Institut für Nutzpflanzenwissenschaften und Ressourcenschutz (INRES)
- Allgemeine Bodenkunde und Bodenökologie -

Digital soil mapping using survey data and soil organic carbon dynamics in semi-arid Burkina Faso

Inaugural-Dissertation

zur Erlangung des Grades

Doktor der Agrarwissenschaftn
(Dr. agr.)

der

Landwirtschaftichen Fakultät

der Rheinischen Friedrich-Wilhelms-Universität Bonn

von

Kpadé Ozias Laurentin Hounkpatin

aus

Cotonou, Benin

Bonn 2018

Angefertigt mit Genehmigung der Landwirtschaftlichen Fakultät der Universität Bonn

Referent: Prof. Dr. Wulf Amelung

Koreferent: Prof. Dr. Mathias Becker

Tag der mündlichen Prüfung: 12 September 2017

Abstract

With computer-assisted geostatistics and data mining methods, digital soil mapping (DSM) offers new possibilities for providing soil spatial information for data scarce areas such as West Africa. Such information could also be essential for understanding tropical soil organic carbon (SOC) sequestration potentials and dynamics. However, the level of accuracy depends on the statistical model selected, the choice of which is not clear from the first for such environments. Moreover, for datasets with imbalanced soil orders, prediction of reference soil groups (RSG) using a DSM approach often biased towards the majority soil order class. I hypothesized that (i) statistical models, which are able to handle both linear and unlinear patterns in data, will provide higher prediction accuracy than those geared towards linear patterns, (ii) pruning the major soil group - the Plinthosols - will result in increased prediction accuracy of the minor RSG, (iii) sites with savannah (SA) and related RSG will present larger SOC stocks than cropland (CR), however, (iv), with land use change (LUC) also the Plinthosols are prone to rapid SOC losses from bulk soil and primarily from coarse particle-size fractions.

To test these hypotheses, I sampled sites within both CR and SA across different RSG in the Dano catchment. For the DSM of soil properties (sand, silt, clay, CEC, SOC, N) in the topsoil (0 - 30 cm), four statistical prediction models – multiple linear regression (MLR), random forest regression (RF), support vector machine (SVM), stochastic gradient boosting (SGB) – were used and compared. To reduce the risk that the spatial prediction of the RSG was biased by the majority class – the Plinthosols – I used a data pruning approach, accounting for 80 %, 90 % and standard deviation core range of the Plinthosols data, respectively, while cutting off all data points belonging to the outer range. Random Forest was used as a robust data mining method along with its recursive feature elimination option to evaluate the performance of these different data subsets. The final assessment of SOC stocks was conducted by considering its variation in CR and SA and in various RSG at different depths. The spatial distribution of SOC stocks as well as the main related factors were then again elucidated using Random Forest. For understanding the temporal dynamics of SOC storage, I investigated a false chronosequence of Plinthosols that had been converted from SA to CR at a duration between 0 and 29 years.

For the DSM of soil properties, results showed from the performance statistics that the machine learning techniques (RF, SVM, SGB) performed marginally better than the MLR, with the RF providing in most cases the highest accuracy. The lower performance of the MLR is attributed to its failure in accounting for non-linear relationships between response and predictor variables. The 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.

Upon the data pruning, the best predictions were observed when removing all PT points lower than 5 % and higher than 95 % of the cumulative percentage of the most important variable (wetness index). Modelling was then conducted solely with terrain

and spectral parameters (TSP) with optimal predictors resulting from RF recursive feature elimination. The resulting prediction model provided a substantial agreement to observation, with a kappa value of 0.57 along with a 35 % increase in prediction accuracy for Cambisols, 16 % for Stagnosols and 7 % for Gleysols. The SAGA wetness index (S.Wet.Ind) was the most important variable driving the RSG suggesting that the humidity regime is a key discriminatory element among the RSG.

The SOC stock distribution in the topsoil revealed a slightly larger SOC stock in the savannah sites (41.4 t C ha^{-1}) than in the cropland (39.1 t C ha^{-1}). Contrastingly, in the subsoil, a significant difference ($p < 0.05$) was observed between the CR recording a larger SOC stock of 40.2 t C ha^{-1}, while the subsoil of the SA sites contained only 26.3 t C ha^{-1}, on the average. Among the RSG, the Gleysols located at lower elevation positions revealed the largest SOC stocks over 0 - 30 cm (44 t C ha^{-1}) and 100 cm depth (86.6 t C ha^{-1}). Silt was the most abundant soil particle in the topsoil and was identified by the RF model as the most important factor related to the spatial distribution of the SOC stock, probably via its influence on soil moisture preservation and SOC storage via aggregation. Precipitation was found as the major factor related to subsoil SOC stock distribution. As the subsoils were also enriched in clay, the vertical transport of SOC rich sediments under tropical heavy rains likely accompanied major soil forming process in the landscape.

The LUC in the chronosequence Plinthosols triggered losses in SOC stock of 24 t C ha^{-1} from the upper 10 cm and 49 t C ha^{-1} from the upper 30 cm. Thus, about 66 % (0 - 10 cm; $p < 0.01$) and 55 % (0 - 30 cm; $p < 0.01$) of the initial stock in the native vegetation had been released after 29 years of cultivation. Also, subsoil was found to be vulnerable to LUC, with SOC losses amounting on average to 0.7 to 19.5 t C ha^{-1} from the 30 - 100 cm depth interval. Losses of SOC occurred from all particle-size fractions with a mean residence time of SOC generally decreasing with increasing equivalent diameter of the particle-size fraction. In this study, I could not confirm Fe oxides as key factor influencing SOC stock stabilization, because only an average of 16 % of the total SOC stock were apparently bound to Fe.

In summary, DSM at local scale using RF with remote sensing data resulted in reasonable prediction accuracy for a large array of soil properties and RSG within a highly heterogeneous landscape. Data pruning proved to be efficient in a context where a RSG belonging to a wide range of terrain parameters overlapped with those related to only few RSG units. The SOC stocks as quantified in the present study reinforce the view that the semi-arid ecosystems of West Africa still offer an opportunity for carbon sequestration and these results represent a baseline for future modelling of SOC dynamics in the region. LUC from natural savannah to permanent cropland, however, affects both topsoil and subsoil SOC though the latter is scarcely considered in the impact analysis of LUC in Africa.

Kurzfassung

Mit Hilfe Computer-basierter Methoden der Geostatistik und Datenbankauswertung bietet die digitale Bodenkartierung (digital soil mapping, DSM) neue Möglichkeiten zur Bereitstellung räumlicher Bodeninformationen für Regionen wie West Afrika, in denen solche Informationen nicht oder nur teilweise vorhanden sind. Diese Informationen können auch wichtig sein für die Abschätzung der Speicherkapazität und -dynamik von organischem Kohlenstoff (soil organic carbon, SOC) in tropischen Böden. Allerdings hängt die Genauigkeit vom gewählten statistischen Modell ab, dessen richtige Wahl für solche Umweltbedingungen anfangs nicht klar ist. Darüber hinaus ist die Vorhersage von Bodentypen (reference soil groups, RSG) durch digitale Bodenkartierung auf Grundlage von Datensätzen mit ungleich verteilten Bodentypen oft beeinflusst durch einen einzelnen dominanten Bodentyp. Meine Hypothesen sind, dass (i) statistische Modelle, die mit linearen und nicht-linearen Mustern in Datensätzen umgehen können, bessere Genauigkeiten bei der Vorhersage erreichen als die Modelle, die auf lineare Muster ausgerichtet sind, (ii) das statistische Beschneiden der Daten des dominanten Bodentyps (Plinthosol, PT) zu einer erhöhten Vorhersagegenauigkeit der anderen Bodentypen führt, (iii) Böden an Savanne-Standorten (SA) durch größere Bodenkohlenstoffvorräte charakterisiert sind als Böden unter Ackerland (cropland, CR), und (iv) mit einer Landnutzungsänderung (land use change, LUC) von Savanne zu Ackerland auch Plinthosole zu einem schnellen Verlust an organischem Bodenkohlenstoff neigen, und zwar insbesondere in den gröberen Fraktionen der partikulären organischen Substanz.

Um diese Hypothesen zu testen, habe ich im Dano-Einzugsgebiet Standorte mit den Landnutzungen CR und SA und verschiedenen RSGs beprobt. Für die digitale Bodenkartierung der Bodeneigenschaften (Sand, Schluff, Ton, CEC, SOC, N-gesamt) im Oberboden (0 - 30 cm) wurden vier statistische Vorhersagemodelle genutzt und verglichen: multiple linear regression (MLR), random forest regression (RF), support vector machine (SVM), stochastic gradient boosting (SGB). Um das Risiko zu reduzieren, dass die Vorhersage der RSGs von der dominanten Klasse (Plinthosols) beeinflusst wird, wurde ein statistischer Ansatz zum Beschneiden der Daten genutzt. Dabei wurden die unteren und oberen 5 % und 10 % sowie die Bereiche außerhalb der Standardabweichung der Plinthosol-Daten beschnitten, so dass nur die Daten innerhalb der genannten Grenzen genutzt wurden. Random Forest wurde als robuste Methode zur Datenauswertung genutzt. Die letztendliche Einschätzung der Kohlenstoffvorräte wurde unter Berücksichtigung ihrer Variation in CR- und SA-Flächen und in verschieden RSGs in unterschiedlicher Tiefe vorgenommen. Die räumliche Verteilung der Kohlenstoffvorräte und der damit zusammenhängenden Faktoren wurde dann erneut durch Random Forest und MLR erklärt. Um die zeitliche Dynamik von SOC-Vorräten zu verstehen, wurde eine falsche Chronosequenz von Plinthosolen untersucht, deren Nutzung sich von SA zu CR über unterschiedliche Zeiträume (0 – 29 Jahre) geändert hat.

In Bezug auf die digitale Bodenkartierung der Bodeneigenschaften zeigte sich, dass die machine learning techniques (RF, SVM, SGB) geringfügig besser abschneiden als MLR, wobei RF in den meisten Fällen die höchste Genauigkeit erreichte. Das

schlechtere Abschneiden von MLR liegt wahrscheinlich daran, dass es nicht-lineare Beziehungen zwischen Ergebnisvariablen und Einflussvariablen nicht wiedergeben kann. Die Satellitendaten, die während der Phase des Pflügens oder der frühen Pflanzentwicklung (z.B. Mai, Juni) aufgenommen wurden, stellten sich als wichtigste spektrale Prädikatoren heraus, während Geländehöhe, Temperatur und Niederschlag wichtige Gelände-/Klimavariablen bildeten.

Im Hinblick auf das Beschneiden der Daten wurden die besten Vorhersagen erreicht, wenn alle PT-Punkte kleiner als 5 % und größer als 95 % des kumulativen Anteils der wichtigsten Variable (wetness index) entfernt wurden. Die Modellierung wurde dann nur mit Geländeparametern und spektralen Parametern (terrain and spectral parameter, TSP) durchgeführt und zwar mit optimalen Prädiktoren aus der RF-Regression. Das daraus resultierende Modell zeigte eine gute Übereinstimmung von Vorhersage und tatsächlicher Beobachtung; der Kappa-Wert erreichte dabei 0.57 und die Vorhersagegenauigkeit stieg an um 35 % für Cambisols, 16 % für Stagnosols und 7 % für Gleysols. Der SAGA wetness indes (S.Wet.Ind) war für die Vorhersage der RSGs die wichtigste erklärende Variable. Das Feuchteregime kann also als diskriminierendes Schlüsselelement zwischen den RSGs angesehen werden.

Die SOC-Vorräte im Oberboden waren an Savanne-Standorten (41.4 t C ha^{-1}) leicht höher als an Ackerstandorten (39.1 t C ha^{-1}). Im Gegensatz dazu waren im Unterboden die SOC-Vorräte bei CR signifikant höher (40.2 t C ha^{-1}) als bei SA (26.3 t C ha^{-1}). Unter den RSGs zeigen Gleye, die in niedrigeren Geländelagen zu finden sind, die größten SOC-Vorräte in 0 - 30 cm (44 t C ha^{-1}) und 0 - 100 cm Tiefe (86.6 t C ha^{-1}). Schluff war die am meisten verbreitete Korngröße im Oberboden und wurde vom RF-Modell als wichtigster Faktor für die räumliche Verbreitung der SOC-Vorräte identifiziert; dieses ist wahrscheinlich zurückzuführen auf den positiven Einfluß dieser Korngröße auf die Wasserhaltefähigkeit und auf die Aggregierung organo-mineralischer Partikel. Der Niederschlag bildete den wichtigsten Faktor für die Verteilung der SOC-Vorräte im Unterboden. Da der Unterboden oft durch eine Tonanreicherung geprägt war, kann der vertikale Transport von kohlenstoffreichen Sedimenten bei tropischem Starkregen hier als ebenfalls wichtiger bodenbildender Prozess angesehen werden.

Der Landnutzungswandel hin zu Ackerland führte bei den untersuchten Plinthosolen zu SOC-Verlusten von 24 t C ha^{-1} in den oberen 10 cm und 49 t C ha^{-1} in den oberen 30 cm. So wurden ca. 66 % (0 - 10 cm; $p < 0.01$) und 55 % (0 - 30 cm; $p < 0.01$) des anfänglichen Kohlenstoffs unter natürlicher Vegetation durch 29 Jahre landwirtschaftlicher Nutzung freigesetzt. Auch der Unterboden war anfällig für Landnutzungsänderungen mit SOC-Verlusten von 0.7 bis 19.5 t C ha^{-1} in 30 - 100 cm Tiefe. Verluste an SOC wurden in allen Korngrößenfraktionen des partikulären Humus beobachtet, wobei die mittlere Verweildauer bei den gröberen Fraktionen abnahm. In dieser Studie konnten die Fe-Oxide nicht als Schlüssel zur Stabilisierung von Kohlenstoffvorräten bestätigt werden, da nur 16 % der Vorräte an Fe-Oxide gebunden waren.

Zusammenfassend kann gesagt werden, dass die digitale Bodenkartierung mit Hilfe von RF und Fernerkundungsdaten akzeptable Vorhersagegenauigkeiten für eine große Bandbreite an Bodeneigenschaften und RSGs innerhalb einer sehr heterogenen Landschaft ermöglicht. Es stellte sich heraus, dass das Beschneiden der Daten dann effizient ist, wenn eine RSG, die zu einer weiten Spannweite von Geländeparametern Beziehungen aufweist, sich mit solchen Parametern überschneidet, die nur mit wenigen RSG-Einheiten zusammenhängen. Die hier quantifizierten SOC-Vorräte unterstreichen, dass die semi-ariden Ökosysteme West-Afrikas immer noch eine Möglichkeit zur Speicherung von Kohlenstoff bieten und dass die Ergebnisse eine Grundlage für die weitere Modellierung der SOC-Dynamik in der Region darstellen. Der Landnutzungswandel von Savanne zu permanenter Ackernutzung beeinflusst Kohlenstoff im Ober- und Unterboden, obwohl letzterer selten bei der Analyse von Auswirkungen des Landnutzungswandels in Afrika berücksichtigt wird.

Table of Contents

List of tables

List of figures

List of abbreviations

asl	: elevation above sea level
CEC	: cation exchange capacity
CR	: cropland
DCB	: dithionite-citrate-bicarbonate
DSM	: digital soil mapping
LU	: land use
LUC	: land use change
ME	: mean error
MIRS	: mid-infrared spectroscopy
MLR	: multiple linear regression
OK	: ordinary kriging
POM	: particulate organic matter
RF	: Random Forest
RF_rfe	: Random Forest with recursive feature elimination
RFIDW	: Random Forest in combination with Inverse Distance Weighting
RFOK	: Random Forest in combination with Ordinary Kriging
RFR	: Random Forest Regression
RMSE	: root mean square error
RMSECV	: root mean square error of cross validation
RMSEP	: root mean square error of prediction
RPD	: ratio of performance to deviation
RSG	: reference soil groups
SD	: standard deviation
SOC	: soil organic carbon
SP	: spectral parameters
TP	: terrain parameters
TSP	: terrain and spectral parameters
WASCAL	: West African Science Service Center on Climate Change and Adapted Land Use

I.

General introduction

1. Rationale

Soils are vital resources for food production, water control and chemical recycling, biodiversity and habitat, providing platform for human activities, supplying raw materials as well as preserving cultural heritage (Blum, 2005). However, human activities via agriculture, grazing, deforestation and other land use such as building of roads or new facilities have affected soil ability to provide its ecosystem services. About 83 % of the land surface is reported by Sanderson et al. (2002) to be affected by human beings with 40 % transformed into agricultural land (Foley et al., 2005) and the remainder used for settlements and other non-farming purpose (Ellis et al., 2010). Estimation indicates that since 1850, about 6 million km^2 of tropical forest/woodland and 4.7 million km^2 of savannas/grasses/steppes have been transformed into farming land (Ramankutty and Foley, 1999). For example, FAO (2004) indicated that the cropland area increased over a period of 40 years (1961 – 2000) in Africa in response to population growth. As Africa population is expected to rise up to 4 billion by the end of the century (UN, 2015), the pressure on soil resources will be rising.

In sub Saharan Africa the increasing human pressure on soil resources has resulted in severe land degradation with issues related to soil erosion, salinity, reduction of organic matter, increase in CO_2 and its feedback on climate change (Tully et al., 2015). Recent evidences showed that the decline in soil fertility is prevalent in West African croplands as a result of population pressure (Grinblat et al., 2015). Nevertheless, for accurately addressing the degree of land degradation, spatial information on soils and soil properties are required for land evaluation. Spatial soil information as represented in soil maps is beneficial for farmers, scientists and policy maker in identifying priority areas and for sound and objective decision making. However, for management decisions at plot or small catchment level the available maps are too coarse and finer resolution is required. Moreover, maps from traditional surveys are mostly qualitative, labor intensive, time consuming and costly (Taghizadeh-Mehrjardi et al., 2015), and thus in most cases also obsolete (Kilasara, 2010).

Recent advances in remote sensing and information systems have paved the way for digital soil mapping (DSM), which couples soil point data with statistically correlated

auxiliary data (McBratney et al., 2003). This approach overcomes the limitation of the traditional mapping method by reducing tremendously both the workload involved as well as the related costs (Giasson et al., 2015).The coupling of point and auxiliary data is carried out by using (geo-) statistical classification or regression models. The auxiliary data include the soil forming factors as described by Jenny (1941). In DSM, these factors are mostly derived from digital elevation models (DEM) and existing parent material, climate, land use or vegetation maps. Further advances are foreseen with the availability of satellite data with high spatial resolution such as RapidEye to improve mapping accuracy (Forkuor, 2014) at a given location in the landscape. Particularly, the combination of the covariates derived from the DEM with optical and radar imagery data has great potential for improving prediction accuracy for a targeted soil property or soil class. This may be of special relevance for West Africa, where there is only scarce soil information at a finer scale.

Soil organic carbon (SOC) is a key indicator for assessing land degradation or soil improvement processes. The COP21 convention in Paris pointed out the relevancy of the sequestration of SOC as an important strategy to mitigate climate change (Rhodes, 2016). SOC is essential for soil fertility and productivity, being involved in most soil functions such as storage of nutrients and water, soil biological activity and structural stability (Holmes et al., 2015). Maintaining SOC is thus necessary for a soil to fulfill primary ecosystem services, especially in West Africa, where natural soil fertility and fertilizer input are low (Doraiswamy et al., 2007). To assess SOC sequestration potentials, however, we again need quantitative data on spatial and temporal carbon stocks, both locally and at national scale. Usually, the SOC stocks vary across the landscape and with related variations in climate (Albaladejo et al., 2013; Stergiadi et al., 2016), land use and land cover change (Muñoz-Rojas et al., 2015; Xiong et al., 2014), topography (Nadeu et al., 2015), texture (Burke et al., 1989), clay mineralogy (Saidy et al., 2012), sesquioxides (Peng et al., 2015) and soil order (Bruun et al., 2013; Wiesmeier et al., 2012). The influence of these factors on SOC dynamics has been frequently investigated in temperate climates; however, the understanding of these interactions for tropical low input agricultural systems is still limited.

Though interest for SOC and controlling factors rose in the last decades, most of the studies focused on the topsoil (30 cm). Subsoil carbon, although equaling atmospheric carbon in amount, is typically neglected in models of soil fertility and carbon balances. Batjes (1996) indicated that about 50 % of the SOC is located below 20 cm depth. Fontaine et al. (2007) showed that subsoil carbon is readily decomposable upon addition of a fresh C-source, suggesting that excluding subsoil carbon from our regard might have been overhasty. Therefore, any small change in the subsoil carbon stock will have a significant impact on the global C budget (Don et al., 2007). Since the tropical subsoil carbon consists mainly of intermediate and passive soil organic matter pools (Lützow et al., 2008), it offers great potential as carbon sink. Consequently, quantification of the SOC stock in the subsoil is vitally important for an accurate evaluation of the sequestration ability of the highly weathered and deep tropical soils.

Monitoring changes in SOC stocks with time should likely include pools of different SOC stability, since overall response rates may be too slow and thus ignored when this monitoring is based on bulk SOC analyses only (e.g., Powlson et al., 1987; Skjemstad et al., 2004). Classically, the identification of such pools involved the fractionation of SOC according to particle size, density or a combination thereof. Particulate organic matter (POM) has been considered as fairly labile pool of soil organic matter (SOM) in many studies as it is more sensitive to land use change (LUC) than bulk SOC, due to its rapid depletion after conversion of soils under natural vegetation to arable cropland (Besnard et al., 1996; Chan, 2001). Monitoring POM should thus also help for scaling changes in land degradation in the context of conversion from natural vegetation to cropland.

2. State of the art

2.1. From digitized soil map to digital soil mapping

Soil mapping played major role in human history as already in 4000 years BP the book Yugong reported on a different distribution of soils in nine provinces of China (Gong et al., 2003). In that period, soils were mapped based on soil properties such as soil fertility, soil color, soil texture, soil moisture and vegetation. The early scientific soil maps in Germany, France, Austria, the Netherlands, and Belgium from the 1850s and

1860s were constructed from concepts grounded in agrogeology (Hartemink et al., 2013). The early soil information was used mostly for military ends or taxation and land assessment purposes (Krupenikov and Tedrow, 1994).

Until the 19th century, only geologic and physiographic factors were considered for soil map delineations. As V.V. Dokuchaev supplemented climate and vegetation to the geologic and physiographic factors in the late 19th century, a full soil-landscape paradigm was introduced (Brown, 2005). From then, soils were considered as a function of parent material, climate, organisms, relief and time. This concept is captured by the fundamental soil state-factor equation developed by Jenny (1941):

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

where S stands for soil, cl for climate (cl), o for organisms, r for relief, p for parent material and t for time. This equation offered the conceptual framework for understanding the important parameters affecting soil variability at global and local scale all over the world.

2.1.1. Conventional soil mapping and drawbacks

Most of the national soil maps in West Africa and in the world were established using the traditional mapping approach. The traditional method for soil mapping mainly involved the use of aerial photography, geology, topographic maps and field observations (profile) for the prediction of areas having the same soil class (Malone, 2012). It has been reported that less than 0.001 % of the mapped area is actually subject to direct observation (Burrough et al., 1971). The map establishment is based on a conceptual understanding of the soil forming processing in a particular area by one or many surveyors. Most of the existing conventional maps are class type and are made up of polygons standing for the soil map units (Scull et al., 2003). Within each unit, the distribution of the soils in the landscape is represented with its internal variation but often lacking is the explicit description of its spatial pattern (Omuto et al., 2013). The traditional approach has been criticized both in its method as well as in its output represented by the resulting map.

In conventional soil mapping, rules and models for the prediction of soil class or soil properties are tacit knowledge of the soil surveyor mainly and are in most cases only expressed in mapping legend. This results in the impossibility to produce map uncertainties, which is critical for map users. Moreover, map polygons are assumed to contain homogeneous soil properties or soil class and each polygon boundary suggest a sharp transition in the distribution of these properties or soil class (Heuvelink, Gerard B. M. and Huisman, 2000). However, this conventional approach labelled as the double-crisp model by Burrough et al. (1997) failed to incorporate the continuous spatial variability of both soil properties and soil-forming processes. Thus, soil maps resulting from traditional approach are mostly produced at coarse scale (Towett, 2013) and cannot be used for decision making at a finer scale. Additionally, traditional soil mapping is often too costly and time demanding, especially in developing countries, and it hardly works for remote places. Furthermore, the representation of map units in polygons makes its integration in existing earth resources difficult, because these are in a grid based format (DEM, satellite imagery) (Malone, 2012). To address all these issues, a new paradigm in soil mapping emerged, which is called digital soil mapping (DSM).

2.1.2. Digital soil mapping

The advancement in computer science and statistical methods led to the use of geo-information technology such as remote sensing data and digital elevation model (DEM) for the description of soil variability in a more continuous and quantitative approach (Heuvelink and Webster, 2001). This new paradigm correlates soil class/soil properties with selected environment covariates data; it is based on statistical models in order to predict these soil class or soil properties at unknown locations. Building on the soil state-factor equation developed by Jenny (1941), McBratney et al. (2003) introduced the conceptual framework for DSM referred to as "scorpan."

$$S_c = f(s, c, o, r, p, a, n) \quad \text{or} \quad S_p = f(s, c, o, r, p, a, n) \qquad \text{(I-2)}$$

where S_c is soil class and S_p is a soil attribute or property, s: soils, other attributes of the soil at a point, c : climate, o = organisms (vegetation, fauna, or human activity), r :

relief (topography), p : parent material (lithology), a : age, n = spatial location, f : function or soil spatial prediction function (SSPF) model.

The DSM implementation basically involves three steps (Omuto et al., 2013): (1) input data provision, (2) classification and regression methods, (3) map production and its validation.

2.1.2.1 Data input for digital soil mapping

The input for digital soil mapping represents the soil forming factors in the scorpan equation. These data consist in soil sampling, soil legacy data and ancillary data (McBratney et al., 2003). Soil surveys are generally carried out either in the traditional way or based on statistical sampling and soil samples are collected and subsequently laboratory analysis are made to assess target soil properties. This information is then used as attribute in the scorpan equation to predict soil class or other soil properties. When soil attributes cannot be accessed from direct soil survey, the required information is to be derived from existing data bases such as soil legacy data, local soil surveys, profile and auger description, or laboratory analysis carried out on samples collected from the field. Particularly soil legacy data have been discussed extensively in many studies, and remain the most important input for DSM especially in many developing countries (Minasny et al., 2012; Sulaeman et al., 2013).

The ancillary data used as input for DSM models represent various soil forming factors. They are environmental covariates data, which are mostly derived from DEM (e.g. altitude, slope, curvature), remote sensing data (e.g. Landsat ETM surface reflectance and imagery) as well as from geological maps standing for parent material and climate (temperature, precipitation) (Malone et al., 2016; Stoorvogel et al., 2009). Typically, the soil point data are overlaid over these environmental georeferenced data layers to extract the values at each point of the landscape.

2.1.2.2 Classification and regression methods

Many function or soil spatial prediction function (SSPF) have been developed and used for digital soil mapping with the advance in computer science and statistics.

These functions enable the estimation of the unknown value of the targeted variable at a certain location. Originally, before soil factors could become quantitatively available, only geospatial models were used for mapping (McBratney et al., 2011). These include trend surface (Grunwald, 2006), nearest neighbours (Mansuy et al., 2014), inverse distance weighting (Robinson and Metternicht, 2006), and splines (Burrough and McDonnell, 1998; Laslett et al., 1987).

Geostatistics with at its core the kriging method have been used for soil mapping for decades with early application by Burgess and Webster (1980). Later on, many other works focused on discussing theoretical and practical application of geostatistics for soil science such as Oliver (1987), Goovaerts (1999) and Webster and Oliver (2007). One of the most fundamental laws in geostatistics is the first law of geography stating that objects that are closer are more similar than objects that are far apart. The spatial variation is described using a semivariogramm, which is half the expected squared difference between values of the targeted variable at two locations. The variogram, which is the representation of the semivariogramm as a function of distance, measures the spatial auto-correlation of soil properties in a certain landscape by the formula (Webster and Oliver, 2007) :

$$\gamma(h) = \frac{1}{2m(h)} \sum_{i=1}^{m(h)} \{s(x_i) - s(x_i + h)\}^2 \text{ (I-3)} \qquad \text{(I-3)}$$

with $\gamma(h)$ is the average semi-variance of the soil property, m(h) is the number of pairs of observation separated by the lag h, s is the value of the property S, x is the coordinate of the point.

Based on that principle, many kriging methods have developed with mainly two approaches: univariate kriging and mutltivariate kriging. The univariate (only one variable used) interpolation embraces techniques such as simple kriging, ordinary kriging, block kriging, factorial kriging, indicator kriging, disjunctive kriging (Li and Heap, 2008). These techniques evolve to more complex ones where other variables co-related to the one being predicted are also considered in the perspective of getting

higher prediction accuracy. Among the multivariate interpolation techniques fall the following: co-kriging, universal kriging, kriging with an external drift, principal component kriging, multivariate factorial kriging, indicator kriging. These techniques are well documented by Li and Heap (2008) and Li and Heap (2014). A major advantage of these geostatistical models is the possibility to provide a quantitative measure of uncertainty (Goovaerts, 1999), while the requirement for larger size dataset for setting a reliable model is a constraint for area with low availability data (Burrough et al., 1971).

Combinations of non-geostatistical and geostatistical methods are also used either for classification or for regression purposes. Such combined methods as referred to by Li and Heap (2014) in general build a primary model between the target variable and selected Jenny´s soil forming factors as explanatory variable. Some kriging techniques are then applied on the residuals to produce an uncertainty map, which is ultimately added to the initial model fit map to generate the final output. Among these mixed methods, the categories are regression kriging, linear mixed model, and trend surface analysis combined with kriging, as well as soil classification combined with other interpolation methods, just to name a few.

In DSM, many SSPFs following the scorpan models are used. These are prediction functions which have been generally presented by McBratney et al. (2000) and McBratney et al. (2003) and more extensively discussed by Hastie et al. (2011). They broadly include either linear methods or machine learning algorithms. Mostly linear models used for DSM are multiple linear regression (Meersmans et al., 2008; Selige et al., 2006), partial least square regression (Amare et al., 2013), principal component regression and partial least square (McBratney et al., 2003), linear discriminant analysis (McBratney et al., 2003), as well as generalized linear models (McKenzie and Ryan, 1999). The term machine learning refers to a broad variety of models meant for pattern analysis in data, also known as data mining, and making data-driven predictions (Witten and Frank, 2005). They became extremely popular as relationship between soil attributes and the scorpan factors are complex, poorly understood and most likely not linear (Povak et al., 2014). Examples of machine learning algorithms in

soil science include support vector machines (Were et al., 2015), neural networks (Behrens et al., 2005; Malone et al., 2009), generalized additive models (Poggio et al., 2013) and decision trees (DT) (Quinlan, J. Ross, 1986).

DT use a hierarchical top-down approach by dividing the data recursively into branch-like divisions, which individually captures a variability in the target variable (McBratney et al., 2003). These divisions are structured as an inverted tree having a root node, as well as a set of internal and terminal nodes (leaf node). The split at each inner node is based on decision rules that affect instances uniquely to child node, with each of the leaf node having a target (regression tree) or a class value (classification tree). Advantages for using DT include their capacity to handle numerical and categorical data without any assumption to probability distribution, computational efficiency, as well as their robustness against nonlinearity and overfitting (Heung et al., 2014).

Most popular DT algorithms include C4.5/SEE5 (Adhikari et al., 2014), as well as Classification and Regression Trees (CART), which build single trees (Breiman et al., 1984). However, the latter are reported to build unstable decision trees, which could bias prediction (Timofeev, 2004). To enhance the prediction accuracy in DSM using DT, methods have been introduced that generate multiple models through iteration, and which ultimately cumulate them to provide the final estimate. McBratney et al. (2003) classified the DTs into two groups: bootstrap aggregating (or bagging) and boosting. Bootstrap aggregating is an iterative process sampling into the training set with replacement, which is the basis for the widely used Random Forest algorithm (Grimm et al., 2008; Hengl et al., 2015; Reza Pahlavan Rad, Mohammad et al., 2014; Wiesmeier et al., 2011).

Boosting functions make predictions by growing new trees based on the information of previously grown trees in an attempt to reduce prediction errors (Yang et al., 2016). Recently, a number of novel hybrid methods have been introduced for DSM consisting in the combination of some machine learning with either Inverse Distance Weighting (IDW) or ordinary kriging (OK). Key examples include the combination of (Li and Heap, 2014): (i) support vector machine with OK or IDW, (ii) RF with IDW or OK

(RFIDW, RFOK) (iii) general regression neural network with IDW or OK , (iv) boosted decision tree (BDT) (Li et al., 2012) with Inverse Distance Weighing or Ordinary Kriging. These methods function in a similar pattern to regression kriging with the application of Inverse Distance Weighing or Ordinary Kriking on the residuals of the model fit. The purpose is to capture any spatial autocorrelation of the residuals for high prediction accuracy of the targeted variable.

2.1.2.3 Validation for map quality assessment

The output map generated by the SSPFs is not free of errors, and the quantification of these errors is relevant for both soil properties and soil class predictions. For the former, the Root Mean Square Error of prediction (RMSE) is mostly reported in literature (Were et al., 2015). For soil class maps, the accuracy assessment is carried out by determining user´s and producer´s accuracy but most importantly the Kappa statistics (Lark, 1995). Malone (2012) reported three main approaches for validation in DSM. These approaches consist in: (1) holding back a proportion of the dataset as an independent set for testing the map accuracy; (2) cross validation with leave-one-out procedures for eliminating one value for parameter estimation, or for multiple values as n-fold-cross validation. With the leave-one-out scheme, one observation is left out while the remaining are used to fit the model. The left out observation is later used to evaluate the accuracy of the model. The same process is carried out again until all the observations are taken into account. The n-fold-cross validation rather divides the whole dataset in a n subset (fold), and the cross validation procedure is carried out on these n subset. The last approach resort to additional sampling using either randomized or probability sampling design. However, when dataset is large enough, validation based on independent set is carried out especially for legacy soil data.

2.2. Instances and feature selection

The handling of large datasets for digital soil mapping can become complex when there are many relevant predictors (features) and soil samples (instances). This issue is commonly designated as high dimensional data evaluation with large amounts of features and instances (Sutha and Tamilselvi, 2015). In such data, not all the features and instances are relevant for the classification or regression operation, because they

partly also contain redundant and noisy information. The latter reduce the learning performance and prediction accuracy (Lagacherie and Holmes, 1997; Schmidt et al., 2008). Avoiding this problem requires a pre-processing step; and two main branches in statistical learning research then address this issue: instance selection (Liu and Motoda, 2001) and feature selection (John et al., 1994). Feature selection consists in singling out a feature subset as small as possible and in reducing multi-collinearity. Instance selection deals with the reduction of the dataset by filtering out irrelevant cases without losing useful information.

There are three main feature selection algorithms available for consideration: the Filter, Wrapper and Hybrid Method (Sutha and Tamilselvi, 2015). The Filter method selects a feature subset only by focusing on the characteristics of the predictors, which is done independently of any mining algorithm. In contrast, the Wrapper method requires the latter for the selection. The Hybrid method uses both inherent characteristics and mining algorithm for determining the best feature or instance subset. When working with large datasets, the Filter method is mostly preferred due to high computation efficiency. The algorithms, which are used for feature selection, are classified into Supervised Learning Algorithms (Le Song et al., 2012; Weston et al., 2003), Unsupervised Learning Algorithms (Handl and Knowles, 2006) and Semi-supervised Learning Algorithms (Doquire and Verleysen, 2011), which combined the former two.

In supervised learning, features are selected based on their ability in separating data into different classes, called class-based separation. Unsupervised feature selection removes irrelevant features by identifying similarity or correlation measures between the features. The latter approach was considered in the present study for removing redundant features as affected by multicollinearity. Though decision trees are reported to be robust to correlated features, the interpretation of the most important feature can be biased when the variables involved are subject to multicollinearity (Kuhn, 2008). Genuer et al. (2010) also reported that variable importance may be overestimated when highly correlated variables are used.

There are many instance selection methods available from other research fields (reviewed by Olvera-López et al., 2010). Their application in DSM is not all that extensive. The first scientists that investigated instance selection in a DSM for soil classification were Moran and Bui (2002). These authors compared two random sampling methods over all soil classes for their training dataset. Schmidt et al. (2008) instead carried out instance selection on single soil classes in order to evaluate the output of the different sample distribution using proportional stratified random sampling and disproportional stratified random sampling schemes. Proportional stratified random sampling takes into account the frequency distribution of each soil class in the entire dataset, while the disproportional approach used the same number of instances for all classes.

Challenges may arise in the application of the disproportional approach when the size of the smallest class is too low for decision trees to accurately learn from the inherent pattern. Also, the number of instances of the smaller class in the available dataset, such as soil legacy data, may affect the distribution of the remaining soil classes when proportional stratified random sampling is performed. Qi (2004) introduced a different approach for instance selection, which was based on fitted histograms of the features. However, this approach is difficult to implement when dealing with many features, which have to be distinctly considered (Schmidt et al., 2008) unless the feature space is reduced (feature selection) and unless the most important feature is chosen for instance selection. The latter scheme has been investigated for noise reduction in the present study on an imbalanced dataset.

2.3. Soil organic carbon

Soils are the major terrestrial sink of carbon with great potential to counteract the adverse effect of global warming (Singh and Lal, 2005). The soil carbon stock amounts to 2157-2293 Gt C with about 67 % existing as SOC (Lal, 2004). About 50 % of this SOC stock are stored in the topsoil (30 cm) making the subsoil also as relevant sink for carbon (Batjes, 1996). The amount of SOC at a given site is the result of the dynamic equilibrium of gain and loss processes directed by different factors (Lal, 2004). These factors vary from climate, topography, soil properties, microbial biomass

and land use (Albaladejo et al., 2013; Jobbágy and Jackson; Jobbágy and Jackson, 2000; Ladd et al., 2013), which are mainly the factors previously mentioned by Jenny (1941). Land use change (LUC) affects SOC stocks and can result in either a sequestration or a release of CO_2 with subsequent impacts on global warming (Houghton, 2003). Carbon sequestration is, however, of crucial importance as SOC affects many soil functions and ecosystem services.

2.3.1. Land use change impact on SOC

About two scenarios of LUC are reported in literature based on whether it leads to SOC depletion or SOC accumulation. One scenario consists in LUC from pasture or native savannah/forest to plantation or to cropland which adversely affects SOC levels (McDonagh et al., 2001; Murty et al., 2002). The size and magnitude of the impact of the anthropogenic influence through agricultural use on the SOC status in soils is complex and determined by various variables, such as land use type, crop type, organic and inorganic fertilizer use, cultivation intensity and history etc. Soil cultivation is characterized by annual cropping with the necessary soil tillage, which disrupts soil aggregate and accelerate the decomposition of organic materials (Wei et al., 2014b). Consequently, SOC contents and stocks decline rapidly and then stabilize after a certain period of time following a land-use change (Don et al., 2011).

The second scenario consists in reversing land degradation due to LUC with former depletion of SOC level through conversion of cropland to grassland or forest (Guo and Gifford, 2002a; Smith, 2008, 2008) as well as via change from conventional tillage to no-tillage cultivation (Amado, Telmo Jorge Carneiro et al., 2006). The latter processes mostly results in C accumulation, though usually not the level formerly found in native ecosystems due to inefficient C accrual in the subsoil (Preger et al., 2010). While these processes are being studied worldwide, little is known on C losses and C sequestration rates in soils typical for Western Africa, such as Plinthosols. Moreover, very few studies included the subsoil into the monitoring of C loss and sequestration rates (Mobley et al., 2015; Olson et al., 2014; Steinmann et al., 2016). Part of the present study focused on the former scenario of SOC losses related to LUC from initial savannah to cropland and subsequent effect on SOC dynamics. These studies also

included different SOC pools that are considered to be functionally homogeneous (Besnard et al., 1996; Degryze et al., 2004).

2.3.2. Qualitative characterization of SOC

The SOC consists in of a variety of compounds of different chemical structure and turnover rate. For SOC turnover modelling, in general, three pools ranging from labile or active pool, intermediate pool and inert or resistant pool, are distinguished (Lützow et al., 2007). The labile or active pool is made up of microbial biomass, fresh plant and root derived elements as well as some microbial residues with a faster (weeks to years) turnover time (Schwendenmann et al., 2007). The intermediate pool refers to refractory plant debris and mineral associated SOC with a longer turnover time ranging from 10 to more than 100 years, while the inert or resistant pool is composed of highly humified compounds if not of black carbon with turnover times in the order of 10^3 years (Parton et al., 1987; Schwendenmann et al., 2007; Trumbore, 1997).

With advancing SOM decomposition, it may be generally assumed that SOC is transferred gradually from the active pool into either CO_2 or more stable pools; various stabilizing processes may account for this but often only a small fraction of fresh organic material ends up in the more stable pools (Derrien and Amelung, 2011). Because each pool has its own pattern of reaction in regard to LUC, considering the changes in specific SOC pools is more effective for indicating early responses of SOM to LUC than bulk SOC (Lützow et al., 2007). Consequently, the functional SOC pools are to be quantified and characterized for a thorough understanding in SOC change patterns due to LUC. Mostly, physical soil fractionation is used for that purpose (e.g., Christensen, 1992; 1996) as shortly annotated below.

2.3.3. SOC fractionation and Chronosequence

SOC fractionation for qualitative analysis can be carried out by using either physical (aggregation, density, size) and/or chemical (solubility, mineralogy) methods (Lützow et al., 2007; Stockmann et al., 2013). Aggregate fractionation uses dry or wet sieving, slaking as well as (ultrasonic) dispersion to separate free SOC from protected SOC that is incorporated within various secondary organomineral complexes. The free SOC

is considered as the active pool and is occluded in the macroaggregate (> 250 mm) while the protected pool is either incorporated in microaggregate (< 250 mm) or termed as intermediate pool or in the clay microstructures (<20 mm) representing the passive pool. The density fractionation differentiates between light fraction (active pool) and heavy fraction (intermediate and passive pool). The light fraction relates to SOC that is not firmly bound to soil minerals while the heavy fraction forms the organomineral complexed compounds (Tisdall and Oades, 1982; Golchin et al., 1994; Lützow et al., 2007). Because the latter pool incorporates both intermediate and passive pool it has been reported as being very heterogeneous by Lützow et al. (2007). Moreover, Six et al. (2000) pointed out microaggregate stabilization within macroaggregate with different dynamics for the respective related SOC. Using wet sieving, they distinguished coarse intra-aggregate particulate organic matter (iPOM) in macroaggreagte while fine iPOM was identified in microaggregate within macroaggreagte. The former has a faster turnover rate compared to the former which is more stable with longer residence time.

As aggregates are so-called secondary particles, separating them into apparent primary particles describes the turnover of SOM at different bonding partners (Christensen, 1992).The particle size fractionation is based on the concept that the status of the SOC dynamics is related to the particle sizes characterized by different decay rate (Moni et al., 2012). Particulate organic material (POM), which is mainly made up of pieces of plant residues, is considered as a labile pool with turnover rate ranging from months to a few years (Besnard et al., 1996; Chan, 2001). POM is the first pool to be affected by LUC and as such is a better indicator of the impact of land use and climate on soil properties than bulk SOC (Ashagrie et al., 2005; Liang et al., 2012). POM is either free or incorporated in aggregate (Cambardella and Elliott, 1993; Christensen, 1992). Based on aggregate and particle size, POM measurement is carried out by considering the coarse (250–2000 µm), medium (53–250 µm) and fine (<53 µm) fractions (Amelung and Zech, 1999; Cambardella and Elliott, 1993; Chefetz et al., 2002). The POM C content and turnover are different in these fractions and are affected by the silt and clay particles level in the soil (Dalal and Mayer, 1986). These fractions are suitable for evaluating the impact of LUC on POM over time.

Evaluating the degree of soil degradation at a given site requires long term data as one time measurement of soil properties such as SOC can be misleading. Measurements are mostly attached to the time at which measures were taken. Farmers' activities and land use management, however, can vary among seasons and years causing fluctuation and variabilities in soil properties (Zingore et al., 2007). Long-term data thus focus on specific plots over years results in order to derive much more accurate data related to alteration in soil properties over time (Tully et al., 2015). Alternatively, the space-for-time approach, i.e., using land-use chronosequences, allows to analyze temporal changes of chemical or physical soil attributes under real-farm practice (Hartemink, 2006). As long-term experimental farms in Western Africa are largely missing, I used this false chronosequence approach for evaluating SOC stock changes after conversion of natural savannah to permanent cropland.

2.4. Objectives

Soil information translated in soil maps and knowledge on soil carbon dynamics provide data to support both policy making and strategies for ensuring food security and sustainable production. As the creation of soil maps by traditional soil surveys are costly and time consuming, new approaches came into focus that speed up and accelerate soil mapping such as DSM. For the implementation of DSM, research priorities are among others: using appropriate model and covariates for a particular landscape in the perspective of better prediction accuracy, solving high data dimension problems, and dealing with soil legacy data subject to imbalance issues. The advancement in statistical models and the availability of a large array of topographical as well as spectral data offer the possibility to investigate ways to tackle some of these issues. Such approaches are of particular importance for soil landscapes in Western Africa, which are sometimes difficult to access, and where experienced field soil scientists are not necessarily abundant. The overall goal of this study thus was to investigate soil properties and soil reference groups mapping within an old, Plinthosol

landscape, using state of the art methodology. To better understand the dynamics of SOC within this region, I additionally sampled a cultivation chronosequence.

Specifically this study focused of the following research questions:

(i) To which degree are novel statistical methods suited for high resolution mapping of soil properties in tropical environment using remote sensing data?

I hypothesize that statistical models which are able to handle both linear and unlinear pattern in data will provide higher prediction accuracy than those geared towards linear pattern. To verify this hypothesis, I compared the performance of multiple linear regression (MLR) to three machine learning methods such as random forest regression, support vector machine and stochastic gradient boosting. I used high resolution optical imagery (RapidEye and Landsat) along with topographical variables for predicting six soil properties (sand, silt, clay, CEC, SOC and N). The model performances were investigated using cross validation for internal assessment while independent datasets were considered for external evaluation.

(ii) Does the application of instance selection using a data pruning approach improve the prediction accuracy of reference soil groups with a dataset subject to severe imbalance?

I hypothesize that pruning the major soil group - the Plinthosols – will result in increased prediction accuracy of the minor reference soil groups. For this purpose, I carry out a data pruning by considering different core range of the Plinthosol data while cutting off all data points belonging to the outer range. This resulted in different training subsets for predicting the reference soil groups using a wide range of remote sensing variables. The evaluation of the various set was carried out by using Random Forest (RF) along with a recursise feature selection for optimal covariate identification. The specifical and mutual contribution of spectral and topographical variables in predicting the reference soil groups was also assessed.

(iii) How does the topsoil (0 - 30 cm) and subsoil (30 - 100 cm) carbon stock vary among different land use and reference soil groups and which main factors affect their respective spatial distribution?

I hypothesize that natural vegetation and associated reference soil groups will have higher carbon stock compared to cropland with the topographical variables being the main factor affecting the spatial distribution of carbon stock irrespective of the depth. For this question, I firstly determined the amount of carbon stock in both topsoil and subsoil in cropland and savannah as well as in five reference soil groups (Cambisols, Gleysols, Lixisols, Plinthosols, Stagnosols). The identification of the driving factors for both topsoil and subsoil SOC stock as well as their respective spatial distribution were investigated using the RF and linear regression as statistical models.

(iv) To which extent does the land use change from natural savannah to cropland system affect the amount of total soil organic carbon and particulate organic matter in Plinthosols? I consider that continuous cultivation in initial savannah land will result in the reduction of both total soil organic carbon and particulate organic matter in the Plinthosols. To verify this hypothesis, I followed a chronosequence approach by sampling fields with known cropping time in the past as well as undisturbed savannah lands which were used as control. I carry out some physical soil fractionation resulting in different size of particulate organic matter (POM). Additionally, the role of iron oxide as a potential stabilizing agent was also investigated.

This PhD thesis was prepared within the framework of the Working Package 2.5 “Soil carbon dynamics, soil fertility and soil degradation under climate and land use change” as part of the West African Science Service Center on Climate Change and Adapted Land Use (WASCAL) project which is funded by the German Research Foundation (BMBF).

II.

Material and methods

1. Study area

The study took place in the Dano district (Lat. 11°8'56.57"N; Long. 3°3'36.45"W), which is part of the Ioba province in the southwestern part of Burkina Faso. Specifically, it mainly focused on the catchment delineated by WASCAL (West African Science Service Center on Climate Change and Adapted Land Use). WASCAL is a large-scale project aiming at enhancing the resilience of human and environmental systems to climate change and increased variability in the West African region. The WASCAL catchment in Dano covers a total area of 580 km^2. An intensive soil sampling was carried out in the sub-catchment which is about one-quarter of the bigger watershed (Fig. II-1). The elevation ranges between 250 and 504 m above sea level (asl) with a mean average of 295 m asl. The relief is relatively flat with an average slope of 0.2 %.

The climate consists in a mono-modal (single peak) rainy season with a mean annual rainfall ranging between 900 and 1200 mm $year^{-1}$. The mean annual temperature varies between 20.1 and 38.4 °C. The lithology is characterized by the dominance of partly volcanic formations from the middle precambrian period and consists in a great proportion of andesic rocks with massive texture, basalt, diabase, gabbro and quartz-rich andesites. The soils of the study area are mostly sandy to sandy loam in surface while sandy clay, clay loam to clayey in the subsoil similarly to the vast majority of the soils in the Ioba province (Hamidime, 2003). They are characterized by a high stone content and low water holding capacity.

The vegetation of the area belongs to the Sudanian domain with woody, arboraceous or scrubby savannah, abundant in perennial grasses (Schmengler, 2010). Hills and higher slope areas are often covered with thick vegetation. However, a great proportion of this vegetation has been converted into croplands with the practice of short or long fallowing systems. Where long fallowing occurred, it was difficult to distinguish it from natural savannah vegetation. Therefore as in the study carried out by Yira et al. (2016) in the same area, long fallowing system and natural savannah are categorized as savannah. Cultivation is mostly rain fed and farming takes place on a small scale with low input (Callo-Concha et al., 2012b) especially regarding fertilizer.

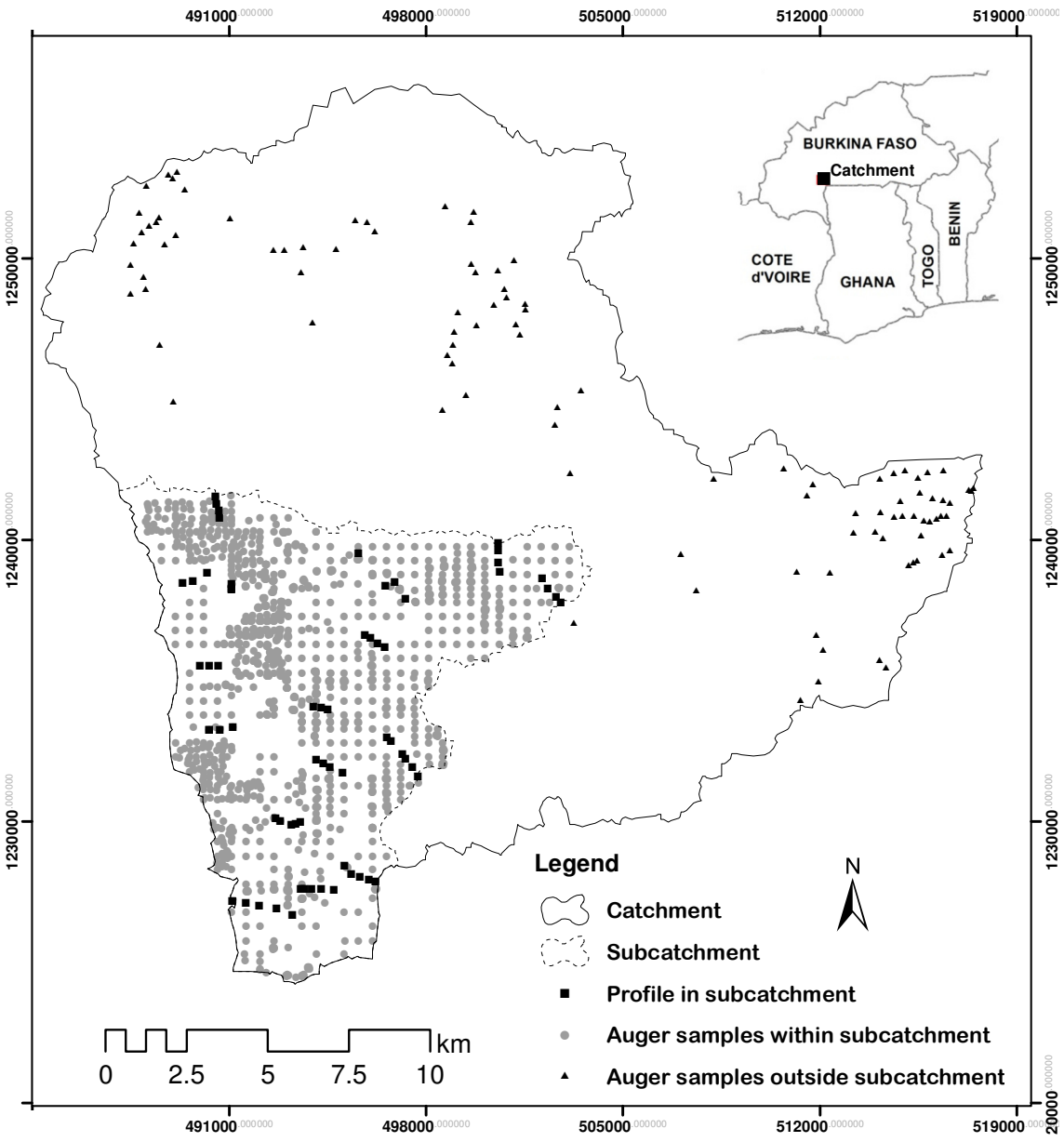

Fig. II-1: Map of the Dano catchment and locations of soil sampling

2. Soil sampling

Soil sampling was carried out in the sub-catchment based on homogeneous units derived from existing soil and land use maps as well as a 90 meter resolution digital elevation model provided by the Shuttle Radar Topography Mission (SRTM). A total of 70 soil profiles were excavated up to 1 m where possible along 16 transects from August to October 2012. For each profile, four soil cores (100 cm^3) were taken per

horizon for the determination of the bulk density (BD). These samples were dried at 105 °C for 24 hours in the oven and corresponding weight were taken for the BD. Each sample was then grounded and sieved for the measurement of the weight of stone content (SC). Moreover, some composite soil samples were collected from each soil horizon for further laboratory analysis resulting in a total of 195 samples with 71 and 124 samples respectively for the A and B horizon.

To account for spatial variability, an intensive auger grid sampling was carried out from August to October 2012 and from August to October 2013 over the entire study area. At each auger point, composite samples as well core samples (4 replicates for BD) were taken but only from the topsoil (A horizon). About 1305 augering composite samples were collected in total with 1203 samples within the subcatchment and 102 samples outside the subcatchment (Fig. II-1). Soil horizon description and soil classification were based on the World Reference Base for soil resources (IUSS et al., 2006).

Apart from Chapter III which focused on samples within and outside the subcatchment, all the remaining chapters are related to the subcatchment. However, the last chapter (Chapter VI) considered samples which were taken from some fields still located in the Dano district but outside the catchment defined by WASCAL.

3. Soil analysis

The composite samples were dried at 40 °C in the oven and sieved to ≤ 2 mm. These samples were analyzed for texture (sand, silt and clay content), pH, cation exchange capacity (CEC), dithionite-extractable Fe oxide (Fe_{DCB}), SOC and N. These parameters were determined following the procedure described by Reeuwijk (2006).

- Texture: The texture analysis was carried out based on a combined wet sieving (sand fraction) and pipette method (silt and clay).

- pH: The pH was determined using a digital pH meter (Orion Star, Thermo Fisher Scientific Inc., Waltham, USA) in suspension of soil in distilled water.

- Cation exchange capacity (CEC): the CEC was obtained from an extraction with chloride of potassium and subsequent micro distillation and titrimetry.

- Fe_{DCB} : the soil samples were treated with the dithionite-citrate-bicarbonate (DCB) for the measurement of the dithionite-extractable Fe (Fe_{DCB}). The Fe_{DCB} content was determined by inductively coupled plasma optical emission spectrometer (ICP-OES).

- C and N: the dried and sieved samples were further milled for C and N analysis. The C and N contents was determined by elemental analysis (ISO 10694, 1995; ISO 13878, 1998) after dry combustion.

4. Determination of SOC stocks

The SOC stock (t C ha^{-1}) was determined by the product of C content, the thickness at a particular depth and the bulk density in each depth along the soil profile. The bulk density was computed by dividing the weight of the oven-dry soil by the volume of the soil cores (Hartge and Horn, 1989). Each quantified bulk density was corrected for the coarse particle content (> 2 mm) which was mainly made up of plinthites. No CaCO3 was found in the collected soil samples. Therefore, the SOC stocks were obtained based on the following equation (II-1):

$$\mathrm{SOC}_{stock} = SOC_i \times \mathrm{BD}_i \times \mathrm{T}_i \times \left(1 - \frac{\mathrm{CP}_i}{100}\right) \qquad \text{(II-1)}$$

where SOCi is the organic carbon concentration (%) of the fine earth (<2 mm) at depth i, BDi is the bulk density (g/cm3) of the fine earth at depth i, Ti is the thickness (cm) of each sample at depth i, and CPi is the coarse-particle content (volume percentage of the fraction >2 mm) at depth i.

5. POM fractionation

The physical fractionation of SOM pools was conducted by two-step ultrasonic dispersion and wet sieving as conducted by Christensen (1992), modified by Amelung and Zech (1999). In brief, 30 g fine earth (< 2 mm) were gently sonicated (60 J ml^{-1}) so that microaggregates were preserved from disruption. The coarse sand fraction (2000–250 µm, POM1) was separated by wet sieving and the filtered remainder was

sonicated a second time (440 J ml^{-1}). The intermediate (250–53 µm, POM2) and silt sized fractions (53-20 µm, POM3) were then separated by wet sieving. The obtained particle-size fractions were dried at 40°C for 24 h before C measurement through elemental analysis (vario MICRO cube, Elementar Analysesysteme GmbH, Hanau, Germany), according to ISO 10694:1995. The concentration of mineral-bound SOM (< 20 µm) was calculated by subtracting the C concentrations of the POM fractions from those of bulk SOC (nonPOM). Regarding potential C losses during fractionation we consider them as minimal as tested by Lobe et al. (2001).

6. Procedure for spectroscopy measurement

The spectra measurement was carried out by inserting 20 mg of the profile samples into microplates and compacted it with a plunger to get a level and plain surface in five replicates. The Bruker Tensor 27 equipped with an automated high throughput device (Bruker HTS-XT) was used to create the spectra. This extension is equipped with a liquid N2-cooled mercury-cadmium telluride (MCT) detector. The spectra recording were done using the OPUS/LAB software within the range of 8000 to 600 cm^{-1} (1250-16700 nm) with resolution of 4 cm^{-1} for each run. This software provides the most representative spectra upon applying the principal component analysis (PCA) and about 50 % of the corresponding profile samples were chosen for laboratory analysis. About 100 profile samples from the subcatchment were conventionally analysed to get the ground truth data while the remaining samples were predicted for SOC, N, CEC and sand, silt and clay fraction.

For each soil parameter, a cross validation method was conducted 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. (2008) (Tab. II-1). The models were optimized with the OPUS QUANT by considering several data processing methods and spectral ranges combination. The data pre-processing consisted in the Multiplicative Scatter Correction method (pH, CEC, silt fraction) and a combination of First derivative and multiplicative Scatter Correction method (SOC, N, Sand and Clay fraction).

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). Only models exhibiting good predictive ability (RPD>2) or close to that (RPD 1.7-2.0) (Albrecht et al., 2008) were used to make predictions for the remaining samples (Tab. II-1). As seen in Tab. II-1, 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.

Tab. II-1: Statistical parameters of the mid infrared spectroscopy-partial least squares regression prediction models (n = 100 samples)

Parameters	Full cross-validation				Test-set validation (V=10 %)			
	$R^{2\,(\%)}$	RMSECV	RPD	Slope	R^2 (%)	RMSEP	RPD	Slope
Sand (%)	70.5	6.8	1.8	0.7	80.9	5.7	2.5	0.7
Silt (%)	75.8	4.9	2	0.8	88.2	3.9	3	0.8
Clay (%)	77.6	6.2	2.1	0.8	80.6	5.5	2.4	0.8
CEC (cmolc kg^{-1})	75.6	3.6	2	0.8	90.5	3.2	3.6	0.8
SOC (%)	95.3	0.1	4.6	0.9	92.2	0.2	3.6	0.9
Nitrogen (%)	85.5	0	2.6	0.9	85.7	0	3	0.8

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

7. Modelling using Random Forest

The random forest analysis for both regression and classification was conducted using the "*Random Forest*" (RF) function as implemented in the RF package (Breiman, 2001) of the R software (R core Team). RF belongs to the family of ensemble machine learning algorithms that predicts a response 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 (Schapire et al., 1998)) 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 (Gislason et al., 2006). 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 (Statnikov et al., 2008). 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 (Kuhn, 2015) 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 (Breiman, 2001). Two importance measures - mean decrease in accuracy (MDA) and mean decrease in impurity (MDI) are frequently computed. To calculate MDA (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, the MDI represents the total decrease in node impurity from splitting on a predictor in the tree construction process, averaged over all trees. For regression, the node impurity is measured by the residual sum of squares (Breiman, 2001). 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 (Breiman, 2001) :

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

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 for regression analysis is expressed as follows:

$$\mathrm{Var} = 1 - \frac{MSE_{OOB}}{Var_{resp}}. \qquad \text{(II-3)}$$

where Var_{resp} is the total variance of the response variable computed with n as divisor (rather than $n - 1$).

III.

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

Modified on the basis of

Gerald Forkuor*, Ozias K.L. Hounkpatin*, Gerhard Welp, Michael Thiel. (2017). PLoS ONE 12(1): e0170478. doi:10.1371/journal.pone.0170478

* Gerald Forkuor and Ozias K.L. Hounkpatin equally contributed to the data collection, data analysis and interpretation as well as the writing of the manuscript of this chapter. As this section was done in joint efforts, the results are part of the present PhD thesis.

1. Introduction

Sustainable land use and optimal soil management require accurate and detailed spatial soil information. In West Africa, where land degradation and loss in soil fertility has been reported by numerous studies (Bationo et al., 2007; Lahmar et al., 2012; Vågen et al., 2005), such information is increasingly required by governments and development partners to aid in improving land management (Sachs et al., 2010). 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 sub-region. This is even crucial for the sustainable use of the soil resources particularly in the context of climate change (Niang et al., 2014).

Digital soil mapping, which includes secondary (non-soil) data sources into the mapping process, has been identified as a potential means of providing soil spatial information (Arrouays et al., 2014; Mulder et al., 2011; Summers et al., 2011). However, recent digital mapping initiatives on the continent (e.g. African Soil Information Service - http://africasoils.net/) (Hengl et al., 2015) and at national scales (e.g. (Akpa et al., 2014)) have used remote sensing 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. Moreover, the success of digital soil mapping is to a large extent dependent on the availability, quality and timing of remote sensing data acquisitions (Blasch et al., 2015). 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 remote sensing data for accurate soil property prediction.

This study consists in a digital soil mapping effort that integrated high spatial resolution multi-temporal RapidEye and Landsat imagery together with ASTER Global DEM terrain derivatives to determine their suitability for improving the availability and accuracy of spatial soil information in rural African landscapes. In that regard, 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) (Grimm et al., 2008; Ließ et al., 2016; Stevens et al., 2012; Wiesmeier et al., 2011) were explored to ascertain the most suitable method for high resolution remote sensing data in the study region. 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?

2. Materials and methods

2.1. Study area (see section II. 1)

2.2. Soil sampling and analysis

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. For soil analysis for texture, CEC, SOC and N see section II. 2 and section II. 3).

2.3. Spectroscopic measurement (See section II. 6)

2.4. Covariate data

2.4.1. 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) (Tyc et al., 2005), while Landsat has eleven spectral channels (Irons et al., 2012) 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 (Tab. III.1) were used in the analysis. Images from both sensors were atmospherically corrected using the ENVI ATCOR

software (Richter and Schläpfer, 2012). 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). Tab. III.1 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 (Ray et al., 2004).

Tab. III-1: Spectral bands of satellite images used and definitions of soil and vegetation indices

Sensor	No. of Bands	Band number, names and abbreviations					
		1	2	3	4	5	6
RapidEye	5	Blue (B)	Green (G)	Red (R)	Red edge (RdE)	Near infrared (NIR)	-
Landsat	6*	Blue (B)	Green (G)	Red (R)	Near infrared (NIR)	Shortwave infrared 1 (SWIR 1)	Shortwave infrared 2 (SWIR 2)

Spectral indices			
Name of Index	Formula	Index property	Reference
Brightness Index (BI)	$((R^2 + G^2 + B^2)/3)^{0.5}$	Average reflectance magnitude	(Ray et al., 2004)
Saturation Index (SI)	$(R - B)/(R + B)$	Spectral slope	(Ray et al., 2004)
Hue Index (HI)	$(2 * R - G - B)/(G - B)$	Primary colors	(Ray et al., 2004)
Coloration Index (CI)	$(R - G)/(R + G)$	Soil color	(Ray et al., 2004)
Redness Index (RI)	$R^2/(B * G^3)$	Hematite content	(Ray et al., 2004)
Normalized Difference Vegetation Index	$(NIR - R)/(NIR + R)$	Health and amount of vegetation	(Huete et al., 2002)

* Spectral bands used in this study

2.4.2. Terrain and climatic variables

Terrain variables (Tab. III-2) were extracted from the 30 m resolution ASTER GDEM (http://asterweb.jpl.nasa.gov/GDEM.ASP). Although previous studies have shown that the 90 m resolution SRTM DEM (Farr and Kobrick, 2000) has a superior absolute accuracy than ASTER GDEM (Forkuor and Maathuis, 2012), the latter was selected

for this study due to its superior spatial resolution. Although 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 (Hijmans et al., 2005a).

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. Tab III-2 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.

Tab. III-2: Terrain and climatic variables considered in this studyTerrain and climatic variables considered in this study

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

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

2.5. Models

2.5.1. 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:

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

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 (R core Team) 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.

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) (Venables and Ripley, 2013). In this study, the "stepAIC" function as implemented in the "MASS" package (Venables and Ripley, 2013) 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. Tab. III-3 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 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.

Tab. III-3: 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

2.5.2. Random Forest Regression (RFR)

For background information on RFR see section II-7.

2.5.3. 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 (Shrestha and Shukla, 2015; Stevens et al., 2012). 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. (Hastie et al., 2011). 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 (Kuhn, 2015). 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.

2.5.4. Stochastic gradient boosting (SGB)

Stochastic gradient boosting (SGB; (Friedman, 2001, 2002)) 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 (Ridgeway, 2008). 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 (Kuhn, 2015). For each soil property, parameters with the lowest error metric (root mean square error) were used for the final model.

2.6. 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. II-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 (Kuhn, 2015) . 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. II-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 (Muñoz and Felicísimo, 2004). 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 equations III-2-3). 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 (Willmott and Matsuura, 2005), its dependence on scale makes it difficult to calculate a model's error in percentage terms. The sMAPE (Makridakis and Hibon, 2000), 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 (Hastie et al., 2011).

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

$$sMAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|O_i - P_i|}{(O_i + P_i)/2} \quad \text{(III-3)}$$

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

3. Results and Discussion

3.1. Model performance

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

3.1.1. 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. Tab. III-4 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 (Malone et al., 2016; Wiesmeier et al., 2014). 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 (Grimm et al., 2008; Wiesmeier et al., 2014) or only spectral data (Coleman et al., 1991; Ray et al., 2004).

Tab. III-4: Internal model validation based on 80 % training data (all Spectral and topographic/climate predictors)

Model	Sand (%)		Silt (%)		Clay (%)		CEC (cmolc kg^{-1})		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

Tab. III-4 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 (Bricklemyer et al., 2007; Zakaria and Shabri, 2012).

3.1.2. Assessment based on independent validation samples

Tab. III-5 and Tab. III-6 present model performance statistics for the external validation inside (20 % of the dataset) and outside the small catchment, respectively (see Fig. II-1). Here, the symmetric mean absolute percentage error (sMAPE) (equation III-3) 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 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, Tab. III-6 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 (Muñoz and Felicísimo, 2004; Odeha et al., 1994; Selige et al., 2006). 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 (Bricklemyer et al., 2007; Hahn and Gloaguen, 2008a; Wålinder, 2014). In general, many studies reported RFR as providing better predictions compared to SVM (Fassnacht et al., 2014; Ließ et al., 2016; Ma et al., 2016; Siegmann and Jarmer, 2015). However, Were et al. (Were et al., 2015) found SVM as best predictor for the spatial distribution of SOC stock compared to RFR. Rossel and Behrens (2010) reported RFR as having better prediction accuracy compared to SGB, while Hitziger and Ließ (2014) 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 Tab. III-5 and Tab. III-6 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 poorly 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 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 (Barnes and Baker, 2000) 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 (Odeha et al., 1994; Ray et al., 2004; Thomasson et al., 2001), apparently to reduce the effect of varying surface characteristics.

Based on their study within a 350 ha demonstration farm in Arizona, Barnes and Baker (Barnes and Baker, 2000) 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 (Callo-Concha et al., 2012b; Forkuor, 2014). The mode and time of land preparation (e.g. tractor, bullocks, manual) (Kamara et al., 2009), nutrient application (e.g. fertility) (Bationo et al., 1998) and water management strategy (Douxchamps et al., 2012) 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 (Rossel et al., 2006; Thomasson et al., 2001).

Tab. III-5: External validation in small catchment based on 20 % testing data with spectral data and terrain/climatic variables

Model	Sand (%)		Silt (%)		Clay (%)		CEC (cmolc kg^{-1})		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

Tab. III-6: External validation based on 102 samples outside the small catchment with spectral data and terrain/climatic variables

Model	Sand (%)		Silt (%)		Clay (%)		CEC (cmolc kg^{-1})		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

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 (Reeves and Smith, 2009), the MIRS-PLSR calibration models in the present study (Tab. II-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 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 %) (McCarty and Reeves, 2006; Terhoeven-Urselmans et al., 2010). Though uncertainty propagation analysis as carried out by Brodský et al. (Brodský et al., 2013) 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 (R^2 > 80 %, RPD ≥ 2). These results indicated that the calibrations were consistent especially for SOC, CEC, N and silt (R^2 > 85 %, RPD ≥ 3). In their study, Brodský et al. (Brodský et al., 2013) 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.

3.2. Variable importance and temporal window for acquisition of RS data

The five top spectral and terrain/climatic variables which contributed most to the accuracy of digital soil mapping in the studied watershed are discernible from Tab. III-7. Though RFR generally provided better predictions, variable ranking from the MLR model was included in the table for comparison purposes. The data in Tab. III-7 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. (Hengl et al., 2015) also recorded elevation as the most important variable influencing SOC contents of topsoil in Africa. Wang et al. (Wang and Ge, 2012) 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 (Hengl et al., 2015; Wang and Ge, 2012). 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.

Tab. III-7: First five predictors that were highly significant for RFR (based on "IncNodePurity" importance measure) and MLR analysis

Model	Rank	Sand (%)	Silt (%)	Clay (%)	CEC (cmolc kg^{-1})	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 Tab. III-1) 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.

Tab. III-7 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 (Tab. III-7). 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.

Tab. III-7 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. (Liao et al., 2013) 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 (Ray et al., 2004).

3.3. 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. III-1). However, along the western border of the study area, medium to higher values of clay, CEC, SOC and N are observed 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 (Wålinder, 2014; Siegmann and Jarmer, 2015). 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.

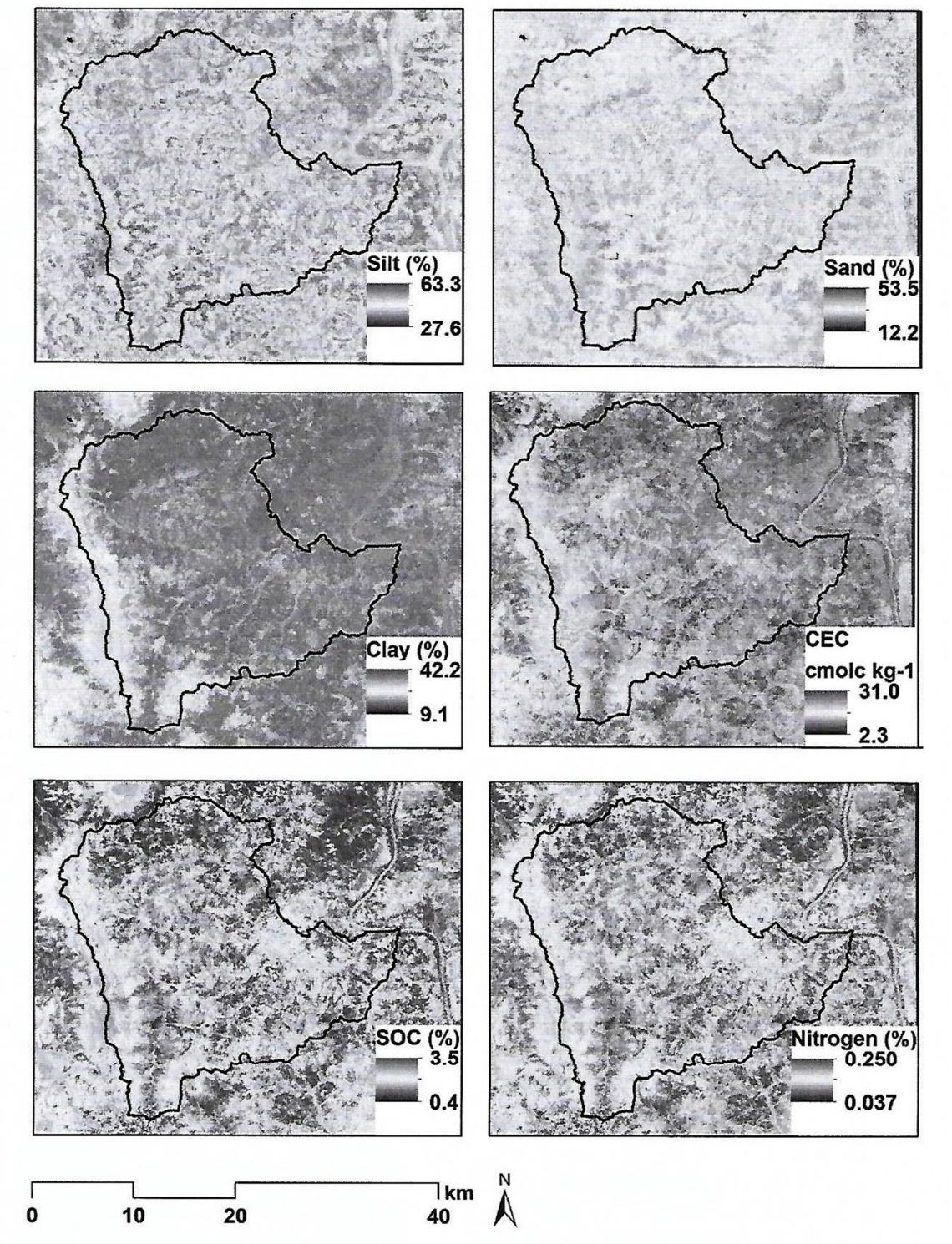

Fig. III-1: Spatial distribution of sand, silt, clay, cation exchange capacity (CEC), soil organic carbon (SOC) and total nitrogen (N) in the topsoil of the studied watershed

4. 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.

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.

IV.

Predicting reference soil groups using legacy data: a data pruning and random forest approach for tropical environment (Dano catchment, SW Burkina Faso)

Modified on the basis of

Ozias K. L. Hounkpatin, Karsten Schmidt, Felix Stumpf, Gerald Forkuor, Thorsten Behrens, Thomas Scholten, Wulf Amelung, Gerhard Welp (2017). Scientific Reports.

Submitted manuscript

1. Introduction

Soils are key asset for sustainable living conditions on earth as their functions are related to food and biomass production, water control and chemical recycling, platform provision for human activities, supply of raw materials and the offering of habitat for soil biodiversity (Blum, 2005). Though soil importance is generally acknowledged, farmers, decision makers as well as the scientific community often lack adequate and timely spatial soil information to address land degradation issues. Various initiatives such as the GlobalSoilMap.net project are currently working to overcome the previous challenges in order to provide up-to-date and relevant soil information in Africa using modern techniques (Sanchez et al., 2009). Being a time- and cost-effective alternative to classical soil surveys, digital soil mapping (DSM; McBratney et al., 2003) – also called soil-landscape modelling (Gessler et al., 1995) and predictive soil mapping (Scull et al., 2003) – is a subset of pedometrical research using geo-statistics and data mining methods to spatially predict soil classes or soil properties based on existing soil and environmental covariate data.

When mapping soil taxonomy units, the quantitative relationship between a certain class unit and the soil formative environmental factors is supposed to be unique as soil classes are different from each other. However, in complex soil-landscapes, the individual features of certain soil classes overlap in space, which is particularly difficult for correct DSM with imbalanced datasets (Gopi et al., 2016). Ideally, balanced datasets are required for decision trees algorithms to produce better classification (Ertekin et al., 2007). However, DSM mostly focuses on soil legacy data whose sampling design might not provide such ideal scheme for post hoc analysis (Mayr et al., 2010), especially for data scarce countries like in tropical areas. Generally, for datasets with uneven class size, the classification model, which is generated from decision trees (DT) algorithm, biases towards the majority class (Ertekin et al., 2007).

This section addresses a digital soil mapping approach to classify reference soil groups in a tropical environment using a large dataset with Plinthosols (PT) as the dominant group. I used Random Forest (RF) as robust data mining method (Schmidt et al., 2014)

to evaluate the performance of different data subsets to deal with class imbalances (e. g. Schmidt et al., 2008) and noise within the dataset. The various RF-models were trained on a detailed covariate set including terrain and multispectral predictors. Though the issue of class imbalance has been acknowledged in many studies dealing with soil classification, to my knowledge, no such method has been applied for legacy soil data from a tropical semi-arid environment. This approach being considered, I hypothesized that: (1) instance selection on the majority soil group would improve the performance of the RF models and result in a better classification of the minority soil groups, (2) integrating spectral bands and indices along with environmental covariates would have greater impacts on RF classification performance compared to their unique contribution.

2. Materials and methods

2.1. Study area (see section II. 1)

2.2. Soil Sampling (see section II. 2)

2.3. Reference soil groups

Six soil classes were encountered in the Dano catchment and were described based on the WRB as follows: Cambisols, Gleysols, Lixisol, Leptosols, Plinthosols and Stagnosols. The Cambisols are young soils with incipient soil formation with beginning horizon differentiation demonstrated by changes in colour, structure or carbonate content. Gleysols refer to water influenced soils which are saturated with groundwater for long enough periods to develop a characteristic “gleyic colour pattern” made up of reddish, brownish or yellowish colours at ped surfaces and/or in the upper soil layer(s), along with greyish/bluish colours inside the peds and/or deeper in the soil. Stagnosols are also water influenced soils characterized by a perched water table showing redox processes caused by surface water due to periodical wetting; they are mottled in the topsoil and subsoil, with or without concretions and/or bleaching. Lixisols consist of strongly weathered soils in which clay has been removed from an eluvial horizon down to an argic subsurface horizon that has low activity clays and a moderate to high base saturation level. Leptosols include very shallow soils over hard rock or very calcareous material, but also deeper soils that are extremely gravelly

and/or stony. Plinthosols point to soils that contain 'plinthite', i.e. an iron rich, humus-poor mixture of kaolinitic clay with quartz and other materials that change irreversibly to a hardpan or to irregular aggregates on exposure to repeated wetting and drying. For more detailed description refer to IUSS et al. (2006).

2.4. Geospatial and spectral variables

To provide a wide range of different environmental covariates dealing with the state factor equation, a set of predictors was delineated (95 variables, Tab. IV-1 & Tab. III-1), which were compiled from different sources with ArcGIS 10.3.1 (Environmental Systems Research Institute, ESRI Inc., Redlands, CA) and SAGA GIS (System for Automated Geoscientific Analyses). About 45 of these variables are terrain attributes (Tab. IV-1), 45 are spectral bands and indices (Tab. III-1) while the remaining data (Tab. IV-2) relate to land use, parent material, geormorphology, and climate (temperature and precipitation).

The terrain attributes were derived from a SRTM (Shuttle Radar Topography Mission) DEM with a 90 m resolution (Jarvis et al., 2008). For land use data, the map generated by Forkuor (2014) covering the study area was used. The parent material allocated to each sampling location was extracted using a geological map (1/ 100 000) of Burkina Faso made by Hottin and Ouedraogo (1992). A geormorphological map (1/ 100 000) from the National Soil Office was considered (Bureau National des sols, 2000). Climatic data include mean annual temperature (Temp) and annual precipitation (Prep) at 1 km resolution from the worldclim datasets (Hijmans et al., 2005b) .

For the spectral data see section III. 2.4.1. Finally all datasets were resampled to a spatial resolution of 90 m.

Tab. IV-1:Terrain attributes used as predictors for soil mapping

Variables	Abbreviation	Unit
Distance to stream ArcGis	Dist.stream	m
Relief intensity ArcGis	Ri	m/m^2
Potential drainage density ArcGis	Pdd	km/km^2
Elevation ArcGis	Elevation	m
Slope ArcGis	Slope.per	%
Maximum Slope SAGA	Slope.maxT	°
Steepest slope SAGA	steepest.slope	°
Flow direction ArcGis/SAGA	A.Flow.d/S.Flow.d*	-
Flow accumulation ArcGis/SAGA	A.Flow.A/S.Flow.A	-
Profile curvature ArcGis	A.Profile.cur/S.Profile.curv	° m^{-1}
Curvature ArcGis	A.curv	m^{-1}
Plan curvature ArcGis	A.Plan.curv/S.Plan.curv	° m^{-1}
General curvature SAGA	S.Gen.curv	° m^{-1}
Total curvature SAGA	S.totalcuv	° m^{-1}
Min curvature SAGA	S.min.curv	° m^{-1}
Max curvature SAGA	S.max.cuv	° m^{-1}
Horizontal curvature SAGA	S.Hor.curv	° m^{-1}
Cross curvature SAGA	S.cross.curv	° m^{-1}
Flow line curvature SAGA	S.Flow.line.curv	° m^{-1}
Catchment Area Rectangle SAGA	S.CA.Rec	m^2
Catchment Area Parallel SAGA	S.CA.Par	m^2
Catchment Area SAGA	S.CA	m^2
Aspect ArcGis/SAGA	A.Asp/S.Asp	-
Eastness	sine.Asp	°
Northness	cose.Asp	°
Slope Length factor SAGA	LS.Factor	m
Topographic Wetness Index ArcGis/SAGA	A.TWI/S.TWI	-
Topographic Wetness Index SAGA	S.TWI	-
SAGA Wetness Index SAGA	S.Wet.Ind	-
Vertical Flow Distance SAGA	Verti.Flow.dist	m
Vertical distance to a network Channel SAGA	Verti.dist.Net	m
Terrain ruggedness SAGA	Terr.Rugg	
Topographic position index SAGA	Topo.Posi.Ind	
Protection index SAGA	Prot.Index	-
Overland flow distance SAGA	Overland.Flow.dist	m
Mass Balance index SAGA	Mass.Bal.ind	-
Horizontal flow distance SAGA	S.HF.dist	m
Convergence Index SAGA	S.convg.ind	-

Channel base index SAGA	S.Chanbase.ind	-	

Tab. IV-2: Land use, lithology, geomorphology units and descriptive statistics for climate variables

	Elements	Area (km^2)	Area (%)
Land use units	Cropland	58.18	34.54
	Savannah	90.24	55.22
	Water	0.46	0.30
	Bare areas	4.43	2.86
	Urban areas	1.24	0.80
Lithology units	Granodiorites and undifferentiated tonalites	0.20	0.13
	Acid Metavolcanites and pyroclastites	14.49	9.35
	Volcano sedimentary rocks	111.78	72.12
	Neutral to alkaline Metavolcanites	28.53	18.41
Geomorphology units	Lateritic ridge	23.18	14.96
	Rocky ridge	4.24	2.74
	Plateau	15.71	10.14
	Upper slope glacis	12.05	7.78
	Middle slope glacis	15.67	10.11
	Alluvial levee	0.38	0.25
	Inland valleys	17.09	11.03
	Peripheral depression	66.65	43.00

	Climate variable	
Statistics	Temperature (°C)	Precipitation (mm)
min	27.22	775.83
max	27.92	810.83
median	27.63	794.17
sd	0.13	8.53

2.5. Modelling with Random Forest

(see section II. 7 for background information about Random Forest)

For the present study, 1000 trees were built and the number of features at each split was defined based on the ten-fold cross-validation tuning procedure with the Classification and Regression Training (Caret) package in R software (Kuhn, 2015).

Though RF is quite robust towards multicollinearity, the presence of highly correlated covariates can lead to biased interpretation as they carry the same information (Kuhn, 2008). Moreover, Genuer et al. (2010) reported that the variable importance based on the mean decrease in classification accuracy is overestimated for highly correlated variables. For model prediction, the feature space was reduced in two ways. Firstly by computing a correlation matrix for the terrain attribute predictors and identifying the minimal set of predictors that can be removed using a specific threshold. This was carried out using the classification and regression training (Caret) package (Kuhn, 2015) in R 3.1.2. A specific threshold of 0.70 was set and the predictor most involved in the pairwise correlations was removed.

Secondly, recursive feature elimination (Kuhn and Johnson, 2013) function of the classification and regression training (Caret) package (Kuhn, 2015) was used to select among all the variables an optimal set of parameters for classification. Recursive feature elimination works by establishing a classification model using all the available predictors, then proceeds to rank these predictors by order of importance, and next discards the predictors of the lowest importance. It replicates the same process till either the reach of a specific threshold or when only one predictor is left (Brungard et al., 2015). The RF modelling was then carried out using covariate predictors retained based on the correlation matrix (RF) and also by using an optimal set of predictors resulting from recursive feature elimination (RF_rfe).

To assess the influence of the different spectral and terrain variables on soil class prediction, a different combination was carried out for running the models: (1) only the spectral parameters (SP), (2) only the terrain parameters (TP) and (3) both terrain and

spectral parameters (TSP). However, the Litho, Geo, LU and Prep attributes were used along the terrain attributes.

2.6. Experimental design: data pruning

The field observation of this study revealed the Plinthosols as the dominant reference soil group with about 73 % of the grand total percentage (Tab. IV-3). As general assumption, the possibility of a potential overestimation of this particular soil class was envisaged as is often the case for such kind of big datasets with imbalance related issues. The first step in the present study was therefore to test this hypothesis by running the model with the entire dataset. In a second step, data pruning was carried out as a method to tackle the potential prominence of the majority class in the feature space once the latter hypothesis revealed true. For this purpose, a set of data pruning experiments was conducted by defining a set of data core ranges (CR).

The different pruning operations were carried out based on the RF variable importance measurement expressed by the mean decrease in classification accuracy. The latter follows the rationale that when values of a variable at a particular node are randomly permuted, this variable is supposedly absent from the model. The difference in the classification accuracy before and after the permutation of the values of the predictor variable, i.e. after considering and excluding this predictor variable, is used as a measure of variable importance (Strobl et al., 2008). These computations are conducted tree by tree till the whole random forest is constructed (Liaw and Wiener, 2002). This results in the discrimination between essential and inessential variables. The most important variable is the one with the highest contribution to model accuracy and with the greatest impact in the feature space, driving the overall classification. Consequently, the most important variable – the wetness index - was used to determine the data core range for the pruning operation of the Plinthosols.

The data pruning experiments were carried out by defining a set of 80 % (80 % CR) and 90 % (90%CR) core range of the Plinthosol data as well as a standard deviation (σ) based (SDCR) core range while cutting off all data points belonging to the outer range. These core ranges were set by (i) calculating the density distribution of the wetness index as revealed by the RF model, (ii) calculating the cumulative percentage

by dividing the cumulative frequency by the total number of observations (n), then multiplying it by 100 (the last value being equal to 100 %), (iii) cutting off all data points belonging to the outer ranges of a chosen data core range, i.e. for defining, e.g., a 80 % core range (Fig. IV-2), all points lower than 10 % and higher than 90 % of the cumulative percentage were cut off. Similarly, a core range based on the standard deviation (σ) of the values (about 68% core range) of the wetness index was defined (Fig. IV-3). For that purpose, values lower than "μ - σ" (with μ being the arithmetic mean of the driving variable) as well as values higher than "μ + σ" were cut off. The standard deviation based core range (SDCR) was then set by considering data values within one standard deviation of the mean (mathematically, μ ± σ).

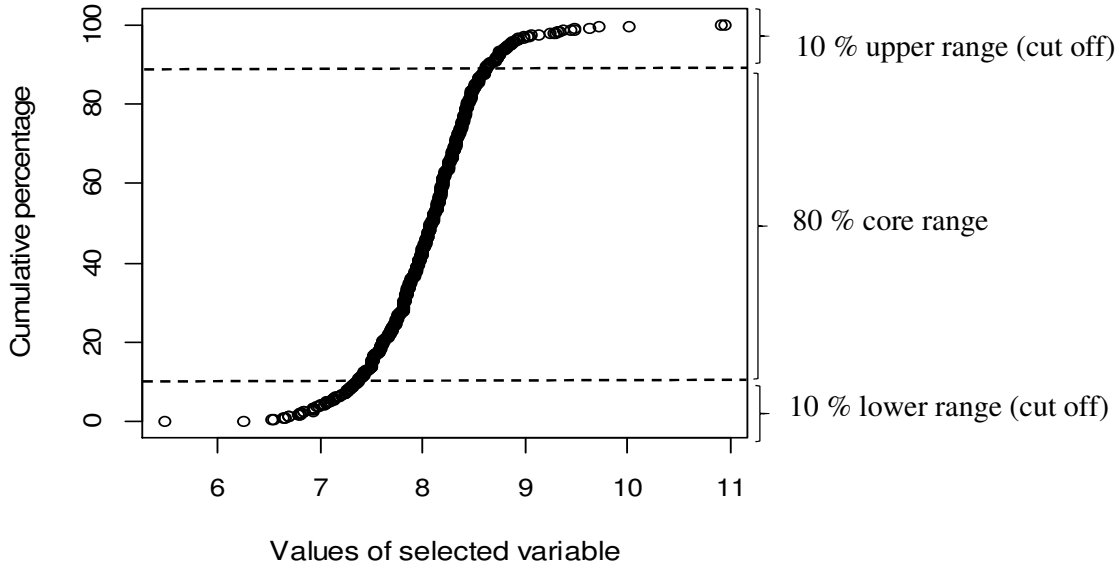

Fig. IV-1: Core range definition of the Plinthosol dataset based on the cumulative percentage of the density distribution of the driving variable (wetness index)

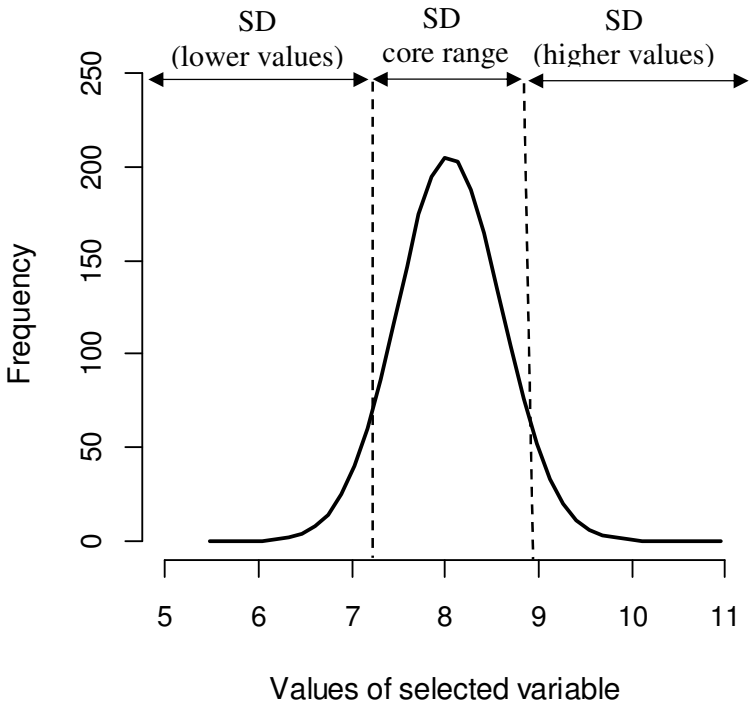

Fig. IV-2: Core range definition of the Plinthosol dataset based on the standard deviation of the values of the driving variable (wetness index)

Finally, a total of four different datasets were defined: (1) entire dataset with all the Plinthosols (AllPT), (2) a 90 % core range (90%CR) of the PT dataset, (3) a 80 % core range (80%CR) of the PT dataset by cutting off all points lower than 10 % and higher than 90 % of the cumulative percentage, (4) a SD core range (SDCR) of the PT dataset by pruning values lower and higher than "$\mu - \sigma$"and "$\mu + \sigma$" respectively. Each dataset was used to train a RF model along with the different categories of predictors: spectral parameters (SP), terrain parameters (TP), terrain plus spectral parameters (TSP).

Tab. IV-3: Count (n) and frequencies (%) of the reference soil groups in the Dano catchment

Reference soil groups	n	Percentage of grand total (%)
Cambisols (CM)	86	6.68
Gleysols (GL)	141	10.95
Leptosols (LP)	22	1.71
Lixisols (LX)	59	4.58
Plinthosols (PT)	645	73.45
Stagnosols (ST)	34	2.64

2.7. Model validation and map comparison

The dataset was split with 80 % used for training and 20 % for validation. The different pruning was carried out on the train set obtained from the split. These pruned dataset (80%CR, 90%CR and SDCR) were evaluated over the same validation data initially obtained from the split. The classification accuracy was based on the Kappa index. The Kappa value ($\varkappa$) gives the level of accuracy for a particular classification due to chance agreement (Congalton and Green, 2008). This is particularly important when dealing with unbalanced class data as a class having larger distribution would result in higher classification accuracy. A $\varkappa$ value of 0 was considered as a random classifier, 1 as perfect classification, 0.80 as strong agreement, between 0.4 and 0.8 as substantial agreement and below 0.4 as poor agreement (Congalton and Green, 2008). The kappa value was computed as follows:

$$\varkappa = (Pr(a) - Pr(e)) / (1 - Pr(e)) \quad \text{(IV-1)}$$

with Pr(a) : relative observed agreement, Pr(e) : hypothetical probability of chance agreement, and $\varkappa$: the kappa index value.

3. Results

3.1. Terrain attribute selection

The minimal set of predictors finally retained for modelling after computing the correlation matrix amounted to 50 variables. Selected predictors consisted of 19 DEM attributes, 22 spectral data as well as lithology, geomorphology, land use, and precipitation. Using the so-called scorpan function (McBrantney et al., 2003), the analyses included the: (1) soil attributes (s) represented by the spectral band and indices like redEdge, Hue Index (HI), Coloration Index (CI), Redness Index (RI), Brightness Index (BI), Near-infrared (NIR), Shortwave-infrared (SWIR), Saturation Index (SI); (2) precipitation as climatic (c) element, (3) indices for vegetation and human activity (o) such as normalized difference vegetation index (NDVI), land use, (4) terrain (r) variables and (5) lithology as proxy for parent material (p). The optimal subset of covariate predictors resulting from the recursive feature elimination approach returned eight variables, namely: wetness index, elevation, distance to stream to network, protection index, precipitation, near infrared and shortwave infrared.

3.2. Model performances with different data treatments

The performance of the RF models was assessed for different data experiments consisting of the entire dataset (AllPT) and the pruned dataset (i.e., 80%CR, 90%CR and SDCR) based on: (1) OOB errors of the different RF models, and (2) the independent validation samples (prediction accuracy of the independent sample set and Kappa values). The data pruning was carried out based on the SAGA wetness index, since this parameter had been identified as contributing most to RF performance in classification accuracy even with RF models based on recursive feature elimination (Fig. IV-6).

3.2.1. Assessment based on the OOB errors

The OOB errors varied with the different combinations of dataset and category of variable (Tab. IV-4). The highest OOB errors were recorded for the prediction based on spectral parameters, ranging from 28.7 % to 32.7 %. The lowest OOB errors were obtained with the terrain parameters (20.0 % to 22.4 %) and with the terrain plus spectral parameters (20.1 % to 22.6 %). Increasing the level of pruning was generally followed by increasing OOB errors for the spectral parameters for both RF and RF_rfe. The OOB errors using the entire data (AllPT) recorded mostly the highest

OOB errors compared to those of the pruned dataset when terrain parameters only or terrain plus spectral parameters were used as predictors. The lowest OBB error (19.6 %) was recorded for the 90%CR dataset associated with terrain plus spectral parameters.

Tab. IV-4: Training set, percentage of Plinthosols (PT) samples removed from the total set, and out of of the bag errors (OOB error) distribution of the different subsets of data

				OOB error (%)		
	Data treatment	n	PT removed (%)	Spectral Parameters	Terrain Parameters	Terrain and Spectral Parameters
RF	AllPT	792	-	28.7	22.4	22.4
	90%CR	743	6.2	29.8	21.7	21.7
	80%CR	694	12.4	32.3	21.3	21.2
	SDCR	667	15.9	33.2	21.2	21.7
RF_rfe	AllPT	792	-	28.7	21.5	22.6
	90%CR	743	6.2	29.6	21.9	19.6
	80%CR	694	12.4	31.4	20.9	20.1
	SDCR	667	15.9	32.7	20.0	20.8

PT: Plinthosols, OOB error: out of the bag error, AllPT: entire dataset, SDPT: dataset with PT pruned based on standard deviation, 15PT: dataset with 15 % of the PT pruned, 25PT: dataset with 25 % of the PT pruned, 30PT: dataset with 30 % of the PT pruned.

3.2.2. Assessment based on independent validation samples

The results of the performance of the RF models based on independent samples are presented in Tab. IV-5 showing the confusion matrix between observed and predicted reference soil groups for the entire dataset (AllTP). The RF and RF_rfe models for the entire dataset displayed a high level of accuracy for the identification of the Plinthosols (95-98 % for RF and 91-96 % for RF_rfe), irrespective of the category of parameters used. Both RF and RF_rfe performed better for the Gleysols and Leptosols when only terrain or terrain plus spectral parameters were considered, with the prediction accuracy being 18-30 % and 50 % greater, respectively, than achieved with the model that was based on spectral parameters only. Cambisols and Stagnosols, however, were in most cases not well predicted (< 35 % prediction accuracy), no

matter which model or category of parameters was chosen. Noteworthy, the classification shows that most of the reference soil groups were misclassified as Plinthosols, again irrespective of the category of model or parameters considered.

Tab. IV-5: Confusion matrix between observed and predicted reference soil groups for the entire dataset

RF

		Predicted (%)					
	Observed	CM	GL	LP	LX	PT	ST
Spectral parameters (AllPT)	CM	**23.5**	5.9	0.0	0.0	70.6	0.0
	GL	0.0	**28.6**	0.0	0.0	71.4	0.0
	LP	0.0	0.0	**25.0**	0.0	75.0	0.0
	LX	0.0	0.0	0.0	**45.5**	54.5	0.0
	PT	0.0	1.6	0.0	0.0	**98.4**	0.0
	ST	0.0	0.0	0.0	0.0	66.7	**33.3**
Terrain parameters (AllPT)	CM	**23.5**	0.0	0.0	0.0	76.5	0.0
	GL	0.0	**60.7**	0.0	0.0	39.3	0.0
	LP	0.0	0.0	**75.0**	0.0	25.0	0.0
	LX	0.0	0.0	0.0	**45.5**	54.5	0.0
	PT	0.0	4.7	0.0	0.0	**95.3**	0.0
	ST	0.0	0.0	0.0	16.7	66.7	**16.7**
Terrain and Spectral parameters (AllPT)	CM	**17.6**	0.0	0.0	0.0	82.4	0.0
	GL	0.0	**57.1**	0.0	0.0	42.9	0.0
	LP	0.0	0.0	**75.0**	0.0	25.0	0.0
	LX	0.0	9.1	0.0	**45.5**	45.5	0.0
	PT	0.0	5.4	0.0	0.0	**93.8**	0.8
	ST	0.0	0.0	0.0	0.0	66.7	**33.3**

RF_rfe

		Predicted (%)					
	Observed	CM	GL	LP	LX	PT	ST
Spectral parameters (AllPT)	CM	**23.5**	5.9	0.0	0.0	70.6	0.0
	GL	0.0	**39.3**	0.0	0.0	60.7	0.0
	LP	0.0	0.0	**25.0**	0.0	75.0	0.0
	LX	0.0	0.0	0.0	**45.5**	54.5	0.0
	PT	0.0	3.1	0.0	0.8	**96.1**	0.0
	ST	0.0	0.0	0.0	0.0	66.7	**33.3**
Terrain parameters (AllPT)	CM	**23.5**	0.0	0.0	0.0	76.5	0.0
	GL	0.0	**57.1**	0.0	0.0	42.9	0.0
	LP	0.0	0.0	**75.0**	0.0	25.0	0.0
	LX	0.0	9.1	0.0	**45.5**	45.5	0.0
	PT	0.0	5.4	0.0	0.0	**94.6**	0.0
	ST	0.0	0.0	0.0	16.7	66.7	**16.7**
Terrain and Spectral parameters (AllPT)	CM	**23.5**	5.9	0.0	5.9	64.7	0.0
	GL	0.0	**60.7**	0.0	0.0	39.3	0.0
	LP	0.0	0.0	**50.0**	0.0	50.0	0.0
	LX	0.0	9.1	0.0	**63.6**	27.3	0.0
	PT	1.6	7.0	0.0	0.0	**91.5**	0.0
	ST	0.0	0.0	0.0	0.0	66.7	**33.3**

Models with (RF_rfe) and without (RF) recursive feature elimination; AllPT: entire dataset, CM: Cambisols, GL: Gleysols, LP: Leptosols, LX: Lixisols, PT: Plinthosols, ST: Stagnosols.

With increasing pruning level, gain in prediction accuracy was observed for most of the different reference soil groups particularly when terrain or terrain plus spectral parameters were used (Fig. IV-4). The RF models based on the recursive feature elimination performed in most cases slightly better than those on a normal run of the RF. For instance, improvement in classification for the Cambisols was observed with the RF_rfe models when using the 80 % (80%CR) and 90 % (90%CR) core range, dataset combined with terrain plus spectral parameters. These Cambisols gained 35 % and 41 % respectively in prediction accuracy compared to the results with the model based on the entire dataset (AllPT). Likewise, with the RF_rfe models, the Gleysols also recorded an increase of 7 % in prediction accuracy with both the 80 % (80%CR) and 90 % (90%CR) core range dataset combined with terrain plus spectral parameters while the standard deviation core range (SDCR) produced an increase of 10 % when associated with the same category of predictors. The highest prediction accuracy for the Lixisols was recorded with the normal RF with 80 % core range (80%CR) and standard deviation core range dataset (SDCR) associated with terrain parameters with an increase of 18 % compared to the results with the entire dataset (AllPT).

The prediction of the Leptosols were greatly improved when both RF and RF_rfe models were run with either terrain only or with terrain plus spectral parameters resulting in an increase of 25 % in prediction accuracy. No other improvement occurred for the Leptosols with the pruned dataset. For the Stagnosols, most of their validation sample points were predicted with 33 % in prediction accuracy except for the RF model based on the standard deviation core range dataset (SDCR) associated with terrain plus spectral parameters. The latter recorded up to 50 % in prediction accuracy. Compared to results from models based on the entire dataset (AllPT), the Plinthosol prediction accuracy dropped generally with increased pruning intensity when using either the terrain parameters only or when the latter were used along the spectral parameters. The RF_rfe model based on the 90%CR dataset associated with the terrain plus spectral parameters recorded a drop of 4.7 % in prediction accuracy for these Plinthosols compared to the results with the entire dataset (AllPT).

The highest kappa value for model based on the entire dataset (AllPT) was found with the RF model associated with the terrain parameters with $\varkappa$=0.51 (Fig. IV-5). Considering the variation of the kappa values ($\varkappa$) in relation to the data treatment, the pruned datasets with models based on the recursive feature elimination (RF_rfe) generally recorded higher Kappa values than the AllPT reference when terrain plus spectral parameters were used as predictors.The combination of the 90%CR and 80%CR dataset (90%CR) with terrain plus spectral parameters (90%CR-TSP) recorded the highest kappa value with respectively $\varkappa$=0.57 and $\varkappa$=0.55. Models run with spectral parameters recorded the lowest kappa values while those conducted with the terrain parameters were improved by recursive feature elimination. However, Fig. IV-5 also shows that the kappa values dropped for most of the models based on the standard deviation core range dataset (SDCR). It is worthy to note that the model based on the 90%CR associated with the terrain plus spectral parameters also recorded the highest kappa value with the lowest OOB errors (19.6 %).

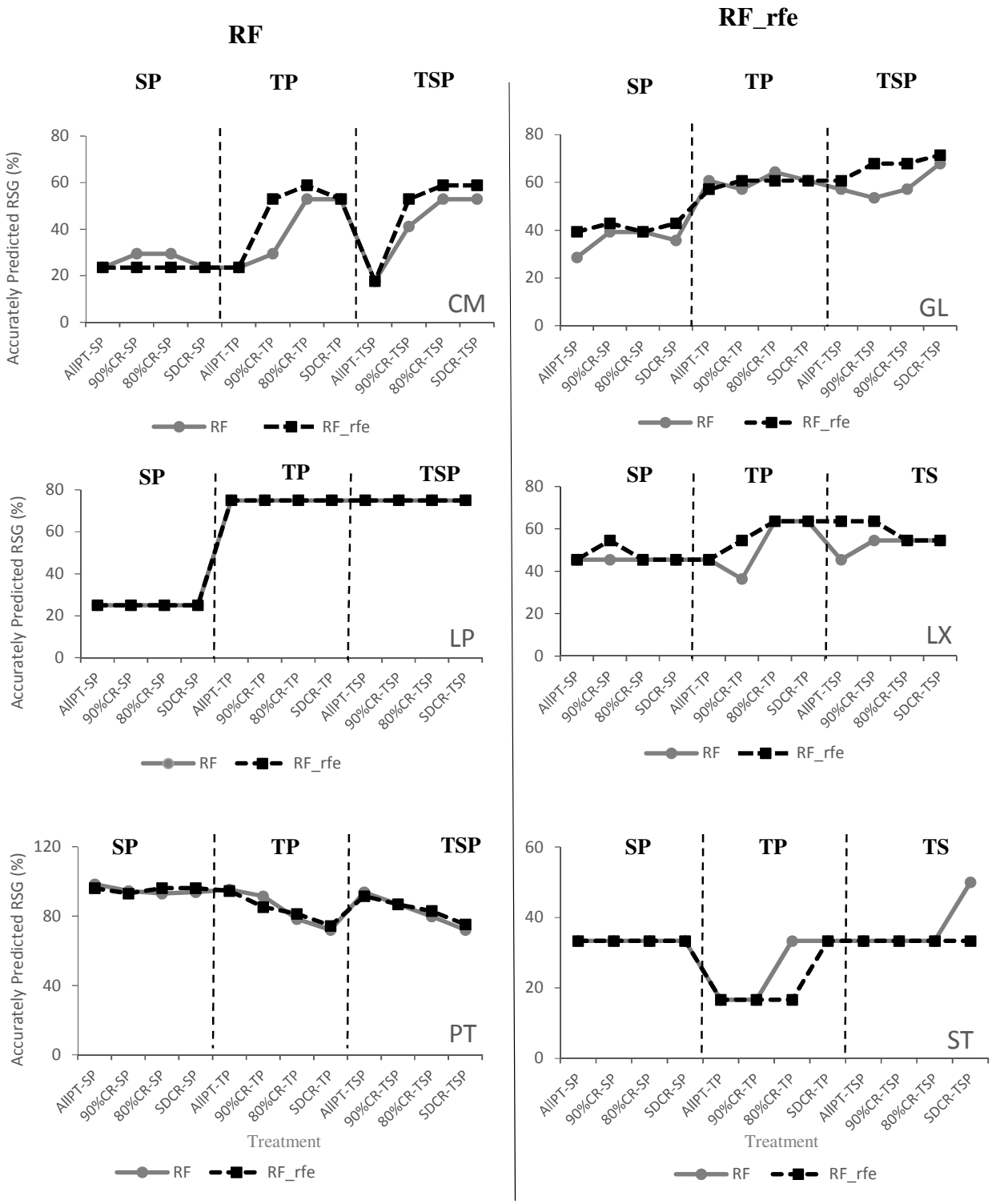

Fig. IV-3: Accurately predicted reference soil groups for different sets of data and covariates

Models with (RF_rfe) and without (RF) recursive feature elimination. CM: Cambisols, GL: Gleysols, LP: Leptosols, LX: Lixisols, PT: Plinthosols, ST: Stagnosols, SP: spectral parameters; TP: topographic parameters, TSP: topographic

and spectral parameters. AllPT: entire dataset including all Plinthosols, AllPT: entire dataset, 90%CR: dataset with 5 % lower and upper range pruning, 80%CR: dataset with 10 % lower and upper range pruning, SDCR: dataset with standard deviation based pruning

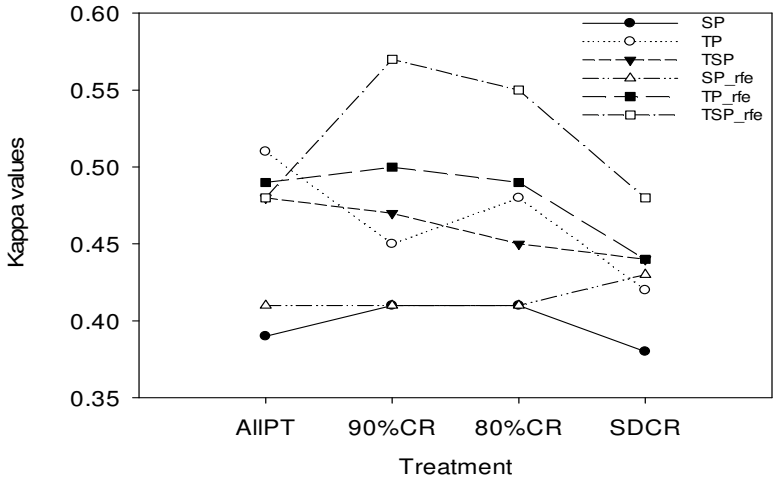

Fig. IV-4: Variation of Kappa values in relation to data treatment
AllPT: entire dataset, 90%CR: dataset with 5 % lower and upper range pruning, 80%CR: dataset with 10 % lower and upper range pruning, SDCR: dataset with standard deviation based pruning, SP_rfe: spectral parameters (SP) with recursive feature elimination, TP_{rfe}: terrain parameters (TP) with with recursive feature elimination, TSP: terrain and spectral parameters (TSP) with recursive feature elimination..

3.3. Prediction of the pruned Plinthosols

Since the models run with the spectral parameters recorded the lowest kappa values, prediction of the pruned Plinthosols were only carried out with the RF and RF_rfe models associated with either terrain parameters only or with terrain plus spectral parameters (Tab. IV-6). Tab. IV-6 shows that none of the models could perfectly predict the Plinthosols, though about half of the models attributed the highest prediction to the Plinthosols. Most of the Plinthosols were predicted as Cambisols (17.7-44.7 %) compared to the remaining RSG while very few were predicted as Leptosols (0-8 %). The highest accurate prediction (> 30 %) varies from 38.8 % (90%CR-TSP of the RF_rfe model) to 71.4 % for the Plinthosols (90%CR of the RF model). All the predictions based on the terrain plus spectral parameters from the

RF_rfe models resulted in higher predictions of the Plinthosols compared to the Cambisols and remaining soil units.

Tab. IV-6: Confusion matrix between observed and predicted reference soil groups for the pruned Plinthosols

		RF						RF_rfe						
		Predicted (%)						Predicted (%)						
	Observed	CM	GL	LP	LX	PT	ST	Observed	CM	GL	LP	LX	PT	ST
Terrain parameters (90%CR)	PT	4.1	12.2	8.2	2.0	**71.4**	2.0	PT	32.7	14.3	2.0	10.2	**26.5**	14.3
Terrain parameters (80%CR)	PT	35.7	12.2	1.0	10.2	**22.4**	18.4	PT	22.4	11.2	1.0	9.2	**42.9**	13.3
Terrain parameters (SDCR)	PT	40.7	10.6	1.6	10.6	**16.3**	20.3	PT	28.5	9.8	0.8	9.8	**34.1**	17.1
Terrain and spectral parameters (90%CR)	PT	44.7	15.4	0.0	7.3	**15.4**	17.1	PT	26.5	16.3	0.0	10.2	**38.8**	8.2
Terrain and spectral parameters (80%CR)	PT	35.7	20.4	1.0	13.3	**21.4**	8.2	PT	17.5	18.6	3.1	5.2	**42.3**	13.4
Terrain and spectral parameters (SDCR)	PT	44.7	15.4	0.0	7.3	**15.4**	17.1	PT	18.7	17.9	2.4	5.7	**43.1**	12.2

Models with (RF_rfe) and without (RF) recursive feature elimination; CM: Cambisols, GL: Gleysols, LP: Leptosols, LX: Lixisols, PT: Plinthosols, ST: Stagnosols; 90%CR: dataset with 5% lower and upper range pruning, 80%CR: dataset with 10% lower and upper range pruning, SDCR: dataset with standard deviation based pruning

3.4. Variable importance

Though many models were considered in the present study with different dataset, results for the variable importance focused only on those which recorded high Kappa values for each category of predictors. Fig. IV-6 presents the variable importance from

models based on (i) the entire dataset associated with terrain parameters (AllPT-TP), (ii) 90 % and 80 % core dataset (90%CR, 80%CR) associated with terrain plus spectral parameters, (iii) the standard deviation core range (SDCR) dataset associated with the spectral parameters. Only the five top variables are presented in the figure

For models based on the entire dataset (AllPT) and on the 90 % and 80 % core dataset (90%CR, 80%CR), the SAGA wetness index (S.Wet.Ind) was ranked as the most important variable driving the reference soil group classification no matter which dataset was used. It was followed by the distance to stream network (Dist.stream) and either by the protection index (degree of local surface convexity or concavity) or elevation. Considering the different reference soil groups, the Gleysols mainly discriminated significantly from the remaining reference soil groups by having the highest moisture level beside the Stagnosols and Lixisols, which also displayed relatively high moisture status (Tab. IV-7). However, the Gleysols differentiated from the latter and from other reference soil groups with the lowest distance to stream network and lowest position in the landscape.

Stagnosols were characterized by the highest moisture level after the Gleysols, and by the highest distance to stream network with a lower protection index. The Lixisols revealed one of the highest moisture level after the Stagnosols, in lower elevation and protection index areas as the Gleysols, but with a higher distance to stream. The moisture distribution along with the distance to stream and elevation also clearly differentiated between the Cambisols and the remaining reference soil groups but particularly it singled out the former from the Leptosols, to which no significant difference was found regarding the protection index. The Leptosols were identified by their lowest soil moisture level as well as by their location at higher elevation and increased slope abundance (higher protection index) along with higher distance to the stream network. The Plinthosols discriminated from all the remaining reference soil groups by their moisture distribution along with the distance to stream for some (Cambisols, Gleysols, Stagnosols) and elevation for others (Leptosols and Lixisols).

The terrain parameters took preeminence over the spectral data considering the 90 % and 80 % core dataset (90%CR, 80%CR) associated with terrain plus spectral

parameters. The shortwave infrared taken in June (June_SWIR2) was listed only at the fifth position after the terrain attributes for the 80 % core range dataset (80%CR) while no spectral data appeared in the five top parameters for the 90 % core range dataset (90%CR). Overall, the contribution of the computed spectral indices was relatively low with soil color (June_CI) coming the fifth position when only spectral parameters were used with the standard deviation core range (SDCR), though the latter provided the highest Kappa value for this particular category of predictor. The results further revealed that the spectral data acquired in June were the most prominent ones for the classification of reference soil groups in the Dano catchment.

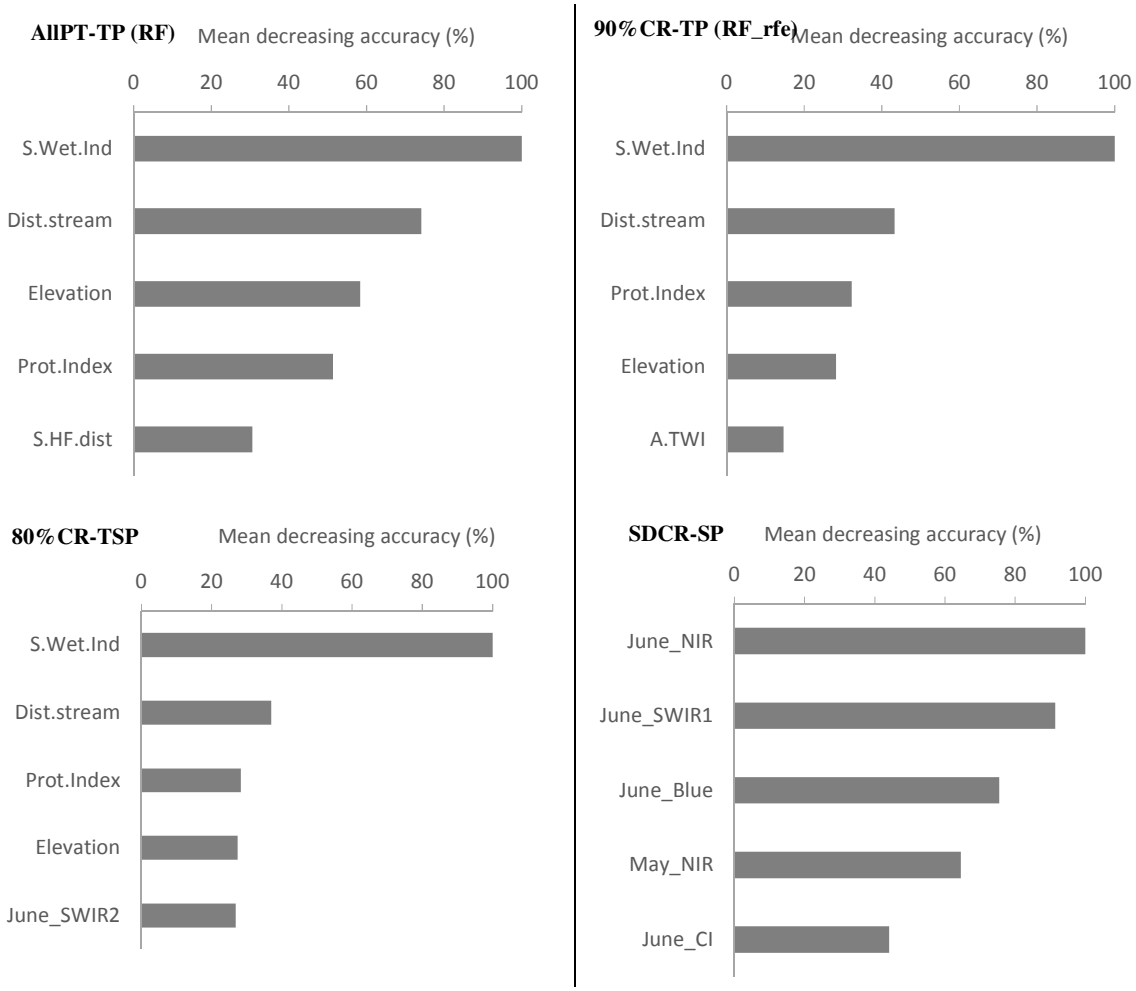

Fig. IV-5: Variable importance for the different data experiments (experiments defined in Tab. IV-1)

Models with (RF_rfe) and without (RF) recursive feature elimination; AllPT-TP: entire dataset including all Plinthosols & topographical parameters (TP), 90%CR-TSP: dataset with 5 % lower and upper range pruning & topographic and spectral parameters (TSP), 80%CR-TSP dataset with 10 % lower and upper range pruning & topographic and spectral parameters (TSP), SDCR-SP: dataset with standard deviation based pruning & spectral parameters, S.Wet.Index: Saga wetness index, Dist.stream: distance to streams, Prot.Index: protection index, S.HF.dist: horizontal flow distance, NIR: near infrared, SWIR: shortwave infrared, CI: coloration

Tab. IV-7: Kruskal–Wallis one-way analysis of variance of the main terrain parameters for the different reference soil groups based on the 90%CR dataset and topographic plus spectral (90%CR-TSP)

RSG (n)	Wetness Index		Distance to stream (m)		Elevation (m)		Protection Index	
	mean	sd	mean	sd	mean	sd	mean	sd
Cambisols (n=69)	7.82^{a}	(±0.68)	647^{a}	(±512)	313^{a}	(±21)	0.03^{a}	(±0.01)
Gleysols (n=113)	8.72^{b}	(±0.71)	242^{b}	(±199)	287^{b}	(±14)	0.02^{b}	(±0.01)
Leptosols (n=18)	6.03^{c}	(±1.29)	857^{c}	(±441)	372^{c}	(±35)	0.06^{ac}	(±0.03)
Lixisols (n=48)	8.26^{d}	(±0.97)	569^{ad}	(±307)	293^{bd}	(±24)	0.02^{bd}	(±0.01)
Plinthosols (n=467)	8.03^{ae}	(±0.4)	747^{cde}	(±515)	309^{e}	(±20)	0.02^{bde}	(±0.01)
Stagnosols (n=28)	8.46^{bdf}	(±0.68)	947^{cf}	(±482)	309^{aef}	(±22)	0.02^{bdef}	(±0.01)

RSG: reference soil group ; letters indicate whether the means are significantly different or not at p=0.05. Same letters stand for no significant difference.

3.5. Spatial distribution of the reference soil groups

The maps (Fig. IV-7) of the RF model based on the entire dataset (AllPT-TP) as well as the RF_rfe model from the standard deviation pruned dataset with spectral parameters (SDCR-SP) reveal an overestimation of the Plinthosols compared to field observation. However, using only spectral data with the entire dataset resulted in many small and isolated spots compared to the continuity and homogeneity of the remaining reference soil groups observed in the map from the AllPT associated with terrain parameters. With the pruned dataset (90%CR, 80%CR from RF_rfe) combined with

terrain and spectral parameters, the remaining soil groups came more into focus. This holds particularly true for the Lixisols and Stagnosols with the maps based on the 90 % and 80 % core dataset (90%CR, 80%CR from RF_rfe) associated with terrain plus spectral parameters.

The soils established on hard rock were classified as Leptosols by all models. Gleysols were predicted in the inland valleys while soils predicted as Cambisols were in general located in the Western part of the study area and mostly predicted in mid-slope regions. Lixisols were mapped in the lower elevation area and spots of Stagnosols were scattered all over the study area, especially in the southern and the eastern part. Plinthosols as the dominant soil group covered most of the landscape but were spatially restricted in the western area where Leptosols and Cambisols were more abundant.

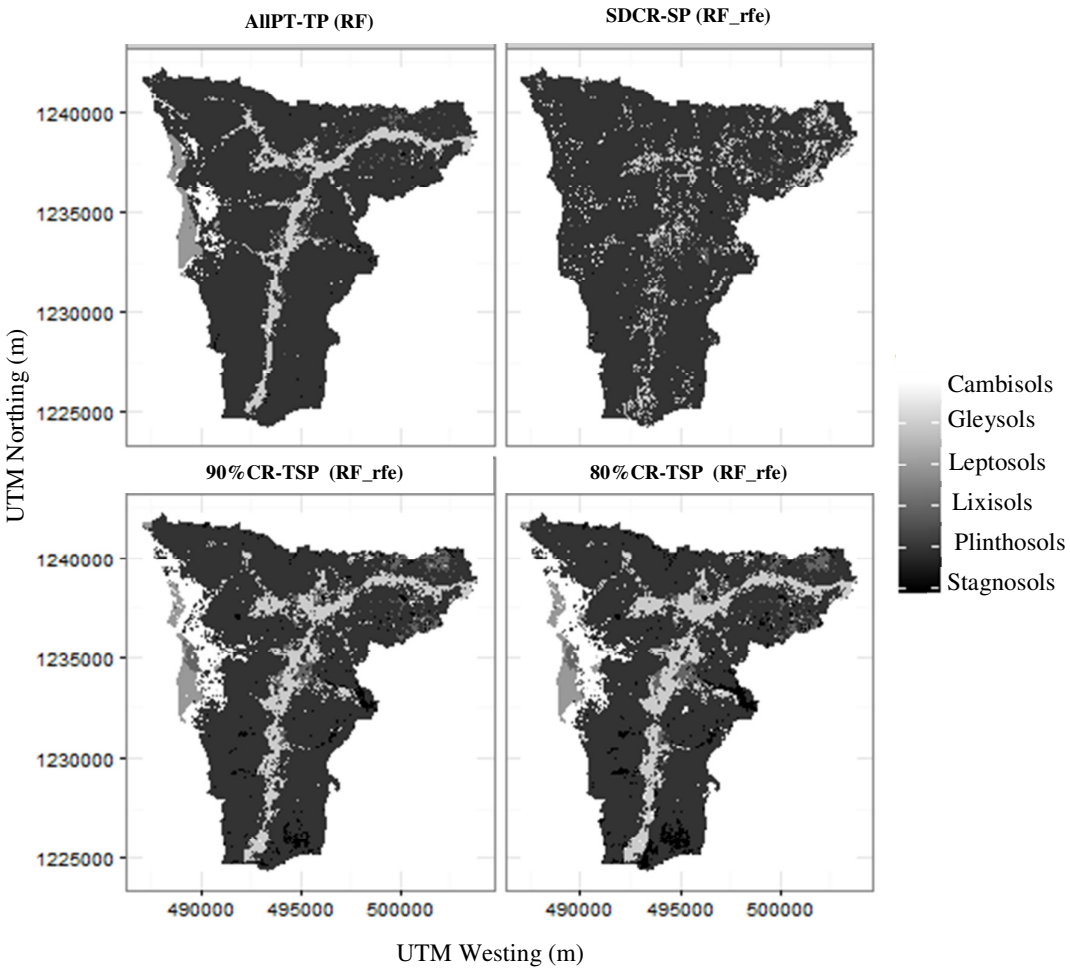

Fig. IV-6: Spatial distribution of the reference soil groups
Models with (RF_rfe) and without (RF) recursive feature elimination; AllPT-TP: entire dataset including all Plinthosols & topographical parameters (TP), 90%CR-TSP: dataset with 5 % lower and upper range pruning & topographic and spectral parameters (TSP), 80%CR-TSP dataset with 10 % lower and upper range pruning & topographic and spectral parameters (TSP), SDCR-SP: dataset with standard deviation based pruning & spectral parameters.

4. Discussion

4.1. Model Performance

The RF and RF_rfe models that were based on the entire dataset (AllPT) resulted in a relatively high OOB error compared with other datasets, with low prediction accuracy for the smaller reference soil groups and an overestimation of the abundance of Plinthosols (Tab. IV-4, IV-5 and Fig. IV-4). As expected, the Plinthosols exercised a stronger influence in the covariate space than other reference soil groups, which can be explained by the higher number of observations of this soil order. As a result, Plinthosols were overestimated while other soil classes were underestimated. When using the pruned dataset, the RF and RF_rfe models were most accurate when using either terrain parameters only, or a combination of the latter with spectral parameters (Fig. IV-4). The OOB errors were lower but revealed similar trends as those reported by Stum (2010) in western Utah, who found OOB errors of 58.9 % when using only DEM variables and 69.1 % when using the Landsat data only, while the combination of both DEM and Landsat data reduced OOB errors globally to 54.2 %. Brungard et al. (2015) reported an OOB error of 52 % when using both DEM and spectral data for reference soil group prediction. Differences in OOB values relative to our study are

due to the fact that OOB solely depends on the training set, which is site and data specific.

With increasing pruning intensity, improvement in prediction occurred for most reference soil groups especially with the random forest based on recursive feature elimination which performed slightly better than the normal run with all the predictors (Fig. IV-4). With the RF_rfe associated with terrain plus spectral parameters, a relatively higher prediction accuracy was observed for the smaller reference soil groups while using the 90 % core dataset. The latter also recorded the lowest OOB error along with the highest kappa value showing substantial agreement between predicted and observed reference soil groups. Consequently, removing all Plinthosol points lower than 5 % and higher than 95 % of the cumulative percentage of the most important variable (wetness index) resulted in slightly better data quality. Actually, the removed points were located in the low frequency range of the wetness index distribution.

Considering the frequency distribution of many predictors, Qi (2004) pointed out that samples from the modal range are more characteristic of a particular soil class than those belonging to the lowest frequencies, which are referred to as potential source of noise. As observed by Schmidt et al. (2008), such an approach is hardly applicable when dealing with many soil covariates since each predictor should be singled out in the analysis. However, focusing on the frequency distribution of the main driving predictor in the present study has proven to be satisfactory with the improvement in prediction accuracy observed with the pruned dataset in general and with the 90%CR dataset in particular. This 90%CR dataset includes the modal range of the wetness index with the outer range (5 %) being cut off. This suggests that the removal of samples beyond the modal range of a major soil reference group could result in an improvement in prediction accuracy, since they are rather a potential source of noise and redundancy due to overlapping information with small soil units.

With increasing pruning intensity, the prediction accuracy for the Plinthosols dropped suggesting that a loss of information for this particular reference soil group occurred with the pruned samples. This remains the main challenge in downsampling as

reported by Visa and Ralescu (2005) as well as by Yan et al. (2015), leaving out samples might result in dropping along some useful instances. The core point is to get a representative subset that is still large enough that losses are minimized but small enough to allow learning algorithms to get the relevant information for prediction (Schmidt et al., 2008). In the present study the results seem to be satisfactory with only 4.7 % drop in prediction accuracy for the Plinthosols by the 90%CR dataset with highest kappa value.

The Kappa values dropped (Fig. IV-5) with most models with prediction based on the standard deviation core range dataset (SDCR). This seems to suggest the SDCR as the pruning limit for the particular dataset of the present study, while revealing pruning between 5 %-10 % as the potential range for model improvement. In fact, pruning beyond the SDCR did not result in further improvements (data not shown). Overall, the kappa values recorded in the present study (0.42 – 0.57) with the terrain and terrain plus spectral parameters are higher than those recorded by Brungard et al. (2015) (< 0.4) who compared eleven machine learning models for predicting soil taxonomic classes in the semi-arid western US. However, as found out in the present study, the authors also point out that models with covariate predictors selected via recursive feature elimination result in higher prediction accuracy.

The different models did not provide a perfect prediction for the pruned Plinthosols which were in some cases classified mainly as Cambisols compared to the remaining smaller units (Tab. IV-6). This suggests that the performance of the different models on unlearned dataset outside their respective core range is limited. Obviously, discriminating the feature space among the reference soil groups for a high prediction accuracy of the Plinthosols was faced with inherent inability to relate from previous learning. Since the removed Plinthosols data were at the outer ends of the distribution of the most important variable (wetness index) it was expected that their prediction would result in high interferences in the feature space among the reference soil groups especially with those having the same values within that range. Since the pattern in these particular datasets was initially unlearned by the models for the Plinthosols, the prediction with the terrain and spectral parameters from the RF_rfe models was

considered as satisfactory. With about 71 % of the Plinthosols rightly predicted (Tab. IV-6), the model based on the 90%CR associated with terrain parameters could have been the best model if not for its low kappa ($\varkappa$=0.45) compared to the RF_rfe models with the terrain and spectral parameters (e.g. 90%CR & 80%CR) recording higher kappa values ($\varkappa$=0.55-0.57). Since the primary concern was the expression of smaller units while minimizing loss of predictive information of the Plinthosols, the results as obtained for the pruned core range dataset sample especially with the optimized predictors via recursive feature elimination pointed out the potential of data pruning to improve classification accuracy as shown by their kappa value. However, the point remains that any prediction of the Plinthosols based on unlearned dataset outside the core range will understandably come out with low to medium prediction accuracy.

Improving the model accuracy as recorded in the present study might require either increasing the number of soil pedon observations for the small classes (Brungard et al., 2015), or the assessment of additional soil features that ameliorate the discrimination between the different reference soil groups. Since a large array of predictors including spectral data were considered in the present study, any further improvement might have to consider different multi- or hyperscale terrain information to account for different spatial scales within one model (Behrens et al., 2010a; Behrens et al., 2010b; Behrens et al., 2014). The present work suggests that already pruning can reduce the overwhelming influence of some dominant reference soil groups, thus better allowing for expressing soil classes of lower occurrence.

4.2. Variable importance and spatial distribution

The terrain attributes drove the classification of the reference soil groups in the Dano catchment (Fig. IV-6). The feature selection algorithms always selected the SAGA wetness index (S.Wet.Ind) followed by the distance to stream network (Dist.stream), the protection index (degree of local surface convexity or concavity), and elevation among the most important terrain attributes. These results are in line with findings of Dobos et al. (2001), who reported an ascendency of terrain attributes such as slope, curvature and potential drainage density over spectral data in temperate climates.

Similarly, Stum (2010) ranked elevation and slope first followed by spectral data. The preeminence of the SAGA wetness index as soil development factor in the Dano catchment suggests that the humidity regime is a key discriminatory element among the reference soil groups. The protection index, distance to stream and elevation may be seen along this line as additional key regulatory parameters for soil moisture and related spatial distribution of the different reference soil groups.

Soils located at lower position and closer to streams, such as Gleysols and Lixisols (Fig. IV-6), had high moisture content than soils located at higher altitude and more far away from the streams, such as Leptosols and Cambisols. As already pointed out by Jenny (1994), soil moisture varies with local variations in topography: soils in depressions (toe-slope) like Gleysols are more humid than upland soils and soils in sloping areas. Also Adhikari et al. (2014) located Gleysols mainly in low slope position or flat areas. Lixisols have been mainly found in lower elevation areas, possibly as result of erosion processes. Gray et al. (2011) allocated Lixisols mainly in near level land or at undulating terrain.

Stagnosols have also high moisture level like the Gleysols, since both originate from water logging processes (IUSS et al., 2006). Stagnosols were generally allocated further away from the streams in relatively flat areas, where water is allowed to stagnate for some time in the year (IUSS et al., 2006). Stagnosols usually develop on a large variety of unconsolidated materials, either on flat or gently sloping areas (IUSS et al., 2006).

Leptosols were found at higher elevation and at larger distance to stream areas. These soils were well predicted by most of the models, since they were established on hard rock on the Ioba mountain , this fitting into the description of the WRB (IUSS et al., 2006). The spatial distribution of these Leptosols was consistent with the finding of Debella-Gilo et al. (2007), who found these soils mainly on hills and at the rocky part of the landscape. The presence of the major part of Cambisols next to the Leptosols might be attributed to erosion and deposition cycles, which are a key element for their distribution in high elevation areas (IUSS et al., 2006). Vasques et al. (2015) also found Cambisols in sloping areas, subject to a more dynamic water flow.

Plinthosols have been found nearly at every position of the landscape, thus occupying a major part of the land. These soils herein developed in level to gently sloping areas with changing groundwater level or stagnating surface water (IUSS et al., 2006). This corresponds to the feature of the study area characterized by a flat and undulating landscape with altitude ranging between 259 and 465 m asl and an average slope gradient of 3.6 % (Schmengler, 2010). Plinthosols are soils characterized by Fe accumulation under hydromorphic conditions. The change in moisture content (wetting and drying) results in the reallocation of dissolved Fe leading to the constitution of Fe poor and Fe rich zones in the soil (Lucas et al., 1992). In the rainy season, mobilization and translocation of Fe2+ ions occurs due to reducing conditions, while the dry season gives place to the oxidation of Fe2+ and precipitation of Fe oxides. As a result, Plinthosols are mainly hydromorphic soils (França et al., 2014), with their formation being greatly affected by soil moisture regime, as also evidenced by the Saga wetness index being the most important variable for the classification of the reference soil groups in the Dano catchment.

The NIR and SWIR spectral data were most prominent when acquired in June (Fig. IV-6) for the classification of reference soil groups in the Dano catchment. This particular period corresponded to the ploughing time. At that time crops were absent or at early stage of development, allowing satellite sensors to directly measure soil reflectance. Nield et al. (2007) reported that Fe rich minerals, which characterize many tropical soils such as Plinthosols, have a strong reflectance in the NIR and Lobell and Asner (2002) pointed out that soil moisture highly affects the NIR and SWIR reflectance. The preceding observations seem to imply that soil moisture and Fe oxide content as captured by soil reflectance provided the main discriminatory elements to differentiate between the different reference soil groups. Since the SWIR relates to soil moisture content as also the case for the Saga Wetness Index, it is obvious that mainly soil moisture controlled the distribution of reference soil groups over the Dano catchment. As the best predictions were found when the pruned data were used in combination with terrain and spectral parameters (TSP), these covariate predictors were assumed to be complementary, i.e., spectral data may only be used for soil taxonomy identification when combined with geomorphological information (see also

Dobos et al., 2001; Stum, 2010). Predicting reference soil groups for digital soil mapping thus heavily relies on concurrent soil-landscape characterization.

5. Conclusion

This study focused on reducing the negative influence of a predominant reference soil group – the Plinthosols – on the spatial prediction of more seldom reference soil groups in tropical environment, here the Dano catchment. For this purpose some ranges of the Plinthosol dataset were cut at different levels of pruning, and re-predicted the digital soil maps based on spectral indices, terrain, and terrain plus spectral parameters using RF modelling with and without recursive feature elimination. When using the entire dataset, lower prediction accuracy was obtained for most of the reference soil groups predicted as Plinthosols. However, increasing pruning intensity resulted in relatively lower OOB errors with subsequent improvement in classification accuracy.

The best prediction was achieved when removing all Plinthosol points lower than 5 % and higher than 95 % of the cumulative percentage of the most important variable (wetness index) and RF modelling conducted solely with terrain and spectral parameters (TSP) with optimal predictors resulting from the recursive feature elimination. This improved classification accuracy by 3 % to 41 % relative to the prediction based on the entire dataset as the pruned samples, potential source of noise and redundant information, were removed. Though terrain parameters proved to be most determinant in the characterization of the landscape for discriminating between the different reference soil groups their combination with spectral bands and indices resulted in better prediction. For this tropical environment, the moisture distribution (SAGA wetness index) was finally identified as the main driving factor for the reference soil group classification in the Dano catchment.

With the ongoing GlobalSoilMap.net initiative in Africa, soil mappings are being carried out using legacy data with some subject to imbalance issues. The pruning as demonstrated in this study can help to improve dataset quality and therewith classification accuracy. This could thus particularly be chosen as suitable alternative when new dense surveys are no viable option for creating soil maps.

IV. Predicting reference soil groups using legacy data: a data pruning and random forest approach

V.

Spatial controls of soil organic carbon stocks in the Sudanian savannah zone of Burkina Faso, West Africa

Modified on the basis of

Ozias K. L. Hounkpatin, Felix Op de Hipt, Aymar Y. Bossa, Gerhard Welp, Wulf Amelung (2017). CATENA.

Submitted manuscript

1. Introduction

Globally, soils contain the largest terrestrial carbon pool on earth. Though subject to regular change, the global amount of carbon in soils is estimated at 2500 Gt, including 1550 Gt of soil organic carbon (SOC) and 950 Gt of soil inorganic carbon (Batjes and Sombroek, 1997; Lal, 2008). As the SOC pool is 3.3 times the size of the atmospheric pool (760 Gt) and 4.5 times the size of the biotic pool (560 Gt) (Lal, 2004), slight changes in soil C cycling may significantly impact the global C cycle. Nevertheless, little is known on the role of tropical soils for these changes, especially not for tropical subsoils.

The ecosystems in West Africa are facing severe degradations due to change in land use from perennial vegetation to cropping, increased cultivation in marginal lands, soil erosion and nutrient mining (Bationo et al., 2007; UNEP, 2006), as well as climate change (Brevik, 2013). Models predicted that as consequence of climate change, soils will convert from carbon sinks to carbon sources (Cox et al., 2000), but prediction uncertainty is large (Cox et al., 2000; Smith, 2008), mainly due to the lack of adequate knowledge on SOC distribution across the landscape. Nowadays, different measures to conserve existing SOC stocks and trap the atmospheric carbon in the soil are being implemented in many areas in Africa and comprise afforestation of degraded lands, agroforestry, application of best agricultural practices and policies (Batjes, 2008). However, data are still lacking on SOC for different agrosystems (Anikwe, 2010) in most African countries. Batjes (2008) even pointed out that an estimation of the current carbon stock should be carried out prior to any focus on carbon change related to land use and climate change.

The variability of carbon stocks in the landscape is associated with the combined action of physical, chemical and biological processes as well as of human land use patterns varying over space and time (Peukert et al., 2012). Generally, this spatial variability is recorded by soil maps, which are key tools for effective land management and modelling. Progress and new development in computer science and statistical methods led to the use of geo-information technology such as remote sensing data and

digital elevation model (DEM) for the digital soil mapping (DSM) of soil properties (Heuvelink and Webster, 2001). The DSM correlates quantitatively environmental covariates standing for soil forming factors and a target variable to be predicted. This correlation is carried out using statistical methods, which build a model used for prediction. The multiple linear regression has been widely used in many studies as a predictive model for the prediction of SOC (Florinsky et al., 2002; Guo et al., 2015; Meersmans et al., 2008). However, soil-landscape relationships are often subject to nonlinear dynamics which might not be captured by MLR (Grimm et al., 2008). Random Forest regression (RF), an ensemble machine learning approach, is reported in literature as being able to overcome this limitation (Hengl et al., 2015; Rad et al., 2014; Wiesmeier et al., 2011). The latter studies indicated the robustness of RF for handling complex and non-linear soil-landscape relationships in DSM.

Potential factors which affect SOC stocks and are used as covariates for DSM, comprise climatic and topographic elements (e.g., mean annual precipitation and temperature, slope etc.), land use, physical soil characteristics (texture, parent material, etc.), and microbial biomass (Albaladejo et al., 2013; Jobbágy and Jackson; Jobbágy and Jackson, 2000; Ladd et al., 2013). Many of these factors have been investigated in various publications across the globe (Albaladejo et al., 2013; Azlan et al., 2011; Bationo et al., 2007; Burke et al., 1989; Chaplot et al., 2010; Jobbágy and Jackson, 2000; Percival et al., 2000). However, these studies mostly focused on surface soil horizons. Yet, more than 50 % of SOC is usually allocated below 20 cm depth (Batjes, 1996). Fontaine et al. (2007) showed that this subsoil carbon is readily decomposable upon addition of a fresh C source, and Fierer et al. (2003) concluded that it is even more sensitive to changes in temperature or nutrient availability than topsoil carbon. But these latter studies have not been performed with tropical soils, which may have specific SOC storage conditions, e.g., due to their special oxide assembly (Feller and Beare, 1997; Kögel-Knabner and Amelung, 2014).

This study was performed in the Sudanian area of Burkina Faso dominated by Plinthosols, i.e., soils with high Fe oxide accrual, particularly in the subsoil. We are not aware that for such soils, nor then for the respective or comparative region, (i) levels and distribution of SOC stocks along with the (ii) interactions between SOC

stock and landscape properties have ever been investigated. Yet, these quantitative data are crucial for the estimation of the local and regional carbon sequestration potential and the participation of developing countries in the Clean Development Mechanism (CDM), mentioned in the Kyoto Protocol as well as the "4 per thousand" initiative launched during the COP21 (Rhodes, 2016). Therefore, this study aimed at estimating the surface and subsoil organic carbon stocks in different land use systems and across various soil orders, as well as assessing the spatial variability of topsoil carbon stocks and underlying factors.

2. Materials and methods

2.1. Study area (see section II. 1)

2.2. Soil Sampling (see section II. 2)

2.3. Soil analysis and mid-infrared prediction (see section II. 3)

2.4. Determination of SOC stocks (see section II. 4)

2.5. Selected variables for explaining SOC stock variability

The variables (Tab. V-1) considered as covariates consist of: terrain attributes, land use, temperature and precipitation, geomorphology and lithology. The terrain attributes were derived from a 90 meter resolution digital elevation model provided by the Shuttle Radar Topography Mission (SRTM). These parameters are clustered into local, regional and combined terrain attributes as defined by Grimm et al. (2008). The parent material (Geo) allocated to each sampling location was derived using a geological map (1/1 000 0000) of Burkina Faso made by Hottin and Ouedraogo (1992). Land use data were collected during the sampling at each location. Climatic data include mean annual temperature (Temp) and annual precipitation (Prep) at 1 km resolution from the worldclim datasets. The climatic data were submitted to bicubic resampling before the extraction of the data.

Moreover, soil properties were also considered as covariates as mentioned in Kumar and Lal (2011) and Were et al. (2015). Soil texture fractions (sand, silt, clay) were considered in addition to the environmental variables. They were derived from interpolated maps using the Ordinary Kriging method. The Ordinary Kriging has been

used in many studies for predicting soil properties at unsampled locations (Zhang and McGrath 2004; Mishra et al. 2009; Chaplot et al. 2010; Were et al. 2015).

The predictors were reduced for the subsoil carbon stock model due to the smaller size of the dataset (n = 70). Feature selection was carried out using the RF recursive feature elimination algorithm of R "caret" Package (Kuhn 2015). The following variables were finally retained for the subsoil carbon stock prediction: elevation, distance to stream, aspect, ruggedness, curvature, catchment area, sand, silt, clay, precipitation and temperature.

Tab. V-1: Selected variables for explaining SOC stocks variability

Group	Parameters	Definition	Abbreviation	Units
Local	Slope	Inclination of the land surface from the horizontal	Slope.per	%
	Slope Length	Distance from origin of overland flow to deposition point	Slope.length	m
	Curvature	Combination of horizontal and vertical curvature	A.curv	m^{-1}
	Maximum Curvature	Maximum Curvature	S.max.cuv	$° m^{-1}$
	Minimum Curvature	Minimum Curvature	S.min.cuv	$° m^{-1}$
	Plan Curvature	Horizontal (contour) curvature	S.Plan.cur	$° m^{-1}$
	Profile Curvature	Vertical rate of change of slope	S.Profile.cur	$° m^{-1}$
	Aspect	Direction the slope faces	A.Asp	°
	Elevation	Vertical distance above sea level	Elevation	m
Regional	Catchment Area	Discharge contributing upslope area	S.CA	m^2
	Distance to stream	Distance to stream network	Dist.stream	m
Combined	Topographic Wetness Index	Ratio of local catchment area to slope	A.TWI	-
	Saga Wetness Index	Ratio of local catchment area to slope	S.Wet.Ind	-
Climatic	Temperature	Temperature	Temp	°C
	Precipitation	Precipitation	Prep	mm
Soil properties	Sand	Sand	Sand	%
	Silt	Silt	Silt	%
	Clay	Clay	Clay	%
Others	Lithology	Lithology	Litho	-
	Geormorphology	Geormorphology	Geo	-
	Land use	Land use	LU	-
	Reference soil group	Reference soil group	rsg	-

2.6. Statistical analysis

Descriptive statistics (means and standard deviation of the mean) were used to characterize the measured values of the variables. Normality of the carbon data was

checked with the Shapiro-Wilk test. The student t test was used for comparison between the SOC stocks of the different land use systems. The Bartlett test for homogeneity of variance was performed due to the unequal size of the data for the soil reference groups (Yu 2011). The significance of the difference in the mean SOC stocks between the reference soil groups was examined by using the Welch ANOVA test, while for multiple means comparisons, the Games-Howell test was performed as carried out in Cornelissen et al. (2001).

2.7. Predictions models

In the present study, MLR and RFR were used as statistical models to predict the spatial distribution of the topsoil SOC stock. MLR is a classical statistical approach to predict the values of a dependent variable (here the SOC stocks) based on a set of independent variables (here the covariates in Table 2). In this study, MLR and MLR were implemented using the R "caret" package (Kuhn, 2015) using tenfold cross validation with 5 repetitions.

For background information on RFR see section II-7.

2.8. Model training and mapping

The topsoil (n = 1239) dataset was split with 70 % of the samples to train the model while 30 % were used as independent validation set. For the subsoil dataset, a split of 80 % was applied. The models derived from the RF for each depth were used to make the respective prediction maps which were corroborated by different validation sets. For the stability and robustness of the models, the different calibrations were carried out based on a 5 time repeated 10-fold cross-validation using the "caret" R Package (Kuhn 2015). The root mean square error (RMSE) of cross validation (RMSECV) as well as RMSE from prediction based the validation set (RSMEPV) were used to assess the model accuracies.

$$RMSE = \left(\frac{1}{n}\sum_{i=1}^{n}(P_i - O_i)^2\right)^{1/2} \qquad \text{(V-1)}$$

where "P" is the predicted value and "O" is the observed/measured value

3. Results and discussion

3.1. Basic soil characteristics

The general soil properties of the different soil profiles for both topsoil (0 - 30 cm) and subsoil (30 - 100 cm) are presented in Tab. V-2. Textural variations occurred among the different soil groups: the Gleysols (GL) were silty and less sandy than the Plinthosols (PT), which peaked in opposite direction. Possibly the latter was caused by pseudo-sand like oxide concretions in the latter, which could not be destroyed completely during conventional texture analyses. The bulk density increased with depth with larger values recorded in the subsoil for both land use systems. Maximum bulk densities were found for the Plinthosol subsoils, which indicated the presence of petroplinthite in some of these profiles. The pH was slightly acidic and comparably similar among land use and reference soil groups at all soil depths. This trend is in line with values reported by Yoni et al. (2005) in Western Burkina Faso.

Tab. V-2: Basic soil characteristics under different land use (mean values with standard deviation (sd))

		N	Sand (%)	Silt (%)	Clay (%)	BD (g cm^{-3})	pH
				0 - 30 cm			
LU	CR	36	$28.1^{a}(\pm 9.1)$	$43.2^{a}(\pm 7.1)$	$28.5^{a}(\pm 10.1)$	$1.4^{a}(\pm 0.1)$	$6.4^{a}(\pm 0.5)$
	SA	34	$29.9^{a}(\pm 12.3)$	$44.8^{a}(\pm 10.5)$	$25.9^{a}(\pm 9.5)$	$1.5^{a}(\pm 0.1)$	$6^{a}(\pm 0.4)$
RSG	CM	8	$25.5^{ac}(\pm 11.3)$	$42.2^{ac}(\pm 6.7)$	$32.2^{a}(\pm 13.6)$	$1.3^{ab}(\pm 0.1)$	$7^{a}(\pm 0.4)$
	GL	12	$19.1^{ba}(\pm 11.3)$	$50.3^{ba}(\pm 9.7)$	$31.7^{a}(\pm 11)$	$1.4^{ac}(\pm 0.1)$	$6.1^{bc}(\pm 0.3)$
	LX	2	$22.6^{ac}(\pm 3.5)$	$55.3^{ac}(\pm 8.6)$	$20.5^{a}(\pm 2.9)$	$1.4^{ac}(\pm 0.001)$	$6.2^{abc}(\pm 0.4)$
	PT	44	$32.8^{c}(\pm 8.9)$	$42^{c}(\pm 8.2)$	$25.2^{a}(\pm 8.4)$	$1.5^{c}(\pm 0.1)$	$6.1^{c}(\pm 0.4)$
	ST	4	$29^{ac}(\pm 10.7)$	$43.9^{ac}(\pm 8.9)$	$27.3^{a}(\pm 9.8)$	$1.4^{ac}(\pm 0.1)$	$6.5^{abc}(\pm 0.4)$
				30 - 100 cm			
LU	CR	36	$21.6^{a}(\pm 6.9)$	$40.7^{a}(\pm 4.8)$	$37.2^{a}(\pm 7.9)$	$2^{a}(\pm 0.7)$	$6.3^{a}(\pm 0.5)$
	SA	34	$22.8^{a}(\pm 5.3)$	$41.8^{a}(\pm 6.2)$	$34.9^{a}(\pm 4.5)$	$2.1^{a}(\pm 0.7)$	$6.1^{a}(\pm 0.4)$
RSG	CM	8	$26.4^{a}(\pm 9.1)$	$39.5^{a}(\pm 2.7)$	$33.7^{a}(\pm 9.9)$	$1.7^{a}(\pm 0.6)$	$6.9^{a}(\pm 0.7)$
	GL	12	$19.7^{a}(\pm 7.5)$	$45.3^{a}(\pm 7.9)$	$34.5^{a}(\pm 5.8)$	$1.6^{a}(\pm 0.1)$	$6.1^{bc}(\pm 0.3)$
	LX	2	$17.9^{a}(\pm 6.1)$	$46^{a}(\pm 6.7)$	$34.4^{a}(\pm 2.3)$	$1.5^{a}(\pm 0.1)$	$6.1^{abc}(\pm 0.2)$
	PT	44	$22.2^{a}(\pm 4.6)$	$40.2^{a}(\pm 4.3)$	$37.1^{a}(\pm 6.3)$	$2.3^{a}(\pm 0.7)$	$6.1^{c}(\pm 0.3)$
	ST	4	$22.9^{a}(\pm 8.5)$	$41.3^{a}(\pm 8.3)$	$35.1^{a}(\pm 4.1)$	$1.8^{a}(\pm 0.8)$	$6.7^{abc}(\pm 0.7)$

LU: land use, CR: cropland, SA: savannah, RSG: Reference soil groups, CM: Cambisols, GL: Gleysols, LX: Lixisols, PT: Plinthosols, ST: Stagnosols, n: number of samples, BD: bulk density. Means followed by the same letters are not significantly different ($p < 0.05$).

3.2. SOC stock in relation to land use and reference soil group

The distribution of the SOC stocks in the different land use systems as well as in the RSG of each specific land use is presented in Tab. V-3. About 73.5 t C ha^{-1} was recorded as the total average of SOC stock in an entire profile (0 - 100 cm) in the Dano catchment with 39 t C ha^{-1} found for the topsoil (0 - 30 cm) and 33.9 t C ha^{-1} for the subsoil (30 - 100 cm), amounting respectively to 53 % and 47 % of the total stock. These results coincide with the findings reported by other authors with Batjes (1996) recording 39 - 70 % of the SOC stock in the first 30 cm while Doetterl et al. (2015) reported about 52 % of SOC stock at the same depth. The total average of SOC stock over 100 cm recorded in the present study is higher than the range estimations of 42 – 45 t C ha^{-1} for West Africa and 64 - 67 t C ha^{-1} reported for Africa (Batjes, 2001); on the other hand, our average value is lower compared to the 82 t C ha^{-1} found by Hien et al. (2003) for the southern Burkina Faso.

In the topsoil, the SOC stock was similar for both land-use systems. The average SOC stocks of the non-cropped sites only slightly exceeded that of the croplands (2.3 t C ha^{-1}; not significant). The lacking significance was due to the Cambisols, which showed significantly larger SOC stocks in the surface soils of the croplands, likely due to former land-degradation or just site preference of the farmers for the better Cambisols. The larger SOC stocks in the surface soils for the other sites under natural vegetation is in line with other studies (Bruun et al., 2013; Singh et al., 2011). A study in Ghana by Boakye-Danquah et al. (2014) reported 22.9 t C ha^{-1} for the topsoil of cultivated area and 49.4 t C ha^{-1} for natural vegetation while Hien et al. (2006) in Burkina Faso recorded between 16 t C ha^{-1} and 25 t C ha^{-1} for cropland soil and 61 t C ha^{-1} for savanna soils. Though the results for the topsoil are in the range of the previous studies carried out in the same region, the margin between the values reported for the two LU systems is quite narrow.

The small difference of SOC stocks between these two land use systems in the Dano catchment suggest a high level of degradation of the sites under savannah, which is subject to overgrazing due to the absence of sufficient grazing areas and the inexistence of straw and silage production (Callo-Concha et al., 2012b). The pressure

on these non-cropped fields is worsened by the presence of migratory herding that add to the local livestock (Gonin and Tallet, 2012). Moreover, the production of a local beer ("dolo") results in the use of about 6400 t of fire wood per year from the native savannah sites; this constitutes also a major source for the degradation of natural resources (Blin and Sidibe, 2012). The sites under savannah may also include old fallow soils, which because of current herding pressure, failed to re-build their carbon stock. Once degraded, it may take decades until SOC stocks in such savannah soils restore (Preger et al., 2010).

One additional peculiarity was the presence of stone lines (Appendix B Fig. X-1) in the croplands, which may have also reduced soil erosion as observed by Schmengler (2010) in the same area. Zougmoré et al. (2004) reported a reduction of runoff by 45 % with the use of stone lines as conservation practice. Therefore, the presence of these stone lines might have contributed to the slowing down of the SOC loss from the cropland.

Intriguingly, significant different C stocks were found for the subsoils that contained more SOC in the cropland than in the savannah sites (Table 4). This SOC storage overcompensated SOC gains in the surface soils, so that significance disappeared on a whole soil profile basis. In part, the larger SOC stocks under cropland may be attributed to the presence of petroplinthite in the subsoil of the savanna soils that were not thus not used for cropping nowadays. In addition, intensive translocation processes in the croplands may have been induced at elevated precipitation events under tropical climate, as formerly reported for the leaching of basic cations into the subsoil (Eze et al., 2014) along with clay and SOC, especially for low acidity soils that also prevailed in our study (Lorenz and Lal, 2005).

Tab. V-3: Soil organic carbon stock in different land use systems and reference soil groups at different depth

LU	RSG	n	0 - 30 cm		30 - 100 cm		0 - 100 cm	
			mean	sd	mean	Sd	mean	sd
CR & SA		70	39	±16.7	33.9	±23.8	73.5	±30.7
CR (t C ha^{-1})		37	39.1^{a}	±16.5	40.2^{a}	±27.9	77.1^{a}	±34.9
SA (t C ha^{-1})		33	41.4^{a}	±17.4	26.3^{b}	±15.9	67.7^{a}	±27.3
CR (t C ha^{-1})	CM	6	40.2^{a}	±12.6	48.7^{a}	30.7	88.9^{a}	±40.5
SA (t C ha^{-1})		2	16.6^{b}	±8.3	20.6^{a}	16.0	37.2^{b}	±7.6
CR (t C ha^{-1})	GL	5	39.9^{a}	±12.2	52.7^{a}	±32	94.4^{a}	±35.4
SA (t C ha^{-1})		7	46.6^{a}	±18.9	35.6^{a}	±15.1	82.5^{a}	±31.2
CR (t C ha^{-1})	LX	1	27.6	.	26.0	.	53.6	.
SA (t C ha^{-1})		1	37.6	.	21.9	.	59.5	.
CR (t C ha^{-1})	PT	22	39.8^{a}	±15	33.7^{a}	±24.5	73.2^{a}	±32.4
SA (t C ha^{-1})		22	42.4^{a}	±16.9	24.6^{a}	±16.3	67.0^{b}	±25.9
CR (t C ha^{-1})	ST	3	9 .0^{a}	±5	54.6^{a}	±42.7	63.6	±46.8
SA (t C ha^{-1})		1	36.7^{b}	.	17.2^{a}	.	54.0	.
CR (t C ha^{-1})	CM	6	40.2^{a}	±12.6	48.7^{a}	±30.7	88.9^{a}	±40.5
	GL	5	40.0^{a}	±12.2	52.7^{a}	±32	92.7^{a}	±38.3
	PT	22	39.8^{a}	±15	33.7^{a}	±24.5	73.2^{a}	±32.4
	ST	3	9.0^{b}	±5	54.6^{a}	±42.7	63.6^{a}	±46.8
SA (t C ha^{-1})	CM	2	16.6^{a}	±8.3	20.6^{a}	±16.5	36.6^{a}	±8.1
	GL	7	46.6^{a}	±18.9	35.6^{a}	±15.1	82.2^{a}	±31.4
	PT	22	42.3^{a}	±16.9	24.6^{a}	±16.3	67.0^{a}	±25.9

LU: land use, RSG: reference soil group, CR: cropland, SA: savannah, n: number of samples. Means followed by the same letters are not significantly different ($p < 0.05$).

Considering the different reference soil groups in the topsoil, the Plinthosols (41.1 t C ha^{-1}) contained more or less as much SOC as the Gleysols (43.8 t C ha^{-1}). The latter also recorded the largest carbon stock over 100 cm depth (86.6 t C ha^{-1}) followed by the Cambisols (75.8 t C ha^{-1}) and the Plinthosols (70.1 t C ha^{-1}) (Appendix B Fig. X-2). The prevalence of SOC in Gleysols might not solely due to limited SOC decomposition under groundwater influence, but could mainly be related to the occurrence of local erosion processes, leading to the transport of SOC rich sediments

from upslope to the lower slope, and thus from other soils into the Gleysols under the combined effect of slope, elevation and heavy tropical rain. Doetterl et al. (2013) reported a significant difference in SOC stocks between erosional and depositional areas due to soil relocation processes and local topographical features. However, with similar SOC stocks in the topsoil between Gleysols, Plinthosols, and Cambisols depositional areas might not correspond only to Gleysols due to the variability of topographic feature across the landscape. On the other hand, the periodic saturation by groundwater reduces oxidation processes in the subsoil.

The Stagnosols of the cropland exhibited the lowest SOC stocks (9 t C ha^{-1}, Tab. V-3). As temporary saturation with water in the stagnosols should normally promote SOC storage rather than distorting it, we attribute this finding firstly to their position at a relatively high position in the landscape favouring vulnerability to soil erosion and secondly to stagnic conditions occurring at a relatively deeper depth regarding the high carbon stock in the subsoil (t C ha^{-1}). Moreover, exposition to a longer cultivation duration with very low input (Bationo and Buerkert, 2001) could also be responsible for the low carbon level of the topsoil but investigation into the land use history is necessary before any sound conclusion. The Stagnosols, exhibiting larger SOC stocks in the subsoil of the croplands, could be taken as additional evidence that for mapping soil C storage the consideration of whole soil profiles is needed.

3.3. Factors affecting the spatial variability of SOC stock

The analysis of variable importance characterizes the influences that different explanatory variables (see Tab. V-1) have on the response variable (here SOC stock). The analysis revealed different preeminent parameters controlling SOC stocks of topsoil (Fig. V-2). Only the top 5 variables are considered in the figure.

The most prominent redictor for the topsoil SOC stock was the silt and sand content followed by the wetness index, elevation and climate variables. Soil texture in general and especially its fine particles (silt and clay) are extensively discussed in literature as important agents accounting for the variance of SOC through adsorption of organic matter (Bationo et al., 2007; Chaplot et al., 2010; Mao et al., 2015; Saiz et al., 2012;

Zhang and Shao, 2014). As recorded in Table 2, the high content of silt in the topsoil makes it the most abundant soil particle involved in potential adsorption. The correlation of wetness index (indicator of soil moisture) and SOC content has been indicated by Kumar (2009) and Zadorova et al. (2014). As hydrological factor, the wetness index affects SOC dynamics at depositional and flat areas where humidity is high resulting in slower decomposition rate (Doetterl et al., 2013). The record of elevation among the prominent variables is in line with findings of Hengl et al. (2015) who also reported it as a major factor affecting SOC stocks in Africa.

Climate variables are widely acknowleged as influential variable for SOC stocks (Doetterl et al., 2013; Manning et al., 2015; Oueslati et al., 2013). Temperature and precipitation distribution affect the production of plant materials and soil fauna activity. Warmer temperatures and wetter conditions would most likely result in higher biomass production and microbial activity. Conversely, a lower heat transfer coupled with lower humidity could result in reduced C decomposition. The dry season of the study area is characterized by higher temperatures with very scarce rainfall which might result in a decrease of bioamass while the rainy season comes with intense and heavy rainfall with subsequent vegetation growth and production of plant material. Though the individual impact of these factors could be explained isolately, it is most likely that due to soil landscape interaction, the amount of carbon stock at a given location is a resultant of their interaction. Precipitation and temperature affect the soil moisture (wetness index) distribution which is in turn infuenced by elevation and soil texture. For example, a higher SOC stock was observed in the topsoil of the Gleysols which were characterized by high moisture and silt content (Tab. V-2) and were located at lower elevation areas.

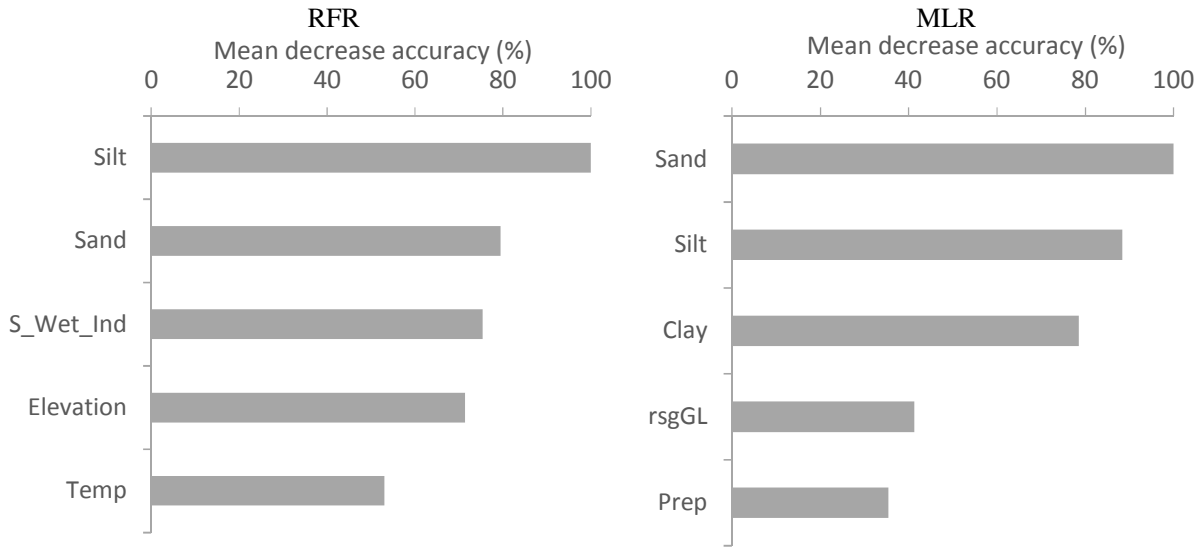

Fig. V-1: Top five variables from the RFR and MLR models for the topsoil (0 – 30 cm)
Wet_ind: wetness index, rsgGL: Gleysols, Temp: Temperature, Prep: Precipitation, RFR: random forest regression, MLR: multiple linear regression

3.4. The spatial distribution of the SOC stock

The spatial distribution pattern of SOC stock in the topsoil (Fig. V-3 A) based on the prediction of RFR and MLR model presents an irregular pattern. There were innumerable patches of small and large SOC stocks across the study area, pointing to a pronounced variability of the SOC stock over small distances though less pronounced on the MLR map. On large scales, elevated SOC stocks in topsoil were observed in the western and south-eastern areas. These areas correspond to the high elevation part of the watershed (Figure 1), with SOC stocks varying between 55 - 65 t C ha^{-1}. The remaining areas displayed low (28 – 40 t C ha^{-1}) to medium (40 – 55 t C ha^{-1}) SOC stocks. Though land use did not come up as key variable for SOC stocks in topsoil, it had an indirect link with elevation, being one of the major influencing factor (Figure 1). In our study area, the density of settlements and adjacent intensively cultivated fields was higher in the lower elevation areas due to the proximity of streams, which provide water for domestic purposes and for the irrigation of crops. Consequently, larger SOC stocks were found in the surface soils that belonged to areas in more

remote and elevated parts of the watershed, which thus exhibited less cultivation intensity and larger areas covered by natural vegetation.

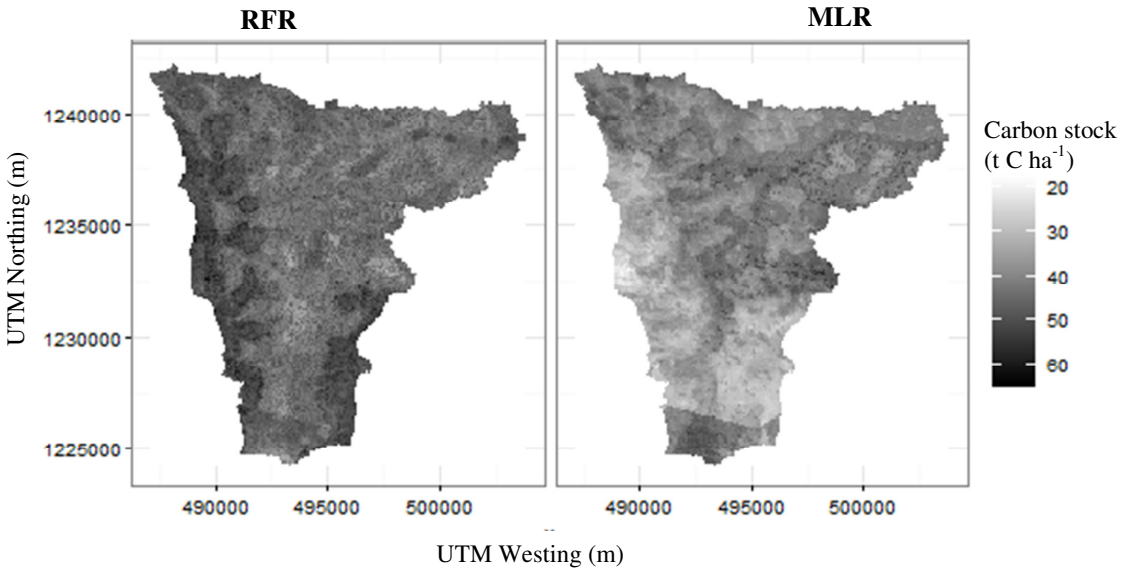

Fig. V-2: Distribution of SOC stock across in the topsoil (0 – 30 cm) based on the RFR and MLR Models. RFR: random forest regression, MLR: multiple linear regression.

3.5. Performance of the RF models

In general, the accuracy of the RFR and MLR prediction models were low (Tab. V-4), though the former performed marginally better than the latter with higher R^2 (13 %) and lower root mean square errors for both cross validation (14.0 t C ha^{-1}) and independent validation (14.2 t C ha^{-1}). This can be attributed to nonlinear pattern in the SOC stock dataset which could not be accounted for by the MLR. Other studies also point out the limitation of MLR to handle nonlinear pattern in dataset hence its lower performance compared to machine learning models such as Random Forest (Hengl et al., 2015; Zhang et al., 2017) . The explained variance as found in the present study could not be improved even when some RSG were removed from the dataset and modelling carried out with the remaining (Appendix B Tab. X-1) though the removal of Plinthosols led to an explained variance of 17 % with both models.

The results of this study regarding the model explained variances are consistent with some existing findings in literature. Grimm et al. (2008) found only 6 % as explained

variance for topsoil and 8 - 25 % for subsoil SOC content after using the Random Forest approach in a tropical island in Panama. Henderson et al. (2005) used a decision tree approach and reported an explained variance of 41 % for topsoil SOC and 24 % for the subsoil. Wiesmeier et al. (2014) analyzed the spatial distribution of SOC stocks and found 52 % of explained variance for the carbon stock based on climate, land use and environmental variables. Schulp and Verburg (2009) and Schulp et al. (2013) reported 21 % to 43 % variance explained for SOC contents and stocks though a wide range of data from soil properties to terrain attributes were used. These authors pointed out that low explained variance for SOC prediction was recorded due to an intrinsic large spatial variability of SOC with the interplay of a large range of factors at local and regional level.

The low explained variance observed in the present study could be attributed to the existence of other environmental and soil parameters affecting SOC stock variability, which have not have been investigated in this study. Such parameters may account for specific soil properties, such as soil structural stability, clay mineralogy, sesquioxide composition, as well as other factors beyond the scope of our design, such as socio-ecological impacts in soil resilience (e.g. Linstädter et al. (2016)). In addition, the root mean square errors obtained in this study is a reflection of errors related to field sampling, laboratory measurement, and statistics as well as random errors. Since all of the soil properties used in the present study were interpolated by ordinary kriging it is evident that related errors translated into the estimation of SOC stock. However, preliminary modelling without these soil properties revealed much lower variances (data not shown) proving them as key variables to be taken into account. Auxiliary data coming from different sources and different scales infer variability in data quality as also pointed out by Were et al. (2015). For example, the resampled lithology file was originally produced at a scale of one-million and as result its distribution on the study area might have been too coarse. Further model improvement would require additional explanatory variables at finer scale with the consideration of multi- or hyper-scale data in order to account for the possibility of SOC stock being subject to factors operating at different levels of scale (Behrens et al., 2010a; Behrens et al., 2010b).

The statistics of the prediction (Tab. V-4) that was based on the validation set showed that the root mean square error of cross validation as well as the root mean square error of prediction (from validation set) for the topsoil from both models were all slightly lower than the standard deviation of the measured values. This points out that the predictions of the models especially from the RFR were as accurate as the training set in spite of the low explained variance. A similar trend had been also recorded by Were et al. (2015).

Tab. V-4: Performance statistics of the RFR and MLR models and general statistics for measured data and SOC stocks of the maps

	R^{2*}	RMSECV	RMSEPV
Statistics for model and validation dataset			
RFR (t C ha^{-1})	13.0	14.0	14.2
MLR (t C ha^{-1})	11.0	14.2	14.8
General statistics for predicted map and measured data			
	Min	Max	Mean (±sd)
RFR predicted data (t C ha^{-1})	27.4	65.1	45.4 (±4.6)
MLR predicted data (t C ha^{-1})	3.0	98.8	44.7 (±6.7)
Measured data (t C ha^{-1})	11.3	79.2	45.5 (±14.9)

RF: random forest, Var_{exp}: explained variance, ME: mean error, RMSECV: root mean square error of cross validation, RMSEP: root mean square error of prediction based on validation set, *explained variance in %.

The general statistics for the measured and predicted SOC stocks for the topsoil maps (Tab. V-4) revealed that the predicted minimum value for the RFR map was larger than the measured one, while the predicted maximum value was lower. The opposite was observed with the MLR whose predictions were larger than the initial range of the measured data. For the RFR, this may be attributed to the fact that the model considered the lowest and highest values of the training data as outliers as also observed by Were et al. (2015). However, the mean SOC stocks measured for the topsoils (45.4 t C ha^{-1}) were very near to the mean SOC stocks predicted from the map (45.7 t C ha^{-1}).

4. Conclusion

This study provided insight into the quantitative status of topsoil (0 - 30 cm) and subsoil (30 - 100 cm) SOC stocks in the Dano catchment in different land use system and across different soil reference groups. Additionally, the driving factors and spatial distribution of the topsoil SOC stock was investigated. RFR and MLR modelling were used as a statistical method for identifying these factors and for mapping the spatial distribution of SOC stocks for the topsoil carbon stock.

The results indicated only a marginal difference between the surface SOC stocks in the savannah and cropland with most of the reference soil groups related to the former recording a slightly larger carbon stock. We attributed these findings to both site preferences by farmers for the better sites selected for cropping, as well as advanced land-use degradation of the savannah land with increasing human grazing pressure.

The topsoil SOC stock variability was primarily affected by soil properties (e.g., silt content) followed by the soil moisture distribution with the wetness index. Sites at higher elevation exhibited elevated SOC stocks in the surface soil. This disentanglement was due to landscape controls on population density and cropping intensity, which both concentrated in the lowlands. RFR performed slightly better than the MLR in predicting the spatial distribution of the topsoil SOC stock, as the latter could not account for the nonlinear association within the data.

Our findings reinforce the view that the semi-arid ecosystems of West Africa still offer a significant opportunity for carbon sequestration to offset ongoing C losses, with the spatial distribution of the topsoil SOC stock driven not only by soil and climate, but also by landscape-specific human pressure on ecosystems.

VI.

Carbon losses from prolonged cropping of Plinthosols in the Dano district (Southwest Burkina Faso)

Modified on the basis of

Kpade. O. L. Hounkpatin, Gerhard Welp, P.B. Irénikatché Akponikpèb, Ingrid Rosendahl, Wulf Amelung. Soil & Tillage Research 175C (2018) pp. 51-61.

Submitted manuscript

1. Introduction

The increase of carbon dioxide in the atmosphere is causing concerns worldwide; hence, recent focus is set on soil carbon sequestration for its mitigation. In fact, it is estimated that soils contain about 2500 gigatons (Gt) of carbon, of which 1550 Gt are SOC (Batjes, 1996; Jobbágy and Jackson, 2000). Tropical soils contain about 26 % of this global SOC inventory and are thus considered as important sources and sinks for carbon dioxide and methane (Batjes, 1996; Batjes, 2004). However, only very few studies acknowledged that the influx of SOC is larger than its efflux particularly in the West African savannah (Ciais et al., 2011). The savannah ecosystems cover about 60 % of tropical Africa (Callo-Concha et al., 2012a). They are characterized by structurally degraded and nutrient depleted soils with poor natural fertility and low fertilizer input (Doraiswamy et al., 2007). Maintaining SOC stocks in these ecosystems is thus mandatory for sustaining essential soil functions such as nutrient and water storage, soil biological activity, and structural stability.

For the African savannah ecosystem, especially in West Africa, several studies revealed a decline in SOC stocks by 20 - 50 % when sites under natural vegetation were converted into cropland (McDonagh et al., 2001; Murty et al., 2002). Most of such SOC losses are reported to occur within the first 20 years (Birch-Thomsen et al., 2007). To understand the underlying mechanisms, however, the monitoring of changes in SOC should include pools of different SOC stability, since overall response rates may be slow and thus ignored when based on bulk SOC analyses only (Powlson et al., 1987; Skjemstad et al., 2004a). A common approach for assessing such pools of different stability has been to fractionate soil into classes of different equivalent particle-size diameter (Christensen, 1992). When done, usually SOC decomposition rates are faster for the sand sized SOM fractions than for the remaining soil (e.g., Balesdent et al., 1988) . Lützow et al. (2008) reported about 50 - 75 % of total organic carbon (TOC) to be associated with the clay fraction, 20 - 40 % with the silt fraction and < 10 % with the sand fraction. The SOC of the latter fraction is frequently named as particulate organic matter (POM), due to its chemical properties matching those of more or less recent plant residues, and because this pool usually responds fast to land-

use change (Besnard et al., 1996; Chan, 2001). Balesdent et al. (1998) reported a 82 % POM-C loss after 35 years of cultivation with 76 % lost in the silt fraction and 53 % in the clay fraction. While most of these studies have been carried out in temperate areas, data on the SOC dynamics after this conversion into low-input agriculture in the West African savannah soil are still sparse (Bruun et al., 2013).

The stability of soil organic matter (SOM) is a major factor that characterizes its mineralization rates, being dependent on various physical, chemical and biological processes. The physicochemical interactions in tropical soils are largely affected by their significant portions in low activity clays (LACs; Barthès et al., 2008) . In contrast to the high activity clay soils (HACs) in temperate climates, LACs have a smaller cation exchange capacity (CEC < 24 cmol($^{+}$) kg^{-1} clay) due to elevated portions of kaolinite, Fe and Al oxides, and hydrous oxides (Juo and Adams, 1984; Powers and Schlesinger, 2002). These oxidic mineral phases, however, may exhibit strong affinity to SOM. While Bationo et al. (2007) pointed to low correlations between the contents of SOC and kaolinite, Feller and Beare (1997) reported that SOC content did not differ significantly between the LACs and HACs. In their study on different tropical soils of Ghana, Brunn et al. (2010) finally refuted the general concept of smectite (i.e., HACs) having higher SOC stabilizing power over kaolinite (i.e., LACs), whose sorption properties are similar to that of oxides (Denef and Six, 2005).

Influences of sesquioxides for stabilization of SOC via organomineral complexes have been discussed in detail by Lützow et al. (2006) and Kögel-Knabner et al. (2008). Beside Al oxides, particularly Fe oxides exhibit a large sorption capacity for SOC compared to other metal oxides (Chorover and Amistadi, 2001; Kaiser and Guggenberger, 2007). And both, Al oxides (e.g., Miltner and Zech, 1998 ; Amelung et al., 2001) as well as the presence of Fe oxides might delay the decomposition rate of SOM (Baldock and Skjemstad, 2000; Kalbitz et al., 2005). Lalonde et al. (2012) and Wagai and Mayer (2007) extracted Fe oxides by a dithionite treatment and concluded that Fe-bound SOM may contribute up to 22 % and 40 % to total SOC content, respectively. Similar estimates for tropical soils are lacking. Such estimates, however,

may be particularly needed for tropical semiarid climates, where beside Ferralsols particularly Plinthosols dominate the soil orders with Fe enrichment, especially on the African continent (Jones et al., 2013).

This study focused on Plinthosols, which are rich in LACs and Fe oxides, and which are the dominating reference soil group in some Sudanian areas of Burkina Faso. Also Lobe et al. (2001) investigated the impact of cultivation duration on SOC pools in the Plinthosols, characterized by soft plinthites. Lobe et al. (2001) focused on the upper 20 cm without specifically addressing the role of Fe oxides on SOC stability. Moreover, their study was carried out in subtropical South Africa with lower rainfall (616 – 663 mm) and temperature (13 - 16°C) compared to the present study. To widen our knowledge on the vulnerability of such widespread soils to arable management, this study focused again on Plinthosols, though with hard plinthite, specifically addressing the role of subsoil and Fe oxides for SOC turnover. Our study thus aimed at (1) investigating at different soil depths, how fast and to what degree Plinthosols with hard plinthites in West Africa are prone to SOC losses when converting native savannah to cropland, (2) assigning these SOC loss rates to different SOC pools (SOC in particle-size fractions), and (3) evaluating the contribution of Fe oxides to SOC stabilization and loss rates.

2. Materials and methods

2.1. Study Area

This study was conducted in the south western part of the Dano district (Dano (11°09´ 45.4´´N, 03°04´34.2´´W) located in the Ioba province, southwest of Burkina Faso (Fig. 1). Refer to section II. 1 for information related to climate, lithology and vegetation.

2.2. Soil Sampling

Soil samples were collected from fields that had been converted from savannah to cropland. Fields with 1, 7, 11, 13, 17, 21, 25, 28, and 29 (Y1 to Y29) years (yr) after cultivation were considered for the present study. About 15 soil profiles were excavated up to 1 m where possible and four soil cores (100 cm3) were collected per

horizon to determine the bulk density (BD). In addition, two profiles were dug and described where cultivation never occurred (0 yr) for control. All the core samples were dried at 105 °C for 24 hours before assessment of the weight of stone content (SC). About 42 soil samples were collected from the A and B horizons for laboratory analysis. However, weighted average of soil properties were considered in the present study for the 0 - 10 cm, 0 - 30 cm and 30 - 100 cm depth.

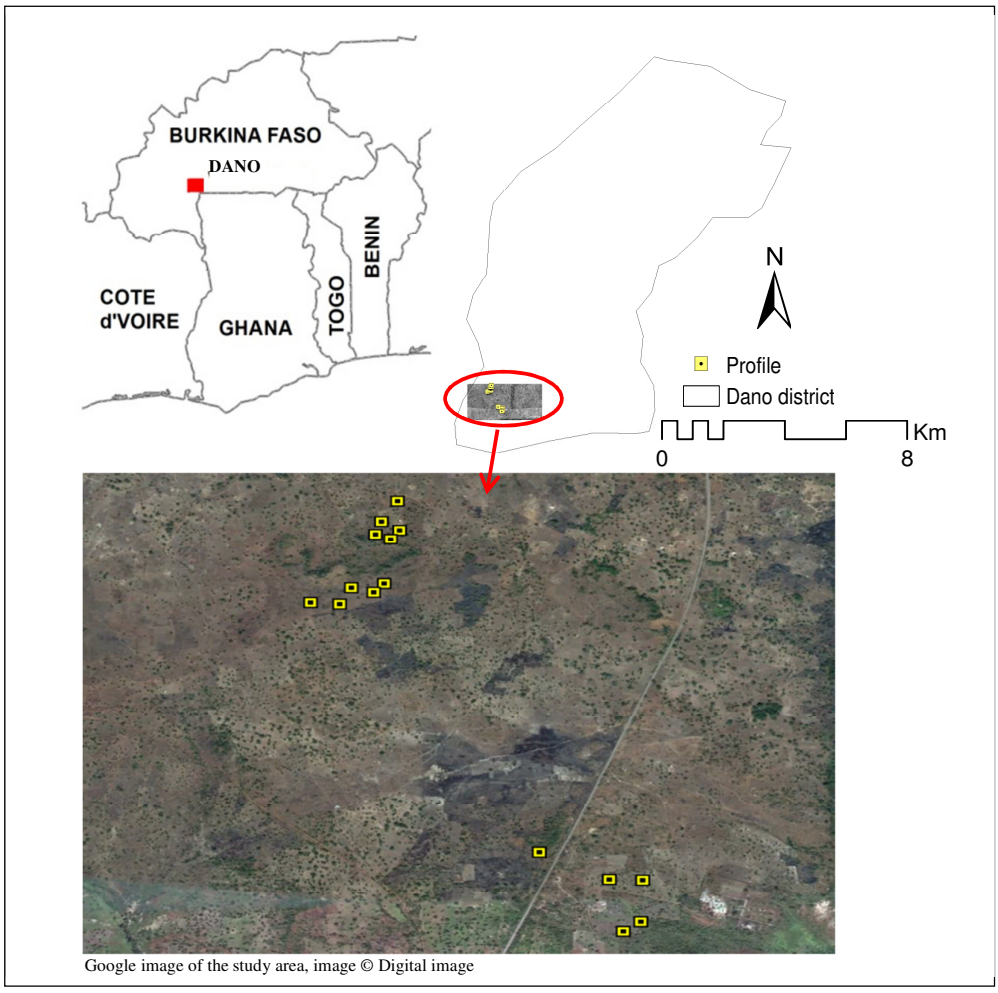

Fig. VI-1: Dano district and profile sampling

2.3. Soil analysis, particle size SOM fractionation

The samples were dried at 40 °C and sieved to 2 mm. For texture analysis and extraction of dithionite-citrate-bicarbonate extractable Fe (Fe_{DCB}) the procedures described by van Reeuwijk (1993) were followed. Total C was determined in ball-milled subsamples after dry combustion with an elemental analyser (Fisons NA 2000).

In 15 topsoil samples (0 - 10 cm) the amount of SOC bound to Fe oxides was estimated by measuring C before and after treating the sample with dithionate-citrate-bicarbonate as described above. The SOC loss was computed considering the initial and the remaining SOC after the DCB treatment.

For the physical fractionation of SOM pools, refer to section II. 5.

2.4. Determination of SOC stocks (see section II. 4)

2.5. Decay model and statistics

The non-linear regression models used by Lobe et al. (2001) and Blécourt et al. (2013) assume that SOC stocks reach a new steady-state equilibrium after converting savannah into cropland. Here, regression fits were tested for both monoexponential and biexponential models. The former assumes a single soil carbon pool (equation VI-1) while the latter considers both a labile and a stable SOC pools (equation VI-2).

$$X_t = X_e + (X_0 - X_e)\exp(-k\,t) \qquad \text{(VI-1)}$$

where X_t is the SOC content / stock at age t, X_e is the SOC content / stock at equilibrium, X_0 is the initial SOC content / stock in the savannah soil (t = 0), and k is a the decay rate constant.

$$X_t = X_1 \exp(-k_1\,t) + X_2 \exp(-k_2\,t) \qquad \text{(VI-2)}$$

where X_t is the SOC content / stock at age t, X_e is SOC content / stock at equilibrium, X_1 is the SOC content / stock of the labile pool, $X_2 = X_0 - X_1$ is the SOC content / stock of the stable pool, k_1 is the decay rate constant per year of the labile pool, k_2 is the decay rate constant per year of the stable pool.

The parameters for the monoexponential model (equation IV-1) and the biexponential model (equation IV-2) were generated by using Regression tool in SigmaPlot 13.0 for Windows (automatic determination of initial parameters, 200 iterations, step size 1, and a tolerance of 1.E-10). The evolution of SOC decay within the different fractions

(POM1, POM2, POM3, nonPOM) were assessed using the same equations. The monoexponential and biexponential models were assessed by carrying out an F-test (Pansu et al., 2004). The mean residence time (MRT) was also computed as the inverse of the exponential constant (and Amelung, 2011) as follows:

$$MRT = 1/k \quad \text{(VI-3)}$$

Based on the biexponential model, the point of kinetic change (t_{kc}) which marks the timing required for the stable pool to dominate the overall losses of SOC (Lobe et al., 2001) was computed. For this purpose,
the first derivative of $X_1 \exp(-k_1 t)$ was equal to that of $X_2 \exp(-k_2 t)$ and t_{kc} (years) was defined as follows:

$$t_{kc} = \frac{lnk_2X_2 - lnk_1X_1}{k_2 - k_1} \quad \text{(VI-4)}$$

A t-tests were carried out to assess the significance between virgin (0 yr) and each cultivated fields for carbon and other soil properties (BD, SC, sand, silt, and clay, FeDCB).

3. Results and discussion

3.1. Physical and chemical soil characteristics

Similar trends were observed for the soil properties in 0 - 10 and 0 - 30 cm (Table 1). For topsoil and subsoil, BD varied from 1.6 g cm^{-3} to 1.7 g cm^{-3} and from 1.5 g cm^{-3} to 2 g cm^{-3}, respectively. Large proportions of petroplinthites in the subsoil of the profiles described in the field Y1, Y7 and Y13 explained the high bulk density of 2 g cm^{-1}. The BD values are similar to those reported by Hien et al. (2006) for the southwestern part of Burkina Faso. In all investigated fields, we found large stone contents (SC > 60 %), mainly consisting of plinthites in both top- and subsoil.

On average, the texture of the topsoil was dominated by the sand fraction (35 %), followed by the silt fraction (33 %) and the clay fraction (31 %). A similar trend was observed in the subsoil with on average 36 %, 35 % and 26 %, respectively, for the sand, silt and clay fraction. The Fe_{DCB} contents ranged from 23.2 g kg^{-1} to 105.5 g kg^{-1} in the topsoil and from 3.6 g kg^{-1} to 77.7 g kg^{-1} in the subsoil. Relatively similar Fe_{DCB} values were recorded by Da Motta and Kämpf (1992) and Osodeke et al. (2005) for the topsoil and subsoil for various soil orders in Brazil and Nigeria respectively. The variability of Fe_{DCB} in relation to the years of cultivation did not follow any clear particular pattern for both topsoil and subsoil.

The topsoil SOC content varied from 9.9 g kg^{-1} to 23.9 g kg^{-1} and mostly decreased with cultivation duration (Table 1). These values are within the range reported by Agbenin and Adeniyi (2005) in Nigeria, Hien et al. (2006) in Burkina Faso, Assize et al. (2013) in Senegal, and Zingore et al. (2005) in Zimbabwe. Lower SOC content in cropland soils compared to natural vegetation is generally admitted in many other studies (Wiesmeier et al., 2013; Yang et al., 2010). The subsoil SOC content was smaller than that of the topsoil in all fields, due to larger direct biomass input into the topsoil as also recorded in other studies (Wang et al., 2014; Zhong and Qiguo, 2001).

Tab. VI-1: Soil physical characteristics, dithionite-citrate-bicarbonate -extractable Fe and SOC content of the chronosequence fields

Year	n	BD ($g\ cm^{-3}$)		SC (%)		Sand (%)		Silt (%)		Clay (%)		Fe_{DCB} ($g\ kg^{-1}$)		SOC ($g\ kg^{-1}$)	
0 - 30 cm															
0	2	1.6	(±0.0)	73.2	(±6.9)	35.0	(±1.3)	37.2	(±3.1)	25.4	(±3.8)	72.1	(±21.2)	23.9	(±0.6)
1	2	1.6	(±0.0)	76.4	(±5.3)	37.9	(±10.3)	32.3	(±6.1)	28.2	(±17.7)	92.5	(±1.0)	18.7	(±6.7)
7	2	1.6	(0.0)	76.3	(±1.3)	36.2	(±1.1)	33.4	(±0.2)	28.1	(±1.2)	105	(±13.9)	17.0	(±5.8)
11	1	1.6	-	65.2	-	27.9	-	33.6	-	36.3	-	57.8	-	12.7	.
13	1	1.6	-	78.3	-	50.8	-	26.5	-	22.3	-	23.2	-	13.7	.
17	2	1.6	(±0.0)	69.4	(±8.2)	41.8	(±2.3)	33.5	(±10.2)	22.9	(±11.2)	41.2	(±1.2)	10.9	(±1.2)
21	1	1.6	-	71.3	-	33.4	-	41.3	-	23.5	-	63.0	-	12.3	.
25	1	1.6	-	62.0	-	27.6	-	36.6	-	35.6	-	52.8	-	10.1	.
28	2	1.6	(±0.0)	65.0	(±11.2)	25.3	(±13.8)	29.5	(±0.0)	43.7	(±14.1)	35.8	(±1.5)	10.4	(±0.2)
29	1	1.7	-	70.6	-	31.8	-	23.9	-	43.1	-	39.9	-	9.9	.
30 - 100 cm															
0	2	1.5	(±0.1)	62.0	(±30.6)	38.7	(±15.4)	31.1	(±0.7)	27.8	(±14.6)	39.3	(±29.2)	4.0	(±0.0)
1	2	2.0*	(±0.0)	91.2	(±1.5)	-	-	-	-	-	-	9.7	(±5.1)	2.4	(±0.1)
7	2	2.0*	(±0.0)	93.7	(±1.4)	-	-	-	-	-	-	3.6	(±2.8)	1.0	(±0.1)
11	1	1.5	-	65.2	-	37.1	-	41.9		18.7	-	69.7	-	4.2	.
13	1	2.0	-	89.0	-	36.7	-	42.2		19.2	-	12.0	-	2.6	.
17	2	1.6	(0.1)	78.4	(±3.4)	37.1	(±9.2)	35.8	(±12.3)	25.0	(±4.4)	29.5	(±2.9)	3.9	(±0.1)
21	1	1.5	-	69.1	-	30.1	-	25.2	-	43.0	-	77.7	-	3.8	.
25	1	1.5	-	67.6	-	46.1	-	32.8	-	18.7	-	61.1	-	4.8	.
28	2	1.5	(±0.1)	71.4	(±7.1)	33.8	(±9.8)	33.6	(±2.6)	30.7	(±8.6)	25.3	(±16.8)	5.9	(±0.1)
29	1	1.5	-	70.0	-	32.4	-	37.0	-	29.0	-	58.8	-	2.6	.

n: number of samples, BD: bulk density, SC: stone content, *petroplinthite, - for n=1

3.2. SOC content in the different POM fractions of the topsoil

The topsoil SOC content in the top 10 cm followed the same trend as for the first 30 cm with a general decrease with cultivation duration (Tab. VI-2). The SOC content in the different POM fractions followed the pattern: nonPOM > POM1 > POM3 > POM2 C, irrespective of the duration of cultivation. This trend was consistent with other studies where POM C content was reported to be larger in finer fractions but diluted in coarser ones (Amelung et al., 1998; Christensen, 1996). The nonPOM pool usually contains microbial products as well as decay products from coarser fractions (Amelung et al., 2002; Guggenberger et al., 1994; Lobe et al., 2002). Thus, the

dominance of the nonPOM fraction suggests a high level of microbe-derived, organo-mineral associations in all the Plinthosols.

Tab. VI-2: SOC content in different particle-size fractions of the topsoil (0 - 10 cm; standard deviation in parentheses)

Age	n	POM1-C		POM2-C		POM3-C		nonPOM-C		SOC	
		250-2000µm		50-250µm		20-50µm		<20 µm			
(years)		(g kg^{-1})		(g kg^{-1})		(g kg^{-1})		(g kg^{-1})		(g kg^{-1})	
0	2	5.1	(±1.2)	1.1	(±0.1)	2.7	(±0.5)	27.1	(±0.4)	36.73	(±2.58)
1	2	2.3	(±1.4)	1.0	(±0.6)	1.3	(±0.9)	16.1	(±5.2)	21.90	(±8.49)
7	2	3.4	(±0.4)	0.9	(±0.2)	1.1	(±0.4)	15.6	(±4.0)	19.98	(±6.97)
11	1	1.7	-	0.5	-	1.1	-	12.5	-	15.40	
13	1	2.5	-	0.5	-	1.4	-	11.9	-	15.65	
17	2	2.5	(±1.9)	0.5	(±0.3)	0.7	(±0.3)	11.2	(±4.3)	14.35	(±6.15)
21	1	1.9	-	0.6		1.5	-	12.0	-	16.10	
25	1	1.5	-	0.4		0.9	-	8.8	-	11.30	
28	2	1.0	(±0.1)	0.4	(±0.1)	0.7	(±0.2)	9.9	(±0.7)	11.75	(±2.19)
29	1	0.8	-	0.2	-	0.5	-	7.7	-	9.85	(±2.58)

- for n=1

3.3. Dynamics of SOC stock in bulk soil at different depths in relation to land use duration

The SOC stock expressed relative to the stock in the savannah land are presented in Fig. VI-2 for the topsoil and the entire soil profile respectively. Because the stocks of SOC revealed a similar temporal trend like those of the SOC contents, only the former are presented here to avoid redundancies. The SOC stock relative to the stock in the savannah land declined with increasing land use duration for the considered depth intervals. Yet, the decline was stronger in the topsoil compared to the entire soil profile. This decline was also faster during the first decade of cultivation but slowed down in the remaining years, suggesting a faster SOC stock loss in the initial years of cultivation as also recorded by Lobe et al. (2001), Solomon et al. (2007) and Don et al. (2011).

In Fig. VI-2, the decline of SOC stocks was additionally fitted with exponential equations (see chapter 3.5 for more details). Based on these equations, the SOC stocks were reduced by 66 % ($p < 0.01$) in 0 - 10 cm and by 55 % ($p < 0.01$) in 0 - 30 cm after 29 years of cultivation. This corresponded to a total SOC loss of 24 t C ha^{-1} and 49 t C ha^{-1} within 29 years. A loss of SOC from topsoils after the conversion of native natural vegetation into cropland is a common phenomenon (Coutinho et al., 2014; Paustian et al., 1997). A much stronger loss was recorded by Pardo et al. (2012) in Tanzania with about 50 % loss of SOC stocks after 10 years of cultivation for the upper 0 - 10 cm depth while in the present study about 38 % was recorded for the same cropping duration. Guo and Gifford (2002b) reported 42 % of SOC stock loss after more than 10 years of cultivation for the top 30 cm depth. A smaller decrease in SOC stocks was found by Don et al. (2011) who recorded 25 % loss of SOC stocks after forest conversion into cropland at an average of 36 cm depth and a time since conversion of 22 years. The present findings are larger than the average of SOC stock loss mentioned in the review of Davidson and Ackerman (1993) who reported 30 % loss in average for the top layer (0 - 30 cm) of some tropical soils with land use change from native vegetation into cropland varying between 0.6 and 90 years.

Very few studies extended the monitoring of SOC losses into the subsoil. In the present study, the soils were sampled down to 100 cm, and found that between 13 to 50 % of the average SOC over 0 - 100 cm was stored in the 30 - 100 cm depth interval. With increasing cropping duration, no clear trends were found for subsoil SOC contents (Tab. VI-1), because large contents of rock fragments likely concentrated SOC in the remaining fine earth (Bornemann et al., 2011). For calculation of subsoil SOC stocks, these amounts of rock fragments are accounted for. The final results then showed that SOC losses extended into the subsoil of some of the fields, and, on the average, 0.7 to 19.5 t C ha^{-1} was lost from the 30 - 100 cm depth interval (Appendix C Fig. XI-1). When considering the whole soil profile over 100 cm depth, the SOC stock was reduced by 52 % ($p < 0.01$) after 29 years of cultivation. This is slightly less than reported by Chandran et al. (2009), who found that up to 63 % of SOC was lost over 100 cm after 40 years of cultivation in semiarid soils in India, while a lower value of

24 % of SOC stock loss was reported by Elberling et al. (2003) in semi-arid Senegal for a similar land use duration and soil depth.

Intriguingly, the results further revealed that no steady-state equilibrium was reached after 29 years of cropping, neither for the topsoil, nor for the entire soil profile. Possibly the cropping time in our study must still be considered as being short. It was repeatedly reported that SOC reached a new steady-state equilibrium after 30 to 50 years of land use duration (Arrouays et al., 1995; Balesdent et al., 1988). Lobe et al. (2001) recorded an equilibrium after 34 years of cropping for the SOC content in South African soils (also Plinthosols, though with soft plinthite). For two regions in Kenya, it took 21 and 37 years after steady-state equilibrium was reached (Solomon et al. (2007). In any case, the present data suggest that SOC losss from the Plinthosols will be ongoing.

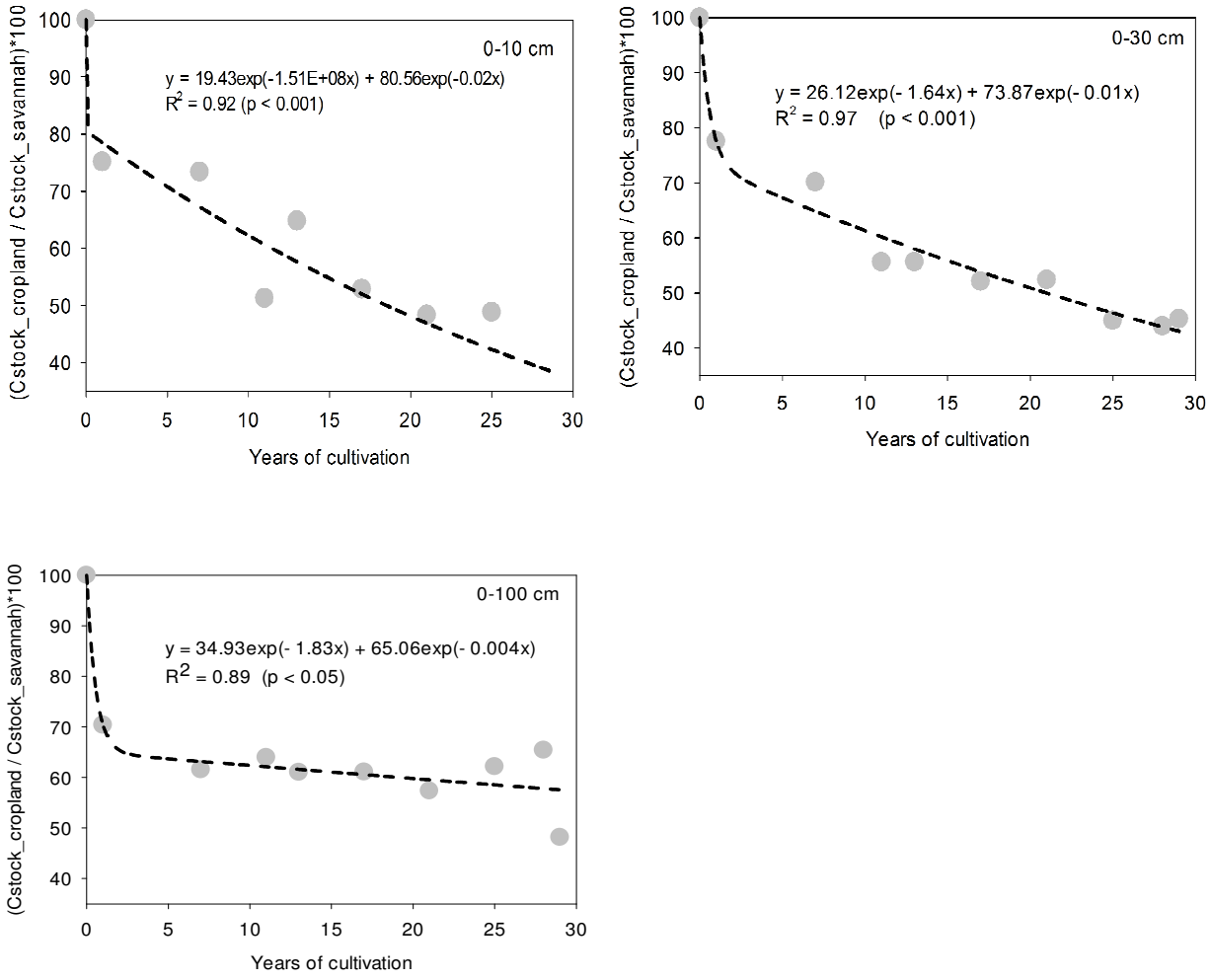

Fig. VI-2: SOC stocks of cropland in relation to SOC stock of savannah soils (in %) for different years of cultivation in the topsoil and entire soil profile

3.4. Dynamics of SOC stock in POM fractions in relation to land use duration for the topsoil

A further insight into the dynamics of the SOC stock loss can be obtained by investigating the pattern of the residual SOC stock ratio in the particle-size fractions. A decline in SOC stock was observed not only in the POM fractions but also in the nonPOM fraction (Fig. VI-3) of the topsoil (0 - 10 cm). Since the equations for SOC losses in POM1 and POM3 were not significant, we further present the variation of SOC stock with land use duration for POM2+POM3 (250 – 20 µm) and all POM -

POM1+POM2+POM3- (2000 µm – 20 µm). Compared to the bulk soil, the SOC losses in the POM fractions followed the same trend with most decline occurring within the first 10 years especially for POM1-C and POM3-C. After 29 years of cultivation, the SOC stock was reduced by 72 % ($p < 0.05$) for POM2-C, 74 % ($p < 0.05$) for POM2+POM3-C and 77 % ($p < 0.05$) for all POM-C. The data are in line with earlier findings that SOC losses mostly originate from the POM fraction. Balesdent et al. (1998), for instance, reported that 82% of SOC in POM was lost after 35 years of cultivation. Losses from the silt fraction were 76%, those from the clay fraction 53%. Besnard et al. (1996) found 43% and 92% POM-C losses, respectively, after 7 years and 35 years of cultivation.

The POM1-C pool contributed relatively more to the SOC losses observed in the bulk soil at 0-10 cm depth compared to POM3-C and POM2-C (Appendix C Fig. XI-2). The POM1 (> 250 µm) which is the coarse sand fraction is considered to be more sensitive to cultivation (Yamashita et al., 2006). We also recorded a large SOC losses for the nonPOM fraction, which amounted to 63% ($p < 0.05$) after 29 years of cultivation. However, the magnitude of the finding for the latter was contradictory in view of literature data (Christensen, 1992; Guimarães et al., 2014) where it is generally reported that SOC exhibits a higher stability with time for the nonPOM fraction. Moreover, when calculating the absolute decline in SOC, it was even larger for the nonPOM following the fact that this fraction initially contained the largest amount of SOC (Tab. VI-2, Fig. XI-2). Also Steinmann et al. (2016) recorded losses of SOC in this fraction as a result of past land uses changes and management in Germany. We thus suggest that nonPOM-C of the studied Plinthosols was more vulnerable to decay than formerly reported, possibly due to a facilitated breaking of soil aggregates that overcame physical stabilization processes (Six et al., 2002).

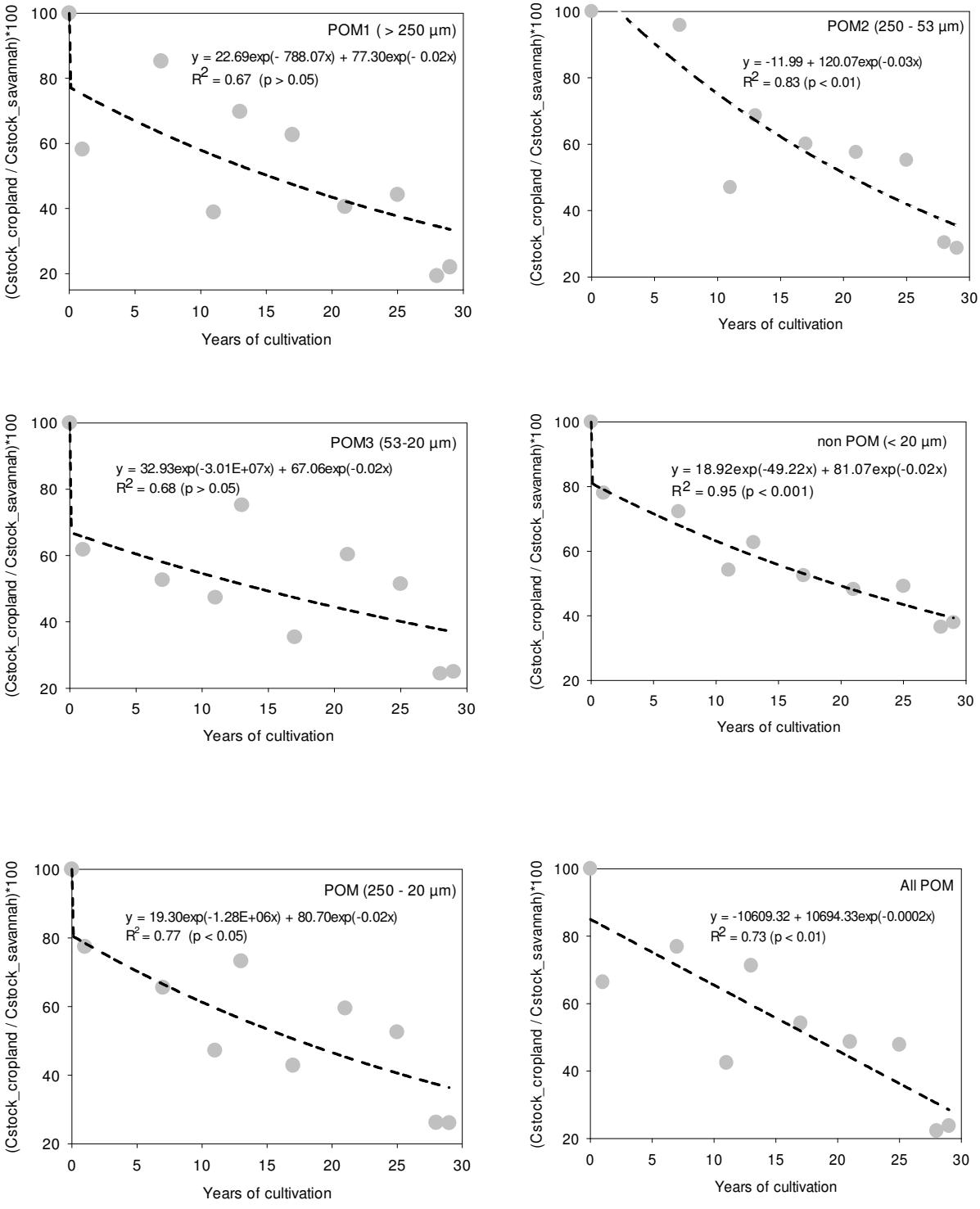

Fig. VI-3: SOC stocks of cropland in relation to SOC stock of savannah in different particulate organic matter (POM) fractions (in %) for different years of cultivation in the topsoil (0 – 10 cm)

3.5. Kinetics of SOC in bulk soil and particle-size fractions

The topsoil SOC stock content and stock as well as the SOC stock of the entire soil profile were fitted with both monoexponential and biexponential models (Tab. VI-3). Only significant models are here reported. Generally, the decline of SOC contents and stocks in the bulk soil was best fitted with a biexponential model, with significant differences to the monoexponential for the SOC content at 0 - 10 cm depth ($p < 0.01$) and for SOC stock at 0 - 30 cm depth ($p < 0.05$).

The mean residence time (MRT) as revealed by the monoexponential models varied with depth for the SOC content and stock. A relative small MRT of 0.93 yr was found for the SOC content at 0 - 10 cm depth while a MRT of 10 yr was recorded for the 0 - 30 cm depth interval. This might indicate that ploughing did not always reach the 30 cm depth but was by incident shallower. If ploughing, for instance, referred to the top 20 cm only, there is only slow turnover of SOM at the 20-30 cm depth interval, thus overall prolonging the MRT at 0 - 30 cm soil depth. The same principle applies to all other ploughing depths below 30 cm.

Intriguingly, a longer MRT was obtained for the top 0 - 10 cm of soil when calculations were performed with SOC stocks instead of SOC contents (Table 3). This finding could be attributed to some compaction in the upper 10 cm of the soil that went along with even larger variability in stone contents (Appendix C Fig. XI-3, XI-4). When the soil is compacted, sampling by volume includes more subsoil, thus diluting SOC concentrations but not stocks. Similarly, rising stone contents may increase carbon saturation (Bornemann et al., 2011) and thus vulnerability of SOC against decay, while not necessarily affecting SOC stocks. Yet, such differences should not be overinterpreted, because fit quality was overall worse than for the bi-exponential model. If using the latter, the MRTs were as short as for the SOC contents.

For the SOC content, the MRT recorded for the monoexponential model for the upper 10 cm was slightly lower than the values reported by Solomon et al. (2007) in Kenya, who, however, assessed SOC loss rates after deforestation and not after converting

savannah to cropland. In contrast, the result for the upper 30cm revealed a MRT that was larger than that recorded by Lobe et al. (2001) in South Africa, who, however, sampled the top 20 cm of soil only (Tab. VI-3). Overall, the MRTs were thus in the range of MRTs reported for other tropical soils, i.e., no specific indication was found that the presence of hard plinthite between 0 - 30 cm soil depth delayed SOC losses at significant scale. In contrast, the MRTs of the topsoil SOC stocks were at least two times lower (i.e., SOC turnover was at least 2 times faster) than that reported for temperate areas by Gregorich et al. (1995) and Wei et al. (2014a), probably due to the warmer climate and more sandy texture favoring faster decomposition. A much lower MRT (< 1) was even recorded by Dalal and Mayer (1986) in the warmer climate area (Riverview, Australia) for a kaolinite dominated sandy loam soil.

The points of kinetic change from the biexponential models revealed that the decline rate for the topsoil SOC content was dominated by the stable pool in less than 1 yr for the upper 10 cm and in less than 2 yr for the upper 30 cm. The same trend was observed for the topsoil SOC stock with the decline rate being dominated by the stable pool within 2 yr. These results suggest that the ability of the soil to release nutrients to plants dropped after two years making the use of fertilizers crucial for subsequent cropping.

The investigation of SOC dynamics in the particle-size fractions confirmed that loss rates were better described with biexponential models, with significant differences to the monoexponential for POM1-C ($p < 0.05$), nonPOM-C ($p < 0.01$) and nonPOM-C stock ($p < 0.01$).

Considering the monoexponential models, the decay rate of SOC related to the various particle-size fractions generally increased from the non POM to the POM1 fraction, as also found in other studies (Balesdent et al., 1988; Balesdent et al., 1998; Lobe et al., 2001). However, contrary to the previous studies the POM2 (250 – 53 µm) fraction recorded the slowest decay rate and longest MRT for both its SOC content and stock. On the one hand, POM2 C represents the intermediate sand fraction (250 – 53 µm) and contains materials at an advanced stage of degradation that could already be occluded in soil aggregates where they might be better protected from decay (Six et al., 2000).

This may explain the lower MRT compared with POM1, but not compared with POM3. It seems thus reasonable to speculate that other factors contributed to the relative long MRT of the POM2-pool. On the one hand this fraction may contain significant amounts of black carbon (the remains from burning events) with low turnover time (Brodowski et al., 2007), on the other hand, also very stable Fe concretions could end in the size range, so that not all SOC in the 250-53 µm fraction is truly POM. The specific role of Fe oxides is thus discussed in the subsequent section.

For the biexponential model, the labile (k_1) pool decreased from the fine fractions to the coarse fractions for the SOC content. These results are contrary to the finding of Lobe et al. (2001) who recorded an increase from clay to the coarse sand fraction. Yet, the sampling depth in both studies is not comparable, in addition, the point of kinetic change t_{kc} was already reached in < 1 year for the upper 10 cm (Table 3). Hence, there are not enough data to truly interpret differences in k values from the labile pool, and it is therefore concluded from the finding that two pools existed with the first one being relevant only for initial SOC losses upon cropping.

Tab. VI-3: Kinetic parameters for the average decline rates of SOC in bulk soil and particle-size fractions as affected by land use duration at different soil depths (results of this study plus literature data)

	Exponential model				Biexponential model					
Site & soil layer	k (yr^{-1})	R^2	t (yr)	MRT (yr)	k_1 (yr^{-1})	MRT1 (yr)	k_2 (yr^{-1})	MRT2 (yr)	R^2	t_{kc} (yr)
Dano Burkina (this study): SOC content (g kg^{-1})										
Bulk soil, 0 - 10 cm	1.07	0.86***	29	0.9	6.43	0.2	0.02	50	0.97***	0.80
Bulk soil, 0 - 30 cm	0.1	0.92***	29	10	1.72	0.6	0.02	50	0.95***	1.90
POM1 (> 250 µm), 0 - 10 cm	2.12	0.63*	29	0.5	33.92	0.0	0.03	33	0.82*	0.20
POM2 (250 - 53 µm), 0 - 10 cm	0.07	0.88***	29	14	0.23	4.3	0.03	33	0.89*	4.26
POM3 (53 - 20 µm), 0 - 10 cm	1.74	0.75**	29	0.6	166	0.0	0.02	50	0.82*	0.05
non POM (< 20 µm), 0 - 10 cm	1.17	0.85**	29	0.8	15633	0.0	0.02	50	0.97***	0
Dano Burkina (this study): SOC stock (t C ha^{-1})										
Bulk soil : 0 - 10 cm	0.07	0.89***	29	14	2.6	0.4	0.03	33	0.94***	1.28
Bulk soil : 0 - 30 cm	0.1	0.93***	29	9.3	1.64	0.6	0.01	100	0.97***	2.12
POM2 (250 - 53 µm)	0.03	0.83***	29	33.3	0.04	25	0.04	25	0.83*	30.37
non POM (< 20 µm)	0.06	0.91***	29	16.7	49.21	0.0	0.02	50	0.96***	0.13
Free State Province, South Africa (Lobe et al., 2001) : SOC content (g kg^{-1})										
Bulk soil, 0 - 20 cm	0.15	0.97	90	6.6	0.23	4.3	0.00	217	0.99	17.1
Coarse sand, 0 - 20 cm	0.4	0.89	90	2.5	0.6	1.7	0.01	100	0.92	8.1
Fine sand, 0 - 20 cm	0.1	0.85	90	10	0.11	9.1	0.00	1429	0.85	46.7
Silt, 0 - 20 cm	0.09	0.97	90	11.1	0.11	9.1	0.00	435	0.98	34
Clay, 0 - 20 cm	0.09	0.97	90	11.1	0.11	9.1	0.00	435	0.97	33.5
Nandi Kenya (Solomon et al., 2007): SOC content (g kg^{-1})										
Bulk soil, 0 - 10 cm	0.16	-	100	6.2	-		-		-	-
Kakamega Kenya (Solomon et al., 2007): SOC content (g kg-1)										
Bulk soil, 0 - 10 cm	0.29	-	103	3.4	-		-		-	-
Pyrenean Piedmont France (Balesdent et al., 1998): SOC content (mg C g^{-1})										
Coarse sand, 0 - 26 cm	0.25	-	40	4	-		-		-	-
Fine sand, 0 - 26 cm	0.18	-	40	5.5	-		-		-	-
Coarse silt, 0 - 26 cm	0.15	-	40	6.7	-		-		-	-
Fine silt, 0 - 26 cm	0.12	-	40	8.3	-		-		-	-
Clay, 0 - 26 cm	0.03	-	40	33.33	-		-		-	-
Ontario, Canada (Gregorich et al., 1995) : SOC stock (t C ha^{-1})										
Bulk soil, 0 - 30 cm	0.03	-	25	33.3	-		-		-	-
Shaanxi China (Wei et al., 2014a): SOC stock (t C ha^{-1})										
Bulk soil, 0 - 10 cm	0.03	-	100	30.3	-		-		-	-

Bulk soil, 0 - 10 cm	0.01	-	100	76.9	-	-	-	-
Riverview, Australia (Dalal & Mayer, 1986) : SOC stock (t C ha^{-1})								
Bulk soil, 0 - 10 cm	1.2	0.87	20	0.8				

-: no data, *: $p \leq 0.05$;**: $p \leq 0.01$;***: $p \leq 0.001$

3.6. Role of Fe oxides for SOC dynamics

To capture the role of Fe oxides for the stabilization of soil organic matter, SOC stocks were analyzed before and after reductive dissolution and subsequent extraction of Fe oxides with DCB. Here, focus was set on the surface Fe enriched (0 - 30 cm) soils (Tab. VI-1). Several studies pointed out that Fe oxides can impede SOC decomposition and reduce SOC losses (Baldock and Skjemstad, 2000; Kalbitz et al., 2005; Poulson et al., 2016). Since Plinthosols are low activity (kaolinitic) clay soils rich in Fe oxides (IUSS et al., 2006), a significant contribution of the latter to SOC stabilization was expected. However, the scatter plot of the SOC stock loss over 29 years against the SOC stock loss due to the DCB treatment (Fig. VI-4) did not yield a significant correlation ($R^2 = 0.0083$, $p > 0.05$).

In our study, about 0.2 % to 48 % with an average of 16 % (± 15 %) of SOC stock were lost after treating the topsoil samples with DCB (Fig. VI-4). Overall, this is consistent with results published by Adhikari and Yang (2015) and Wagai and Mayer (2007) who found about 5 - 44 % and 4 - 37 % (0 - 28 cm depth) of Fe associated SOC respectively. However, for the results from Wagai and Mayer (2007) only one soil order recorded the highest amount of Fe associated SOC (37 %) while less than 25 % of Fe-SOC complexation was observed with the remaining. Out of the 58 to 80 % of the organic matter subject to organomineral complexation, only 2 to 7 % was observed by Basile-Doelsch et al. (2009) to be associated with Fe in some Oxisols at 0 - 20 cm depth in Madagascar. However, Poulson et al. (2016) found an average of 37 % for Fe bound SOC in some US forest soils at 0 – 20 cm depth, which is two times higher than the averaged reported in the present study. The difference might be related to higher initial Fe oxide of the forest soils compared to the cropland of the present study.

Since in our study most of SOC stock was found as nonPOM, intimate association with clay and silt particles was suggested as the main mechanism for SOC stability in the studied Plinthosols rather than specific occlusions into oxides. Though specific measurements of occluded SOC had not been carried out, it was believed that pure Fe concretions would not point at any elevated SOC content if detectable at all. The fast loss of SOC content and stock in the topsoil suggests, however, that the binding of SOC to clay or silt plus clay in the nonPOM fraction is not as stable as in other soils and remains still accessible to decomposition upon continuous cultivation.

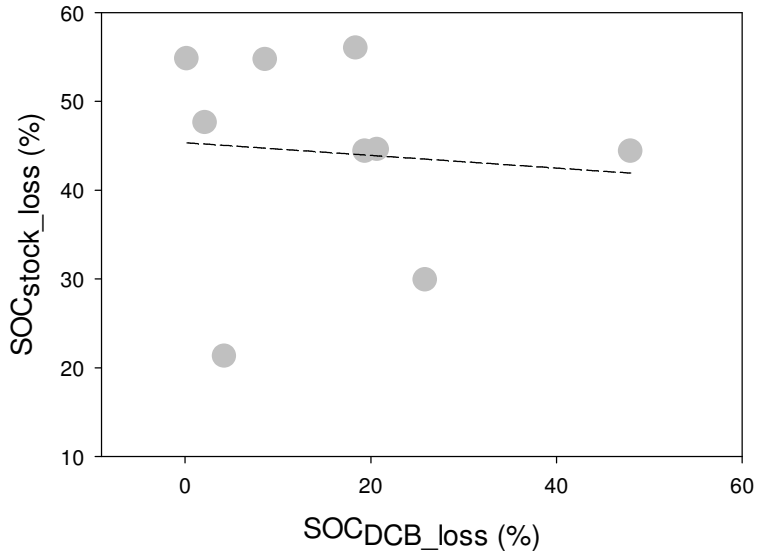

Fig. VI-4: Relation between real SOC stock loss in topsoil (0 - 30 cm) over a period of up to 29 years and SOC stock loss after DCB treatment

4. Conclusion

In the Plinthosols studied here, the conversion of natural vegetation to continuous cultivation resulted in a decline of SOC contents and stocks. Continuous cultivation reduced the SOC stock by 34 %, 45 % and 48 % after 29 years of cultivation in the upper 10 cm, 30 cm and 100 cm, respectively. SOC losses extended also into the subsoil, and, on the average, 0.7 to 19.5 t C ha^{-1} was lost from the 30 - 100 cm depth interval. For the upper 10 cm, the losses occurred from all POM fractions as well as from the nonPOM fraction. However, the SOC loss occurred faster in the coarse sand-sized fraction, which thus exhibited the lowest mean residence time of the studied

fractions. The point of kinetic change, which marks the timing required for the stable pool to dominate the overall losses of SOC, indicated for both topsoil and the dominance of the decline rate by the stable pool in less than 3 years, suggesting that provision of fertilizers would be necessary to raise the productivity of the soils. Our results also suggest that Fe oxides only played a minor role as stabilizing agent for SOC. This points to the formation of silt and clay sized organomineral complexes as well as to the binding of Al oxides as main protection mechanism, i.e., the hypothesis that Fe exerts a major control on SOC losses in these plinthitic soils is refuted. Though the nonPOM fraction is usually associated with a higher stability, the cultivation induced SOC losses from this fraction indicate that it remains vulnerable to decomposition processes when savannah is broken for continuous cropping in these major reference soil groups.

VII.

Synthesis and perspectives

1. Introduction

Land degradation is a major issue nowadays in Sub-Saharan Africa especially in the semi-arid regions where climate conditions (Sivakumar and Stefanski, 2007) and land use pressure (Meshesha et al., 2012) affect soil productivity and livelihood. Addressing land degradation requires having the necessary soil spatial information which is crucial in any land evaluation. As pointed out by Henry et al. (2009), soil preservation or recommended conservation practices cannot be carried out without maps of soil properties and soil groups. One of the major reasons maps are required is the highly spatial variability of soil properties as dissimilarities in values are often recorded within small distances of meters or even decimeters (Wiesmeier et al., 2014).

In addition, management decision at small scales such as plots or small catchment require finer scales maps which are not available as traditional maps are mostly built at a coarse scale. Recent advances in remote sensing and information systems resulted in a new paradigm in soil mapping called “digital soil mapping” (DSM) which couples soil legacy data with some statistically correlated auxiliary data (McBratney et al., 2003). With the increased availability of free high resolution remote sensing data, DSM offered an unique opportunity for map data provision especially in West Africa where dearth of baseline data prevent accurate decision making towards sustainable management practices. For implementing DSM, using adequate models to carry out such correlation and conducting data treatments to remove redundancies and noise due to imbalance data are key determinants for improvement in prediction accuracy (Schmidt et al., 2008).

Generally land degradation adversely affects the soil organic carbon (SOC) which is the key indicator of soil health owing to its major role in most soil functions such as the storage of nutrients and water, soil biological activity and structural stability. Much attention has been given to SOC pools in soils because of its determinant role in the global carbon cycle and its potential for mitigating or aggravating the amount of the greenhouse gases in the atmosphere (Davidson and Janssens, 2006; Liu et al., 2011). In West Africa where natural soil fertility and fertilizer input are low, preserving SOC is of the utmost importance for soil to fulfill key ecosystem services (Doraiswamy et

al., 2007). Though some carbon budget estimates have become available, there is still a lot of uncertainties whether Africa is a carbon source or sink (Valentini et al., 2014) due to data scarcity. A step forwards in reducing these uncertainties require more data in SOC estimation over different soil and land use type in Africa in general and in West Africa in particular.

The SOC content and stock vary at different point of the landscape resulting from the interplay of various factors that determine its amount in time and space. Thus, various studies have been carried out on SOC and its determining factors such as climate (Albaladejo et al., 2013; Stergiadi et al., 2016), land use/cover change (Muñoz-Rojas et al., 2015; Xiong et al., 2014), topography (Nadeu et al., 2015), sesquioxides (Peng et al., 2015) and soil type (Bruun et al., 2013; Wiesmeier et al., 2012). With the interplay of these factors, SOC reaches equilibrium values depending on the type of systems and locations. However, the equilibrium is adversely affected when natural areas are cleared and converted into cropping land (McDonagh et al., 2001; Murty et al., 2002). Such conversion is reported to be followed by a decline in SOC and analysis include pools of different SOC stability, since overall response rates may be slow and thus ignored when based on bulk SOC analyses only (Skjemstad et al., 2004b). Moreover, most studies only focused on surface soil horizons while more than 50 % of SOC stock is usually allocated below 20 cm depth (Batjes, 1996). Achieving the Kyoto protocol requires the assessment of stocks of SOC in different land use and soil type at different depth which is an essential step towards evaluating the sequestration potential of a land. Additionally, a good understanding of factors affecting carbon dynamics is necessary for the development of adequate management strategies.

Land degradation assessment and accurate conservation decision by farmers, scientists and policy makers require spatial and temporal distribution of both soil properties and soil groups which can be made available with new statistical techniques related to digital soil mapping. Key information of soil health indicator such as SOC and its dynamics with land use change are also crucial for sound management practices and for the computation of future climate scenarios as well as the identification of the

potential for C sequestration or emission. My objectives were therefore to : (i) assess the use of finer spatial and temporal resolution optical imagery along with topographical variables to improve the prediction accuracy in DSM of some soil properties, (ii) to evaluate the impact of different data pruning methods as a mean for improving data quality in the prediction accuracy of some reference soil groups (iii) determine the amount, distribution and driving factors of SOC stock in different soil groups and land use, (iv) to investigate the impact of land use change on soil SOC content and stock along a cultivation chronosequence.

2. Summary of the results

(i) 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

Spatial soil information is crucial for environmental modelling, risk assessment and decision making. The availability and 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 ability of Remote Sensing data in improving knowledge of local scale soil information in West Africa have not been fully explored. This study was conducted to assess the use of high spatial resolution satellite data (RapidEye and Landsat), terrain/climatic data and laboratory analyzed soil samples to map the spatial distribution of six soil properties – silt, sand, 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 used and compared. A cross validation was carried out for internal validation while the predictions were validated against an independent set of soil samples considering the modelling and an extrapolation area.

Results showed from the performance statistics that the machine learning techniques performed marginally better than the MLR, with the RFR providing in most cases the highest accuracy. 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. In view of the increased availability of freely available Remote Sensing data (e.g. Landsat, SRTM, Sentinels), soil information at local and regional scales in data poor

regions such as West Africa can be improved with relatively little financial and human resources.

(ii) Predicting reference soil groups in the Dano catchment (Southwest Burkina Faso) using legacy data: data pruning and random forest approach

Digital soil mapping uses quantitative correlations between a set of covariates and a target variable to be predicted. However, predicting taxonomic classes could be challenging when a major soil class belonging to a wide range of covariates overlaps with those related to smaller class units. The extent to which different data pruning methods which result in different subsets of the majority class could lead to an increase in prediction accuracy by using Random Forest (RF) was investigated. The Random Forest modelling was conducted either with (RF_rfe) or without (RF) recursive feature elimination. The methods were applied for digital mapping of some reference soil groups in the Dano catchment (Burkina, West Africa), using a large soil dataset in which the Plinthosols were the major soil class. In total, four datasets were used including the entire dataset (AllPT) and the pruned dataset consisting respectively of 80 %, 90 % and standard deviation core range of the Plinthosols data while cutting off all data points belonging to the outer range. The Plinthosol samples which were removed by pruning were latter predicted using the models developed for the respective train dataset. For the entire dataset (AllPT) as well as for each data subset, three groups of covariates consisting in (i) terrain parameters (TP), (ii) spectral parameters (SP) and (iii) terrain and spectral parameters (TSP) were considered for the prediction of the reference soil group (RSG).

No matter the Random Forest models, the predictions based on AllPT revealed an overestimation of the Plinthosols, which reduced the prediction accuracy of the remaining reference soil groups. This overestimation was independent of the group of covariates considered. However, about 3 to 41 % improvement in prediction accuracy was recorded when using different pruned datasets for the identification of reference soil groups. The best prediction was attained when removing all Plinthosol points

lower than 5 % and higher than 95 % of the cumulative percentage of the most important variable (wetness index) and modelling conducted solely with terrain and spectral parameters (TSP) with optimal predictors resulting from the RF_rfe. The resulting prediction model provided a substantial agreement to observation, with a kappa value of 0.57 along with a 35 % increase in prediction accuracy for Cambisols, 7 % for Gleysols and 16 % for Stagnosols. The pruned Plinthosol samples recorded a prediction accuracy varying between 15 % and 71 %. When combined, the terrain parameters took preeminence over the spectral bands and indices with the SAGA wetness index, a proxy for soil moisture distribution, being the most important variable contributing to the quality of the RF model. This study thus points to the potential of using data pruning to reduce the influence of a predominant reference soil group on the spatial prediction of smaller soil units in tropical environment.

(iii) Spatial controls of soil organic carbon stocks in the Sudanian savannah zone of Burkina Faso, West Africa

The ability to project and to mitigate the impacts of climate change is closely related to the evaluation of soil organic carbon (SOC) stocks across different types of land use and soil groups. Therefore, this study aimed at estimating the surface (0 - 30 cm) and subsoil (30 – 100 cm) organic carbon stocks in different land use systems and across various soil groups. A further aim was to assess the spatial variability of SOC stocks and factors affecting its distribution. About 70 soil profiles were considered along with additional auger (1205 samples) sampling to account for spatial variation in both cropland (CR) and savannah (SA). Mid-infrared spectroscopy and partial least-squares analysis were used as a fast and low-cost technique to handle the large amount of samples for the SOC content estimation. The machine learning technique Random Forest Regression (RFR) and multiple linear regression (MLR) were used for modelling the surface SOC stocks topsoil (0 - 30 cm). The covariates considered include topographic, texture along with climatic data used as surrogate for soil forming factors for model calibration. The prediction maps produced by the calibrated models were validated by an independent dataset.

Overall, about 53 % of the carbon stock over 1 m depth was held in the upper 30 cm and is proned to release upon non-sustainable management practices Only a marginal difference was recorded between the topsoil SOC stock in SA soils (41.4 t C ha^{-1}) and cropland soils (39.1 t C ha^{-1}). For the subsoil, a significant difference ($p < 0.05$) was observed for the SOC stock between the CR recording about 40.2 t C ha^{-1} and the SA with 26.3 t C ha^{-1}. Over 0 - 30 cm and 100 cm depth, Gleysols (44 t C ha^{-1} and 86.64 t C ha^{-1} respectively) located at lower elevation position stored the highest amount of SOC stock. The topsoil SOC stock variability was primarily affected by the silt content followed by the wetness index. Both RFR and MLR estimated mean top- SOC stocks of the catchment fairly well, with RFR being superior to MLR in terms of lower statistical error metrics. These findings reinforce the view that the semi-arid ecosystems of West Africa still offers a significant opportunity for carbon sequestration and these results represent a baseline for future carbon dynamics modelling in the region.

(iv) Carbon losses from prolonged arable cropping of Plinthosols in Southwest Burkina Faso

The conversion of natural ecosystems into agricultural land affects the atmospheric CO_2 concentration whose increase contributes to global warming. This study aimed at assessing these effects in Plinthosols, which are characterized by large contents of Fe oxides that are usually known to protect SOC from rapid decay. For that purpose, Plinthosols were sampled down to one meter (if feasible) that had been converted from native savannah into cropland 0 to 29 years ago in the Dano district (Southwest Burkina Faso). Beside the assessment of SOC stocks, the proportion of SOC remaining after Fe oxide removal was determined as well as its distribution among the following particle-size classes: 2000 - 250 µm (coarse sand-sized SOC; POM1), 250 µm – 53 µm (fine-sand-sized SOC; POM2), 53 µm – 20 µm (very fine sand-sized SOC; POM3), and < 20 µm (nonPOM).

The extent of change in SOC stock was found to vary with depth and the age of the cropland. A decrease in SOC stock of 24 t C ha^{-1} and 49 t C ha^{-1} were recorded for the upper 10 cm and 30 cm indicating that about 66 % ($p < 0.01$) and 55 % ($p < 0.01$)

of the initial stock in the native vegetation had been released respectively after 29 years of cultivation. SOC losses extended also into the subsoil, and, on the average, 0.7 to 19.5 t C ha^{-1} was lost from the 30 - 100 cm depth interval. About 52 % ($p < 0.01$) of SOC stock loss was recorded for the upper 100 cm after 29 years. Losses of SOC occurred in all soil fractions with mean residence time generally increasing with particle size. The Fe oxide was found to play a minor role as stabilizing agent as only 16 % (± 15 %) in average of the SOC stock was lost after treating the samples with dithionite-citrate-bicarbonate (DCB). Though most carbon was found as nonPOM, indicating that organo-mineral associations are a key parameter for carbon stabilization, its depletion with increasing cultivation duration suggests that the destruction of aggregates in these fields increased the vulnerability of this pool to microbial degradation. The loss rates of SOC were thus similar to those reported for other soil types, i.e., plinthite formation played only a minor role in stabilizing the remaining SOC from decomposition.

3. Synthesis

This study was motivated by the need to evaluate the impact of different category of covariates and statistical methods for DSM at catchment level as well as to investigate the SOC dynamics along a false chronosequence. The results of Chapter III, IV and V pointed out the potential of the application of DSM in predicting soil properties and reference soil groups. The resulting maps revealed the spatial variability of soil properties and reference soil groups while the models also provided insight into the key variables affecting their respective distribution. The question whether soils in the Dano catchment have potential or would function as a source or sink for carbon was elucidated in Session V and VI.

Sustainable land use and management require high resolution spatial information on soil properties for accurate decision and knowledge-based policies. The combination of high spatial resolution satellite (RapidEye and Landsat) along with terrain/climatic data resulted in better prediction accuracy of soil properties by the RF models. In assessing the models' performance inside and outside the the small catchment (modelling area), the performance statistics revealed that the machine learning

techniques provided marginal improvement in the different prediction. The lower performance of the MLR is attributed to its failure in accounting for non-linear relationships between response and predictor variables. The size and heterogeneity of the landscapes with varying surface characteristics due to various farm management practices and terrain attributes introduces complex relationships in the environmental variables which cannot be captured fully by linear models (Selige et al., 2006; Smith et al., 2013). Consequently, recommendation goes for non-parametric models such as RFR, support vector machines (SVM) and neural networks which were found superior to MLR for heterogeneous landscape (Hahn and Gloaguen, 2008b; Wålinder, 2014). However, for more homogeneous areas MLR is likely to provide good prediction accuracy.

For the high resolution mapping of the soil properties, the spectral data especially those acquired during ploughing or early crop development stages (e.g. May, June) were found to be the most important predictors in contrary to the trend observed for the RSG prediction. These findings indicate the strong impact of optimal timing for RS data acquisition for predicting soil properties. A timely acquired RS data along with terrain/climatic variables would therefore contribute in better prediction accuracy when models able to handle non-linear relationships are considered.

Predicting reference soil groups with a dataset subject to imbalance issues led to an overestimation of the dominant soil groups represented by the Plinthosols. The observed noises were due to the Plinthosols belonging to a wide range of predictors also shared by the smaller soil units. Only the pruned dataset with RF models including at least the terrain attributes resulted in a better expression of the smaller soil units in the corresponding maps. Consequently, pruning the majority class - the Plinthosols - by different methods while using Random Forest (RF) to evaluate the various datasets proved to be an efficient way for improving the prediction accuracy. This indicates that for areas where alternatives such as increasing the soil pedons with soil groups having lower observations (Brungard et al., 2015) would be costly and time consuming, pruning could be considered as a possible option.

Considering the variables used as surrogates for soil forming factors, only the combination of both terrain attributes and spectral data resulted in better prediction of the RSG along with the pruned dataset. However, the terrain attributes took preeminence over the spectral variables for the distribution of the RSG. Though the latter contradict the results of Scull et al. (2005), it confirms the finding of Dobos et al. (2001) and Stum (2010) who also recorded terrain attribute as playing the major role for discriminating soil units. The SAGA wetness index was the most prominent along with distance to stream, protection index and elevation for the top four variables. As outlined in section I, the SAGA wetness index coming as top variables suggests soil moisture distribution as the key factor discriminating among the RSG while the remaining top variables are playing a regulatory role. The RFR models then classifyied wet soil in low elevation and distance to stream area (Gleysols) and the dry soil (Leptosols) on high elevation and distance to stream areas while the remaining soil groups occupy intermediate position between these two groups. This spatial distribution of the different RSG is in agreement with expected soil–landscape relationships as described in the IUSS et al. (2006) and also confirmed by other studies assessing decision tree model ability for predicting soil classes (Brungard et al., 2015; Taghizadeh-Mehrjardi et al., 2012). In summary, the majority class data pruning resulted in an increase of prediction accuracies of the smaller soil units while using Random Forest (RF) as robust method to evaluate the various sets.

The quantification of soil organic carbon (SOC) stocks is of global concern as soils constitute the major C pool and could turn out as substantial sinks or sources for atmospheric CO_2. The results presented in Chapter V established that the SOC stocks are primarily (53 %) located in the topsoil (0 – 30 cm) which is within the range reported by Batjes (1996). The lower SOC stock in the topsoil of the CR confirmed the adverse effect of cultivation along with the removal of biomass which is not available for the built-up of organic matter in the soil. The significantly higher SOC stock in the subsoil of the CR was quite surprising but might be attributed to the relocation of SOC content and clay from the topsoil to lower layers under the heavy rains of the tropics. With the bare soil surface of the CR, the intensity of the impact of rainfall is expected to be higher compared to SA with higher amount of material being relocated.

The distribution of the SOC stock in the different RSG revealing Gleysols located at lower elevation area as having the highest amount suggest the impact of erosion processes with transport of sediment from higher location to lower areas. The silt content followed by sand, wetness index, elevation and temperature were found to be the five top variables. Since topography affects soil properties, the elevation determines the spatial distribution of silt and wetness index which as mentioned earlier is an indicator of soil moisture. SOC is related to silt via the physical and chemical protection it provides (Feller and Beare, 1997; Jones, 1973; McGrath and Zhang, 2003) while soil moisture distribution which depends on precipitation affects decomposition processes along with soil temperature (Parton et al., 1993).

In general, the accuracy of the prediction models were low though the RFR performed marginally better than the MLR with higher R^2 (13 %) and lower error metrics. The low explained variances are due to intrinsic high spatial variability of SOC with the interplay of complex and large range of factors at local and regional level which might not have been fully captured in the present study. For example, elements such as clay mineralogy and sesquioxides were not considered in the models. Moreover, errors related to field sampling, laboratory measurement, statistics as well as random errors could also play a role. However, other studies also recorded lower accuracies varying from 6 % to 43 % (Grimm et al., 2008; Henderson et al., 2005; Schulp et al., 2013; Schulp and Verburg, 2009) resulting mainly from the high spatial variability of the SOC. It is obvious that more investigation are required to improve the accuracy of DSM in highly heterogeneous landscape located in semi-arid tropical area.

The assessment of the impacts of LUC on SOC content and stock in the Plinthosols revealed a general decline with increasing land use duration (Chapter VI) for both topsoil (0 – 30 cm) and subsoil (30 – 100 cm). The study highlights that SOC in subsoil can also be disrupted as a result of LUC contrary to the general trend considering it as inert and insensitive. The topsoil labile fraction (POM) is more vulnerable to LUC as also recorded in previous studies (Liang et al., 2012; Yang et al., 2009). However, the fine sand fraction POM2 recording a smaller turnover rate with subsequent higher MRT is contrary to previous studies suggesting the existence of either chemically resistant material or of some organic coating protecting from

degradation (Christensen, 1992). The consideration of the functional group composition of these fractions could help shed further light for such pattern in SOC dynamics.

Many studies have pointed out the role of sesquioxides as key element affecting the stability of SOC (Barthès et al., 2008; Dalal and Bridge, 1996; Guggenberger and Haider, 2002). However, the results of this study (Chapter VI) could not establish Fe containing sesquioxides as major stabilizing agent of the carbon stock for the topsoil (0 – 30 cm). No correlation could be established between the SOC stock loss after DCB treatment with Fe oxide content. The high SOC stock observed in the nonPOM fraction (fine silt plus clay), showed that organo-mineral associations are the key parameter for carbon stabilization. However, ternary OC-Fe oxides-Silt plus Clay association could also be involved (Wagai and Mayer, 2007) alongside the metal oxides and clay (Silt plus Clay) individual contribution but this requires further investigation.

4. Outlook

Though the results of the present study offer indications that DSM of soil properties and reference soil groups has great potential in providing soil information at local level in data poor regions such as West Africa, the prediction accuracy of the different models still have to be improved. High inherent spatial variability in soil properties and the heterogeneity of the landscape are major reasons advanced for such performances of the models. However, prediction accuracy of the models could be increased by: (1) carrying out land surface segmentation (Drăguţ and Dornik, 2016) for the creation of homogeneous strata based on the identified most important variables – elevation, wetness index, distance to stream –spectral data of June- using Random Forest as model for prediction, and (2) by considering multi- or hyperscale terrain information to account for different spatial scales within one model (Behrens et al., 2010b; Behrens et al., 2010a; Behrens et al., 2014).

The present study only evaluated the impact of LUC on topsoil POM fractions while there is more and more evidence that subsoil POM C could also be affected (Sheng et al., 2015) but little is still known about the magnitude of the response of subsoil POM

for LAC soils in the tropical semi-arid regions. Further study could therefore quantify the extent of the impact of LUC for such deeper soil layers. While results from the present study also revealed high SOC in the nonPOM fraction (fine silt plus clay), the role of Fe containing sesquioxides in SOC stabilization was found to be poor. A further step would be to specifically assess possible stabilization processes including a direct assessment of the amount of SOC associated with Fe oxide and Al oxide and the amount held by ternary OC-Fe oxides- Clay plus Silt association along with clay occluded SOC. The purpose would be to find out whether the stability of SOC is more related to physical protection within stable aggregates or sorption to clay particles or to ternary OC-Fe oxides-Silt plus Clay association or whether multiple protective mechanisms are involved.

The results of this study can also be considered as a baseline work for modelling activities regarding SOC prediction coupled with climate change scenarios in the Dano catchment. Using false chronosequence approach with the remaining soil groups apart from the Plinthosols, the SOC pattern for the next 50 to 100 years under different climatic scenarios of the West Africa semi-arid regions can be further explored.

VIII.

References

References

Adhikari, D., Yang, Y., 2015. Selective stabilization of aliphatic organic carbon by iron oxide. Scientific reports 5.

Adhikari, K., Minasny, B., Greve, M.B., Greve, M.H., 2014. Constructing a soil class map of Denmark based on the FAO legend using digital techniques. Geoderma 214, 101–113.

Akpa, S.I.C., Odeh, I.O.A., Bishop, T.F.A., Hartemink, A.E., 2014. Digital mapping of soil particle-size fractions for Nigeria. Soil Science Society of America Journal 78 (6), 1953–1966.

Albaladejo, J., Ortiz, R., Garcia-Franco, N., Navarro, A., Almagro, M., Pintado, J., Martínez-Mena, M., 2013. Land use and climate change impacts on soil organic carbon stocks in semi-arid Spain. J Soils Sediments 13 (2), 265–277.

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.

Amado, Telmo Jorge Carneiro, Bayer, C., Conceição, P.C., Spagnollo, E., De Campos, Ben-Hur Costa, Da Veiga, M., 2006. Potential of carbon accumulation in no-till soils with intensive use and cover crops in southern Brazil. Journal of Environmental Quality 35 (4), 1599–1607.

Amare, T., Hergarten, C., Hurni, H., Wolfgramm, B., Yitaferu, B., Selassie, Y.G., 2013. Prediction of soil organic carbon for Ethiopian highlands using soil spectroscopy. ISRN Soil Science 2013.

Amelung, W., Lobe, I., Du Preez, C. C., 2002. Fate of microbial residues in sandy soils of the South African Highveld as influenced by prolonged arable cropping. European Journal of Soil Science 53 (1), 29–35.

Amelung, W., Miltner, A., Zhang, X., Zech, W., 2001. Fate of microbial residues during litter decomposition as affected by minerals. Soil Science 166 (9), 598–606.

Amelung, W., Zech, W., 1999. Minimisation of organic matter disruption during particle-size fractionation of grassland epipedons. Geoderma 92, 73–85.

Amelung, W., Zech, W., Zhang, X., Follett, R.F., Tiessen, H., Knox, E., Flach, K.-W., 1998. Carbon, nitrogen, and sulfur pools in particle-size fractions as influenced by climate. Soil Science Society of America Journal 62 (1), 172–181.

Anikwe, M., 2010. Carbon storage in soils of Southeastern Nigeria under different management practices. Carbon Balance and Management 5 (1), 5.

Arrouays, D., Balesdent, J., Mariotti, A., Girardin, C., 1995. Modelling organic carbon turnover in cleared temperate forest soils converted to maize cropping by using 13C natural abundance measurements. Plant Soil 173 (2), 191–196.

Arrouays, D., McKenzie, N., Hempel, J., de Forges, A., McBratney, A.B., 2014. GlobalSoilMap: basis of the global spatial soil information system. CRC press.

Ashagrie, Y., Zech, W., Guggenberger, G., 2005. Transformation of a Podocarpus falcatus dominated natural forest into a monoculture Eucalyptus globulus plantation

at Munesa, Ethiopia: soil organic C, N and S dynamics in primary particle and aggregate-size fractions. Agriculture, Ecosystems & Environment 106 (1), 89–98.

Azlan, A., Aweng, E.R., Ibrahim, C.O., 2011. The Correlation Between Total Organic Carbon (TOC), Organic Matter and Water Content in Soil Collected from Different Land Use of Kota Bharu, Kelantan. Journal of Applied Sciences Research 7 (7), 915.

Baldock, J.A., Skjemstad, J.O., 2000. Role of the soil matrix and minerals in protecting natural organic materials against biological attack. Organic Geochemistry 31 (7), 697–710.

Balesdent, J., Besnard, E., Arrouays, D., Chenu, C., 1998. The dynamics of carbon in particle-size fractions of soil in a forest-cultivation sequence. Plant Soil 201 (1), 49–57.

Balesdent, J., Wagner, G.H., Mariotti, A., 1988. Soil organic matter turnover in long-term field experiments as revealed by carbon-13 natural abundance. Soil Science Society of America Journal 52 (1), 118–124.

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

Barthès, B.G., Kouakoua, E., Larré-Larrouy, M.-C., Razafimbelo, T.M., de Luca, Edgar F, Azontonde, A., Neves, C.S., de Freitas, Pedro L, Feller, C.L., 2008. Texture and sesquioxide effects on water-stable aggregates and organic matter in some tropical soils. Geoderma 143 (1), 14–25.

Basile-Doelsch, I., Brun, T., Borschneck, D., Masion, A., Marol, C., Balesdent, J., 2009. Effect of landuse on organic matter stabilized in organomineral complexes: a study combining density fractionation, mineralogy and δ 13 C. Geoderma 151 (3), 77–86.

Bationo, A., Buerkert, A., 2001. Soil organic carbon management for sustainable land use in Sudano-Sahelian West Africa. Nutrient Cycling in Agroecosystems 61 (1-2), 131–142.

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.

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.

Batjes, N.H., 1996. Total carbon and nitrogen in the soils of the world. European Journal of Soil Science, 47, 15 1 - 163.

Batjes, N.H., 2004. Soil carbon stocks and projected changes according to land use and management: a case study for Kenya. Soil Use and Management 20 (3), 350–356.

Batjes, N.H., 2008. Mapping soil carbon stocks of Central Africa using SOTER. Geoderma 146 (1–2), 58–65.

Batjes, N.H., Sombroek, W.G., 1997. Possibilities for carbon sequestration in tropical and sub-tropical soils. Global Change Biology 3, 161–173.

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

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.

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

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

Besnard, E., Chenu, C., Balesdent, J., Puget, P., Arrouays, D., 1996. Fate of particulate organic matter in soil aggregates during cultivation. European Journal of Soil Science 47 (4), 495–503.

Birch-Thomsen, T., Elberling, B., Fog, B., Magid, J., 2007. Temporal and spatial trends in soil organic carbon stocks following maize cultivation in semi-arid Tanzania, East Africa. Nutrient Cycling in Agroecosystems 79 (3), 291–302.

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.

Blécourt, M. de, Brumme, R., Xu, J., Corre, M.D., Veldkamp, E., 2013. Soil carbon stocks decrease following conversion of secondary forests to rubber (Hevea brasiliensis) plantations. PLoS ONE 8 (7), e69357.

Blin, J., Sidibe, S., 2012. Caractérisation et amélioration d'un foyer de cuisson de" dolo" équipé d'un brûleur à huile végétale (jatropha).

Blum, W.E.H., 2005. Functions of soil for society and the environment. Reviews in Environmental Science and Bio/Technology 4 (3), 75–79.

Bornemann, L., Herbst, M., Welp, G., Vereecken, H., Amelung, W., 2011. Rock fragments control size and saturation of organic carbon pools in agricultural topsoil. Soil Science Society of America Journal 75 (5), 1898–1907.

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.

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

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

Brevik, E., 2013. The Potential Impact of Climate Change on Soil Properties and Processes and Corresponding Influence on Food Security. Agriculture 2013 3, 398–417.

Bricklemyer, R.S., Lawrence, R.L., Miller, P.R., Battogtokh, N., 2007. Monitoring and verifying agricultural practices related to soil carbon sequestration with satellite imagery. Agriculture, Ecosystems & Environment 118 (1), 201–210.

Brodowski, S., Amelung, W., Haumaier, L., Zech, W., 2007. Black carbon contribution to stable humus in German arable soils. Geoderma 139 (1), 220–228.
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.
Brown, D.J., 2005. A historical perspective on soil-landscape modeling. Environmental Soil-Landscape Modeling: Geographic Information Technologies and Pedometrics, 61–103.
Brungard, C.W., Boettinger, J.L., Duniway, M.C., Wills, S.A., Edwards Jr. , Thomas C., 2015. Machine learning for predicting soil classes in three semi-arid landscapes. Geoderma 239–240, 68–83.
Bruun, T.B., Elberling, B., Neergaard, A., Magid, J., 2013. Organic carbon dynamics in different soil types after conversion of forest to agriculture. Land Degradation & Development 26 (3), 272–283.
Bureau National des sols, 2000. Etude morphologiaue des provinces de la Bougouriba et du Ioba, Echelle 1/100 000, Rapport technique n° 121, Burkina Faso.
Burgess, T.M., Webster, R., 1980. Optimal interpolation and isar1thmic mapping of soil properties. European Journal of Soil Science 31 (2), 315–331.
Burke, I.C., Yonker, C.M., Parton, W.J., Cole, C.V., Schimel, D.S., Flach, K., 1989. Texture, Climate, and Cultivation Effects on Soil Organic Matter Content in U.S. Grassland Soils. Soil Sci. Soc. Am. J. 53 (3), 800–805.
Burrough, P.A., Beckett, P.H., Jarvis, M.G., 1971. The relation between cost and utility in soil survey (i–iii) 1. Journal of Soil Science 22 (3), 359–394.
Burrough, P.A., McDonnell, R.A., 1998. Principles of Geographic Information Systems: Spatial Information Systems and Geostatistics. Oxford University Press, New York.
Burrough, P.A., van Gaans, Pauline FM, Hootsmans, R., 1997. Continuous classification in soil survey: spatial correlation, confusion and boundaries. Geoderma 77 (2), 115–135.
Callo-Concha, D., Gaiser, T., Ewert, F., 2012a. Farming and cropping systems in the West African Sudanian Savanna. WASCAL research area: Northern Ghana, Southwest Burkina Faso and Northern Benin. Econstor ZEF Working Paper Series 100, 1–49.
Callo-Concha, D., Gaiser, T., Ewert, F., 2012b. Farming and cropping systems in the West African Sudanian Savanna. WASCAL research area: Northern Ghana, Southwest Burkina Faso and Northern Benin. No. 100. ZEF Working Paper Series, 2012.
Cambardella, C.A., Elliott, E.T., 1993. Methods for physical separation and characterization of soil organic matter fractions. Geoderma 56 (1-4), 449–457.
Chan, K.Y., 2001. Soil particulate organic carbon under different land use and management. Soil Use and Management 17 (4), 217–221.
Chandran, P., Ray, S.K., Durge, S.L., Raja, P., Am Nimkar, Bhattacharyya, T., Pal, D.K., 2009. Scope of horticultural land-use system in enhancing carbon

sequestration in ferruginous soils of the semi-arid tropics. Current science 97 (7), 1039.

Chaplot, V., Bouahom, B., Valentin, C., 2010. Soil organic carbon stocks in Laos: spatial variations and controlling factors. Global Change Biology 16 (4), 1380–1393.

Chefetz, B., Tarchitzky, J., Deshmukh, A.P., Hatcher, P.G., Chen, Y., 2002. Structural characterization of soil organic matter and humic acids in particle-size fractions of an agricultural soil. Soil Science Society of America Journal 66 (1), 129–141.

Chorover, J., Amistadi, M.K., 2001. Reaction of forest floor organic matter at goethite, birnessite and smectite surfaces. Geochimica et Cosmochimica Acta 65 (1), 95–109.

Christensen, B.T., 1992. Physical fractionation of soil and organic matter in primary particle size and density separates, in: , Advances in soil science. Springer, pp. 1–90.

Christensen, B.T., 1996. Carbon in primary and secondary organomineral complexes. Structure and organic matter storage in agricultural soils, 97–165.

Ciais, P., Bombelli, A., Williams, M., Piao, S.L., Chave, J., Ryan, C.M., Henry, M., Brender, P., Valentini, R., 2011. The carbon balance of Africa: synthesis of recent research studies. Philosophical transactions - Royal Society. Mathematical, Physical and engineering sciences 369 (1943), 2038–2057.

Coleman, T.L., Agbu, P.A., Montgomery, O.L., Gao, T., Prasad, S., 1991. Spectral band selection for quantifying selected properties in highly weathered soils. Soil Science 151 (5), 355–361.

Congalton, R.G., Green, K., 2008. Assessing the accuracy of remotely sensed data: principles and practices. CRC press.

Coutinho, H.L.C., Noellemeyer, E., Carvalho Balieiro, F. de, Piñeiro, G., Fidalgo, E.C.C., Martius, C., da Silva, C.F., 2014. 21 Impacts of Land-use Change on Carbon Stocks and Dynamics in Central-southern South American Biomes: Cerrado, Atlantic Forest and Southern Grasslands. Soil Carbon: Science, Management and Policy for Multiple Benefits 71, 243.

Cox, P.M., Betts, R.A., Jones, C.D., Spall, S.A., Totterdell, I.J., 2000. Acceleration of global warming due to carbon-cycle feedbacks in a coupled climate model. Nature 408 (6809), 184–187.

Dalal, R.C., Bridge, B.J., 1996. Aggregation and organic matter storage in sub-humid and semi-arid soils. Structure and organic matter storage in agricultural soils, 263–307.

Dalal, R.C., Mayer, R.J., 1986. Long term trends in fertility of soils under continuous cultivation and cereal cropping in southern Queensland. II. Total organic carbon and its rate of loss from the soil profile. Soil Research 24 (2), 281–292.

Davidson, E.A., Ackerman, I.L., 1993. Changes in soil carbon inventories following cultivation of previously untilled soils. Biogeochemistry 20 (3), 161–193.

Davidson, E.A., Janssens, I.A., 2006. Temperature sensitivity of soil carbon decomposition and feedbacks to climate change. Nature 440.

Degryze, S., Six, J., Paustian, K., Morris, S.J., Paul, E.A., Merckx, R., 2004. Soil organic carbon pool changes following land-use conversions. Global Change Biology 10 (7), 1120–1132.

Denef, K., Six, J., 2005. Clay mineralogy determines the importance of biological versus abiotic processes for macroaggregate formation and stabilization. European Journal of Soil Science 56 (4), 469–479.

Derrien, D., Amelung, W., 2011. Computing the mean residence time of soil carbon fractions using stable isotopes: impacts of the model framework. European Journal of Soil Science 62 (2), 237–252.

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.

Doetterl, S., Stevens, A., van Oost, K., Quine, T.A., van Wesemael, B., 2013. Spatially-explicit regional-scale prediction of soil organic carbon stocks in cropland using environmental variables and mixed model approaches. Geoderma 204, 31–42.

Don, A., Schumacher, J., Freibauer, A., 2011. Impact of tropical land-use change on soil organic carbon stocks – a meta-analysis. Global Change Biology 17 (4), 1658–1670.

Don, A., Schumacher, J., Scherer-Lorenzen, M., Scholten, T., Schulze, E.-D., 2007. Spatial and vertical variation of soil carbon at two grassland sites—implications for measuring soil carbon stocks. Geoderma 141 (3), 272–282.

Doquire, G., Verleysen, M. (Eds.), 2011. Graph Laplacian for semi-supervised feature selection in regression problems. Springer, 248-255.

Doraiswamy, P.C., McCarty, G.W., Hunt Jr, E. R., Yost, R.S., Doumbia, M., Franzluebbers, A.J., 2007. Modeling soil carbon sequestration in agricultural lands of Mali. Agricultural Systems 94 (1), 63–74.

Douxchamps, S., Ayantunde, A.A., 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).

Drăguţ, 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.

Elberling, B., Touré, A., Rasmussen, K., 2003. Changes in soil organic matter following groundnut–millet cropping at three locations in semi-arid Senegal, West Africa. Agriculture, Ecosystems & Environment 96 (1), 37–47.

Ellis, E.C., Klein Goldewijk, K., Siebert, S., Lightman, D., Ramankutty, N., 2010. Anthropogenic transformation of the biomes, 1700 to 2000. Global Ecology and Biogeography 19 (5), 589–606.

Ertekin, S., Huang, J., Bottou, L., Giles, L. (Eds.), 2007. Learning on the border: active learning in imbalanced data classification. ACM, 127-136.

Eze, P.N., Udeigwe, T.K., Meadows, M.E., 2014. Plinthite and its associated evolutionary forms in soils and landscapes: a review. Pedosphere 24 (2), 153–166.
FAO, 2004. FAOSTAT, Food and Agriculture Organization of the United Nations. [Available online at http://apps.fao.org.].
Farr, T.G., Kobrick, M., 2000. Shuttle Radar Topography Mission produces a wealth of data. Eos, Transactions American Geophysical Union 81 (48), 583–585.
Fassnacht, F.E., 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.
Feller, C., Beare, M.H., 1997. Physical control of soil organic matter dynamics in the tropics. Geoderma 79 (1–4), 69–116.
Fierer, N., Allen, A.S., Schimel, J.P., Holden, P.A., 2003. Controls on microbial CO2 production: a comparison of surface and subsurface soil horizons. Global Change Biology 9 (9), 1322–1332.
Florinsky, I.V., Eilers, R.G., Manning, G.R., Fuller, L.G., 2002. Prediction of soil properties by digital terrain modelling. Environmental Modelling & Software 17 (3), 295–311.
Foley, J.A., DeFries, R., Asner, G.P., Barford, C., Bonan, G., Carpenter, S.R., Chapin, F.S., Coe, M.T., Daily, G.C., Gibbs, H.K., 2005. Global consequences of land use. Science 309 (5734), 570–574.
Fontaine, S., Barot, S., Barre, P., Bdioui, N., Mary, B., Rumpel, C., 2007. Stability of organic carbon in deep soil layers controlled by fresh carbon supply. Nature 450 (7167), 277–280.
Forkuor, G., 2014. Agricultural Land Use Mapping in West Africa Using Multi-sensor Satellite Imagery. University of Wuerzburg: Wuerzburg, Germany, p.191.
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.
Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. Annals of Statistics, 1189–1232.
Friedman, J.H., 2002. Stochastic gradient boosting. Computational Statistics & Data Analysis 38 (4), 367–378.
Genuer, R., Poggi, J.-M., Tuleau-Malot, C., 2010. Variable selection using random forests. Pattern Recognition Letters 31 (14), 2225–2236.
Giasson, E., Caten, A.t., Bagatini, T., Bonfatti, B., 2015. Instance selection in digital soil mapping: a study case in Rio Grande do Sul, Brazil. Ciência Rural 45 (9), 1592–1598.
Gislason, P.O., Benediktsson, J.A., Sveinsson, J.R., 2006. Random forests for land cover classification. Pattern Recognition Letters 27 (4), 294–300.
Golchin, A., Oades, J.M., Skjemstad, J.O., Clarke, P., 1994. Soil structure and carbon cycling. Soil Research 32 (5), 1043–1068.

Gong, Z., Zhang, X., Chen, J., Zhang, G., 2003. Origin and development of soil science in ancient China. Geoderma 115 (1), 3–13.
Gonin, A., Tallet, B., 2012. Changements spatiaux et pratiques pastorales: les nouvelles voies de la transhumance dans l'Ouest du Burkina Faso. Cahiers Agricultures 21 (6), 448–454.
Goovaerts, P., 1999. Geostatistics in soil science: state-of-the-art and perspectives. Geoderma 89 (1), 1–45.
Gopi, S.C., Suvarna, B., Padmaja, T.M., 2016. High Dimensional Unbalanced Data Classification Vs SVM Feature Selection. Indian Journal of Science and Technology 9 (30).
Gregorich, E.G., Monreal, C.M., Ellert, B.H., 1995. Turnover of soil organic matter and storage of corn residue carbon estimated from natural 13C abundance. Canadian journal of soil science 75 (2), 161–167.
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.
Grinblat, Y., Kidron, G.J., Karnieli, A., Benenson, I., 2015. Simulating land-use degradation in West Africa with the ALADYN model. Journal of Arid Environments 112, 52–63.
Grunwald, S. (Ed.), 2006. Environmental soil-Landscape modeling: Geographic information technologies and pedometrics: What do we really know about the space-time continuum of soil-landscapes? CRC Press.
Guggenberger, G., Christensen, B.T., Zech, W., 1994. Land-use effects on the composition of organic matter in particle-size separates of soil: I. Lignin and carbohydrate signature. European Journal of Soil Science 45 (4), 449–458.
Guggenberger, G., Haider, K.M., 2002. Effect of mineral colloids on biogeochemical cycling of C, N, P and S in soil. IUPAC SERIES ON ANALYTICAL AND PHYSICAL CHEMISTRY OF ENVIRONMENTAL SYSTEMS 8, 267–322.
Guo, L.B., Gifford, R.M., 2002a. Soil carbon stocks and land use change: a meta analysis. Global Change Biology 8 (4), 345–360.
Guo, L.B., Gifford, R.M., 2002b. Soil carbon stocks and land use change: a meta analysis. Global Change Biology 8 (4), 345–360.
Guo, P.-T., Li, M.-F., Luo, W., Tang, Q.-F., Liu, Z.-W., Lin, Z.-M., 2015. Digital mapping of soil organic matter for rubber plantation at regional scale: An application of random forest plus residuals kriging approach. Geoderma 237, 49–59.
Hahn, C., Gloaguen, R., 2008a. Estimation of soil types by non linear analysis of remote sensing data. Nonlinear Processes in Geophysics 15 (1), 115–126.
Hahn, C., Gloaguen, R., 2008b. Estimation of soil types by non linear analysis of remote sensing data. Nonlinear Processes in Geophysics 15 (1), 115–126.
Hamidime, S., 2003. Etude de la végétation ligneuse associée aux lieux de cultes du terroir du village de Djikologo en pays Dagara (province de Ioba): rapport de stage de deuxième année de l'Institut de Développement Rural. : rapport de stage de

deuxième année de l'Institut de Développement Rural. Bobo Dioulasso (BKF) ; Bobo Dioulasso : IDR ; IRD, 2003, 45 p. multigr. Mém. ing., IDR.

Handl, J., Knowles, J., 2006. Feature subset selection in unsupervised learning via multiobjective optimization. International Journal of Computational Intelligence Research 2 (3), 217–238.

Hartemink, A.E., 2006. Assessing soil fertility decline in the tropics using soil chemical data. Advances in Agronomy 89, 179–225.

Hartemink, A.E., Krasilnikov, P., Bockheim, J.G., 2013. Soil maps of the world. Geoderma 207, 256–267.

Hartge, K., Horn, R., 1989. Die physikalische Untersuchung von Böden. Enke Verlag, Stuttgart.

Hastie, T., Tibshirani, R.J., Friedman, J.H., 2011. The elements of statistical learning: data mining, inference, and prediction. Springer.

Henderson, B.L., Bui, E.N., Moran, C.J., Simon, D.A., 2005. Australia-wide predictions of soil properties using decision trees. Geoderma 124 (3), 383–398.

Hengl, T., Heuvelink, Gerard B. M, Kempen, B., Leenaars, Johan G. B., Walsh, M.G., Shepherd, K.D., Sila, A., MacMillan, R.A., Mendes de Jesus, Jorge, Tamene, L., Tondoh, J.E., 2015. Mapping soil properties of Africa at 250 m resolution: Random Forests significantly improve current predictions. PLoS ONE 10 (6), e0125814 EP -.

Henry, M., Valentini, R., Bernoux, M., 2009. Soil carbon stocks in ecoregions of Africa. Biogeosciences Discussions 6 (1), 797–823.

Heung, B., Bulmer, C.E., Schmidt, M.G., 2014. Predictive soil parent material mapping at a regional-scale: a random forest approach. Geoderma 214, 141–154.

Heuvelink, G.B., Webster, R., 2001. Modelling soil variation: past, present, and future. Geoderma 100 (3), 269–301.

Heuvelink, Gerard B. M., Huisman, J.A., 2000. Choosing between abrupt and gradual spatial variation. Quantifying Spatial Uncertainty in Natural Resources: Theory and Applications for GIS and Remote Sensing, 111–117.

Hijmans, R.J., Cameron, S.E., Parra, J.L., Jones, P.G., Jarvis, A., 2005a. Very high resolution interpolated climate surfaces for global land areas. International Journal of Climatology 25 (15), 1965–1978.

Hijmans, R.J., Cameron, S.E., Parra, J.L., Jones, P.G., Jarvis, A., 2005b. Very high resolution interpolated climate surfaces for global land areas. International Journal of Climatology 25 (15), 1965–1978.

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.

Holmes, A., Müller, K., Clothier, B., Deurer, M., 2015. Carbon sequestration in kiwifruit orchard soils at depth to mitigate carbon emissions. Communications in soil science and plant analysis 46 (sup1), 122–136.

Hottin, G., Ouedraogo, O.F., 1992. Carte Géologique du Burkina Faso. - 2. édition, (B.M.G.B.). Échelle: 1: 1.000.000.

Houghton, R.A., 2003. Revised estimates of the annual net flux of carbon to the atmosphere from changes in land use and land management 1850–2000. Tellus B 55 (2), 378–390.

Huete, A., Didan, K., Miura, T., Rodriguez, P.E., Gao, X., Ferreira, L.G., 2002. Overview of the radiometric and biophysical performance of the MODIS vegetation indices. Remote Sensing of Environment 80 (1-2), 195–213.

Irons, J.R., Dwyer, J.L., Barsi, J.A., 2012. The next Landsat satellite: The Landsat data continuity mission. Remote Sensing of Environment 122, 11–21.

ISO 10694:1995. Soil quality – Determination of organic and total carbon after dry combustion (elemental analysis). Beuth, Berlin.

IUSS, ISRIC, FAO, 2006. World reference base for soil resources-a framework for international classification, correlation and communication. World Soil Resources, Report 103 FAO, Rome, Italy.

Jenny, H., 1941. Factors of soil formation: A system of quantitative pedology, 281 pp. McGraw-Hill, New York.

Jobbágy, E.G., Jackson, R., 2000. The vertical distribution of soil organic carbon and its relation to climate and vegetation. Ecological Applications 10 (2), 423.

John, G.H., Kohavi, R., Pfleger, K. (Eds.), 1994. Irrelevant features and the subset selection problem, 121-129.

Jones, A., Breuning-Madsen, H., Brossard, M., Dampha, A., Deckers, J., Dewitte, O., Gallali, T., Hallett, S., Jones, R., Kilasara, M., 2013. Soil atlas of Africa. European Commission.

Jones, M.J., 1973. The organic matter content of the savanna soils of West Africa. Journal of Soil Science 24 (1), 42–53.

Juo, A.S., Adams, F., 1984. Chemistry of LAC soils. Low Activity Clay (LAC) Soils, 37.

Kaiser, K., Guggenberger, G., 2007. Sorptive stabilization of organic matter by microporous goethite: sorption into small pores vs. surface complexation. European Journal of Soil Science 58 (1), 45–59.

Kalbitz, K., Schwesig, D., Rethemeyer, J., Matzner, E., 2005. Stabilization of dissolved organic matter by sorption to the mineral soil. Soil Biology and Biochemistry 37 (7), 1319–1331.

Kamara, A.Y., Ekeleme, F., Chikoye, D., Omoigui, L.O., 2009. Planting date and cultivar effects on grain yield in dryland corn production. Agronomy Journal 101, 91–98.

Kilasara, M. (Ed.), 2010. Selection and use of soil characteristics in digital soil mapping in Tanzania.

Kögel-Knabner, I., Amelung, W., 2014. Dynamics, chemistry, and preservation of organic matter in soils. Reference Module in Earth Systems and Environmental Sciences, from Treatise on Geochemistry 12, 157-215I.

Kögel-Knabner, I., Guggenberger, G., Kleber, M., Kandeler, E., Kalbitz, K., Scheu, S., Eusterhues, K., Leinweber, P., 2008. Organo-mineral associations in temperate

soils: Integrating biology, mineralogy, and organic matter chemistry. Journal of Plant Nutrition and Soil Science 171 (1), 61–82.
Krupenikov, I.A., Tedrow, J.C., 1994. History of Soil Science from Its Inception to the Present. Soil Science 158 (4), 301.
Kuhn, M., 2008. Building predictive models in R using the caret package. Journal of Statistical Software 28 (5), 1–26.
Kuhn, M., 2015. Caret: classification and regression training. Astrophysics Source Code Library 1, 5003.
Kuhn, M., Johnson, K., 2013. Applied predictive modeling 26. Springer.
Kumar, N., 2009. Investigating the potentiality of regression kriging in the estimation of soil organic carbon versus the extracted result from the existing soil map, 99 pp.
Ladd, B., Laffan, S.W., Amelung, W., Peri, P.L., Lucas C. R. Silva, Gervassi, P., Bonser, S.P., Navall, M., Sheil, D., 2013. Estimates of soil carbon concentration in tropical and temperate forest and woodland from available GIS data on three continent. Global Ecology and Biogeograpgy 22, 461–469.
Lagacherie, P., Holmes, S., 1997. Addressing geographical data errors in a classification tree for soil unit prediction. International Journal of Geographical Information Science 11 (2), 183–198.
Lahmar, R., Bationo, B.A., Lamso, N.D., 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.
Lal, R., 2004. Soil carbon sequestration impacts on global climate change and food security. Science 304 (5677), 1623–1627.
Lal, R., 2008. Carbon sequestration. Philosophical Transactions of the Royal Society B: Biological Sciences 363 (1492), 815–830.
Lalonde, K., Mucci, A., Ouellet, A., Gélinas, Y., 2012. Preservation of organic matter in sediments promoted by iron. Nature 483 (7388), 198–200.
Lark, R.M., 1995. Components of accuracy of maps with special reference to discriminant analysis on remote sensor data. International Journal of Remote Sensing 16 (8), 1461–1480.
Laslett, G.M., McBratney, A.B., Pahl, P., Hutchinson, M.F., 1987. Comparison of several spatial prediction methods for soil pH. Journal of Soil Science 38 (2), 325–341.
Le Song, Smola, A., Gretton, A., Bedo, J., Borgwardt, K., 2012. Feature selection via dependence maximization. Journal of Machine Learning Research 13 (May), 1393–1434.
Li, J., Heap, A.D., 2008. A review of spatial interpolation methods for environmental scientists, 137–145.
Li, J., Heap, A.D., 2014. Spatial interpolation methods applied in the environmental sciences: A review. Environmental Modelling & Software 53, 173–189.
Liang, Q., Chen, H., Gong, Y., Fan, M., Yang, H., Lal, R., Kuzyakov, Y., 2012. Effects of 15 years of manure and inorganic fertilizers on soil organic carbon

fractions in a wheat-maize system in the North China Plain. Nutrient Cycling in Agroecosystems 92 (1), 21–33.
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.
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.
Liu, H., Motoda, H., 2001. Data reduction via instance selection, in: , Instance selection and construction for data mining. Springer, pp. 3–20.
Liu, Z., Shao, M., Wang, Y., 2011. Effect of environmental factors on regional soil organic carbon stocks across the Loess Plateau region, China. Agriculture, Ecosystems & Environment 142 (3), 184–194.
Lobe, I., Amelung, W., Du Preez, C. C., 2001. Losses of carbon and nitrogen with prolonged arable cropping from sandy soils of the South African Highveld. European Journal of Soil Science 52 (1), 93–101.
Lobe, I., Du Preez, C. C., Amelung, W., 2002. Influence of prolonged arable cropping on lignin compounds in sandy soils of the South African Highveld. European Journal of Soil Science 53 (4), 553–562.
Lorenz, K., Lal, R., 2005. The depth distribution of soil organic carbon in relation to land use and management and the potential of carbon sequestration in subsoil horizons. Advances in Agronomy 88, 35–66.
Lützow, M.v., Kögel-Knabner, I., Ekschmitt, K., Matzner, E., Guggenberger, G., Marschner, B., Flessa, H., 2006. Stabilization of organic matter in temperate soils: mechanisms and their relevance under different soil conditions–a review. European Journal of Soil Science 57 (4), 426–445.
Lützow, M. von, Kögel-Knabner, I., Ekschmitt, K., Flessa, H., Guggenberger, G., Matzner, E., Marschner, B., 2007. SOM fractionation methods: relevance to functional pools and to stabilization mechanisms. Soil Biology and Biochemistry 39 (9), 2183–2207.
Lützow, M. von, Kögel-Knabner, I., Ludwig, B., Matzner, E., Flessa, H., Ekschmitt, K., Guggenberger, G., Marschner, B., Kalbitz, K., 2008. Stabilization mechanisms of organic matter in four temperate soils: development and application of a conceptual model. Journal of Plant Nutrition and Soil Science 171 (1), 111–124.
Ma, W., Tan, K., Du, P. (Eds.), 2016. Predicting soil heavy metal based on Random Forest model. IEEE, 4331-4334.
Makridakis, S., Hibon, M., 2000. The M3-Competition: results, conclusions and implications. International journal of forecasting 16 (4), 451–476.
Malone, B.P., 2012. Practicable methodologies for delivering comprehensive spatial soils information. Dissertation. The University of Sydney, p. 265.
Malone, B.P., Jha, S.K., Minasny, B., McBratney, A.B., 2016. Comparing regression-based digital soil mapping and multiple-point geostatistics for the spatial extrapolation of soil data. Geoderma 262, 243–253.

Malone, B.P., McBratney, A.B., Minasny, B., Laslett, G.M., 2009. Mapping continuous depth functions of soil carbon storage and available water capacity. Geoderma 154 (1), 138–152.

Manning, P., Vries, F.T., Tallowin, J.R.B., Smith, R., Mortimer, S.R., Pilgrim, E.S., Harrison, K.A., Wright, D.G., Quirk, H., Benson, J., 2015. Simple measures of climate, soil properties and plant traits predict national-scale grassland soil carbon stocks. Journal of Applied Ecology 52 (5), 1188–1196.

Mansuy, N., Thiffault, E., Paré, D., Bernier, P., Guindon, L., Villemaire, P., Poirier, V., Beaudoin, A., 2014. Digital mapping of soil properties in Canadian managed forests at 250m of resolution using the k-nearest neighbor method. Geoderma 235, 59–73.

Mao, D.H., Wang, Z.M., Li, L., Miao, Z.H., Ma, W.H., Song, C.C., Ren, C.Y., Jia, M.M., 2015. Soil organic carbon in the Sanjiang Plain of China: storage, distribution and controlling factors. Biogeosciences 12 (6), 1635–1645.

Mayr, T., Rivas-Casado, M., Bellamy, P., Palmer, R., Zawadzka, J., Corstanje, R., 2010. Two methods for using legacy data in digital soil mapping, in: , Digital Soil Mapping. Springer, pp. 191–202.

McBratney, A.B., Minasny, B., MacMillan, R.A. and Carre, 2011. Digital soil mapping, in: P.M. Huang, Y. Li, M.E. Sumner (Eds.), Handbook of Soil Sciences: Properties and Processes. CRC Press, Boca Raton, FL, 37:1-43.

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

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

McCarty, G.W., Reeves, J.B., 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.

McDonagh, J.F., Thomsen, T.B., Magid, J., 2001. Soil organic matter decline and compositional change associated with cereal cropping in southern Tanzania. Land Degradation & Development 12 (1), 13–26.

McGrath, D., Zhang, C., 2003. Spatial distribution of soil organic carbon concentrations in grassland of Ireland. Applied Geochemistry 18 (10), 1629–1639.

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

Meersmans, J., Ridder, F. de, Canters, F., Baets, S. de, van Molle, M., 2008. A multiple regression approach to assess the spatial distribution of Soil Organic Carbon (SOC) at the regional scale (Flanders, Belgium). Geoderma 143 (1), 1–13.

Meshesha, D.T., Tsunekawa, A., Tsubo, M., 2012. Continuing land degradation: cause–effect in Ethiopia's Central Rift Valley. Land Degradation & Development 23 (2), 130–143.

Miltner, A., Zech, W., 1998. Carbohydrate decomposition in beech litter as influenced by aluminium, iron and manganese oxides. Soil Biology and Biochemistry 30 (1), 1–7.

Minasny, B., Malone, B., McBratney, A.B. (Eds.), 2012. Digital Soil Assessments and Beyond. Taylor & Francis Group, London.

Mobley, M.L., Lajtha, K., Kramer, M.G., Bacon, A.R., Heine, P.R., Richter, D.D., 2015. Surficial gains and subsoil losses of soil carbon and nitrogen during secondary forest development. Global Change Biology 21 (2), 986–996.

Moni, C., Derrien, D., Hatton, P.-J., Zeller, B., Kleber, M., 2012. Density fractions versus size separates: does physical fractionation isolate functional soil compartments? Biogeosciences 9, 5181–5197.

Moran, C.J., Bui, E.N., 2002. Spatial data mining for enhanced soil map modelling. International Journal of Geographical Information Science 16 (6), 533–549.

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

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

Muñoz-Rojas, M., Jordán, A., Zavala, L.M., De la Rosa, D, Abd-Elmabod, S.K., Anaya-Romero, M., 2015. Impact of land use and land cover changes on organic carbon stocks in Mediterranean soils (1956–2007). Land Degradation & Development 26 (2), 168–179.

Murty, D., Kirschbaum, Miko U. F., Mcmurtrie, R.E., Mcgilvray, H., 2002. Does conversion of forest to agricultural land change soil carbon and nitrogen? A review of the literature. Global Change Biology 8 (2), 105–123.

Nadeu, E., Quiñonero-Rubio, J.M., Vente, J. de, Boix-Fayos, C., 2015. The influence of catchment morphology, lithology and land use on soil organic carbon export in a Mediterranean mountain region. Catena 126, 117–125.

Niang, I., Ruppel, O. C., Abdrabo, M. A., Essel, A., Lennard, C., Padgham, J., Urquhart, P., 2014. Africa, in: Barros, V.R., Field, C.B., Dokken, D.J., Mastrandrea, M.D., Mach, K.J., Bilir, T.E., Chatterjee, M., Ebi, K.L., Estrada, Y.O., Genova, R.C., Girma, B., Kissel, E.S., Levy, A.N., MacCracken, S., Mastrandrea, P.R., White, L.L. (Eds.), 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 University Press, Cambridge, United Kingdom and New York, NY, USA, pp. 1199–1265.

Odeha, I.O., McBratney, A.B., Chittleborough, D.J., 1994. Spatial prediction of soil properties from landform attributes derived from a digital elevation model. Geoderma 63 (3), 197–214.

Oliver, M.A., 1987. Geostatistics and its application to soil science. Soil Use and Management 3 (1), 8–20.

Olson, K.R., Al-Kaisi, M., Lal, R., Lowery, B., 2014. Examining the paired comparison method approach for determining soil organic carbon sequestration rates. Journal of Soil and Water Conservation 69 (6), 193A-197A.

Olvera-López, J.A., Carrasco-Ochoa, J.A., Martínez-Trinidad, J.F., Kittler, J., 2010. A review of instance selection methods. Artificial Intelligence Review 34 (2), 133–143.

Omuto, C., Nachtergaele, F., Rojas, R.V., 2013. State of the Art Report on Global and regional Soil Information: Where are we? Where to go? Food and Agriculture Organization of the United Nations, p. 81, 81 pp.

Oueslati, I., Allamano, P., Bonifacio, E., Claps, P., 2013. Vegetation and topographic control on spatial variability of soil organic carbon. Pedosphere 23 (1), 48–58.

Pansu, M., Bottner, P., Sarmiento, L., Metselaar, K., 2004. Comparison of five soil organic matter decomposition models using data from a 14C and 15N labeling field experiment. Global Biogeochemical Cycles 18 (4).

Pardo, M.T., Almendros, G., Zancada, M.C., López-Fando, C., González-Vila, F.J., 2012. Cultivation-induced effects on the organic matter in degraded southern African soils. Communications in soil science and plant analysis 43 (3), 541–555.

Parton, W.J., Schimel, D.S., Cole, C.V., Ojima, D.S., 1987. Analysis of factors controlling soil organic matter levels in Great Plains grasslands. Soil Science Society of America Journal 51 (5), 1173–1179.

Parton, W.J., Scurlock, J.M., Ojima, D.S., Gilmanov, T.G., Scholes, R.J., Schimel, D.S., Kirchner, T., Menaut, J., Seastedt, T., Garcia Moya, E., 1993. Observations and modeling of biomass and soil organic matter dynamics for the grassland biome worldwide. Global Biogeochemical Cycles 7 (4), 785–809.

Paustian, K., Levine, E., Post, W.M., Ryzhova, I.M., 1997. The use of models to integrate information and understanding of soil C at the regional scale. Geoderma 79 (1-4), 227–260.

Peng, X., Yan, X., Zhou, H., Zhang, Y.Z., Sun, H., 2015. Assessing the contributions of sesquioxides and soil organic matter to aggregation in an Ultisol under long-term fertilization. Soil and Tillage Research 146, 89–98.

Percival, H.J., Parfitt, R.L., Scott, N.A., 2000. Factors Controlling Soil Carbon Levels in New Zealand Grasslands Is Clay Content Important? Soil Sci. Soc. Am. J. 64 (5), 1623–1630.

Peukert, S., Bol, R., Roberts, W., Macleod, C.J.A., Murray, P.J., Dixon, E.R., Brazier, R.E., 2012. Understanding spatial variability of soil properties: a key step in establishing field-to farm-scale agro-ecosystem experiments. Rapid Communications in Mass Spectrometry 26 (20), 2413–2421.

Poggio, L., Gimona, A., Brewer, M.J., 2013. Regional scale mapping of soil properties and their uncertainty with a large number of satellite-derived covariates. Geoderma 209, 1–14.

Poulson, S.R., Dynes, J.J., McBeth, J.M., 2016. Iron-bound organic carbon in forest soils: quantification and characterization. Biogeosciences 13 (16), 4777.

Povak, N.A., Hessburg, P.F., McDonnell, T.C., Reynolds, K.M., Sullivan, T.J., Salter, R.B., Cosby, B.J., 2014. Machine learning and linear regression models to predict catchment-level base cation weathering rates across the southern Appalachian Mountain region, USA. Water Resources Research 50 (4), 2798–2814.

Powers, J.S., Schlesinger, W.H., 2002. Relationships among soil carbon distributions and biophysical factors at nested spatial scales in rain forests of northeastern Costa Rica. Geoderma 109 (3), 165–190.

Powlson, D.S., Prookes, P.C., Christensen, B.T., 1987. Measurement of soil microbial biomass provides an early indication of changes in total soil organic matter due to straw incorporation. Soil Biology and Biochemistry 19 (2), 159–164.

Preger, A.C., Kösters, R., Du Preez, C. C., Brodowski, S., Amelung, W., 2010. Carbon sequestration in secondary pasture soils: a chronosequence study in the South African Highveld. European Journal of Soil Science 61 (4), 551–562.

Qi, F., 2004. Knowledge Discovery from Area-Class Resource Maps: Data Preprocessing for Noise Reduction. Transactions in GIS 8 (3), 297–308.

Quinlan, J. Ross, 1986. Induction of decision trees. Machine Learning 1 (1), 81–106.

R core Team. R: A Language and Environment for Statistical Computing [Internet]. Vienna, Austria; 2015. Available: http:/www.r-project.org/.

Rad, M.R.P., Toomanian, N., Khormali, F., Brungard, C.W., Komaki, C.B., Bogaert, P., 2014. Updating soil survey maps using random forest and conditioned Latin hypercube sampling in the loess derived soils of northern Iran. Geoderma 232 (234), 97–106.

Ramankutty, N., Foley, J.A., 1999. Estimating historical changes in global land cover: Croplands from 1700 to 1992. Global Biogeochemical Cycles 13 (4), 997–1027.

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.

Reeuwijk, V.L., 2006. Procedures for soil analysis. 7th edition. Technical Report 9. Wageningen, Netherlands, ISRIC – World Soil Information.

Reeves, J.B., Smith, D.B., 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.

Reza Pahlavan Rad, Mohammad, Toomanian, N., Khormali, F., Brungard, C.W., Bayram Komaki, C., Bogaert, P., 2014. Updating soil survey maps using random forest and conditioned Latin hypercube sampling in the loess derived soils of northern Iran. Geoderma 232 (97-106), 232–234.

Rhodes, C.J., 2016. The 2015 Paris climate change conference: COP21. Science progress 99 (1), 97–104.

Richter, R., Schläpfer, D., 2012. Atmospheric / Topographic Correction for Satellite Imagery : ATCOR-2/3 User Guide [Internet]. Wil, Switzerland: ReSe Applications Schläpfer.

http://www.dlr.de/eoc/Portaldata/60/Resources/dokumente/5_tech_mod/atcor3_manual_2012.pdf.
Ridgeway, G., 2008. gbm: Generalized Boosted Regression Models. http://www.saedsayad.com/docs/gbm2.pdf.
Robinson, T.P., Metternicht, G., 2006. Testing the performance of spatial interpolation techniques for mapping soil properties. Computers and Electronics in Agriculture 50 (2), 97–108.
Rossel, R., Behrens, T., 2010. Using data mining to model and interpret soil diffuse reflectance spectra. Geoderma 158 (1-2), 46–54.
Rossel, R., Walvoort, D.J., McBratney, A.B., Janik, L.J., Skjemstad, J.O., 2006. Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties. Geoderma 131 (1), 59–75.
Sachs, J., Remans, R., Smukler, S., Winowiecki, L., Andelman, S.J., Cassman, K.G., Castle, D., DeFries, R., Denning, G., Fanzo, J., 2010. Monitoring the world's agriculture. Nature 466 (7306), 558–560.
Saidy, A.R., Smernik, R.J., Baldock, J.A., Kaiser, K., Sanderman, J., Macdonald, L.M., 2012. Effects of clay mineralogy and hydrous iron oxides on labile organic carbon stabilisation. Geoderma 173–174, 104–110.
Saiz, G., Bird, M.I., Domingues, T., Schrodt, F., Schwarz, M., Feldpausch, T.R., Veenendaal, E., Djagbletey, G., Hien, F., Compaore, H., Diallo, A., Lloyd, J., 2012. Variation in soil carbon stocks and their determinants across a precipitation gradient in West Africa. Global Change Biology 18 (5), 1670–1683.
Sanchez, P.A., Ahamed, S., Carré, F., Hartemink, A.E., Hempel, J., Huising, J., Lagacherie, P., McBratney, A.B., McKenzie, N.J., de Lourdes Mendonça-Santos, Maria, 2009. Digital soil map of the world. Science 325 (5941), 680–681.
Sanderson, E.W., Jaiteh, M., Levy, M.A., Redford, K.H., Wannebo, A.V., Woolmer, G., 2002. The Human Footprint and the Last of the Wild The human footprint is a global map of human influence on the land surface, which suggests that human beings are stewards of nature, whether we like it or not. BioScience 52 (10), 891–904.
Schapire, R.E., Freund, Y., Bartlett, P., Lee, W.S., 1998. Boosting the margin: A new explanation for the effectiveness of voting methods. Annals of Statistics, 1651–1686.
Schmengler, A.C., 2010. Modeling soil erosion and reservoir sedimentation at hillslope and catchment scale in semi-arid Burkina Faso, Bonn, 150 pp.
Schmidt, K., Behrens, T., Daumann, J., Ramirez-Lopez, L., Werban, U., Dietrich, P., Scholten, T., 2014. A comparison of calibration sampling schemes at the field scale. Geoderma 232, 243–256.
Schmidt, K., Behrens, T., Scholten, T., 2008. Instance selection and classification tree analysis for large spatial datasets in digital soil mapping. Geoderma 146 (1), 138–146.

Schulp, C.J.E., Verburg, P.H., Kuikman, P.J., Nabuurs, G.-J., Olivier, J.G.J., Vries, W. de, Veldkamp, T., 2013. Improving national-scale carbon stock inventories using knowledge on land use history. Environmental management 51 (3), 709–723.

Schulp, E., Verburg, P.H., 2009. Effect of land use history and site factors on spatial variation of soil organic carbon across a physiographic region. Agriculture, Ecosystems & Environment 133 (1), 86–97.

Schwendenmann, L., Pendall, E., Potvin, C., 2007. Surface soil organic carbon pools, mineralization and CO2 efflux rates under different land-use types in Central Panama, in: , Stability of Tropical Rainforest Margins. Springer, pp. 107–129.

Scull, P., Franklin, J., Chadwick, O.A., 2005. The application of classification tree analysis to soil type prediction in a desert landscape. Ecological Modelling 181 (1), 1–15.

Scull, P., Franklin, J., Chadwick, O.A., McArthur, D., 2003. Predictive soil mapping: a review. Progress in Physical Geography 27 (2), 171–197.

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.

Sheng, H., Zhou, P., Zhang, Y., Kuzyakov, Y., Zhou, Q., Ge, T., Wang, C., 2015. Loss of labile organic carbon from subsoil due to land-use changes in subtropical China. Soil Biology and Biochemistry 88, 148–157.

Shrestha, N.K., 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.

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.

Singh, B.R., Lal, R., 2005. The potential of soil carbon sequestration through improved management practices in Norway. Environment, Development and Sustainability 7 (1), 161–184.

Sivakumar, M.V.K., Stefanski, R., 2007. Climate and land degradation—an overview, in: , Climate and Land Degradation. Springer, pp. 105–135.

Six, J., Elliott, E.T., Paustian, K., 2000. Soil macroaggregate turnover and microaggregate formation: a mechanism for C sequestration under no-tillage agriculture. Soil Biology and Biochemistry 32 (14), 2099–2103.

Skjemstad, J.O., Spouncer, L.R., Cowie, B., Swift, R.S., 2004a. Calibration of the Rothamsted organic carbon turnover model (RothC ver. 26.3), using measurable soil organic carbon pools. Soil Research 42 (1), 79–88.

Skjemstad, J.O., Spouncer, L.R., Cowie, B., Swift, R.S., 2004b. Calibration of the Rothamsted organic carbon turnover model (RothC ver. 26.3), using measurable soil organic carbon pools. Soil Research 42 (1), 79–88.

Smith, P., 2008. Land use change and soil organic carbon dynamics. Nutrient Cycling in Agroecosystems 81 (2), 169–178.

Smith, P.F., Ganesh, S., Liu, P., 2013. A comparison of random forest regression and multiple linear regression for prediction in neuroscience. Journal of neuroscience methods 220 (1), 85–91.

Solomon, D., Lehmann, J., Kinyangi, J., Amelung, W., Lobe, I., Pell, A., Riha, S., Ngoze, S., Verchot, L.O., Mbugua, D., 2007. Long-term impacts of anthropogenic perturbations on dynamics and speciation of organic carbon in tropical forest and subtropical grassland ecosystems. Global Change Biology 13 (2), 511–530.

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

Steinmann, T., Welp, G., Wolf, A., Holbeck, B., Große-Rüschkamp, T., Amelung, W., 2016. Repeated monitoring of organic carbon stocks after eight years reveals carbon losses from intensively managed agricultural soils in Western Germany. Journal of Plant Nutrition and Soil Science 179 (3), 355–366.

Stergiadi, M., Perk, Marcel van der, Nijs, T. de, Bierkens, M.F.P., 2016. Effects of climate change and land management on soil organic carbon dynamics and carbon leaching in northwestern Europe. Biogeosciences 13 (5), 1519–1536.

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.

Stockmann, U., Adams, M.A., Crawford, J.W., Field, D.J., Henakaarchchi, N., Jenkins, M., Minasny, B., McBratney, A.B., Courcelles, Vivien de Remy de, Singh, K., Wheeler, I., Abbott, L., Angers, D.A., Baldock, J., Bird, M., Brookes, P.C., Chenu, C., Jastrow, J.D., Lal, R., Lehmann, J., Oâ€™Donnell, A.G., Parton, W.J., Whitehead, D., Zimmermann, M., 2013. The knowns, known unknowns and unknowns of sequestration of soil organic carbon. Agriculture, Ecosystems & Environment 164 (0), 80–99.

Stoorvogel, J.J., Kempen, B., Heuvelink, G.B., Bruin, S. de, 2009. Implementation and evaluation of existing knowledge for digital soil mapping in Senegal. Geoderma 149 (1), 161–170.

Stum, A.K., 2010. Random forests applied as a soil spatial predictive model in arid Utah. All Graduate Theses and Dissertations. Paper 736. htt://digitalcommons.usu.edu/etd/736. Springer.

Sulaeman, Y., Minasny, B., McBratney, A.B., Sarwani, M., Sutandi, A., 2013. Harmonizing legacy soil data for digital soil mapping in Indonesia. Geoderma 192, 77–85.

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.

Sutha, K., Tamilselvi, J.J., 2015. A Review of Feature Selection Algorithms for Data Mining Techniques. International Journal on Computer Science and Engineering 7 (6), 63.

Taghizadeh-Mehrjardi, R., Minasny, B., McBratney, A.B., Triantafilis, J., Sarmadian, F., Toomanian, N., 2012. Digital soil mapping of soil classes using decision trees in central Iran.

Taghizadeh-Mehrjardi, R., Nabiollahi, K., Minasny, B., Triantafilis, J., 2015. Comparing data mining classifiers to predict spatial distribution of USDA-family soil groups in Baneh region, Iran. Geoderma 253, 67–77.

Terhoeven-Urselmans, T., Vagen, T.-G., Spaargaren, O., Shepherd, K.D., 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.

Thomasson, J.A., Sui, R., Cox, M.S., Al–Rajehy, A., 2001. Soil reflectance sensing for determining soil properties in precision agriculture. Transactions of the ASAE 44 (6), 1445.

Timofeev, R., 2004. Classification and regression trees (CART) theory and applications. Humboldt University, Berlin.

Tisdall, J.M., Oades, J., 1982. Organic matter and water-stable aggregates in soils. Journal of Soil Science 33 (2), 141–163.

Towett, E.K., 2013. Prediction of soil properties for agricultural and environmental applications from infrared and X-ray soil spectral properties. University of Hohenheim.

Trumbore, S.E., 1997. Potential responses of soil organic carbon to global environmental change. Proceedings of the National Academy of Sciences 94 (16), 8284–8291.

Tully, K., Sullivan, C., Weil, R., Sanchez, P., 2015. The State of Soil Degradation in Sub-Saharan Africa: Baselines, Trajectories, and Solutions. Sustainability 7 (6), 6523–6552.

Tyc, G., Tulip, J., Schulten, D., Krischke, M., Oxfort, M., 2005. The RapidEye mission design. Acta Astronautica 56 (1), 213–219.

UN, 2015. Department of Economic and Social Affairs, Population Division (2015). World Population Prospects: The 2015 Revision, Key Findings and Advance Tables. Working Paper No. ESA/P/WP.241.

UNEP, 2006. Global Environment Outlook 3.: http://www.grida.no/geo/geo3/english/149.htm.

Vågen, T., Lal, R., Singh, B.R., 2005. Soil carbon sequestration in sub-Saharan Africa: a review. Land Degradation & Development 16 (1), 53–71.

Valentini, R., Arneth, A., Bombelli, A., Castaldi, S., Cazzolla Gatti, R., Chevallier, F., Ciais, P., Grieco, E., Hartmann, J., Henry, M., 2014. A full greenhouse gases budget of Africa: synthesis, uncertainties, and vulnerabilities. Biogeosciences 11, 381–407.

van Reeuwijk, L.P., 1993. Procedures for Soil Analysis. International Soil Reference and Information Centre (ISRIC). Wageningen. Netherlands.

Venables, W.N., Ripley, B.D., 2013. Modern applied statistics with S-PLUS. Springer Science & Business Media.

Wagai, R., Mayer, L.M., 2007. Sorptive stabilization of organic matter in soils by hydrous iron oxides. Geochimica et Cosmochimica Acta 71 (1), 25–35.

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

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.

Webster, R., Oliver, M.A., 2007. Geostatistics for environmental scientists. John Wiley & Sons.

Wei, X., Huang, L., Xiang, Y., Shao, M., Zhang, X., Gale, W., 2014a. The dynamics of soil OC and N after conversion of forest to cropland. Agricultural and Forest Meteorology 194, 188–196.

Wei, X., Shao, M., Gale, W., Li, L., 2014b. Global pattern of soil carbon losses due to the conversion of forests to agricultural land. Scientific reports 4.

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

Weston, J., Elisseeff, A., Schölkopf, B., Tipping, M., 2003. Use of the zero-norm with linear models and kernel methods. Journal of Machine Learning Research 3 (Mar), 1439–1461.

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.

Wiesmeier, M., Barthold, F., Spörlein, P., Geuß, U., Hangen, E., Reischl, A., Schilling, B., Angst, G., Lützow, M. von, Kögel-Knabner, I., 2014. Estimation of total organic carbon storage and its driving factors in soils of Bavaria (southeast Germany). Geoderma Regional 1, 67–78.

Wiesmeier, M., Spörlein, P., Geuß, U., Hangen, E., Haug, S., Reischl, A., Schilling, B., Lützow, M. von, Kögel-Knabner, I., 2012. Soil organic carbon stocks in southeast Germany (Bavaria) as affected by land use, soil type and sampling depth. Global Change Biology 18 (7), 2233–2245.

Willmott, C.J., 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.

Witten, I.H., Frank, E., 2005. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann.

Xiong, X., Grunwald, S., Myers, D.B., Ross, C.W., Harris, W.G., Comerford, N.B., 2014. Interaction effects of climate and land use/land cover change on soil organic carbon sequestration. Science of The Total Environment 493, 974–982.

Yang, R.-M., Zhang, G.-L., Liu, F., Lu, Y.-Y., Yang, F., Yang, F., Yang, M., Zhao, Y.-G., Li, D.-C., 2016. Comparison of boosted regression tree and random forest models for mapping topsoil organic carbon concentration in an alpine ecosystem. Ecological Indicators 60, 870–878.

Yang, Y., Guo, J., Chen, G., Yin, Y., Gao, R., Lin, C., 2009. Effects of forest conversion on soil labile organic carbon fractions and aggregate stability in subtropical China. Plant Soil 323 (1-2), 153–162.

Yira, Y., Diekkrüger, B., Steup, G., Bossa, A.Y., 2016. Modeling land use change impacts on water resources in a tropical West African catchment (Dano, Burkina Faso). Journal of Hydrology 537, 187–199.

Zadorova, T., Žížala, D., Penížek, V., Čejková, Š., 2014. Relating extent of colluvial soils to topographic derivatives and soil variables in a Luvisol sub-catchment, Central Bohemia, Czech Republic. Soil & Water Res 9 (2), 47–57.

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

Zhang, P., Shao, M., 2014. Spatial Variability and Stocks of Soil Organic Carbon in the Gobi Desert of Northwestern China. PLoS ONE 9 (4), e93584.

Zingore, S., Murwira, H.K., Delve, R.J., Giller, K.E., 2007. Influence of nutrient management strategies on variability of soil fertility, crop yields and nutrient balances on smallholder farms in Zimbabwe. Agriculture, Ecosystems & Environment 119 (1), 112–126.

Zougmoré, R., Ouattara, K., Mando, A., Ouattara, B., 2004. Rôle des nutriments dans le succès des techniques de conservation des eaux et des sols (cordons pierreux, bandes enherbées, zaï et demi-lunes) au Burkina Faso. Science et changements planétaires/Sécheresse 15 (1), 41–48.

X.

Appendix A

Supporting information to section IV

Tab. IX-1: Selected variables for Random Forest modelling

N°	Environmental variables	
	Variable	Abbreviation
1	Aspect ArcGis	A.Asp
2	Flow accumulation ArcGis	A.Flow.A / S.Flow.A
3	Flow direction ArcGis	A.Flow.d / S.Flow.d
4	Plan curvature ArcGis	A.Plan.curv / S.Plan.curv
5	Topographic Wetness Index ArcGis/SAGA	A.TWI / S.TWI
6	Northness	cose.Asp
7	Distance to stream ArcGis	Dist.stream
8	Elevation ArcGis	Elevation
9	Protection index SAGA	Prot.Index
10	Catchment Area Parallel SAGA	S.CA.Par
11	Flow line curvature SAGA	S.Flow.line.curv
12	Horizontal flow distance SAGA	S.HF.dist
13	SAGA Wetness Index SAGA	S.Wet.Ind
14	Total curvature SAGA	Sa_totalcuv
15	Terrain ruggedness SAGA	Terr.Rugg
16	Geomorphology	Geo
17	Lithology	Litho
18	Land use	LU
19	Precipitation	Prep
	Spectral variables and indices	
	Variable	Acquisition period
20	RI, SI, HI, NDVI, redEdge	March
21	RI, SI, BI, CI, HI, NIR	April
22	RI, SI, BI, CI, HI, NIR	May
23	Blue, CI, HI, NIR, SWIR1	June

HI: Hue Index, CI: Coloration Index, RI: Redness Index, BI: Brightness Index, NIR: Near infra red, SWIR: Shortwave infra red, SI: Saturation Index, NDVI : Normalized Difference Vegetation Index

Tab. IX-2: Confusion matrix between observed and predicted reference soil groups for the core range dataset with (RF_rfe) and without (RF) recursive feature elimination using the spectral parameters

RF

Spectral parameters (90%CR)	Predicted (%)					
Observed	CM	GL	LP	LX	PT	ST
CM	**29.4**	5.9	0.0	5.9	58.8	0.0
GL	0.0	**39.3**	0.0	0.0	60.7	0.0
LP	0.0	0.0	**25.0**	0.0	75.0	0.0
LX	0.0	0.0	0.0	**45.5**	54.5	0.0
PT	0.0	4.7	0.0	0.8	**94.6**	0.0
ST	0.0	0.0	0.0	0.0	66.7	**33.3**

Spectral parameters (80%CR)	Predicted (%)					
Observed	CM	GL	LP	LX	PT	ST
CM	**29.4**	5.9	0.0	5.9	58.8	0.0
GL	0.0	**39.3**	0.0	3.6	57.1	0.0
LP	0.0	0.0	**25.0**	0.0	75.0	0.0
LX	0.0	9.1	0.0	**45.5**	45.5	0.0
PT	0.0	4.7	0.8	1.6	**93.0**	0.0
ST	0.0	0.0	0.0	16.7	50.0	**33.3**

Spectral parameters (SDCR)	Predicted (%)					
Observed	CM	GL	LP	LX	PT	ST
CM	**23.5**	5.9	0.0	5.9	64.7	0.0
GL	0.0	**35.7**	0.0	0.0	64.3	0.0
LP	0.0	0.0	**25.0**	0.0	75.0	0.0
LX	0.0	0.0	0.0	**45.5**	54.5	0.0
PT	0.0	3.9	0.0	2.3	**93.8**	0.0
ST	0.0	0.0	0.0	0.0	66.7	**33.3**

RF_rfe

Spectral parameters (90%CR)	Predicted (%)					
Observed	CM	GL	LP	LX	PT	ST
CM	**23.5**	5.9	0.0	0.0	70.6	0.0
GL	0.0	**42.9**	0.0	0.0	57.1	0.0
LP	0.0	0.0	**25.0**	0.0	75.0	0.0
LX	0.0	0.0	0.0	**54.5**	45.5	0.0
PT	0.0	6.2	0.0	0.8	**93.0**	0.0
ST	0.0	0.0	0.0	0.0	66.7	**33.3**

Spectral parameters (80%CR)	Predicted (%)					
Observed	CM	GL	LP	LX	PT	ST
CM	**23.5**	5.9	0.0	5.9	64.7	0.0
GL	0.0	**39.3**	0.0	0.0	60.7	0.0
LP	0.0	0.0	**25.0**	0.0	75.0	0.0
LX	0.0	0.0	0.0	**45.5**	54.5	0.0
PT	0.0	3.1	0.0	0.8	**96.1**	0.0
ST	0.0	0.0	0.0	0.0	66.7	**33.3**

Spectral parameters (SDCR)	Predicted (%)					
Observed	CM	GL	LP	LX	PT	ST
CM	**23.5**	5.9	0.0	5.9	64.7	0.0
GL	0.0	**42.9**	0.0	0.0	57.1	0.0
LP	0.0	0.0	**25.0**	0.0	75.0	0.0
LX	0.0	9.1	0.0	**45.5**	45.5	0.0
PT	0.0	3.9	0.0	0.0	**96.1**	0.0
ST	0.0	0.0	0.0	0.0	66.7	**33.3**

CM: Cambisols, GL: Gleysols, LP: Leptosols, LX: Lixisols, PT: Plinthosols, ST: Stagnosols; 90%CR: dataset with 5% lower and upper range pruning, 80%CR: dataset with 10% lower and upper range pruning, SDCR: dataset with standard deviation based pruning.

Tab. IX-3: Confusion matrix between observed and predicted reference soil groups for the core range dataset with (RF_rfe) and without (RF) recursive feature elimination using the terrain parameters

RF

Terrain parameters	Observed	Predicted (%)					
		CM	GL	LP	LX	PT	ST
(90%CR)	CM	**29.4**	0.0	0.0	0.0	70.6	0.0
	GL	0.0	**57.1**	0.0	0.0	42.9	0.0
	LP	0.0	0.0	**75.0**	0.0	25.0	0.0
	LX	0.0	0.0	0.0	**36.4**	63.6	0.0
	PT	0.8	7.0	0.0	0.8	**91.5**	0.0
	ST	0.0	0.0	0.0	16.7	66.7	**16.7**
(80%CR)	CM	**52.9**	0.0	0.0	0.0	47.1	0.0
	GL	0.0	**64.3**	0.0	7.1	28.6	0.0
	LP	0.0	0.0	**75.0**	0.0	25.0	0.0
	LX	0.0	0.0	0.0	**63.6**	27.3	9.1
	PT	6.2	9.3	0.0	1.6	**78.3**	4.7
	ST	16.7	0.0	0.0	33.3	16.7	**33.3**
(SDCR)	CM	**52.9**	0.0	0.0	0.0	47.1	0.0
	GL	0.0	**60.7**	0.0	7.1	32.1	0.0
	LP	0.0	0.0	**75.0**	0.0	25.0	0.0
	LX	0.0	0.0	0.0	**63.6**	27.3	9.1
	PT	10.9	9.3	0.0	1.6	**72.1**	6.2
	ST	16.7	0.0	0.0	33.3	16.7	**33.3**

RF_rfe

Observed	Predicted (%)					
	CM	GL	LP	LX	PT	ST
CM	**52.9**	0.0	0.0	0.0	47.1	0.0
GL	0.0	**60.7**	0.0	0.0	39.3	0.0
LP	0.0	0.0	**75.0**	0.0	25.0	0.0
LX	0.0	9.1	0.0	**54.5**	27.3	9.1
PT	3.1	8.5	0.0	0.8	**85.3**	2.3
ST	0.0	0.0	0.0	16.7	66.7	**16.7**
CM	**58.8**	0.0	0.0	0.0	41.2	0.0
GL	0.0	**60.7**	0.0	3.6	35.7	0.0
LP	0.0	0.0	**75.0**	0.0	25.0	0.0
LX	0.0	0.0	0.0	**63.6**	27.3	9.1
PT	4.7	7.8	0.0	1.6	**81.4**	4.7
ST	0.0	0.0	0.0	33.3	50.0	**16.7**
CM	**52.9**	0.0	5.9	0.0	41.2	0.0
GL	0.0	**60.7**	0.0	3.6	35.7	0.0
LP	0.0	0.0	**75.0**	0.0	25.0	0.0
LX	0.0	9.1	0.0	**63.6**	18.2	9.1
PT	8.5	9.3	0.0	1.6	**74.4**	6.2
ST	0.0	0.0	0.0	33.3	33.3	**33.3**

CM: Cambisols, GL: Gleysols, LP: Leptosols, LX: Lixisols, PT: Plinthosols, ST: Stagnosols; 90%CR: dataset with 5% lower and upper range pruning, 80%CR: dataset with 10% lower and upper range pruning, SDCR: dataset with standard deviation based pruning.

Tab. IX-4: Confusion matrix between observed and predicted reference soil groups for the core range dataset with (RF_rfe) and without (RF) recursive feature elimination using the terrain and spectral parameters

RF

		Predicted (%)					
	Observed	CM	GL	LP	LX	PT	ST
Terrain and spectral parameters (90%CR)	CM	**41.2**	0.0	0.0	0.0	58.8	0.0
	GL	0.0	**53.6**	0.0	0.0	46.4	0.0
	LP	0.0	0.0	**75.0**	0.0	25.0	0.0
	LX	0.0	18.2	0.0	**54.5**	18.2	9.1
	PT	2.3	7.8	0.8	0.0	**86.8**	2.3
	ST	0.0	0.0	0.0	0.0	66.7	**33.3**
Terrain and spectral parameters (80%CR)	CM	**52.9**	0.0	0.0	0.0	47.1	0.0
	GL	0.0	**57.1**	0.0	0.0	42.9	0.0
	LP	0.0	0.0	**75.0**	0.0	25.0	0.0
	LX	0.0	9.1	0.0	**54.5**	27.3	9.1
	PT	7.0	8.5	0.0	0.0	**79.8**	4.7
	ST	16.7	0.0	0.0	16.7	33.3	**33.3**
Terrain and spectral parameters (SDCR)	CM	**52.9**	0.0	0.0	0.0	47.1	0.0
	GL	0.0	**67.9**	0.0	0.0	32.1	0.0
	LP	0.0	0.0	**75.0**	0.0	25.0	0.0
	LX	0.0	9.1	0.0	**54.5**	27.3	9.1
	PT	12.4	10.9	0.0	0.0	**72.1**	4.7
	ST	16.7	0.0	0.0	16.7	16.7	**50.0**

RF_rfe

	Predicted (%)					
Observed	CM	GL	LP	LX	PT	ST
CM	**52.9**	5.9	0.0	5.9	35.3	0.0
GL	0.0	**67.9**	0.0	3.6	28.6	0.0
LP	0.0	0.0	**75.0**	0.0	25.0	0.0
LX	0.0	9.1	0.0	**63.6**	27.3	0.0
PT	3.1	7.0	0.0	1.6	**86.8**	1.6
ST	0.0	0.0	0.0	0.0	66.7	**33.3**
CM	**58.8**	5.9	0.0	5.9	29.4	0.0
GL	0.0	**67.9**	0.0	3.6	28.6	0.0
LP	0.0	0.0	**75.0**	0.0	25.0	0.0
LX	0.0	9.1	0.0	**54.5**	27.3	9.1
PT	3.9	7.8	0.0	1.6	**82.9**	3.9
ST	0.0	0.0	0.0	0.0	66.7	**33.3**
CM	**58.8**	5.9	0.0	5.9	29.4	0.0
GL	0.0	**71.4**	0.0	3.6	25.0	0.0
LP	25.0	0.0	**25.0**	25.0	25.0	0.0
LX	0.0	0.0	0.0	**54.5**	27.3	18.2
PT	7.0	10.9	0.0	2.3	**75.2**	4.7
ST	0.0	0.0	0.0	0.0	33.3	**66.7**

CM: Cambisols, GL: Gleysols, LP: Leptosols, LX: Lixisols, PT: Plinthosols, ST: Stagnosols; 90%CR: dataset with 5% lower and upper range pruning, 80%CR: dataset with 10% lower and upper range pruning, SDCR: dataset with standard deviation based pruning.

X.

Appendix B

Supporting information to section V

Fig. X-1: Stone line in a field of the Dano catchment

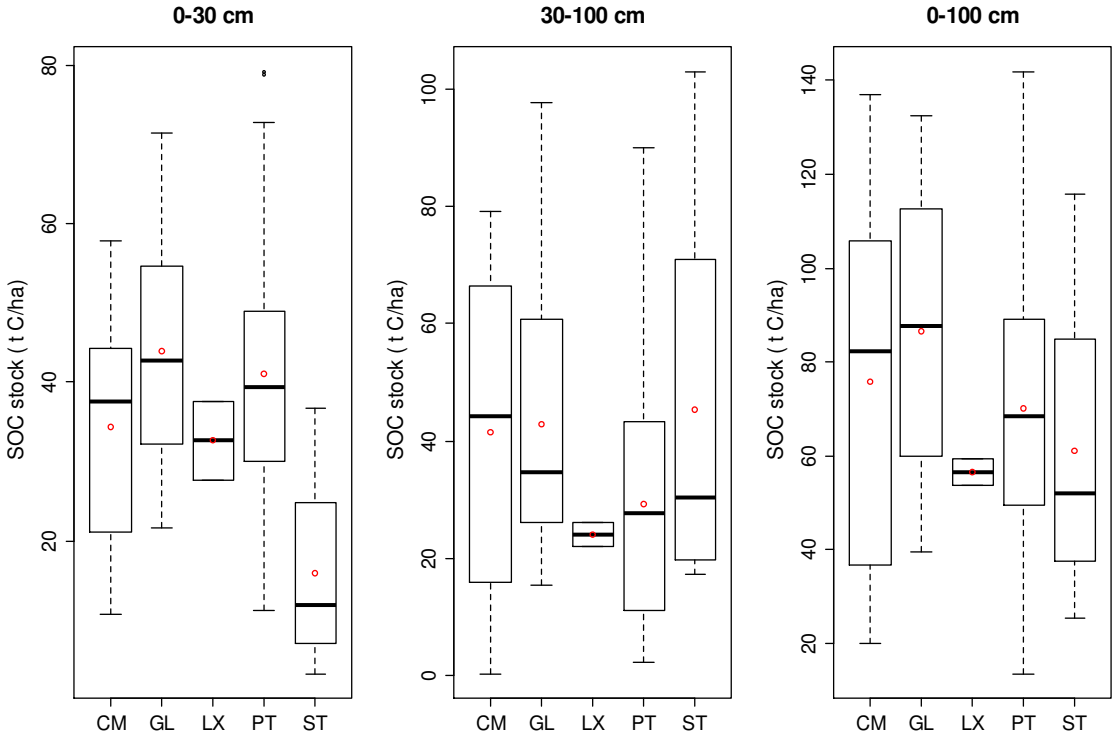

Fig. X-2: SOC stock in different RSG and depths. (CM: Cambisols, GL: Gleysols, LX: Lixisols, PT: Plinthosols, ST: Stagnosols). Lines within the boxes give the median, red circle within the boxes the mean, boxes the 25th and 75th percentile, whiskers the lowest and highest values.

Tab. X-1: Random Forest and multiple linear regression model performance and statistics of toposoil reference soil groups

Dataset	Random Forest			Linear Regression		
	R^{2*}	RMSECV	RMSEPV	R^2	RMSECV	RMSEPV
Entire dataset	13.0	14.0	14.2	11.0	14.2	14.8
Dataset without PT	17.5	13.6	15.8	17.8	14.5	20.8
Dataset without GL&ST	10.2	14.1	13.8	12.6	14.1	28.6
Dataset without CM	13.2	13.9	13.8	9.4	14.4	16.5

PT: Plinthosols, GL: Gleysols, Stagnosols, CM: Cambisols, RMSECV: root mean square error of cross validation, RMSEPV: root mean square error of prediction based on validation set, *explained variance in %.

Tab. X-2: General characteristics of some representative soil profiles

Reference soil group	Horizon	Depth	pH (H_20)	N (%)	C (%)	C (t ha-1)	CEC (cmolc kg^{-1})	BD (g cm^{-3})	SC (%)	Sand (%)	Silt (%)	Clay (%)	Color
Cambisol	Ahp	0-24	7.2	0.1	0.9	28.5	36.32	1.6	58.5	14.8	38.3	46.9	10 YR 3/6
	Bw1	24-38	7.0	0.1	1.0	18.0	35.6	1.6	58.0	16.6	33.5	50.0	2.5 Y 4/6
	Bw2	38-100	8.0	0.0	0.5	47.8	29.6	1.6	20.7	17.9	36.0	46.1	2.5 Y 4/6
Gleysol	Ah	0-31	6.2	0.1	2.0	69.4	20.8	1.1	0.0	3.7	64.8	31.6	10 YR 3/4
	Bl1	31-50	6.1	0.0	0.8	21.8	10.3	1.5	0.0	9.8	66.0	24.3	7.5 YR 4/4
	Bl2	50-100	6.2	0.0	0.6	46.4	10.9	1.6	0.0	9.9	61.6	28.5	7.5 YR 5/6
Lixisol	Ah	0-17	6.0	0.1	0.9	18.4	5.3	1.4	40.0	32.1	47.0	18.2	7.5 YR 4/3
	Bt1	17-37	5.8	0.0	0.5	13.3	6.5	1.5	32.0	16.9	51.9	27.9	5 YR 5/8
	Bt2	37-74	6.0	0.0	0.3	18.0	6.1	1.8	68.7	13.3	50.8	33.2	5 YR 5/8
	Bt3	74-100	5.9	0.0	0.2	6.8	5.6	1.2	10.6	12.9	50.5	33.4	7.5 YR 6/6
Plinthosol	Ahv	0-18	6.6	0.1	1.9	41.2	8.5	1.5	54.6	39.2	46.7	11.6	7.5 YR 4/6
	Btv1	18-56	5.9	0.1	0.6	26.0	6.8	1.6	72.0	29.5	42.1	22.7	2.5 YR 4/6
	Btv2	56-102	5.6	0.0	0.2	12.5	6.8	1.5	47.9	27.4	36.0	34.5	2.5 YR 5/8
Plinthosol	Ahv	0-12	5.9	0.1	1.9	32.1	9.1	1.7	48.7	40.9	43.8	13.5	7.5 YR 4/3
	Bv	Dez 40	6.2	0.1	0.9	34.9	9.0	1.7	48.4	29.8	44.4	29.1	7.5 YR 5/8
Plinthosol	Ahp	0-21	6.8	0.1	1.2	29.3	9.7	1.4	37.8	28.9	51.6	26.8	7.5 YR 5/3
	Bv1	21-41	6.6	0.1	0.9	17.6	10.5	1.4	73.4	17.6	47.5	39.8	7.5 YR 6/4
	Bv2	41-69	6.5	0.1	0.8	25.4	10.9	1.6	74.5	17.9	44.7	44.4	10 YR 6/4
Stagnosol	Ah	0-24	6.5	0.1	0.3	7.5	18.7	1.5	57.0	43.9	36.2	21.0	7.5 YR 4/4
	Bg1	24-70	7.4	0.1	1.2	58.0	26.3	1.2	16.1	16.7	50.7	33.2	7.5 YR 5/3
	Bg2	70-100	7.3	0.1	1.6	51.5	28.1	1.2	30.2	4.5	50.2	40.5	10 YR 2/3

XI.

Appendix C

Supporting information to section VI

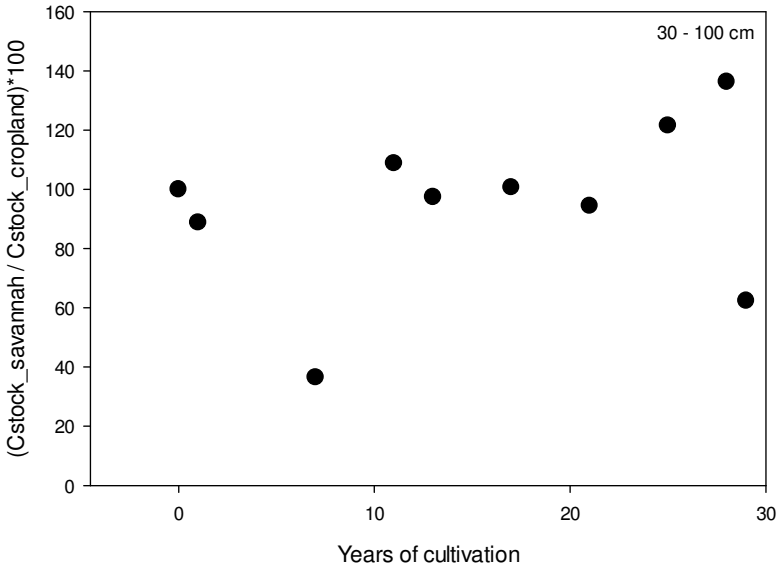

Fig. XI-1: SOC stocks of cropland in relation to SOC stock of savannah soils (in %) for different years of cultivation in the subsoil (30 – 100 cm)

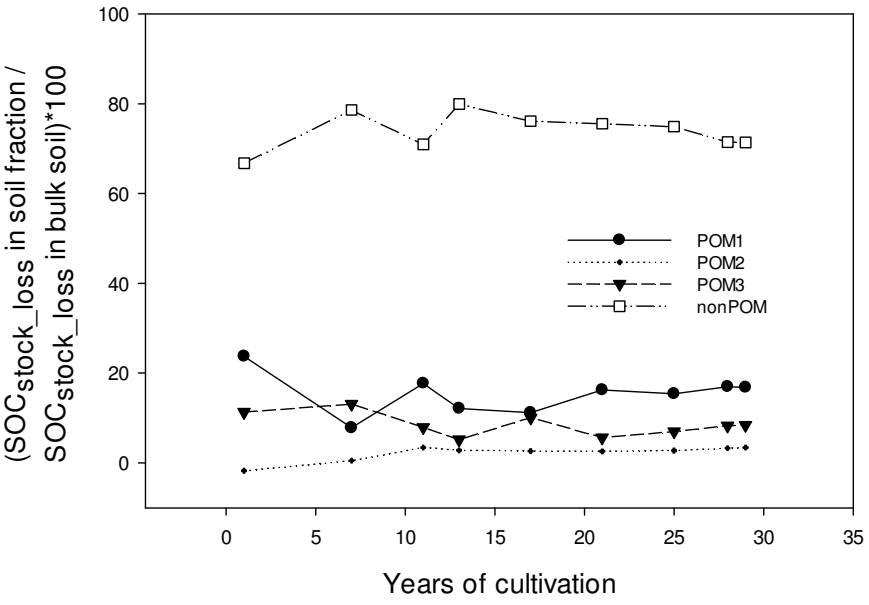

Fig. XI-2: Percentage of residual SOC stock of cropland (in relation to SOC stock of savannah soils) in soil fractions relative to the residual SOC stock in bulk soil (in relation to SOC stock of savannah soils) of the cropland for different years of cultivation in the topsoil (0 – 10 cm)

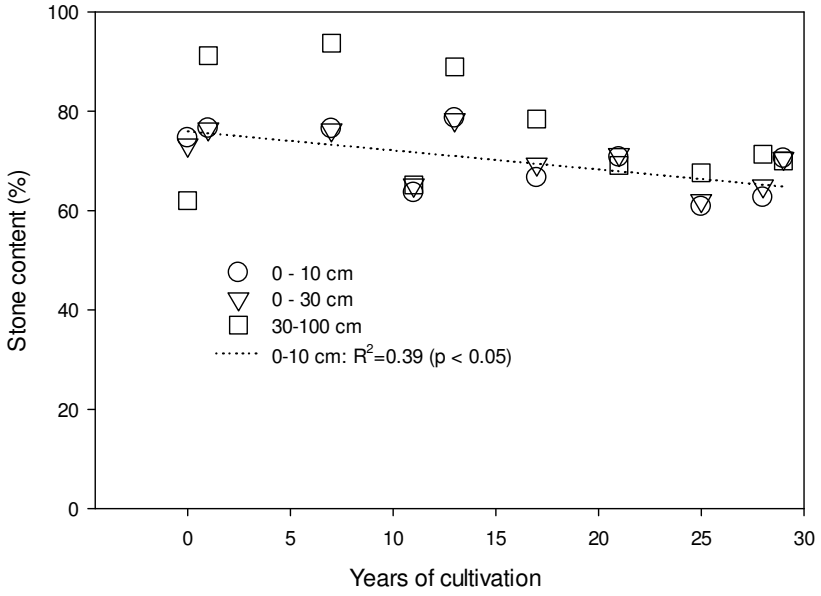

Fig. XI-3: Stone content at different depths in relation to the duration of cultivation

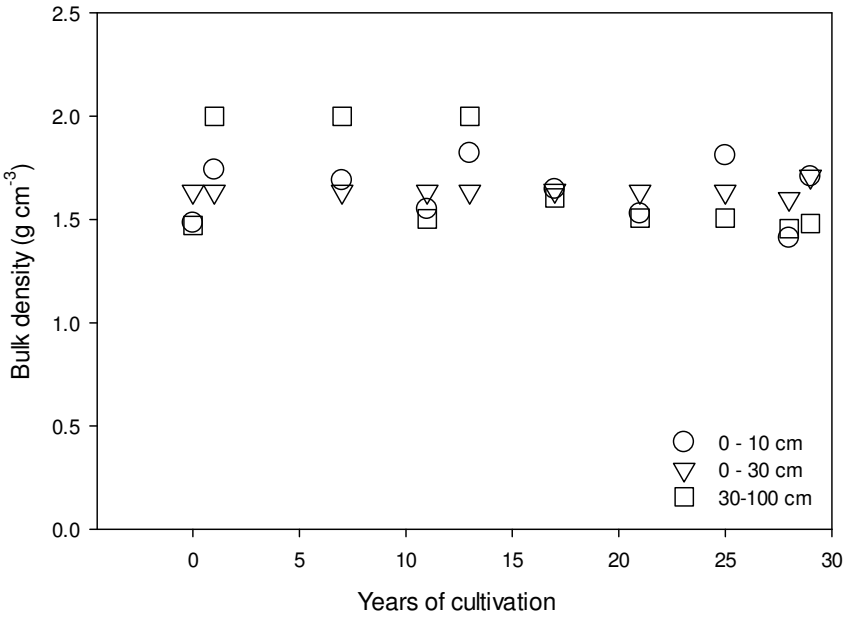

Fig. XI-4: Bulk density at different depths in relation to the duration of cultivation

XII.

Appendix D

The data that form the basis of this dissertation thesis are available in electronic format from the office of INRES-soil science or from myself.

Contact details :

INRES-Bodenwissenschaften
Nußallee 13
D-53115 Bonn
bobo@uni-bonn.de

Kpadé Ozias Laurentin Hounkpatin
INRES-Bodenkunde
Nußallee 13 D-53115 Bonn
hozias@uni.bonn.de