR software [42]. In addition to the original spectral bands, six soil and vegetation indices were calculated for each image. In all, twenty-one spectral bands and twenty-four spectral indices were derived (i.e. six indices for each of the four images). Table 2 provides further details of the spectral bands of RapidEye and Landsat as well as formulae and definitions of the spectral indices calculated. These spectral indices have been found to be useful in digital soil mapping [43].

Terrain and climatic variables. Terrain variables were extracted from the 30 m resolution ASTER GDEM (http://asterweb.jpl.nasa.gov/GDEM.ASP) [30]. Although previous studies have shown that the 90 m resolution SRTM DEM [45] has a superior absolute accuracy than ASTER GDEM [46], the latter was selected for this study due to its superior spatial resolution. Although

Table 2. Spectral bands of satellite images used and definitions of soil and vegetation indices calculated.

Spectral bands							
	Band number	**1**	**2**	**3**	**4**	**5**	**6**
	RapidEye band name	Blue (B)	Green (G)	Red (R)	Red edge (RdE)	Near infra red (NIR)	
	Landsat band name	Blue (B)	Green (G)	Red (R)	Near infrared (NIR)	Shortwave infrared 1 (SWIR 1)	Shortwave infrared 2 (SWIR 2)
Spectral indices							
	Indices			**Formula**		**Index property**	**Reference**
	Brightness Index (BI)			$((R^2 + G^2 + B^2) / 3)^{0.5}$		Average reflectance magnitude	[43]
	Saturation Index (SI)			(R–B) / (R + B)		Spectral slope	[43]
	Hue Index (HI)			(2 * R–G–B) / (G–B)		Primary colors	[43]
	Coloration Index (CI)			(R–G) / (R + G)		Soil color	[43]
	Redness Index (RI)			$R^2 / (B * G^3)$		Hematite content	[43]
	Normalized Difference Vegetation Index (NDVI)			(NIR–R) / (NIR + R)		Health and amount of vegetation	[44]

doi:10.1371/journal.pone.0170478.t002

the 30 m SRTM data has been made freely available, it came at a time that this manuscript was at an advanced development stage. The data was pre-processed to generate a depressionless DEM prior to the calculation of terrain variables. Climatic data (i.e. mean annual precipitation and temperature over 50 years) at 1 km resolution were obtained from worldclim [47].

In order to ensure integration with the RapidEye data, the DEM and climatic variables were resampled to 5 m resolution using the bilinear and bicubic interpolation methods, respectively. Table 3 lists the 29 terrain and climatic variables that were used in this study together with the relevant references. Most derivatives were calculated using the System for Automated Geoscientific Analysis (SAGA) software, while few were calculated with ArcGIS.

Models

Multiple linear regression (MLR). Linear regression models aim at explaining the spatial distribution of a dependent variable by means of a linear combination of predictors (independent variables). In the case of this study, the various soil parameters are considered the dependent variables while the spectral and terrain/climatic variables are the independent variables. Linear regression models generally have the form:

$$y = a + \sum_{i=1}^{n} b_i * x_i \pm \varepsilon_i \quad (1)$$

where "y" is the dependent variable (soil parameter), "x_i"are the predictors, "n" is the number of predictors, "a" is the intercept, "b_i" are the partial regression coefficients and "ε" is the standard error of estimate. The regression equation is used to predict the spatial distribution of the parameter of interest based on the independent variables.

The "lm" function implemented in the R software [61] was used for MLR analysis. A matrix of predictors was developed by superimposing the training samples on the spectral and terrain/climatic spatial layers and extracting the corresponding values. One soil property was modelled at a time as the response (dependent) variable with the developed matrix as the predictors. For each model, the adjusted R^2 and residual standard error were recorded. In addition, the predictors that were significant at 1% significance level were noted.

Table 3. Terrain and climatic variables considered in this study.

Parameters	Definition	Units	Authors
Slope*	Inclination of the land surface from the horizontal	Radians/%	[48]
Steepest slope	Maximal rate of elevation change ingravitational field	radians	[49]
Curvature	Curvature	degree m^{-1}	[6]
General curvature	Combination of horizontal and vertical curvature	degree m^{-1}	[26]
Plan curvature*	Horizontal (contour) curvature	degree m^{-1}	[27]
Maximum curvature	Maximum Curvature	degree m^{-1}	[50]
Minimum curvature	Minimum Curvature	degree m^{-1}	[50]
Total curvature	Curvature of the surface itself	degree m^{-1}	[51]
Parallel curvature	Parallel curvature	degree m^{-1}	[52]
Rectangle curvature	Rectangle curvature	degree m^{-1}	[52]
Flow line curvature	Flow line curvature	degree m^{-1}	[52]
Profile Curvature	Vertical rate of change of slope	degree m^{-1}	[53]
Horizontal curvature	Measure of flow convergence and divergence	degree m^{-1}	[54]
Flow direction*	Path of water flow	-	[55]
Aspect	Direction the slope faces	degree	[53]
Cose Aspect	Direction the slope faces: eastness	Degree	[56]
Sine Aspect	Direction the slope faces: northness	degree	[56]
Elevation	Vertical distance above sea level	m	[53]
Protection index	Extent at which a cell is protected by relief based on the immediate surrounding cell		[52]
Topographic position index	Location higher or lower than the average of their surroundings		[27]
Saga Wetness Index	Ratio of local catchment area to slope	-	[57]
Flow accumulation*	Ultimate flow path of every cell on the landscape grid	-	[58]
Channel network base Level	Channel network base level elevation	m	[59]
Temperature (mean annual)	Temperature	˚C	[60]
Precipitation (mean annual)	Precipitation	mm	[60]

The variables with (*) were calculated in SAGA as well as ArcGIS due to slight differences in the computational algorithms used by the two software packages.

doi:10.1371/journal.pone.0170478.t003

A common limitation of regression models is the problem of multicollinearity, which occurs when there is significant correlation between the predictors. Since the number of predictors identified in this study are many (seventy-four), and there could be high correlation between some of them, a stepwise regression analysis was first conducted to produce uncorrelated predictors needed to model each soil parameter and thereby minimize the problem of multicollinearity. Stepwise regression identifies a subset of predictors based on the statistical significance of the predictors (using stepwise, forward selection, or backward elimination) [62]. In this study, the "stepAIC" function as implemented in the "MASS" package [62] of the R statistical package was used for the stepwise regression. For each soil parameter, a subset of uncorrelated predictors were identified for subsequent analysis. Table 4 presents the number of spectral and terrain/climatic predictors that were eventually used in the MLR for each soil property. On average, less than 50% of the initial predictors were eventually selected for each soil property with the exception of carbon, for which 53% were selected. In order to ensure

Table 4. Number of spectral and terrain/climatic predictors used in modelling each soil parameter.

Data/Parameter	Sand	Silt	Clay	CEC	SOC	Nitrogen
Spectral	17	22	21	12	26	19
Terrain/climatic	9	10	5	13	12	12
Total	26	32	26	25	38	31

doi:10.1371/journal.pone.0170478.t004

comparison with the Random Forest Regression (RFR), the same set of predictors were maintained for the RFR analysis, although it (RFR) does not greatly suffer from the multicollinearity problem.

Random forest regression (RFR). The RFR analysis was conducted using the "*Random Forest*" (RF) function as implemented in the RF package [63] of the R software [61]. RF [64] belongs to the family of ensemble machine learning algorithms that predicts a response (in this case the respective soil parameters) from a set of predictors (matrix of training data) by creating multiple Decision Trees (DTs) and aggregating their results. Each tree in the forest is independently constructed using a unique bootstrap sample of the training data. Whereas other machine learning algorithms (e.g. bagging and boostrapping [65]) use the best split among all predictors for node splitting, RF chooses the best split from a randomly selected subset of predictors. The introduction of this additional randomness decreases the correlation between trees in the forest, and consequently increases accuracy [66]. Additionally, RF requires no assumption of the probability distribution of the target predictors as with linear regression, and is robust against nonlinearity and overfitting, although overfitting may occur in instances where noisy data are being modelled [67]. For RF modelling, parameters requiring tuning such as the number of trees to grow in the forest (ntree) and the number of randomly selected predictor variables at each node (mtry) were set using the grid search method in the R "caret" package [68] using tenfold cross validation with 5 repetitions.

RF optionally provides information on the relative importance of the predictors (variable importance) used in the construction of the forest [63]. Two importance measures—%IncMSE and IncNodePurity—are frequently computed. To calculate %incMSE (increase in mean standard error), each tree is constructed with and without a predictor. Then, the difference between the two cases is averaged over all trees and normalized by the standard deviation of the differences. The second measure (IncNodePurity) represents the total decrease in node impurity from splitting on a predictor in the tree construction process, averaged over all trees. In RFR, the node impurity is measured by the residual sum of squares [63]. RF computes an internal accuracy measure based on the samples that are omitted from the bootstrapped samples used in the tree construction (i.e. out-of-bag, OOB). The accuracy of the model is given by the mean square error (MSE_{OOB}) of the aggregated OOB predictions generated from the bootstrap subset and is computed as follows [63]:

$$\mathrm{MSE}_{OOB} = n^{-1}\sum_{i=1}^{n}(z_i - \hat{z}_i^{OOB})^2 \quad (2)$$

Where "n" is the number of observations, z_i is the average prediction of the ith observation and $\hat{z}_i^{OOB}$ is the average prediction for the ith observation from all trees for which the observation was OOB.

The explained variance is expressed as follows:

$$\mathrm{Var} = 1 - \frac{MSE_{OOB}}{Var_{resp}} \quad (3)$$

For each soil parameter modelled, the MSE_{OOB} explained variance and variable importance measure were recorded for subsequent analysis and discussion.

Support vector machines for regression (SVM). Initially used for classification, the support vector machine (SVM) has been extended for regression with the prediction of soil properties [28,69]. Relying on Kernel functions, input data are plotted into a new hyperspace where separations are performed. The ultimate purpose is to get an optimal hyperspace for data fitting and prediction using the ε-insensitive loss function, which tolerates errors smaller than the constant ε set as a threshold. Detailed information about SVM can be found in Hastie et al. [70]. The determination of the best parameters (bandwidth cost parameter, insensitive loss function,) for tuning the model for each soil parameters was carried out using the grid search method in the R "caret" package [68]. For this purpose, ten random partitions of the training data with five repetitions was carried for leave-one-group-out cross-validation of the model. Parameters resulting in the lowest root mean square error were considered for modelling.

Stochastic gradient boosting (SGB). Stochastic gradient boosting (SGB; [71,72]) is a hybrid method incorporating both boosting and bagging approaches. First, small classification or regression trees are sequentially built from the residuals of the preceding tree (s). Instead of focusing on the full training set, the SGB carries out a boosting by selecting (without replacement) at each step a random sample of the data leading to a gradual improvement of the model. More details related to the background and mathematical functions behind the SGB can be found in Ridgeway [73]. The required parameters for model fitting (interaction depth, shrinkage rate) were set by using the tenfold cross validation with five repetitions also with the R "caret" package [68]. For each soil property, parameters with the lowest error metric (root mean square error) were used for the final model.

Accuracy assessment

The performance of the four models–MLR, RFR, SVM, SGB–in predicting the soil properties was assessed by using 80% of the detailed soil samples in the sub-watershed (which was the focus of the sampling) (Fig 1) for cross validation. A 10-fold cross-validation scheme with 5 repetitions was applied to ensure model stability and reliability using the "caret" R Package [68]. The remaining 20% served as an independent validation dataset. In order to assess the predictive strength of the models outside the sub-watershed (i.e. the core sampled area), all the soil samples outside the sub-watershed (102 samples) (Fig 1) were reserved for the purposes of accuracy assessment and used as a second independent validation dataset.

Though R^2 is a valid statistic for assessing the prediction accuracy of a model, a high R-squared model may not necessarily lead to accurate predictions. This is because the model could systematically and significantly over- and/or under-predict the data at different points along the regression line. An over-fitted model could also lead to poor predictions [74]. It is, therefore, important to evaluate the models with other performance statistics, preferably based on an independent set of observations, to provide additional information on the prediction accuracy of the models.

For each soil parameter, two error statistics—root mean squared error (RMSE) and the symmetric mean absolute percentage error (sMAPE)—were calculated (see Eqs 4 and 5). The two statistics served as the basis for comparing the performance of the two models in predicting the spatial distribution of the different soil properties. Although RMSE is a frequently used statistic in the literature to indicate the average error of a model [75], its dependence on scale makes it difficult to calculate a model's error in percentage terms. The sMAPE [76], on the other hand, provides a percentage-wise error and facilitates a comparison of the accuracy with which each soil property is predicted. The sMAPE (as defined in this paper), however, can

provide unreliable estimates if either observed or forecasted value is negative [70].

$$RMSE = \left[\frac{1}{n}\sum_{i=1}^{n}(P_i - O_i)^2\right]^{1/2} \quad (4)$$

$$sMAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|O_i - P_i|}{(O_i + P_i)/2} \quad (5)$$

where "P" is the predicted value and "O" is the observed/true value.

Results and Discussion

Model performance

The performance of the four models investigated was assessed based on: (1) model internally generated accuracy statistics and (2) independent validation samples.

Assessment based on internal accuracy statistics. This assessment was achieved by comparing the RMSE and the adjusted R^2 (hereinafter referred to as R^2) derived from the four models for the respective soil parameters. Table 5 presents results of the comparison. R^2 ranged between 21 and 53% for MLR, 18 and 53% for RFR, 20 and 51% for SVM and 16 and 51% for SGB. Silt was the only soil parameter that achieved an R^2 of greater than 50% for all models. The other soil parameters recorded relatively lower R^2, with sand, clay, SOC and nitrogen consistently having R^2 below 40%. The generally low R^2 obtained in this study independently of the models can be attributed to a complex interplay and high variability of environmental factors in the studied watershed and surrounding regions [12,77]. High variability in agricultural soil management practices, nutrient application, vegetation cover and climatic factors (temperature, precipitation) are believed to be among the factors that resulted in the low correlations observed. Nonetheless, the range of R^2 values obtained in this study is comparable to other studies that considered only terrain/climatic covariates [26,77] or only spectral data [43,78].

Table 5 shows that RFR performed marginally better than the other models in generating a model for the soil parameters with relatively lower RMSE and higher R^2. The only exception was in the case of sand and clay, where MLR performed better than the RFR recording better error metrics. Generally, the machine learning methods (RF, SVM, SGB) were found to be more accurate than MLR using the RSME of cross validation for assessing model performance [79,80].

Assessment based on independent validation samples. Tables 6 and 7 present model performance statistics for the external validation inside (20% of the dataset) and outside the small catchment, respectively (see Fig 1). Here, the symmetric mean absolute percentage error (sMAPE) (Eq 5) was calculated and used as the basis for comparing the four models. Inside the small catchment, the RFR generally performed better than the other models, achieving the

Table 5. Internal model validation based on 80% training data (All Spectral and topographic/climate predictors).

Model	Sand		Silt		Clay		CEC		SOC		Nitrogen	
	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2
MLR	7.566	0.346	5.940	0.537	6.946	0.212	4.786	0.357	0.546	0.348	0.038	0.352
RFR	7.586	0.342	5.937	0.538	7.022	0.185	4.689	0.383	0.528	0.39	0.038	0.354
SVM	7.592	0.342	6.091	0.519	6.993	0.206	4.889	0.333	0.551	0.341	0.038	0.339
SGB	7.707	0.318	6.094	0.514	7.164	0.162	4.767	0.360	0.539	0.367	0.038	0.339

doi:10.1371/journal.pone.0170478.t005

Table 6. External validation in small catchment based on 20% testing data with spectral data and terrain/climatic variables.

Model	Sand		Silt		Clay		CEC		SOC		Nitrogen	
	RMSE	sMAPE	RMSE	sMAPE	RMSE	sMAPE	RMSE	sMAPE	RMSE	sMAPE	RMSE	sMAPE
MLR	8.482	0.189	5.900	0.107	6.708	0.239	4.787	0.415	0.541	0.285	0.043	0.290
RFR	7.764	0.180	5.708	0.103	6.590	0.242	4.593	0.318	0.512	0.261	0.041	0.273
SVM	8.415	0.188	5.899	0.107	6.667	0.234	4.897	0.394	0.549	0.283	0.043	0.287
SGB	7.954	0.189	5.819	0.107	6.791	0.242	4.562	0.314	0.526	0.272	0.041	0.286

doi:10.1371/journal.pone.0170478.t006

highest prediction accuracy (i.e. 100-sMAPE) for four soil properties (sand, silt, SOC, nitrogen) while SVM and SGB produced the best prediction for clay and CEC, respectively. Prediction accuracies by the RFR model ranged from a low of 68% for CEC to a high of 90% for silt, with an average accuracy of 77%. Compared to the MLR, for example, RFR improved prediction accuracy by 0.9% for sand, 0.4% for silt, 9.7% for CEC, 2.4% for SOC, and 1.7% for N. Generally, SVM and SGB also outperformed the MLR. In assessing the models' performance outside the small catchment, Table 7 reveals that RFR achieved a better prediction accuracy for silt (85%) and clay (52%), SVM for sand (81%) and SOC (53%), and SGB for CEC (60%) and nitrogen (55%) with prediction accuracies of 69%, 85%, and 52%, respectively. The RFR model achieved an average accuracy of 62% for the validation outside the small catchment.

Compared to MLR, the high performance of RFR and the other machine learning models could be due to the existence of a non-linear relationship between soil parameters and the predictors which MLR could not adequately resolve. Although MLR is widely used in statistical predictions, its limitation in handling non-linear relationships between response and predictor variables, especially in heterogeneous landscapes, has been noted in literature [74,81,82]. Non-parametric models such as RFR, SVM and SGB have been found superior to MLR due to their ability to handle non-linear relations and multi-source data [17,80,83]. In general, many studies reported RFR as providing better predictions compared to SVM [29,84–86]. However, Were et al. [87] found SVM as best predictor for the spatial distribution of SOC stock compared to RFR. Rossel et al. [88] reported RFR as having better prediction accuracy compared to SGB, while Hitziger et al. [89] found the latter superior to the former in soil property prediction. Similarly, SVM and SGB occasionally outperformed RFR in this study. This, and previous results, suggest that no single machine learning algorithm might serve best for every landscape and that many models should be calibrated to identify the most accurate model for prediction.

A comparison of Tables 6 and 7 reveals a general reduction in the predictive accuracy of the models outside the small catchment (which was the focus of sampling), although the magnitude of reduction varies depending on the model and soil property. Taking RFR, for example, the magnitude of reduction in prediction accuracy (i.e. 100-sMAPE) equalled 13% for sand, 4% for silt, 24% for clay, 10% for CEC, 21% for SOC, and 18% for nitrogen. In general, all models performed relatively poor in predicting clay, SOC and nitrogen outside the small catchment, with average accuracy reductions of 28%, 20% and 19%, respectively. On the other hand, the models

Table 7. External validation based on 102 samples outside the small catchment with spectral data and terrain/climatic variables.

Model	Sand		Silt		Clay		CEC		SOC		Nitrogen	
	RMSE	sMAPE	RMSE	sMAPE	RMSE	sMAPE	RMSE	sMAPE	RMSE	sMAPE	RMSE	sMAPE
MLR	17.341	0.547	9.350	0.157	11.804	0.548	5.597	0.469	0.847	0.505	0.059	0.496
RFR	14.115	0.314	8.713	0.146	10.623	0.478	4.891	0.415	0.765	0.472	0.053	0.457
SVM	20.257	0.193	9.106	0.153	14.738	0.566	5.669	0.448	0.750	0.471	0.057	0.488
SGB	15.184	0.341	8.846	0.148	10.875	0.497	4.960	0.398	0.759	0.476	0.051	0.454

doi:10.1371/journal.pone.0170478.t007

performed well in predicting silt and CEC outside the small catchment, showing minimal accuracy reductions of 4% and 7%, respectively. These results suggest that the accuracy of extrapolating soil predictions outside the sampled area may differ depending on the soil property as well as on the non-comparability of the small catchment with regard to surface, land use and other characteristics.

Despite these differences, the accuracies achieved in the external validation can be assumed to be reasonably good considering the heterogeneity and size of the watershed in this study. Barnes and Baker [14] noted that the use of multi-spectral data for predicting the spatial distribution of soil properties can achieve optimal results when the study is conducted in an area with uniform soil surface characteristics. Consequently, several of such studies have been conducted at plot level or on relatively small watersheds [19,43,81], apparently to reduce the effect of varying surface characteristics.

Based on their study within a 350 ha demonstration farm in Arizona, Barnes and Baker [14] found that variations in surface characteristics such as crop residue, soil moisture and row orientation between fields limited the accuracy with which soil properties were mapped. These differences in surface characteristics may have influenced the results of this analysis, considering that the study area is an agricultural watershed populated by smallholder farmers who use diverse farm management practices [32,34]. The mode and time of land preparation (e.g. tractor, bullocks, manual) [90], nutrient application (e.g. fertility) [91] and water management strategy [92] can differ to a high degree from field to field due to availability of labour, crops to be cultivated or farm inputs utilized. Model calibrations based on samples from such localized and highly variable conditions can limit its predictive capacity outside the sampled areas [19,93].

Limited accuracy could also be related to potential error propagation from the MIRS models to the maps. Digital soil mapping based on mid infrared spectroscopy—partial least squares regression (MIRS-PLSR) prediction models might be affected by uncertainties at varying level of the mapping process such as spectra collection, model building and resulting prediction. Due to the heterogeneity of the landscape both in the small catchment and even more in the bigger catchment all the spectral variability might not have been covered resulting in possible feedback on the accuracy of MIRS-PLSR prediction models. Based on the classification of MIRS models by Reeves and Smith [94], the MIRS-PLSR calibration models in the present study (Table 1) range from models with very high predictive ability as for SOC (R^2 = 95%, RPD = 4.6) to models with high (R^2 = 85%, RPD = 2.6) to medium predictive ability (R^2 = 70–77%, RPD = 1.8–2.1) respectively for Nitrogen and for the remaining soil properties (CEC, sand, silt and clay). In some other studies, MIRS provided better prediction models for SOC, N, CEC ($R^2 > 0.77$) compared to clay, silt and sand (R^2 = 0.22–73%) [95,96]. Though uncertainty propagation analysis as carried out by Brodský et al. [97] was out of the scope of the present study, the error metrics from the test set validation provided satisfactory evidence on the predictive ability of the MIRS-PLSR models (R2 > 80%, RPD $\geq$ 2). These results indicated that the calibrations were consistent especially for SOC, CEC, N and silt (R2 > 85%, RPD $\geq$ 3). In their study, Brodský et al. [97] found PLSR (with visible and near infrared) to cause lower uncertainties in the final map compared to uncertainty originating from ordinary kriging used as mapping model. Based on the sMAPE, the RFR and remaining machine learning models displayed quite satisfactory accuracy from the prediction of MIRS-PLSR models. This is obviously to their ability to handle both linear and non-linear patterns in dataset.

Variable importance and temporal window for acquisition of RS data

The 5 top spectral and terrain/climatic variables which contributed most to the accuracy of digital soil mapping in the studied watershed are discernible from Table 8. Though RFR

Table 8. First five predictors that were highly significant for RFR (based on "IncNodePurity" importance measure) and MLR analysis.

Model	Rank	Sand	Silt	Clay	CEC	SOC	Nitrogen
MLR	1	june_SWIR2	june_SWIR2	june_NIR	june_SWIR2	Elevation	Elevation
	2	june_green	June_RI	June_RI	May_RI	prep	March_NDVI
	3	June_CI	may_red	may_blue	may_RE	march_NIR	march_NIR
	4	may_green	june_red	June_SI	June_BI	March_NDVI	march_green
	5	April_HI	June_BI	June_CI	june_red	june_SWIR1	March_CI
RFR	1	june_SWIR2	June_RI	june_NIR	june_SWIR2	june_red	june_NIR
	2	may_NIR	May_SI	June_RI	june_blue	june_NIR	June_SI
	3	june_green	june_SWIR1	june_blue	May_RI	Elevation	Elevation
	4	May_SI	june_SWIR2	june_SWIR1	March_NDVI	June_SI	march_green
	5	may_green	May_CI	temp	june_red	June_BI	may_red

The names of the spectral predictors (see Table 2) here are a concatenation of the month of satellite acquisition and a spectral channel or indice. For example, "May_BI" represents the brightness index calculated from the May RapidEye image. prep: precipitation, temp: temperature.

doi:10.1371/journal.pone.0170478.t008

generally provided better predictions, variable ranking from the MLR model was included in the table for comparison purposes. The data in Table 8 reveal that both models include elevation in the list of the five most significant predictors for SOC and N while the other soil parameters had only spectral predictors. The only exception was for clay for which the RFR recorded also temperature among its driving factors while the MLR also displayed precipitation as key factor following elevation.

Similar to the findings of this study, Hengl et al. [11] also recorded elevation as the most important variable influencing SOC contents of topsoil in Africa. Wang et al. [98] found that elevation and slope, along with soil clay and water contents, were among the most significant factors affecting SOC and N variability. Terrain/climatic variables are reported to have control on soil water status, dynamics of plant litter mineralisation as well as erosion and deposition processes [11,98]. The influence of elevation on predicting SOC and N, for example, can be related to corresponding variations in soil temperature as well as the intensity of cultivation which is higher in the lower areas as compared to the higher areas because of accessibility.

Table 8 reveals that generally, satellite images acquired in June and May were the most important in developing a model for predicting the soil properties under consideration. Spectral bands of the June Landsat image consistently came up as important predictors for the soil properties. The prominence of June and May images can partly be explained by the coincidence with the ploughing period or early stages of crop development when the soils of most agricultural plots are exposed. This allows satellite sensors to directly measure soil reflectance; hence, a good correlation between laboratory processed soil samples and satellite derived spectral reflectance is possible. The March imagery was the most important spectral predictor for SOC and N in MLR and was listed also for CEC and N in RFR (Table 8). March and April are the hottest months in the studied watershed, thus the prominence of the March imagery could be attributed to a higher loss of biomass with consequent higher mineralisation rate and SOC input.

Table 8 further reveals that the shortwave infrared (SWIR) and near-infrared (NIR) channels of Landsat, as well as soil specific indices like brightness, redness and saturation index were important spectral predictors in developing the respective models. The importance of the SWIR and NIR channels in this analysis confirms the findings of other studies. Liao et al. [99] used Landsat ETM bands as covariates in modelling soil textural properties (sand, silt, clay) and found that NIR (band 4) and SWIR (band 5, band 7) had a significant correlation with the

analysed soil properties and explained most of their variability. Soil specific spectral indices were also found useful in digital soil mapping by other studies [43].

Maps of the spatial distribution of the soil properties

In our study, the spatial distribution of soil properties does not display a clear pattern of hot and cold spot areas for all soil properties, but rather a patchy distribution (Fig 2). However, along the western border of the study area, medium to higher values of clay, CEC, SOC and N are observed as evidenced by a continuous yellowish band and reddish spots on the boundary line while the proportions of silt, on the contrary, recorded their lowest values in these areas. These zones correspond to the most elevated terrain where natural vegetation is prominent and accessibility is difficult for farming activities. This suggests a higher net primary production providing the input for nitrogen and carbon whose stability is reinforced by a higher clay content resulting in a higher CEC. It is widely acknowledged that SOC input is higher where substantial net primary productivity deposit occurs [83,84]. The remaining areas of lower elevation are settlement zones and cultivated areas and consequently displayed relatively medium (yellowish areas) and lower values (greenish areas) for the soil properties with some spots of high values in certain places.

Improving predictive accuracy with remote sensing data

The results of the present study point out the potential of terrain and spectral data for soil property mapping at high resolution at local scales. However, the prediction accuracy of the different models, though satisfactory, requires further improvement. The high intrinsic spatial fluctuation in soil properties, the heterogeneity of the landscape as well as highly variable management practices by farmers are among the potential sources of noise affecting model performance. Improving prediction accuracy might require partitioning the landscape into relatively homogeneous areas based on the covariates used as predictors. The latter could be carried out by considering land surface segmentation [100] or the conditioned Latin hypercube (cLHS) [101] using the key variables driving each soil property. Moreover, though the present study considered a vast array of predictors including topographical and spectral data, a single analysis scale was applied as generally carried out in DSM. However, some studies point out the fact that the spatial distribution of soil properties is subject to various factors operating at different levels of scale [56,102]. Therefore, multi- or hyper-scale data (terrain and spectral information) which account for different spatial scales might further improve prediction as recorded in other studies such as those of Behrens et al. [103] and Miller et al. [104]. The latter authors, however, did not focus on the stratification of the land surface which in addition to multi- or hyper-scale approaches have great potential in ameliorating prediction accuracy. Further investigations are therefore required in that regards.

Conclusion

Accurate and detailed spatial soil information is essential for environmental modelling, risk assessment and decision making. This study explored the use of high spatial resolution satellite (RapidEye and Landsat) and terrain/climatic data as well as laboratory analysed soil samples to map the spatial distribution of six soil properties–sand, silt, clay, CEC, SOC and N–in a 580 km^2 agricultural watershed in south-western Burkina Faso. Four statistical prediction models–multiple linear regression (MLR), random forest regression (RFR), support vector machine (SVM), stochastic gradient boosting (SGB)–were tested and compared. Internal validation was conducted by cross validation while the predictions were validated against an independent set of soil samples considering the modelling area and an extrapolation area.

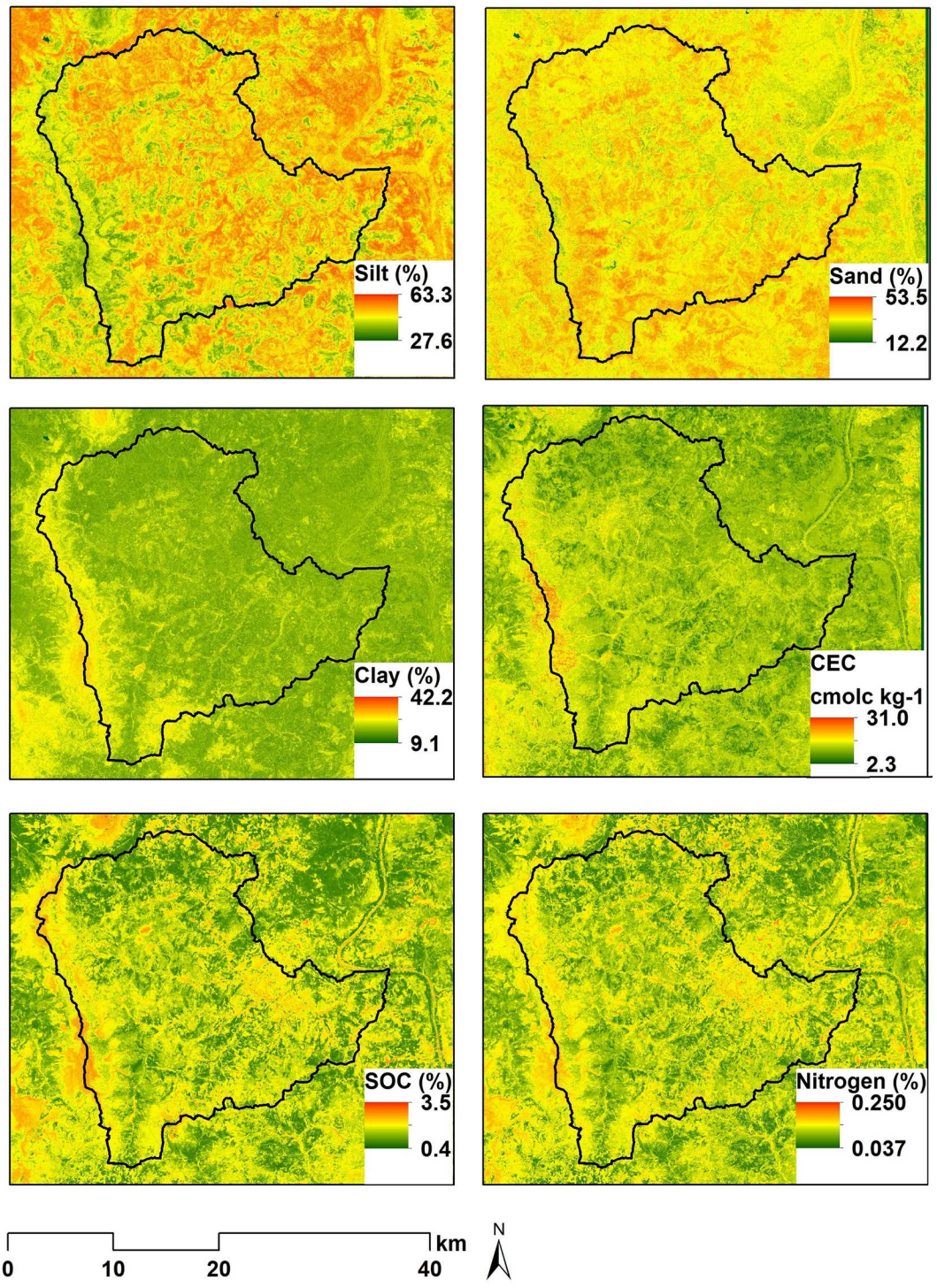

Fig 2. Spatial distribution of sand, silt, clay, cation exchange capacity (CEC), soil organic carbon (SOC) and total nitrogen (N) in the topsoil of the studied watershed.

doi:10.1371/journal.pone.0170478.g002

Results indicate that the RFR performed marginally better than the remaining models at modelling stage for most soil properties except for sand and clay for which MLR offered a better predictive ability. However, the RFR achieved a higher performance statistics for the

external validations in the considered areas but not for all soil properties in the extrapolated area. Beyond the modelling area, the SVM better predicted SOC while SGB performed better for CEC and N.

The machine learning algorithms performed generally better than the MLR for the prediction of soil properties at unsampled locations. Inability of MLR to handle non-linear relationships between dependent and independent variables is believed to be the source of this limitation. Prediction accuracies from the RFR model ranged from 68% for CEC to 89% for silt.

These prediction accuracies can be deemed to be reasonable, considering the high variability in farm management practices and environmental variables in the studied watershed. Satellite data acquired during ploughing or early crop development stages (e.g. May, June) were found to be the most important spectral predictors while elevation, temperature and precipitation came up as prominent terrain/climatic variables in predicting soil properties. The shortwave and near infrared channels of Landsat8 as well as soil specific indices of redness, coloration and saturation were prominent spectral channels.

The accuracies obtained in this study are promising for future local scale digital soil mapping efforts in data poor regions such as West Africa, considering the increasing availability of free high resolution remote sensing data. The use of remote sensing data can reduce soil sampling efforts and therefore reduce soil mapping costs. Further research is, however, required on the effect of high variability in farm management practices and environmental variables on the accuracy of digital soil maps. In addition, the potential of land surface stratification and multi- or hyper-scale analysis approaches in improving prediction accuracy are worth investigating.

Supporting Information

S1 Dataset. Soil properties data in the small and big catchment.
(7Z)

S2 Dataset. Soil properties data and code for R statistical software.
(7Z)

S1 File. Shapefiles of the data points in the small and big catchment.
(7Z)

Acknowledgments

The authors would like to acknowledge: (1) the German Federal Ministry of Education and Research (BMBF) for providing financial support for conducting this research, (2) the RapidEye Science Archive (RESA) team for providing RapidEye images, (3) the United States Geological Surveys for the use of Landsat data and (4) the West African Science Service Centre on Climate Change and Adapted Land use (WASCAL) for providing logistical support during fieldwork.

Author Contributions

Conceptualization: GF OKLH.

Data curation: GF OKLH.

Formal analysis: GF OKLH.

Investigation: GF OKLH.

Methodology: GF OKLH.

Project administration: GF OKLH.

Resources: GW MT.

Software: GF OKLH.

Supervision: GW MT.

Validation: GF OKLH.

Visualization: GF OKLH.

Writing – original draft: GF OKLH.

Writing – review & editing: GF OKLH GW MT.

References

1. Vågen T, Lal R, Singh BR (2005) Soil carbon sequestration in sub-Saharan Africa: a review. Land Degradation & Development 16 (1): 53–71.
2. Bationo A, Kihara J, Vanlauwe B, Waswa B, Kimetu J (2007) Soil organic carbon dynamics, functions and management in West African agro-ecosystems. Agricultural Systems 94 (1): 13–25.
3. Lahmar R, Bationo BA, Lamso ND, Guéro Y, Tittonell P (2012) Tailoring conservation agriculture technologies to West Africa semi-arid zones: building on traditional local practices for soil restoration. Field Crops Research 132: 158–167.
4. Sachs J, Remans R, Smukler S, Winowiecki L, Andelman SJ, Cassman KG et al. (2010) Monitoring the world's agriculture. Nature 466 (7306): 558–560. doi: 10.1038/466558a PMID: 20671691
5. Niang I, Ruppel, O. C., Abdrabo, M. A., Essel A, Lennard C, Padgham J, Urquhart P (2014) Africa. In: Barros VR, Field CB, Dokken DJ, Mastrandrea MD, Mach KJ et al., editors. Climate Change 2014: Impacts, Adaptation, and Vulnerability. Part B: Regional Aspects. Contribution of Working Group II to the Fifth Assessment Report of the Intergovernmental Panel of Climate Change. Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press. pp. 1199–1265.
6. Dobos E, Montanarella L, Nègre T, Micheli E (2001) A regional scale soil mapping approach using integrated AVHRR and DEM data. International Journal of Applied Earth Observation and Geoinformation 3 (1): 30–42.
7. Behrens T, Scholten T (2006) Digital soil mapping in Germany—a review. Journal of Plant Nutrition and Soil Science 169 (3): 434–443.
8. Mulder VL, Bruin S de, Schaepman ME, Mayr TR (2011) The use of remote sensing in soil and terrain mapping—A review. Geoderma 162 (1): 1–19.
9. Summers D, Lewis M, Ostendorf B, Chittleborough D (2011) Visible near-infrared reflectance spectroscopy as a predictive indicator of soil properties. Ecological Indicators 11 (1): 123–131.
10. Arrouays D, McKenzie N, Hempel J, de Forges A, McBratney AB (2014) GlobalSoilMap: basis of the global spatial soil information system: CRC press.
11. Hengl T, Heuvelink, Gerard B. M, Kempen B, Leenaars, Johan G. B., Walsh MG, Shepherd KD et al. (2015) Mapping soil properties of Africa at 250 m resolution: Random Forests significantly improve current predictions. PLoS ONE 10 (6): e0125814 EP -. doi: 10.1371/journal.pone.0125814 PMID: 26110833
12. Malone BP, Jha SK, Minasny B, McBratney AB (2016) Comparing regression-based digital soil mapping and multiple-point geostatistics for the spatial extrapolation of soil data. Geoderma 262: 243–253.
13. Metternicht GI, Zinck JA (2003) Remote sensing of soil salinity: potentials and constraints. Remote Sensing of Environment 85 (1): 1–20.
14. Barnes EM, Baker MG (2000) Multispectral data for mapping soil texture: possibilities and limitations. Applied Engineering in Agriculture 16 (6): 731.
15. Saadat H, Bonnell R, Sharifi F, Mehuys G, Namdar M, Ale-Ebrahim S (2008) Landform classification from a digital elevation model and satellite imagery. Geomorphology 100 (3): 453–464.
16. Ehsani AH, Quiel F (2009) A semi-automatic method for analysis of landscape elements using Shuttle Radar Topography Mission and Landsat ETM+ data. Computers & Geosciences 35 (2): 373–389.
17. Hahn C, Gloaguen R (2008) Estimation of soil types by non linear analysis of remote sensing data. Nonlinear Processes in Geophysics 15 (1): 115–126.

18. Scudiero E, Skaggs TH, Corwin DL (2014) Regional scale soil salinity evaluation using Landsat 7, western San Joaquin Valley, California, USA. Geoderma Regional 2: 82–90.
19. Thomasson JA, Sui R, Cox MS, Al–Rajehy A (2001) Soil reflectance sensing for determining soil properties in precision agriculture. Transactions of the ASAE 44 (6): 1445.
20. Gomez C, Rossel R, McBratney AB (2008) Soil organic carbon prediction by hyperspectral remote sensing and field vis-NIR spectroscopy: An Australian case study. Geoderma 146 (3): 403–411.
21. Ben-Dor E, Chabrillat S, Demattê JA, Taylor GR, Hill J, Whiting ML, Sommer S (2009) Using imaging spectroscopy to study soil properties. Remote Sensing of Environment 113: S1, S38–S55.
22. Fabiyi OO, Ige-Olumide O, Fabiyi AO (2013) Spatial analysis of soil fertility estimates and NDVI in south-western Nigeria: a new paradigm for routine soil fertility mapping. Research Journal of Agriculture and Environmental Management 2 (12): 403–411.
23. Akpa SIC, Odeh IOA, Bishop TFA, Hartemink AE (2014) Digital mapping of soil particle-size fractions for Nigeria. Soil Science Society of America Journal 78 (6): 1953–1966.
24. Blasch G, Spengler D, Itzerott S, Wessolek G (2015) Organic matter modeling at the landscape scale based on multitemporal soil pattern analysis using RapidEye data. Remote Sensing 7 (9): 11125–11150.
25. Forkuor G, Conrad C, Thiel M, Landmann T, Barry B (2015) Evaluating the sequential masking classification approach for improving crop discrimination in the Sudanian Savanna of West Africa. Computers and Electronics in Agriculture 118: 380–389.
26. Grimm R, Behrens T, Marker M, Elsenbeer H (2008) Soil organic carbon concentrations and stocks on Barro Colorado Island—Digital soil mapping using Random Forests analysis. Geoderma 146 (1–2): 102–113.
27. Wiesmeier M, Barthold F, Blank B, Kögel-Knabner I (2011) Digital mapping of soil organic matter stocks using Random Forest modeling in a semi-arid steppe ecosystem. Plant Soil 340 (1–2): 7–24.
28. Stevens A, Miralles I, van Wesemael B (2012) Soil organic carbon predictions by airborne imaging spectroscopy: comparing cross-validation and validation. Soil Science Society of America Journal 76 (6): 2174–2183.
29. Ließ M, Schmidt J, Glaser B (2016) Improving the spatial prediction of soil organic carbon stocks in a complex tropical mountain landscape by methodological specifications in machine learning approaches. PLoS ONE 11 (4): e0153673. doi: 10.1371/journal.pone.0153673 PMID: 27128736
30. Yilma T (2006) Modeling farm irrigation decisions under rainfall risk in the White-Volta basin of Ghana: a tool for policy analysis at the farm-household level: Cuvillier-Verlag, p.164. 164 p.
31. Fujisada H, Bailey GB, Kelly GG, Hara S, Abrams MJ (2005) Aster dem performance. IEEE transactions on Geoscience and Remote Sensing 43 (12): 2707–2714.
32. Callo-Concha D, Gaiser T, Ewert F (2012) Farming and cropping systems in the West African Sudanian Savanna. WASCAL research area: Northern Ghana, Southwest Burkina Faso and Northern Benin: ZEF Working Paper Series.
33. Bureau National des sols (BUNASOL) (2000) Etude morphologiaue des provinces de la Bougouriba et du Ioba, Echelle 1/100 000.
34. Forkuor G (2014) Agricultural Land Use Mapping in West Africa Using Multi-sensor Satellite Imagery: University of Wuerzburg: Wuerzburg, Germany, p.191.
35. Reeuwijk VL (2006) Procedures for soil analysis. 7th edition. Technical Report 9. Wageningen, Netherlands, ISRIC–World Soil Information.
36. Janik LJ, Skjemstad JO, Shepherd KD, Spouncer LR (2007) The prediction of soil carbon fractions using mid-infrared-partial least square analysis. Soil Research 45 (2): 73–81.
37. Bornemann L, Welp G, Brodowski S, Rodionov A, Amelung W (2008) Rapid assessment of black carbon in soil organic matter using mid-infrared spectroscopy. Organic Geochemistry 39 (11): 1537–1544.
38. Bellon-Maurel V, McBratney A (2011) Near-infrared (NIR) and mid-infrared (MIR) spectroscopic techniques for assessing the amount of carbon stock in soils—Critical review and research perspectives. Soil Biology and Biochemistry 43 (7): 1398–1410.
39. Albrecht R, Joffre R, Le Petit J, Terrom G, Périssol C (2008) Calibration of chemical and biological changes in cocomposting of biowastes using near-infrared spectroscopy. Environmental Science & Technology 43 (3): 804–811.
40. Tyc G, Tulip J, Schulten D, Krischke M, Oxfort M (2005) The RapidEye mission design. Acta Astronautica 56 (1): 213–219.
41. Irons JR, Dwyer JL, Barsi JA (2012) The next Landsat satellite: The Landsat data continuity mission. Remote Sensing of Environment 122: 11–21.

42. Richter R, Schläpfer D (2012) Atmospheric / Topographic Correction for Satellite Imagery: ATCOR-2/3 User Guide [Internet]. Wil, Switzerland: ReSe Applications Schläpfer. Available: http://www.dlr.de/eoc/Portaldata/60/Resources/dokumente/5_tech_mod/atcor3_manual_2012.pdf.
43. Ray S, Singh J, Das G, Panigraphy S (2004) Use of high resolution remote sensing data for generating site specific soil management plan. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences (35): 127–132.
44. Huete A, Didan K, Miura T, Rodriguez PE, Gao X, Ferreira LG (2002) Overview of the radiometric and biophysical performance of the MODIS vegetation indices. Remote Sensing of Environment 80 (1–2): 195–213.
45. Farr TG, Kobrick M (2000) Shuttle Radar Topography Mission produces a wealth of data. Eos, Transactions American Geophysical Union 81 (48): 583–585.
46. Forkuor G, Maathuis B (2012) Comparison of SRTM and ASTER derived digital elevation models over two regions in Ghana-Implications for hydrological and environmental modeling: INTECH Open Access Publisher.
47. Hijmans RJ, Cameron SE, Parra JL, Jones PG, Jarvis A (2005) Very high resolution interpolated climate surfaces for global land areas. International Journal of Climatology 25 (15): 1965–1978.
48. Allen DE, Pringle MJ, Bray S, Hall TJ, Reagain PO, Phelps D et al. (2013) What determines soil organic carbon stocks in the grazing lands of north-eastern Australia. Soil Research 51 (8): 695–706.
49. Travis MR, Elsner GH, Iverson WD, Johnson CG (1975) VIEWIT: Computation of seen areas, slope, and aspect for land-use planning. Gen. Tech. Rep. PSW-GTR-11. Berkeley, CA: Pacific Southwest Research Station, Forest Service, U.S. Department of Agriculture: 70 p.
50. Wood J (1996) The geomorphological characterisation of digital elevation models: University of Leicester (United Kingdom), p.465.
51. Blaga L (2012) Aspects regarding the significance of the curvature types and values in the studies of geomorphometry assisted by GIS. Annals of the University of Oradea, Geography Series/Analele Universitatii din Oradea, Seria Geografie 22 (2).
52. Yokoyama R, Shirasawa M, Pike RJ (2002) Visualizing topography by openness: a new application of image processing to digital elevation models. Photogrammetric Engineering and Remote Sensing 68 (3): 257–266.
53. Davy MC, Koen TB (2013) Variations in soil organic carbon for two soil types and six land uses in the Murray Catchment, New South Wales, Australia. Soil Research 51 (8): 631–644.
54. Florinsky IV (2012) Digital terrain analysis in soil science and geology: Academic Press.
55. Kitchingman A, Lai S (2004) Inferences on potential seamount locations from mid-resolution bathymetric data. Seamounts: Biodiversity and Fisheries. Fisheries Centre Research Report 12 (5): 7–12.
56. Behrens T, Zhu A-X, Schmidt K, Scholten T (2010) Multi-scale digital terrain analysis and feature selection for digital soil mapping. Geoderma 155 (3): 175–185.
57. Böhner J, Selige T (2006) Spatial prediction of soil attributes using terrain analysis and climate regionalisation. Gottinger Geographische Abhandlungen 115: 13–28.
58. Xiong X, Grunwald S, Myers DB, Kim J, Harris WG, Comerford NB (2014) Holistic environmental soil-landscape modeling of soil organic carbon. Environmental Modelling & Software 57: 202–215.
59. Vogel S, Märker M (2010) Reconstructing the Roman topography and environmental features of the Sarno River Plain (Italy) before the AD 79 eruption of Somma–Vesuvius. Geomorphology 115 (1): 67–77.
60. Page KL, Dalal RC, Pringle MJ, Bell M, Dang YP, Radford B, Bailey K (2013) Organic carbon stocks in cropping soils of Queensland, Australia, as affected by tillage management, climate, and soil characteristics. Soil Research 51 (8): 596–607.
61. R core Team R: A Language and Environment for Statistical Computing [Internet]. Vienna, Austria; 2015. Available: http://www.r-project.org/.
62. Venables WN, Ripley BD (2013) Modern applied statistics with S-PLUS: Springer Science & Business Media.
63. Liaw A, Wiener M (2002) Classification and regression by Random Forest. CR News 2 (3): 18–22.
64. Breiman L (2001) Random Forests. Machine Learning 45 (1): 5–32.
65. Schapire RE, Freund Y, Bartlett P, Lee WS (1998) Boosting the margin: A new explanation for the effectiveness of voting methods. Annals of Statistics: 1651–1686.
66. Gislason PO, Benediktsson JA, Sveinsson JR (2006) Random forests for land cover classification. Pattern Recognition Letters 27 (4): 294–300.

67. Statnikov A, Wang L, Aliferis CF (2008) A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. BioMed Central Bioinformatics 9 (1): 1.

68. Kuhn M (2015) Caret: classification and regression training. Astrophysics Source Code Library 1: 5003.

69. Shrestha NK, Shukla S (2015) Support vector machine based modeling of evapotranspiration using hydro-climatic variables in a sub-tropical environment. Agricultural and Forest Meteorology 200: 172–184.

70. Hastie T, Tibshirani RJ, Friedman JH (2011) The elements of statistical learning: data mining, inference, and prediction: Springer.

71. Friedman JH (2001) Greedy function approximation: a gradient boosting machine. Annals of Statistics: 1189–1232.

72. Friedman JH (2002) Stochastic gradient boosting. Computational Statistics & Data Analysis 38 (4): 367–378.

73. Ridgeway G (2008) gbm: Generalized Boosted Regression Models. Available: http://www.saedsayad.com/docs/gbm2.pdf. Accessed 17 November 2016.

74. Muñoz J, Felicísimo ÁM (2004) Comparison of statistical methods commonly used in predictive modelling. Journal of Vegetation Science 15 (2): 285–292.

75. Willmott CJ, Matsuura K (2005) Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Climate research 30 (1): 79–82.

76. Makridakis S, Hibon M (2000) The M3-Competition: results, conclusions and implications. International journal of forecasting 16 (4): 451–476.

77. Wiesmeier M, Barthold F, Spörlein P, Geuß U, Hangen E, Reischl A et al. (2014) Estimation of total organic carbon storage and its driving factors in soils of Bavaria (southeast Germany). Geoderma Regional 1: 67–78.

78. Coleman TL, Agbu PA, Montgomery OL, Gao T, Prasad S (1991) Spectral band selection for quantifying selected properties in highly weathered soils. Soil Science 151 (5): 355–361.

79. Zakaria ZA, Shabri A (2012) Streamflow forecasting at ungaged sites using support vector machines. Applied Mathematical Sciences 6 (60): 3003–3014.

80. Bricklemyer RS, Lawrence RL, Miller PR, Battogtokh N (2007) Monitoring and verifying agricultural practices related to soil carbon sequestration with satellite imagery. Agriculture, Ecosystems & Environment 118 (1): 201–210.

81. Odeha IO, McBratney AB, Chittleborough DJ (1994) Spatial prediction of soil properties from landform attributes derived from a digital elevation model. Geoderma 63 (3): 197–214.

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

83. Wålinder A (2014) Evaluation of logistic regression and random forest classification based on prediction accuracy and metadata analysis: Linnaeus University: Sweden, p.44.

84. Siegmann B, Jarmer T (2015) Comparison of different regression models and validation techniques for the assessment of wheat leaf area index from hyperspectral data. International Journal of Remote Sensing 36 (18): 4519–4534.

85. Ma W, Tan K, Du P, editors (2016) Predicting soil heavy metal based on Random Forest model: IEEE. 4331–4334 p.

86. Fassnacht FE, Hartig F, Latifi H, Berger C, Hernández J, Corvalán P, Koch B (2014) Importance of sample size, data type and prediction method for remote sensing-based estimations of aboveground forest biomass. Remote Sensing of Environment 154: 102–114.

87. Were K, Bui DT, Dick ØB, Singh BR (2015) A comparative assessment of support vector regression, artificial neural networks, and random forests for predicting and mapping soil organic carbon stocks across an Afromontane landscape. Ecological Indicators 52: 394–403.

88. Rossel R, Behrens T Using data mining to model and interpret soil diffuse reflectance spectra. Geoderma 158 (1–2): 46–54.

89. Hitziger M, Ließ M (2014) Comparison of three supervised learning methods for digital soil mapping: Application to a complex terrain in the Ecuadorian Andes. Applied and Environmental Soil Science 2014: 12.

90. Kamara AY, Ekeleme F, Chikoye D, Omoigui LO (2009) Planting date and cultivar effects on grain yield in dryland corn production. Agronomy Journal 101: 91–98.

91. Bationo A, Lompo F, Koala S (1998) Research on nutrient flows and balances in West Africa: state-of-the-art. Agriculture, Ecosystems & Environment 71 (1): 19–35.

92. Douxchamps S, Ayantunde AA, Barron J (2012) Evolution of agricultural water management in rainfed crop-livestock systems of the Volta Basin. CPWF R4D Working Paper Series 04. Colombo, Sri Lanka: CGIAR Challenge Program for Water and Food (CPWF).

93. Rossel R, Walvoort DJ, McBratney AB, Janik LJ, Skjemstad JO (2006) Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties. Geoderma 131 (1): 59–75.

94. Reeves JB, Smith DB (2009) The potential of mid-and near-infrared diffuse reflectance spectroscopy for determining major-and trace-element concentrations in soils from a geochemical survey of North America. Applied Geochemistry 24 (8): 1472–1481.

95. McCarty GW, Reeves JB (2006) Comparison of near infrared and mid infrared diffuse reflectance spectroscopy for field-scale measurement of soil fertility parameters. Soil Science 171 (2): 94–102.

96. Terhoeven-Urselmans T, Vagen T-G, Spaargaren O, Shepherd KD (2010) Prediction of soil fertility properties from a globally distributed soil mid-infrared spectral library. Soil Science Society of America Journal 74 (5): 1792–1799.

97. Brodský L, Vašát R, Klement A, Zádorová T, Jakšík O (2013) Uncertainty propagation in VNIR reflectance spectroscopy soil organic carbon mapping. Geoderma 199: 54–63.

98. Wang X, Ge L (2012) Evaluation of filters for ENVISAT ASAR speckle suppression in pasture area. Proceedings of the ISPRS Annals of the XXII ISPRS Congress-Photogrammetry, Remote Sensing and Spatial Information Sciences. Melbourne; 2012. pp. 341–346.

99. Liao K, Xu S, Wu J, Zhu Q (2013) Spatial estimation of surface soil texture using remote sensing data. Soil Science and Plant Nutrition 59 (4): 488–500.

100. Drăguț L, Dornik A (2016) Land-surface segmentation as a method to create strata for spatial sampling and its potential for digital soil mapping. International Journal of Geographical Information Science 30 (7): 1359–1376.

101. Stumpf F, Schmidt K, Behrens T, Schönbrodt-Stitt S, Buzzo G, Dumperth C et al. (2016) Incorporating limited field operability and legacy soil samples in a hypercube sampling design for digital soil mapping. Journal of Plant Nutrition and Soil Science 179 (4).

102. Behrens T, Schmidt K, Zhu AX, Scholten T (2010) The ConMap approach for terrain-based digital soil mapping. European Journal of Soil Science 61 (1): 133–143.

103. Behrens T, Schmidt K, Ramirez-Lopez L, Gallant J, Zhu A-X, Scholten T (2014) Hyper-scale digital soil mapping and soil formation analysis. Geoderma 213: 578–588.

104. Miller BA, Koszinski S, Wehrhan M, Sommer M (2015) Impact of multi-scale predictor selection for modeling soil properties. Geoderma 239: 97–106.