# Variability of physical, chemical and hydraulic parameters in soils of Mt. Kilimanjaro across different land uses

Dissertation

submitted to the
Faculty of Biology, Chemistry and Geosciences of the
University of Bayreuth
to obtain the degree of
Dr. rer. nat

by
Anna Kühnel
born December 19, 1984

Bayreuth, September 2014

Die vorliegende Arbeit wurde in der Zeit von August 2010 bis September 2014 in Bayreuth in der Abteilung Bodenphysik unter Betreuung von Herrn Professor Dr. Bernd Huwe angefertigt.

Vollständiger Abdruck der von der Fakultät für Biologie, Chemie und Geowissenschaften der Universität Bayreuth genehmigten Dissertation zur Erlangung des akademischen Grades eines Doktors der Naturwissenschaften (Dr. rer. nat).

Dissertation eingereicht am: 22.09.2014
Zulassung durch die Promotionskommision: 08.10.2014
Wissenschaftliches Kolloquim: 24.04.2015

Amtierender Dekan Prof. Dr. Rhett Kempe

Prüfungsausschuss:
Prof. Dr. Bernd Huwe (Erstgutachter)
Prof. Dr. Michael Hauhs (Zweitgutachter)
Prof. Dr. Egbert Matzner (Vorsitz)
Prof. Dr. Hans Keppler

Drittgutacher: Prof. Dr. Wulf Amelung

# Table of Contents

# List of Figures

# List of Tables

## Summary

The heterogeneity of soils is a key to biological processes, carbon turnover and water storage. Climate change and anthropogenic land use changes often decrease important functioning of soil systems. In order to understand the soil complexity, sophisticated technologies to measure soil properties and processes at high spatial resolutions are needed. This study addresses the heterogeneity of various soil physical, chemical and hydraulic properties in different land uses at Mt. Kilimanjaro and its implications for carbon and water storage of the soils. The objectives were to apply visible to near infrared diffuse reflectance spectroscopy in-situ to gain information about the intact soil.

The first three studies address the challenges that arise from in-situ spectral measurements. We could show that regression models calibrated with a regional spectral database fail to predict clay and carbon content from in-situ soil spectra. The incorporation of additional samples that were scanned in the field into a spectral calibration database increased prediction accuracies. As the collection of soil information with traditional laboratory methods is time demanding, often only a limited amount of additional samples are available. Therefore, we used the synthetic minority oversampling technique and demonstrated its potential for generating new soil spectra, which can be used to balance a calibration database regarding in-situ spectral characteristics. Including these new spectra into calibration models improved prediction accuracies substantially. Based on these findings we propose a framework for modelling with limited in-situ spectra.

Consequently we applied this framework for the prediction of soil organic carbon, nitrogen, clay, silt and sand content in soil profiles in a high spatial resolution with predicted values every 3 cm. Soil hydraulic properties lack a direct physical basis, which could be reflected in the soil spectra. Furthermore, the creation of a spectral database would require a huge effort. Therefore, we used pedotransfer functions to predict soil hydraulic properties in the soil profiles.

Soils of the four different land uses, homegarden (a traditional agroforestry system at Mt. Kilimanjaro), coffee plantation, maize field and savannah were thus thoroughly described regarding physical, chemical and hydraulic parameters. Soil heterogeneity of the less intensively managed land uses, homegarden and savannah was much higher, than those of coffee plantation and maize field. With our sampling design, it is unfortunately difficult to differentiate between pedogenic and land use effects on the soil properties. However, with our proposed framework, additional soil information can be derived rapidly to characterise further sites.

In a further study, using basic soil properties derived from spectral measurements, we could improve spatial predictions for microbial biomass in two savannah plots

at a spatial scale of several meters. Soil microbial biomass was strongly related to organic C and N content of the soil. We could further show, that its spatial distribution is related to vegetation and surface morphology.

In order to better understand implications of land use changes for the whole Mt. Kilimanjaro, further research regarding the soils of the forest zones is needed. As spectral characteristics of the volcanic soils in the forest differ from those in the lower zones of the mountain, these should be studied comprehensively. A characterisation of the water and carbon storage potential of the different soils of Mt. Kilimanjaro would then be possible.

We conclude, that by integrating visible to near infrared spectroscopy into additional prediction methods like geostatistics or pedotransfer functions, various soil physical, chemical, biological and hydraulic parameters can be derived rapidly and accurately.

## Zusammenfassung

Die Vielschichtigkeit des Bodengefüges ist der Schlüssel zu biologischen Prozessen, Kohlenstoff-Umsätzen und zu der Fähigkeit des Bodens Wasser zu speichern. Klimawandel und anthropogene Landnutzungsänderungen führen oft zu einer verringerten Funktionsfähigkeit des Bodens. Um die Vielschichtigkeit des Bodens zu verstehen, werden anspruchsvolle Methoden benötigt, die es erlauben Bodeneigenschaften und Prozesse in hoher räumlicher Auflösung zu messen. Die vorliegende Arbeit beschäftigt sich mit der Heterogenität von verschiedenen bodenphysikalischen, -chemischen und - hydraulischen Eigenschaften und deren Auswirkungen auf die Kohlenstoff- und Wasserspeicherfähigkeit der Böden in verschiedenen Landnutzungen am Kilimandscharo. Ziel war es, die Spektroskopie im sichtbaren und nahen Infrarot direkt im Feld zu nutzen, um Informationen über den intakten Boden zu erlangen.

Die erste Studie beschäftigt sich mit den Herausforderungen von in-situ Spektralmessungen. Wir konnten zeigen, dass Regressionsmodelle, die mit Hilfe einer regionalen spektralen Datenbank kalibriert wurden, an der Vorhersage des Ton- und Kohlenstoffgehalts aus in-situ Spektren scheitern. Die Einarbeitung von direkt im Feld gewonnenen, zusätzlichen Proben in die Datenbank verbesserte die Vorhersagegenauigkeiten. Da es zeitaufwändig ist, zusätzliche Bodeninformationen mit Hilfe von traditionellen Labormethoden zu gewinnen, stehen oft nur geringe zusätzliche Proben zur Verfügung. Deswegen nutzten wir die 'synthetic minority oversampling technique', eine Methode mit der synthetische Daten generiert werden können, und zeigten die Möglichkeiten mit dieser Methode neue Bodenspektren zu erstellen. Diese zusätzlichen Bodenspektren können dann dazu genutzt werden, einen Datensatz hinsichtlich der in-situ Eigenschaften der Spektren auszugleichen. Das Einbeziehen dieser neuen Spektren in Regressionsmodelle erhöhte die Vorhersagegenauigkeit wesentlich. Auf Grundlage dieser Erkenntnisse stellen wir einen Ansatz zur Modellierung mit in-situ Spektren vor.

Anschließend nutzten wir diesen Ansatz für die Vorhersage von organischem Kohlenstoff-, Stickstoff-, Ton-, Schluff- und Sandgehalten in hoher räumlicher Auflösung (alle 3cm) in Bodenprofilen. Bodenhydraulischen Eigenschaften fehlt es jedoch an einer direkten physikalischen Basis, die im Spektrum des Bodens sichtbar wäre. Außerdem wäre das Erstellen einer Spektraldatenbank mit hydraulischen Eigenschaften ein sehr großer Arbeitsaufwand. Um diese hydraulischen Eigenschaften in den Bodenprofilen vorherzusagen, nutzten wir deshalb 'pedotransfer' Funktionen.

Böden aus den vier verschiedenen Landnutzungen, Homegarden (ein traditionelles Agroforstsystem am Kilimandscharo), Kaffeeplantage, Maisfeld and Savanne,

wurden somit ausführlich hinsichtlich ihrer physikalischen, chemischen und hydraulischen Eigenschaften charakterisiert. Die Heterogenität des Bodens in den weniger intensiv gemanagten Landnutzungen Homegarden und Savanne war sehr viel höher als die des Maisfeldes und der Kaffeeplantage. Aufgrund des Probenahmedesigns ist es leider schwierig zwischen pedogenen und landnutzungsbedingten Einflüssen auf die Bodeneigenschaften zu unterscheiden. Mit dem vorgestellten Ansatz ist es jedoch möglich, zusätzliche Bodeninformationen sehr schnell zu erfassen und weitere Flächen zu untersuchen.

In einer anschließenden Studie konnten wir die räumliche Vorhersage der mikrobiellen Biomasse auf zwei Savannenstandorten verbessern, indem wir Informationen über zusätzliche Bodenparameter nutzten. Diese Parameter wurden vorher mit Hilfe von Spektralmessungen abgeleitet. Die mikrobielle Biomasse korrelierte mit organischem Kohlenstoff und Stickstoff im Boden. Wir konnten außerdem zeigen, dass ihre räumliche Verteilung mit der Vegetation und der Oberflächenstruktur des Geländes zusammen hängt.

Um die Auswirkungen von Landnutzungsänderungen des gesamten Kilimandscharogebietes besser zu verstehen, sind weitere Untersuchungen der Böden, vor allem in den Waldflächen des Berges erforderlich. Da sich die spektralen Eigenschaften der Vulkanböden stark von denen der unteren Gebiete unterscheiden, sollten diese eingehender untersucht werden. Eine Charakterisierung der unterschiedlichen Böden des Kilimandscharos hinsichtlich ihrer Wasser- und Kohlenstoffspeicherfähigkeit wäre somit möglich.

Durch die Einbindung der Spektroskopie in andere Vorhersagemethoden, wie zum Beispiel Geostatistik oder pedotransfer Funktionen können verschiedene bodenphysikalische, -chemische, -biologische und -hydraulische Eigenschaften schnell und präzise vorhergesagt werden können.

## Acknowledgements

First of all, I want to express my gratitude to my supervisor, Prof. Dr. Bernd Huwe for the possibility to carry out this thesis at the department of Soil Physics, the opportunity to conduct my field studies at Mt. Kilimanjaro, for many discussions on the concept and statistics and for the freedom to develop my own ideas.

I am grateful to all students, helpers and friends, who shared some times at the field station in Nkweseko, especially Gemma, Holle, Thomas, Juliane, Friederike, Judith, Toni, David, Andi, Eva, Alice, Eunyoung, Stefan, Maria, Julia, Marcel, Tim, Insa, Ephraim, William, James, David, Hannes, Jojo, Johannes, Andreas, Christoph and Silke. You made some hard times bearable and other times just really great! In particular I want to thank Johannes Hepp, with who I discussed field work directly in the fieldand who I could further call whenever samples had to be measured immediately! Furthermore I want to acknowledge the great help of all station staff, drivers and field workers, especially of Ayubu, Rafiki, Fortu, Mama Tuma, Augusti, Wilbard, Nelson, J4, Samuel, George and Raymundi. Asanteni sana kwa kunifundishwa kiswahili au kwa kuchimba mashimo na kwa nisaidia kufanya kazi tu.

I also want to thank my colleagues and friends from Bayreuth, especially Steffen, who shared not only hard times writing, but also Glashaus afternoons and evenings and Andreas, who did not only support me with all the field installations, but also provided me with a down to earth view on station live and Frau Wittke, who I could always write to from Tanzania and any administrative problems were suddenly solved (as far as she could influence them). I am thankful to the members of the BayCEER laboratory and the laboratory in Göttingen, who analysed various soil parameters. Thank goes furthermore to all student helpers who did a great job! I am also grateful to Verena, Holle, Manu and Alana for valuable comments and input.

Thanks goes to the German Research Foundation for financing the KiLi-Project (DFG Research-Unit 1246), the speaker and the organisers of the project and the Tanzanian authorities for their support, specifically the Tanzanian Commision for Science and Technology (COSTECH) and the Tanzanian Wildlife Research Institute (TAWIRI).

Then I want to thank my family and Christoph and further the Steinbauer family, who gave me a home wherever and whenever I needed it!

Especially I want to thank Christina Bogner, who supported me with a lot of things in the last years, such as general statistics, writing functions in R, extending my knowledge of Latex, writing concisely and most of all the difference between good and bad science. ;-) Without you, writing this thesis would not have been much fun and your encouragements supported me a great deal.

# Extended Summary

## Introduction

### The importance of soil

Soil is not only the biggest terrestrial carbon (C) pool (Batjes, 1996), but also an important water storage compartment (Rockström et al., 2009) and the most complex and diverse ecosystem of the world (Young et al., 2004).

Soils can be described as a dynamic system, that is formed by parent material, topography, climate, organisms and time (Jenny et al., 1941). The parent material provides the basis for soil formation, topography determines the amount material that is accumulated or eroded, rainfall and temperature determine the rate of mineralization and humification (Zech et al., 1997), organisms influence the soil through accumulation and decomposition of organic matter, bioturbation and root growth (Bot et al., 2005). All these factors interact constantly and change with time (Figure I).

The first meters of soil are most important for human activity and they provide the basis for agriculture. These upper layers of soil, which provide most environmental services are very susceptible to land use changes and climate change. About 1500 $\times$ $10^{15}$ g C is stored in the first meter of the soils globally (Batjes, 1996), which is almost three times the amount of the living vegetation (Eswaran et al., 1993). Whether soils act as a source or sink for $CO_2$, greatly influences global climate (IPCC, 2008). Human activities, such as land use changes can alter the soil C storage and thus might increase $CO_2$ release to the atmosphere. Interactions are however complex and the contribution of different land uses to climate change is not yet fully understood.

The infiltrated rainfall water, which is attached to soil particles and accessible to roots, is the most important water resource for food productions globally, more influential than water from rivers and aquifers (Rockström et al., 2009). Considering the population growth and the global climate change, the demand of water for food production is increasing and many countries will experience water shortage in the next years (Fraiture et al., 2001; IPCC, 2008). Therefore, in addition to the distribution of rainfall, the potential of soils to store water is important for food production and global common welfare. Information on how the water storage potential of soils changes with land use intensification and global change is however scarce.

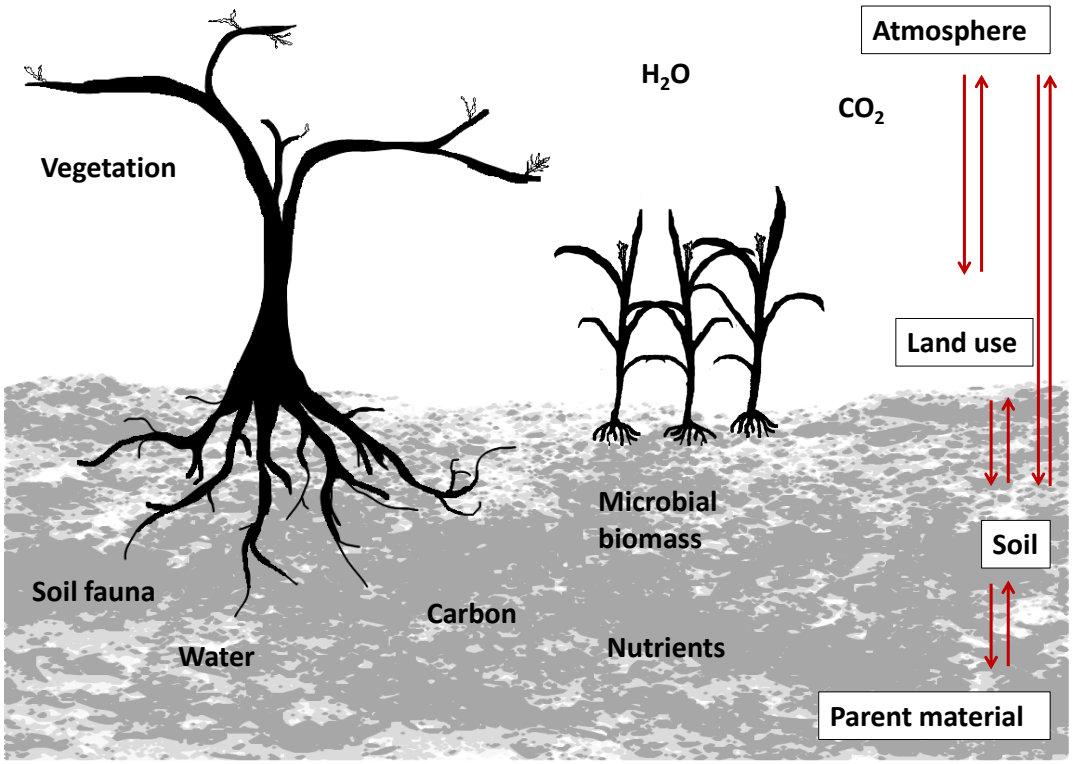



Figure I: Soils are a complex system, with various interactions between the soil and other compartments.

Soils do not only provide the physical habitat for soil biota (Young et al., 2004), but organism are an essential internal component of the soil complex. Relationships between the physical soil structure and soil biota are diverse (Nunan et al., 2003). However, the knowledge of interactions between soil biota and the spatial structure of the soil is limited, as measuring physical and biological characteristics of an intact soil is difficult. Additionally, soil biota exerts direct and indirect effects on plants by stimulating nutrient takeover and nutrient availability or mycorrhization (Hussey et al., 1982; Wardle et al., 2004). These effects influence herbivores and predators (Bot et al., 2005) and thus the development of plant communities, which in turn can regulate soil organisms and the soil structure (Wardle et al., 2004).

Hence, the soil system is complex and heterogeneous. The physical and chemical heterogeneity of soils can influence the variability of gaseous fluxes from or into the soil. Stoyan et al. (2000) reported small scale variability of $CO_2$ from soil respiration in relation to C contents of the soil. Soil moisture dynamics and water movement are influenced by topography, climate and the spatial distribution of soil parameters and vegetation (Ridolfi et al., 2003). Infiltration rates and unsaturated water distributions are further affected by macropore flow, which is highly dynamic and influenced by physical, biological and man-made disturbances (Šimůnek et al., 2003).

A detailed understanding of the spatial distribution and interactions of soil physical, chemical, biological and hydraulic properties and how these properties are altered by land use change are of great concern. The aim of this thesis is to enhance current methodology to quantify properties and interactions of the mentioned different soil characteristics. Focus is on soils of different land uses in sub-Saharan Africa, which is in respect to soil one of the least studied geographical regions.

### Soils and land use change in sub-Saharan Africa

In sub-Saharan Africa, agricultural land cover increased by more than 55% between 1975 and 2000 (Brink et al., 2009). Land cover changes, such as agricultural intensification or deforestation in tropical areas are interactively driven by climate change and land use changes (Lambin et al., 2003). The effect of land use on soils can be exemplified in the mountain regions of Tanzania. Rain forests of mountainous areas are often the primary water source for densely populated rural and urban areas (Bjørndalen, 1992). The Mt. Kilimanjaro area is one of the largest mountain massifs in Tanzania and provides fertile soils, favourable climatic conditions and stable water supply for many people living at the southern slopes of the mountain (Rohr et al., 2003). The growing population, however, induced major land use changes (Misana et al., 2003). Former forested areas in the submontane zone of the mountain were converted into agroforestry systems or coffee plantations. Natural savannah vegetation in the colline zone of Mt. Kilimanjaro was converted to agricultural fields, where maize, cassava or beans are grown (Misana et al., 2003). An intensification in land and water use could result in decreased water availability downstream. Mbonile (2005) highlighted intensive water conflicts in the Pangani River Basin, which receives an essential part of its water from Mt. Kilimanjaro. Detailed knowledge about the soil water storage capabilities of different land uses and how it is affected by global change is urgently needed.

### Measuring soil characteristics using visible and near infrared diffuse reflectance spectroscopy

Visible and near infrared diffuse reflectance spectroscopy (Vis-NIR DRS) is an inexpensive and widely used analytical tool to assess various soil properties simultaneously (Shepherd et al., 2002; Viscarra Rossel et al., 2006; Stenberg et al., 2010, Box 1). Compared to classical laboratory analysis, Vis-NIR DRS measurements are mostly non-destructive, faster and less expensive. It is increasingly used for analyses of a wide range of soil parameters, especially in Africa, where few soil laboratories exist (Shepherd et al., 2007).

**Box 1. Visible to near infrared diffuse reflectance spectroscopy**

Spectroscopy in general describes the interaction between electromagnetic radiation and matter (Miller, 2001). The term diffuse reflectance refers to indirect interactions, as not the direct radiation from an object is measured but the percent of reflectance. Visible to near infrared is the wavelength region of the electromagnetic radiation.

Visible to near infrared diffuse reflectance spectroscopy (Vis-NIR DRS) therefore characterises the reflectance of radiation with wavelength between 350 and 2500 nm from any matter. The chemical and physical structure of matter can thus be inferred by the interactions between radiation and matter. The technique was first developed by Ben-Gera et al. (1968) for the determination of fat and moisture in meat products and has been widely used and extended to other fields since. In soil spectroscopy, Vis-NIR DRS refers to the full range of wavelengths, whereas the wavelength region between 1000 and 2500 nm is often called short wave infrared in remote sensing (Ben-Dor et al., 1994).

**Physical and chemical principles**

Molecules have static properties like the atomic composition, and dynamic properties like molecular rotation and vibration (Miller, 2001). When molecules absorb electromagnetic radiation, they transfer the radiation energy to vibrations. As molecule bonds only absorb radiation with a certain energy, which corresponds to the difference between two vibrational states, characteristic absorption bands with characteristic intensities can be seen.

There are different forms of vibrations, i.e. stretching and bending of the molecular bond, which each show characteristic absorption features (Miller, 2001). These fundamental vibrations and thus absorption bands occur mostly in the mid-infrared region (2500 to 25000 nm) (Hunt, 1977). In Vis-NIR DRS the overtones and combinations of these fundamental vibrations are seen and used for interpretation (Viscarra Rossel et al., 2010).

**Soil spectra**

Soil is a complex mixture consisting mainly of minerals and organic matter with many different substances like acids, lipids, carbohydrates, proteins, lignin and cellulose (Kögel-Knabner, 2002; Baldock, 2007). Typical molecular bonds of organic matter for example are those of C-H, N-H and O-H (Stuart, 2004). Secondary soil minerals typically show absorption features caused by Al-OH or Mg-OH bonds of minerals or the O-H bond of molecular water inside the mineral (Hunt, 1977; Clark et al., 1990). Considering the amount of different molecular bonds within one molecule and furthermore the amount of different substances that contribute to soils, it is not possible to assign clear bands directly to a certain property. Statistical analysing techniques, however, give the possibility to build models, that relate the spectral information, i.e. all absorption features at the given wavelength range, to the required parameter.

Multiple studies confirm the ability of Vis-NIR DRS to derive soil information from spectra of dried and sieved soil samples, often using partial least squares regression (PLSR) (Shepherd et al., 2002; Viscarra Rossel et al., 2006; Vågen et al., 2006; Awiti et al., 2008; Stevens et al., 2013). However, the application of Vis-NIR DRS in the field remains challenging and models calibrated on air-dried spectra often fail to predict soil properties from in-situ spectra (Nocita et al., 2013). Indeed, in-situ spectra differ from those collected on air-dried sieved soil samples by soil moisture, surface roughness and bulk density (Chang et al., 2001; Morgan et al., 2009; Nocita

et al., 2011). One of the most important influences on the in-situ spectra is the varying water content in the field. Nocita et al. (2013) and Rodionov et al. (2014) for example tested an approach to classify samples according to their moisture content and calibrate individual models for different moisture classes.

Another method to improve predictions is to add a few new samples into a larger calibration database (Viscarra Rossel et al., 2009). This approach, often called "spiking", has been widely used for air-dried spectra from a small target area (Shepherd et al., 2002; Brown, 2007; Sankey et al., 2008; Guerrero et al., 2010; Wetterlind et al., 2010), as these new samples might differ greatly in their spectral characteristics. The collection of new samples, however, is time demanding and an alternative is needed to effectively apply Vis-NIR DRS to derive in-situ soil information.

### General objectives

The goal of this thesis is to characterise the soils of different land uses at Mt. Kilimanjaro, regarding their carbon and water storage potentials. Detailed spatial patterns of various soil physical, chemical and hydraulic properties were analysed. In order to obtain these properties at high resolution, the suitability of in-situ visible and near infrared diffuse reflectance spectroscopy was evaluated.

The specific methodological objectives were:

i) to visualize the variability of clay content in-situ (Manuscript 1),

ii) to increase the accuracy of spectral calibration models for the prediction of soil C content from in-situ field spectra (Manuscript 2),

iii) to develop a framework that can be used to predict soil C content with an existing spectral database and limited in-situ spectra (Manuscript 3),

iv) to test if this framework is applicable for other soil physical and chemical parameters, namely clay, silt, sand and nitrogen (N) content (Manuscript 4),

v) to assess the performance of random forest models for the prediction of soil hydraulic properties from the basic soil physical and chemical parameters (Manuscript 4),

vi) to evaluate if basic soil physical and chemical properties, derived from Vis-NIR DRS, can increase the prediction accuracy of microbial biomass with multivariate spatial interpolation techniques (Manuscript 5).

## Material and Methods

### Study area and research project

Mt. Kilimanjaro, the highest mountain in Africa (Figure II), is located in north-east Tanzania, close to the border to Kenya, about 3° south of the equator (Figure III). This ancient shield volcano covers an elevation from 800 m up to 5895 m a.s.l. and an area of about 5000 $km^2$. Its formation began 2.5 million years ago and different phases of volcanic activity followed (Nonnotte et al., 2008). The volcanic soils that consequently formed on the superficial deposits are relatively young and the volcanic origin is still visible, even in soils of the lower zones of the mountain. An overview of soil types of the colline and submontane zones is given in Appendix C.

Figure II: Mt. Kilimanjaro seen from the submontane zone in the south (left) and from the colline zone in the east (right)

Due to its height, several ecosystems differentiated along the elevational gradient. An overview of all the elevational zones is given in Lambrechts et al. (2002) and Misana et al. (2003). This elevational gradient provides optimal conditions to study implications of land use and climate changes and was therefore chosen by the research unit "Kilimanjaro ecosystems under global change: Linking biodiversity, biotic interactions and biogeochemical ecosystem processes (KiLi)". The following work was conducted as part of this research unit. Figure IV highlights how this work connects with other projects of KiLi.

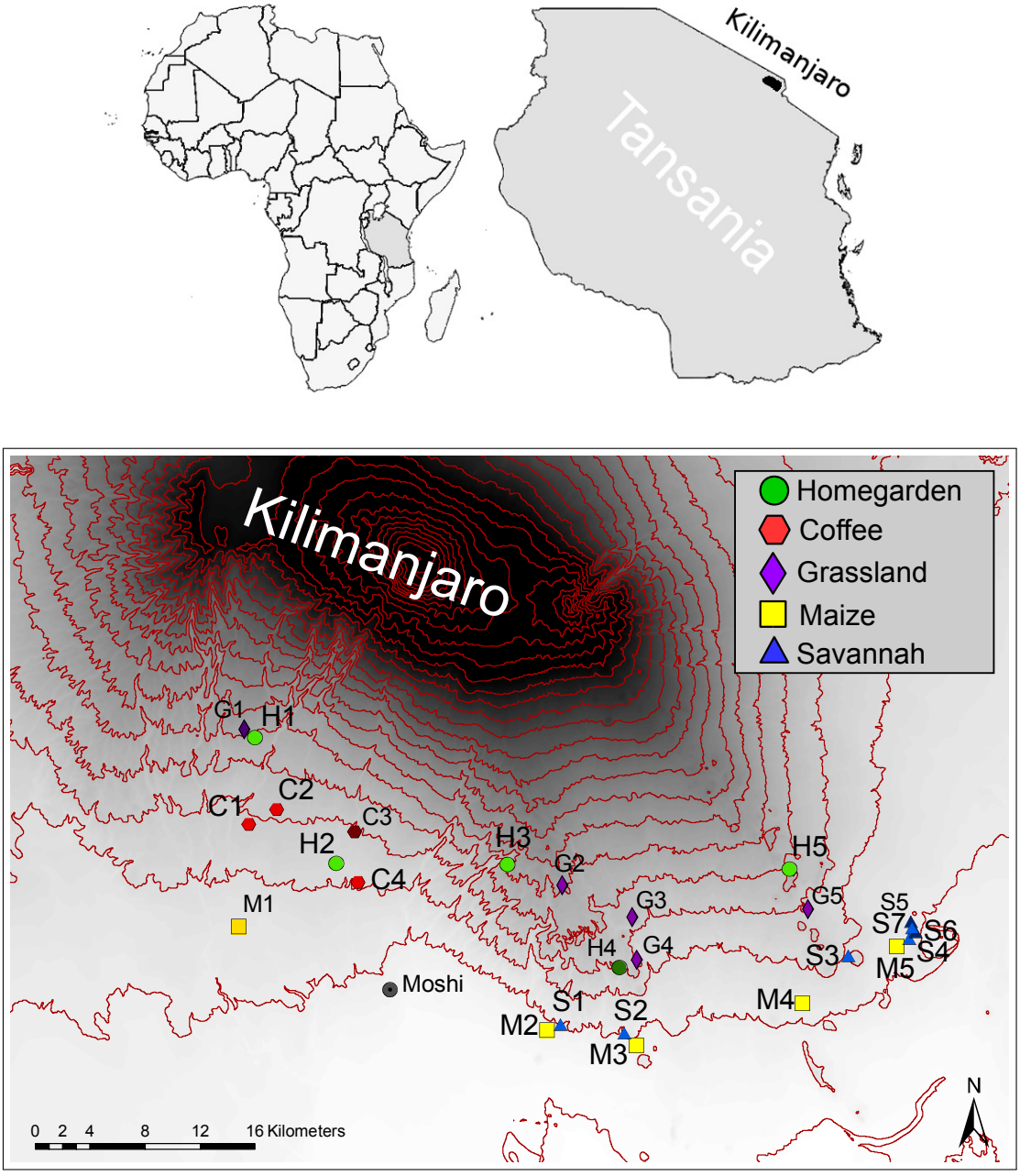



Figure III: Study area and research plots. H = homegarden, C = coffee, G = grassland, M = maize, S = savannah, Source: commons.wikimedia and OpenStreetMap.

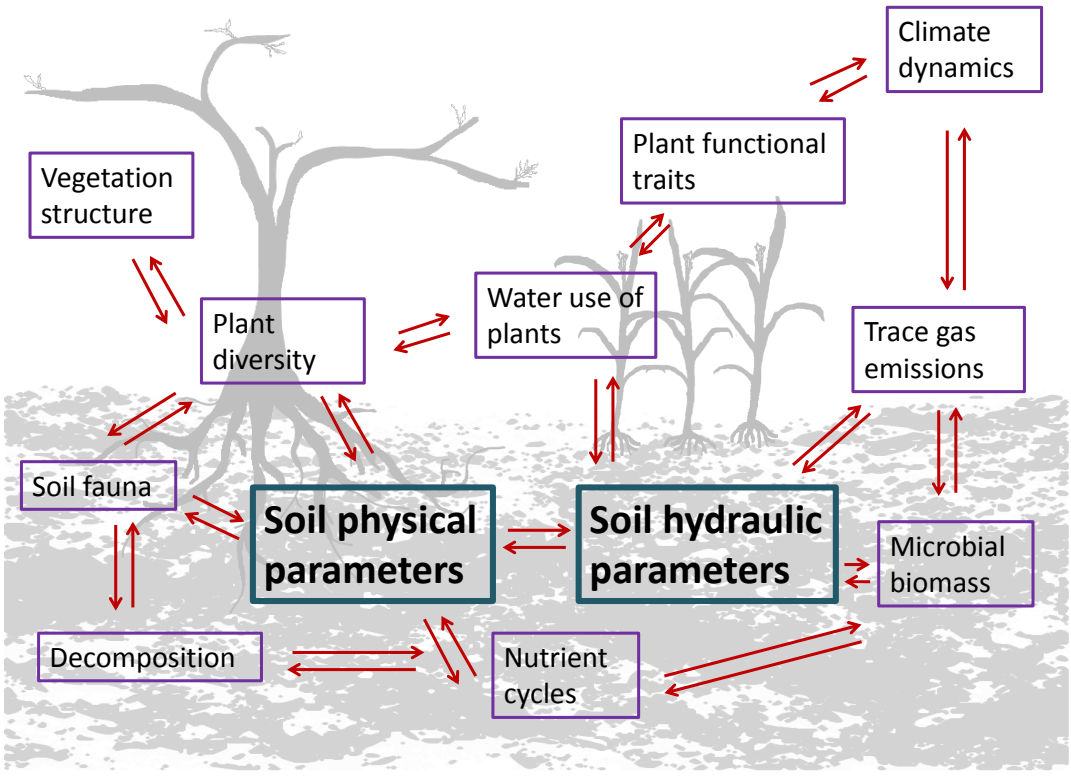



Figure IV: Integration of this dissertation into the research unit KiLi (Kilimanjaro ecosystems under global change). Main focus of this thesis are the soil physical and soil hydraulic parameters across different land uses (indicated in bold). These parameters are linked to various other parameters and processes, of which several are addressed by other members of the research unit. This figure is not exhaustive and only highlights important points from a soil physicists point of view.

The main focus in this thesis is on the land uses of the mountain's submontane (Figure V) and colline zones (Figure VI). At least one plot in each land use was chosen for a detailed soil profile study (Manuscript 1, 3 and 4). The chosen land uses were homegarden, a traditional agroforestry system, coffee plantation and grassland of the submontane zone, and maize field and two different natural savannahs of the colline zone. The respective soil types were Sodic Vertisol, Haplic Vertisol, Haplic Andosol, Vitric Cambisol, Sodic Vertisol and Rendzic Leptosol.

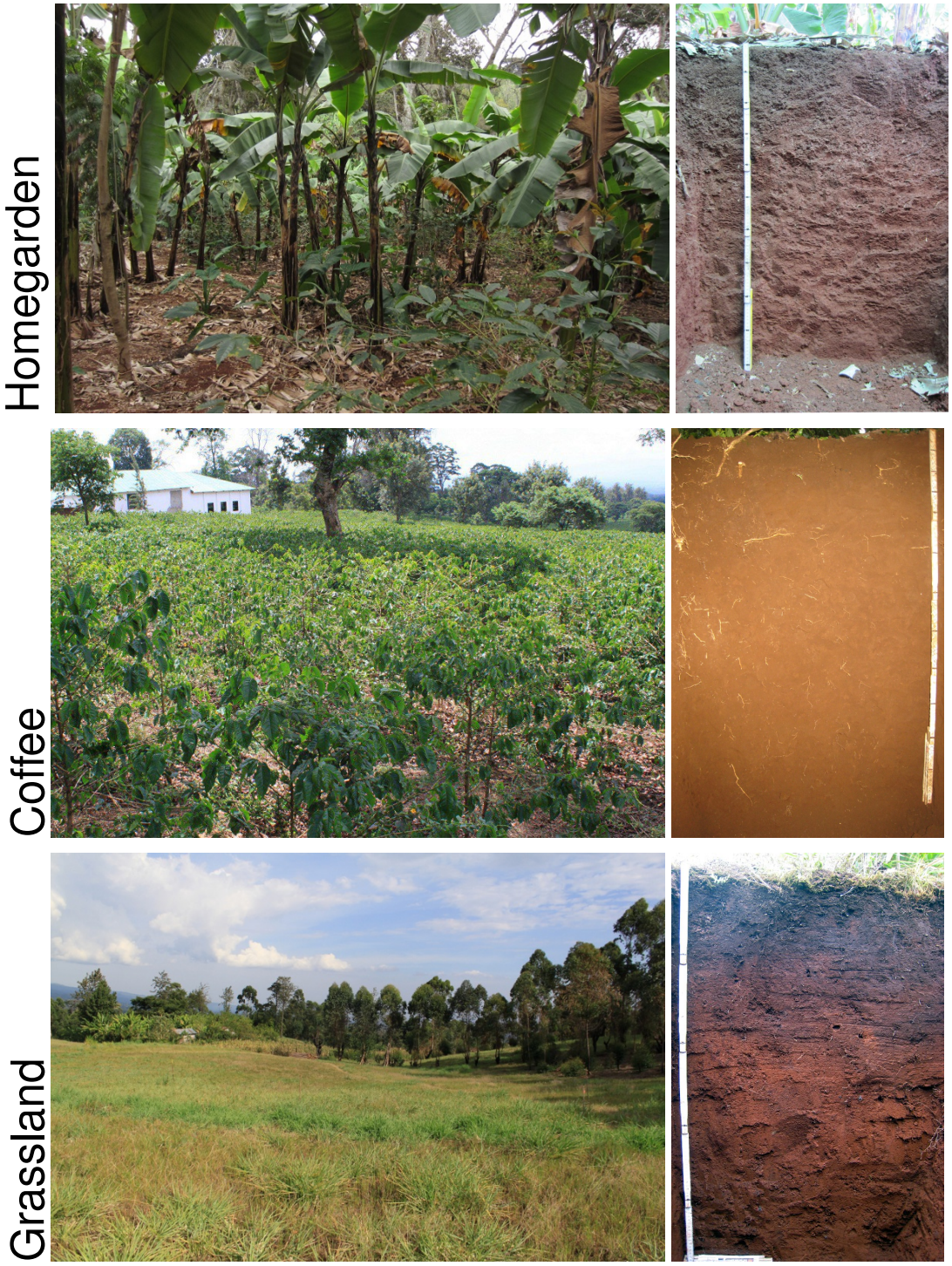



Figure V: Different land uses of the submontane zone of Mt. Kilimanjaro. Homegarden (Sodic Vertisol), coffee plantation (Haplic Vertisol) and grassland (Haplic Andosol).

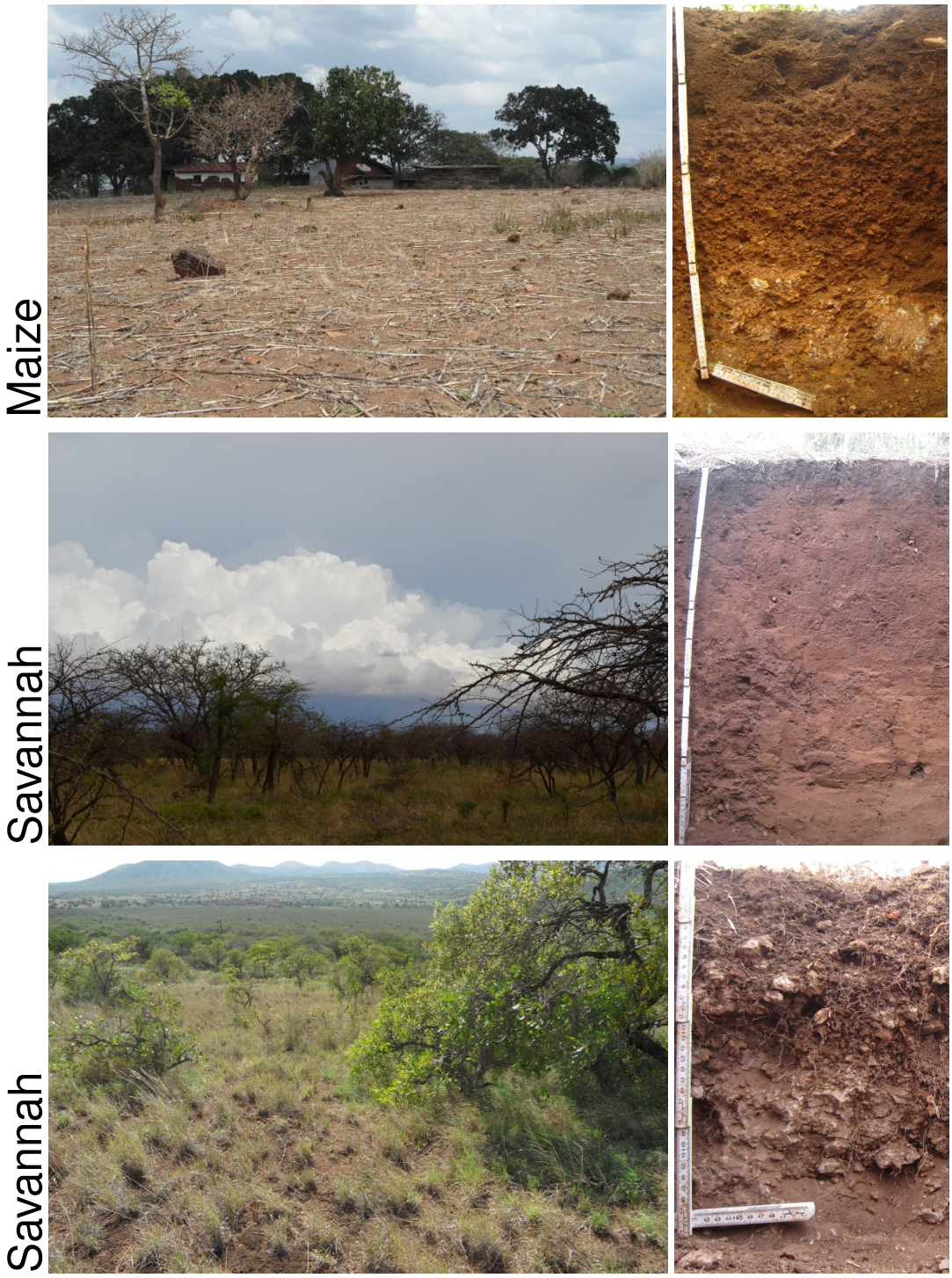



Figure VI: Different land uses of the colline zone of Mt. Kilimanjaro. Maize field (Vitric Cambisol), savannah with plain terrain (Sodic Vertisol), savannah at the slopes (Rendzic Leptosol).

## Measurements

### Spectral database

A regional database of land uses at Mt. Kilimanjaro was established, which included soil spectra, C and N content and soil texture. Samples were collected with a soil auger on 26 sites from different land uses and separated by soil horizon, making a total of 191 soil samples. All samples were dried at 45 ℃ and sieved to < 2mm prior to spectral analysis (Figure VII). This regional spectral database is the basis for the development of prediction models for various soil parameters.

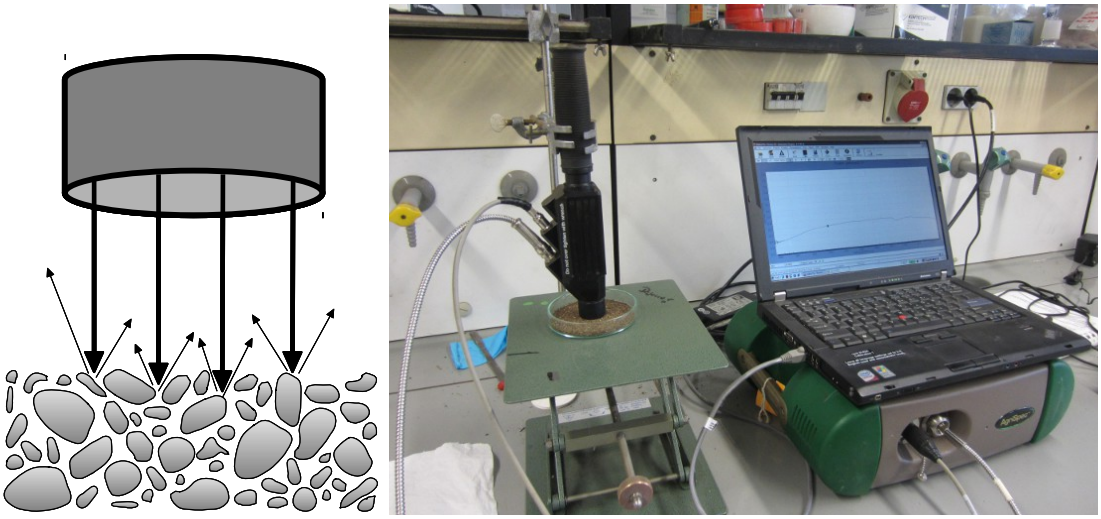


Figure VII: Schematic view of the contact probe of the spectrometer and air-dried measurement in the laboratory. Light with different wavelengths (350 – 2500 nm in visible to near infrared spectroscopy) is emitted to a surface. Depending on the physical and chemical properties of that surface, the individual wavelengths are reflected back to the detectors to a greater or lesser extent.

### Soil profile study (Manuscript 1, 3 and 4)

Covering all land uses, a detailed study of soil profiles was conducted. On six selected plots, a soil pit was dug to a depth of approximately 1 m or until continuous bedrock was reached. A frame of 0.5 m $\times$ 1 m with 3 $\times$ 3 cm segments was placed on one profile wall. Subsequently, each segment was scanned with the contact probe of a visible to near infrared spectrometer, resulting in about 500 in-situ soil spectra per profile (depending on soil depth). Finally, small soil cores reference samples were randomly extracted (Figure VIII). In addition, a classical soil sampling was conducted and bulk soil horizon samples collected.

A model developed with soil spectra from the regional database was now used to predict clay content from the in-situ spectra for four of the six profiles (Manuscript 1).

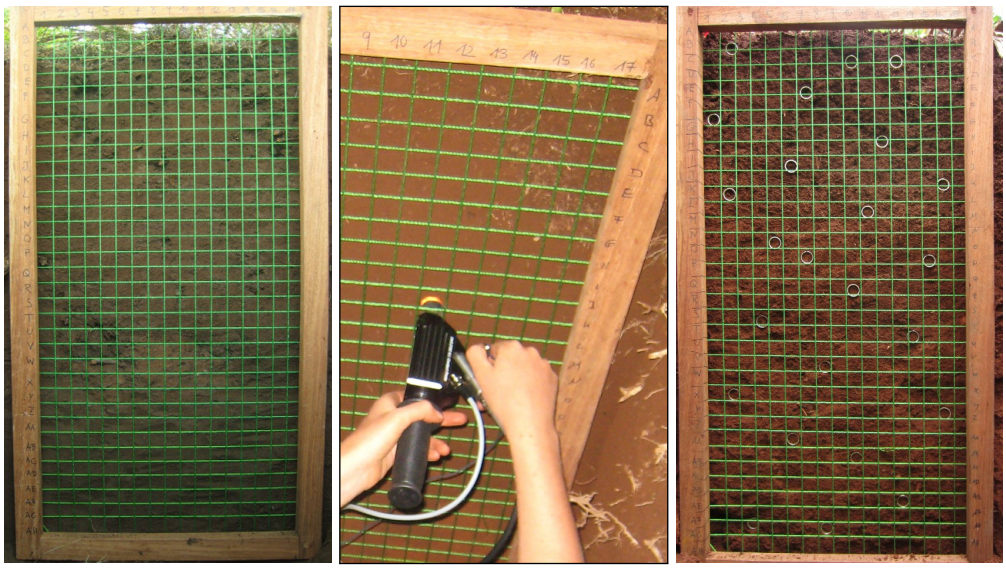


Figure VIII: Soil profile study: sampling frame in soil profile (left), scanning of segments with portable spectrometer (middle), random sampling with small soil cores for reference (right).

In Manuscript 2 we used in-situ spectra of one selected soil profile to test the applicability of the synthetic minority oversampling technique (SMOTE, Box 2). In-situ spectra of all six profiles were further chosen to study the difference between in situ soil spectra and soil spectra from the regional database (Manuscript 3). Furthermore we developed models for the prediction of C from in situ spectra (Manuscript 3). In Manuscript 4, various soil parameters of selected profiles were analysed in every segment of the profiles.

**Box 2. Synthetic minority oversampling technique**

If certain spectral characteristics are rare in a calibration database, properties of these rare spectra are predicted inaccurately. In order to increase the influence of the rare characteristics, the synthetic minority oversampling technique (SMOTE) and its extension for regression (Chawla et al., 2002; Torgo et al., 2013) can be used. New synthetic spectra are generated with the following equation:

$$X_s = X_o + \delta(X_n - X_o) \tag{1.1}$$

where $X_s$ is the synthetic spectra, $X_o$ the original spectra, $X_n$ a randomly chosen neighbour of $X_o$ and $\delta$ a random number between 0 and 1. Including these new spectra in the calibration helps to balance the data and potentially increases prediction accuracy for rare cases. Details on the application on in-situ spectra are described in Manuscript 2.

**Field scale study (Manuscript 5)**

The spatial heterogeneity and distribution of microbial biomass at the field scale was studied with geostatistical interpolations (Manuscript 5, Box 3). The sampling was designed in a hierarchically nested grid on a 15 m x 15 m plot in two savannah systems (Figure IX). Vis-NIR DRS was used to predict the co-variables (soil organic C and N and clay content) as input parameters for multivariate spatial interpolation methods. Maps of microbial biomass were created to visualize its spatial distribution. For more details see Manuscript 5.

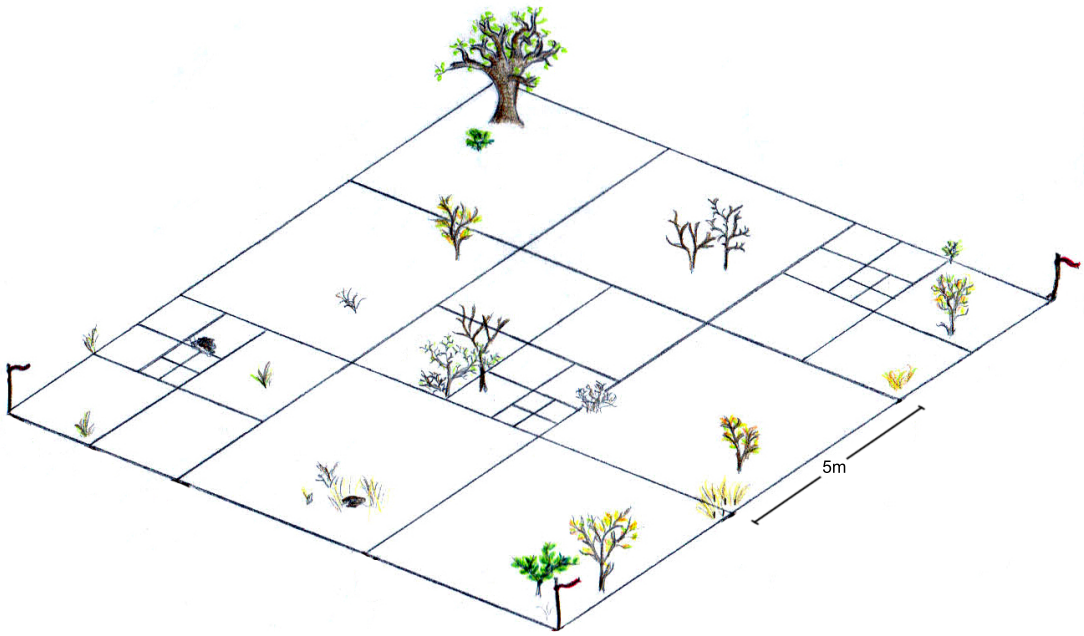



Figure IX: Design of heterogeneity campaign (Manuscript 5) with sampling points at each intersection/corner; 15 x 15m, smallest sampling interval = 0.625m.

### Box 3. Geostatistical interpolations

The spatial autocorrelation of a parameter can be analysed by calculating the semivariances $Y(h)$, dependent on the distance to each other (Matheron, 1963):

$$Y(h) = \frac{1}{2} \cdot \frac{1}{N(h)} \sum_{i=1}^{N(h)} (O(s_i) - O(s_i + h))^2 \tag{1.2}$$

where $N(h)$ is the number of compared point pairs per distance $h$, $O(s_i)$ is the value at the location $s_i$ and $O(s_i + h)$ is the value at the distance $h$ from the location $s_i$. Subsequently, the semivariances are averaged by distance intervals and displayed in the so called empirical variogram. These empirical variograms enable the examination of the spatial structure of a variable.

A variogram model can be fitted to the empirical semivariogram and predictions at unknown locations can be made with the following equation:

$$P(s_0) = \sum_{i=1}^{N} (\Omega_i(s_0) \cdot O(s_i)) \tag{1.3}$$

where $P(s_0)$ stands for the predicted value at location $s_0$, $\Omega$ is the spatial weighting function based on the semivariogram, $O(s_i)$ is the observation at location $s_i$ and $N$ is the number of observations. This procedure is called ordinary kriging (OK) and was developed by Krige (1951) and Matheron (1963) and uses information of only one variable. Any value of that variable at an unknown location is now calculated as a weighted linear combination of measured values at locations $s_i$ ($i$ = 1, 2, ..., $N$). In addition to OK, several multivariate methods have been developed, which include information on additional parameters. More information can be found in Manuscript 5.

## Overview of measured and predicted parameters

| data set | measured | predicted with Vis-NIR DRS | predicted with MI or RF | Manu-script |
|---|---|---|---|---|
| All plots: soil auger samples | C, N, clay, silt, sand | – | – | 1-5* |
| Soil profile study: small soil core reference samples | C, N, clay, silt, sand | – | – | 1-4* |
| Soil profile study: horizons | C, N, clay, silt, sand, $\theta_s$, $\theta_{1.8}$**, $\theta_{4.2}$, *AWC*, $\phi$, *FAC*, $K_{r(1.8)}$**, $\rho_b$ | – | – | 4 |
| Soil profile study: all segments | – | C, N, clay, silt, sand | $\theta_s$, $\theta_{1.8}$, $\theta_{4.2}$, *AWC*, $\phi$, *FAC*, $K_{r(1.8)}$, $\rho_b$ (RF) | 4 |
| Field study: 16 sampling points | C, N | – | – | 5 |
| Field study: 61 sampling points | $C_{mic}$ , $N_{mic}$ , pH | C, N, clay | – | 5 |
| Field study: regular grid | – | – | C, N, clay, pH, $C_{mic}$, $N_{mic}$ (MI) | 5 |

*some studies include only subsets, MI = multivariate interpolation, RF = random forest
**calculated from measured parameters

# Results and Discussion

## Overview of main outcomes

| **Manuscript** | Aims | Results/Conclusion |
|---|---|---|
| 1 | Visualization of small scale variability of clay content | In-situ prediction of clay content from air-dried soil spectra alone is not possible |
| 2 | Increasing the accuracy of Vis-NIR DRS to predict C content for rare data cases | SMOTE can be applied to spectroscopy and increases prediction accuracy of PLSR models for rare data cases |
| 3 | Development of a framework for the prediction of C content from in-situ spectra | Framework for prediction of C content from in-situ soil spectra was successfully implemented |
| 4 | Application of in-situ Vis-NIR DRS to other soil parameters; Prediction of soil hydraulic properties from basic soil parameters; Identification of small scale spatial pattern in different land uses | Accuracy of predictions depends on study site and soil parameter; spatial pattern of soil hydraulic parameters differ between land uses |
| 5 | Prediction of basic soil parameters from Vis-NIR DRS; Evaluation of multivariate spatial interpolation techniques to predict microbial biomass; Characterization of microbial biomass distribution | Multivariate techniques perform better than ordinary kriging; microbial biomass distribution is heterogeneous |

## First steps towards in situ prediction of soil properties by Vis-NIR DRS

In the first study, we visualized the spatial pattern of clay content in soil profiles of four different land uses, namely homegarden, coffee, maize and savannah (Manuscript 1). Previously, a model for the prediction of clay content was developed based on air-dried, sieved soil spectra from the whole Kilimanjaro region. The model itself performed well, with $R^2$ of 0.84 and the results are comparable to other studies (Shepherd et al., 2002; Stenberg, 2010). The prediction accuracy for in-situ samples on the other hand was very poor and only clay content of the

homegarden profile could be predicted with a $RMSE$ of < 5% clay compared to 12 – 30% for the other profiles. The main outcome of Manuscript 1 was therefore, that in-situ prediction of clay content from air-dried soil spectra alone is not possible. Consequently, we didn't analyse the spatial structure of clay content further in this study and worked towards model development instead.

A number of studies confirm that including new data in a spectral database (i.e. "spiking" that database) improves predictions for small target areas (Brown, 2007; Sankey et al., 2008). Shepherd et al. (2002), for instance, showed that when analysing a new area that lacks representative soil samples in the calibration database, adding some samples to the calibration can help.

The first approach to increase prediction accuracy of the spectral models was therefore to add spectra from the target area to the database (Appendix A). This approach was implemented for two savannah sites and three different sampling schemes for soil organic carbon and nitrogen. The prediction accuracy for C improved for almost all sampling schemes when new spectra were added.

Nevertheless, predictions were still not very accurate, especially for models with in-situ spectra. A probable reason could be, that the amount of samples used for spiking was insufficient to have a major influence on the spectral database. Furthermore, the datasets used for spiking in that study were taken under different conditions. Even though the PLSR models now included some similar spectral characteristics from the target site, the specific conditions at the time of scanning were not incorporated. Details of the sampling schemes and prediction accuracies can be found in Appendix A.

Therefore we analysed the spectral characteristics of in-situ spectra and air-dried spectra thoroughly (Manuscript 2 and Manuscript 3). To better identify differences we split the regional database according to land use. We could show that most in-situ soil spectra differ substantially from the air-dried spectra. This is in agreement with other studies and is probably mostly due to moisture content and sample preparation (Lobell et al., 2002; Nocita et al., 2013).

Figure X shows exemplary the spectral characteristics from all in-situ and air-dried spectra from one local database and in addition some moist spectra (illustrated with the first two principal components). The in-situ spectra were clearly separated from the local set. Some moist spectra were close to the in-situ spectra, others however, showed even more diverse characteristics than the air-dried local spectra. This demonstrates clearly, that field moisture is not the only problem, when dealing with in-situ spectra. Additionally, smearing of soil and differences in size and shape can lead to contrasting reflectance behaviour compared to sieved soil samples (Wetzel, 1983; Chang et al., 2001; Vågen et al., 2006; Morgan et al., 2009; Gras et al., 2014).

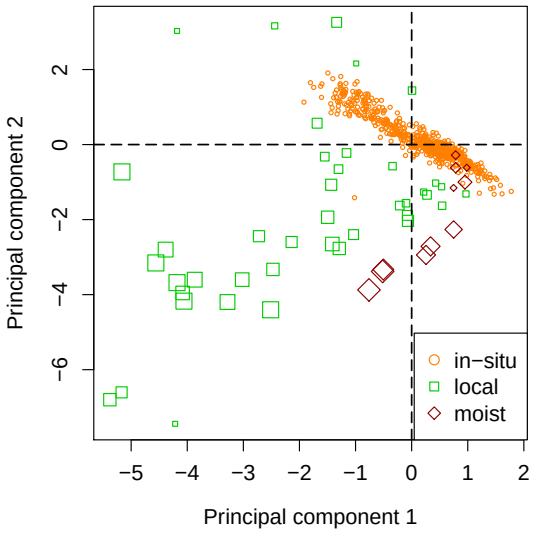



Figure X: Score plot of the first two principal components of in-situ spectra (circles) and air dried spectra of the local homegarden database (squares). Additionally, the projection of moist spectra is shown (diamonds). The symbol size for the local and moist data points is scaled according to their C content.

Exemplary, reflectance spectra of two soil profiles are displayed in Figure XI. Differences between in-situ and air-dried spectra are striking. However, not all wavelengths of the spectra are affected in the same way. The wavelengths around 1400 nm and 1900 nm are altered to a lesser extent, yet inconsistently for A- and B-horizons and for the two sites.

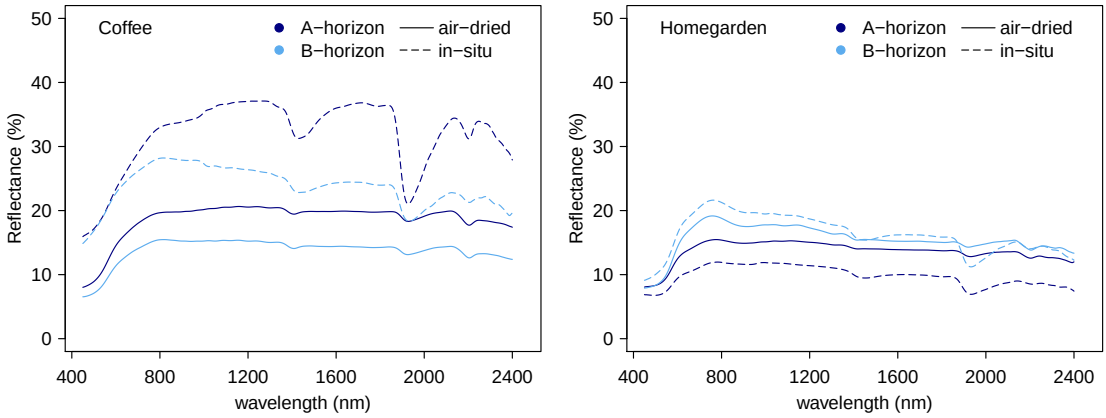



Figure XI: Reflectance spectra of two different soil horizons of a homegarden and a coffee plantation on Mt. Kilimanjaro in the visible and near infrared range; in-situ spectra were taken on the wall of the soil profile, air-dried refers to spectra taken at collected soil auger samples, that have been sieved and dried at 45℃.

### Increasing the accuracy of in-situ predictions for Vis-NIR spectroscopy

Based on these previous findings, we aimed at including the in-situ spectra in the calibration model in order to cover all in-situ spectral characteristics, i.e. soil moisture content, surface conditions and differences due to bulk density. As a large number of additional reference samples would be required for spiking a model in order to gain enough influence of the in-situ samples, the idea of applying synthetic minority oversampling technique (SMOTE) to spectroscopy was born (Manuscript 2). The new spectra, that were generated with SMOTE from the in-situ spectra, were consequently added to the already existing regional or local databases (Manuscript 3).

Using SMOTE, new synthetic spectra can be generated from only a limited amount of reference samples. Spiking a local database with synthetic soil spectra clearly improved predictions of soil organic C for in-situ samples (Manuscript 2 and Manuscript 3). Few other studies working on including in-situ spectra in an air-dried spectral database exist. However, Viscarra Rossel et al. (2009) demonstrated that predictions of clay content from field spectra improved when adding in-situ soil spectra to the calibration database.

We could show that the calibration models learned the relationship between in-situ spectra and C content and were therefore able to predict C content on the in-situ spectra. An additional advantage of spiking the calibration dataset with synthetic spectra is that we can simultaneously incorporate in-situ spectral characteristics and the features from a new site into the calibration model.

We concluded from Manuscript 2 and Manuscript 3, that prediction of C content from in-situ spectra is possible by applying SMOTE to spectroscopy. This approach is especially promising if only a limited amount of field reference data is available.

### Spectral predictions as input for pedotransfer functions

Spectra, that were generated by SMOTE, were included in calibration models for the prediction of several other soil parameters namely clay, silt, sand and nitrogen content (Manuscript 4). However, the prediction accuracy for these parameters differed and were smaller than those of C content for most sites.

The basic soil parameters clay, sand and C content are used in many pedotransfer functions to predict soil hydraulic properties (Berg et al., 1997; Minasny et al., 2011). Using clay, sand and C content of bulk soil horizon samples from the classical sampling of four profiles, we built pedotransfer functions. Although only a limited amount of samples was available, satisfactory predictions for most points in the profiles were achieved. Several physical and hydraulic properties were consequently predicted for all segments of the soil profiles with basic soil parameters

from Vis-NIR DRS. We could show that median predictions in the soil were mostly in line with measured hydraulic parameters.

Combining these two methods enabled us to study small scale variations of several important physical and hydraulic properties in-situ. Further details on how these properties vary between the different land uses can be found in Manuscript 4.

### Spectral predictions as input for geostatistical models

In Manuscript 5 we used basic soil parameters, that were mostly derived from Vis-NIR DRS to increase the spatial prediction accuracy of geostatistical models. We could show, that the SMOTE approach from Manuscript 2 can also be applied to air-dried spectra. In this case, the spectral characteristics from two savannah field sites were lacking in the database. PLSR models that included synthetic spectra generated from air-dried spectra of the field sites performed well. For further details see Appendix D.

Accordingly, using different input parameters derived from Vis-NIR DRS in the multiple spatial interpolation methods to predict microbial biomass improved predictions, compared to the ordinary kriging interpolation. The improvement of prediction accuracy by using multivariate approaches has already been reported in several studies (Bourennane et al., 2000; Mishra et al., 2010; Li et al., 2011; Mishra et al., 2012). Information on which input parameters are useful for the prediction of microbial biomass out of those which can be derived easily, is however limited. In this study we could show, that microbial biomass can be spatially predicted with increased accuracy by including the explanatory variables, C, N, clay content and soil pH (or a combination of only some of these, depending on parameter and site). Details on the pattern and distribution of microbial biomass can be found in Manuscript 5.

The main outcomes of this study are, that C, N, clay content and soil pH were suitable variables to predict spatial relations of soil microbial parameters at small scales in natural savannah ecosystems of East Africa. Most of these parameters can be derived with Vis-NIR DRS. The combination of geostatistical methods with Vis-NIR DRS can thus further reduce laboratory efforts and costs.

## Conclusions

In order to study small scale heterogeneity on the intact soil, a technique to obtain many non-destructive samples is required. Vis-NIR DRS has such potential. However, until now, in-situ spectral measurements have not been reliable, as the differing field conditions influence the soil spectra substantially (Manuscript 1). By applying SMOTE to spectroscopy and thus combining spectra in calibration models, in-situ predictions of C content are possible (Manuscript 2 and 3). This concept is extendible to other soil parameters. Prediction accuracy of the calibration models, however, differs between variables and sites (Manuscript 4). Nevertheless, as many parameter can be derived from the same spectra, multivariate datasets can be created easily (Manuscript 4). These datasets provide optimal input for the acquisition of additional information, for example as input parameters for pedotransfer functions or multivariate interpolation techniques (Manuscript 4 and 5). The combination of different prediction and interpolation techniques thus enables us to gather detailed soil information at various scales on the in-situ soil.

## Record of contributions to the included manuscripts

**Manuscript 1**

**Visualizing small scale variability of soil chemical properties on Mt. Kilimanjaro by Vis-NIR spectroscopy**

| | | |
|---|---|---|
| Anna Kühnel | 75 % | Experimental design, field sampling, laboratory work, analysing the data, writing the manuscript |
| Christina Bogner | 15 % | Discussion on data analysis, comments to improve the manuscript |
| Holger Pabst | 5 % | Field sampling, comments to improve the manuscript |
| Bernd Huwe | 5 % | Discussion on experimental design, comments to improve the manuscript |

**Manuscript 2**

**Predicting with limited data – Increasing the accuracy in Vis-NIR diffuse reflectance spectroscopy by SMOTE**

| | | |
|---|---|---|
| Christina Bogner | 70 % | Idea, model development, analysing the data, writing the manuscript |
| Anna Kühnel | 25 % | Field sampling, laboratory work, data preparation, comments to improve the manuscript |
| Bernd Huwe | 5 % | Comments to improve the manuscript |

**Manuscript 3**

| **In-situ prediction of soil organic carbon by Vis-NIR spectroscopy with limited data** | | |
|---|---|---|
| Anna Kühnel | 47.5 % | Experimental design, field sampling, laboratory work, analysing the data, writing the manuscript |
| Christina Bogner | 47.5 % | Idea, model development, data analysis, writing the manuscript |
| Bernd Huwe | 5 % | Experimental design, comments to improve the manuscript |

**Manuscript 4**

| **Small scale spatial variability of soil hydraulic properties in different land uses at Mt. Kilimanjaro** | | |
|---|---|---|
| Anna Kühnel | 80 % | Experimental design, field sampling, laboratory work, analysing the data, writing the manuscript |
| Christina Bogner | 10 % | Model development, comments to improve the manuscript |
| Bernd Huwe | 10 % | Experimental design, comments to improve the manuscript |

**Manuscript 5**

| **Spatial patterns of microbial biomass and fauna activity in savannah soils at Mt. Kilimanjaro** | | |
|---|---|---|
| Anna Kühnel | 43 % | Experimental design, field sampling, acquisition of spectral data, data preparation and analysis, preparation of manuscript |
| Holger Pabst | 43 % | Field sampling, microbial biomass, data preparation and analysis, preparation of manuscript |
| Juliane Röder | 3 % | Field sampling, provision of bait lamina data, discussions on the results, suggestions to improve the manuscript |
| Christina Bogner | 3 % | Data evaluation, suggestions to improve the manuscript |
| Yakov Kuzyakov | 3 % | Discussions on the results, suggestions to improve the manuscript |
| Bernd Huwe | 7 % | Field sampling, discussion on the results, suggestions to improve the manuscript |

## References


Awiti, A. O., M. G. Walsh, K. D. Shepherd, and J. Kinyamario (Jan. 2008). "Soil condition classification using infrared spectroscopy: A proposition for assessment of soil condition along a tropical forest-cropland chronosequence". In: *Geoderma* 143.1-2, pp. 73–84. DOI: 10.1016/j.geoderma.2007.08.021.

Baldock, J. A. (2007). "Composition and Cycling of Organic Carbon in Soil". In: *Soil Biology*. Ed. by P. Marschner and Z. Rengel. Vol. 10. Springer Berlin Heidelberg, pp. 1–35. DOI: 10.1007/978-3-540-68027-7_1.

Batjes, N. (1996). "Total carbon and nitrogen in the soils of the world". In: *European Journal of Soil Science* 47.2, pp. 151–163. DOI: 10.1111/j.1365-2389.1996.tb01386.x.

Ben-Dor, E. and A. Banin (1994). "Visible and near-infrared (0.4-1.1 [mu]m) analysis of arid and semiarid soils". In: *Remote Sensing of Environment* 48.3, pp. 261–274. DOI: DOI:10.1016/0034-4257(94)90001-9.

Ben-Gera, I. and K. H. Norris (1968). "Direct Spectrophotometric Determination of Fat and Moisture in Meat Products". In: *Journal of Food Science* 33.1, pp. 64–67. DOI: 10.1111/j.1365-2621.1968.tb00885.x.

Berg, M. van den, E. Klamt, L. P. van Reeuwijk, and W. G. Sombroek (1997). "Pedotransfer functions for the estimation of moisture retention characteristics of Ferralsols and related soils". In: *Geoderma* 78.3-4, pp. 161–180. DOI: 10.1016/S0016-7061(97)00045-1.

Bjørndalen, J. E. (1992). "Tanzania's vanishing rain forests – assessment of nature conservation values, biodiversity and importance for water catchment". In: *Agriculture, Ecosystems & Environment* 40.1-4. Biotic Diversity in Agroecosystems, pp. 313–334. DOI: 10.1016/0167-8809(92)90100-P.

Bot, A. and J. Benites (2005). *The importance of soil organic matter: Key to drought-resistant soil and sustained food production*. 80. Food & Agriculture Org.

Bourennane, H., D. King, and A. Couturier (2000). "Comparison of kriging with external drift and simple linear regression for predicting soil horizon thickness with different sample densities". In: *Geoderma* 97.3-4, pp. 255–271. DOI: 10.1016/S0016-7061(00)00042-2.

Brink, A. B. and H. D. Eva (2009). "Monitoring 25 years of land cover change dynamics in Africa: A sample based remote sensing approach". In: *Applied Geography* 29.4, pp. 501–512. DOI: 10.1016/j.apgeog.2008.10.004.

Brown, D. J. (2007). "Using a global VNIR soil-spectral library for local soil characterization and landscape modeling in a 2nd-order Uganda watershed". In: *Geoderma* 140.4, pp. 444–453. DOI: 10.1016/j.geoderma.2007.04.021.

Chang, C.-W., D. A. Laird, M. J. Mausbach, and C. R. Hurburgh (2001). "Near-Infrared Reflectance Spectroscopy-Principal Components Regression Analyses of Soil Properties". In: *Soil Science Society of America Journal* 65.2, pp. 480–490. DOI: 10.2136/sssaj2001.652480x.

Chawla, N. V., K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer (2002). “SMOTE: Synthetic Minority Over-sampling Technique”. In: *Journal of Artificial Intelligence Research* 16, pp. 321–357.

Clark, R. N., T. V. V. King, M. Klejwa, G. A. Swayze, and N. Vergo (1990). “High spectral resolution reflectance spectroscopy of minerals”. In: *Journal of Geophysical Research: Solid Earth* 95.B8, pp. 12653–12680. DOI: 10.1029/JB095iB08p12653.

Eswaran, H., E. Van Den Berg, and P. Reich (1993). “Organic carbon in soils of the world”. In: *Soil science society of America journal* 57.1, pp. 192–194. DOI: 10.2136/sssaj1993.03615995005700010034x.

Fraiture, C. de, D. Molden, U. Amarasinghe, and I. Makin (2001). “PODIUM: Projecting water supply and demand for food production in 2025”. In: *Physics and Chemistry of the Earth, Part B: Hydrology, Oceans and Atmosphere* 26.11-12, pp. 869–876. DOI: 10.1016/S1464-1909(01)00099-5.

Gras, J.-P., B. G. Barthès, B. Mahaut, and S. Trupin (Feb. 2014). “Best practices for obtaining and processing field visible and near infrared (VNIR) spectra of topsoils”. In: *Geoderma* 214–215.0, pp. 126–134. DOI: 10.1016/j.geoderma.2013.09.021.

Guerrero, C., R. Zornoza, I. Gómez, and J. Mataix-Beneyto (2010). “Spiking of NIR regional models using samples from target sites: Effect of model size on prediction accuracy”. In: *Geoderma* 158.1-2, pp. 66–77. DOI: 10.1016/j.geoderma.2009.12.021.

Hunt, G. (1977). “Spectral Signatures of Particulate Minerals in the Visible and Near Infrared”. In: *Geophysics* 42.3, pp. 501–513. DOI: 10.1190/1.1440721.

Hussey, R. and R. Roncadorl (1982). “Vesicular-arbuscular mycorrhizae may limit nematode activity and improve plant growth”. In: *Plant Disease* 66.1, pp. 9–14.

IPCC, ed. (2008). *Climate change 2007: Impacts, adaptation and vulnerability : Working Group II contribution to the Fourth Assessment Report of the IPCC Intergovernmental Panel on Climate Change*. Geneva: Intergovernmental Panel on Climate Change.

Jenny, H. et al. (1941). *Factors of soil formation*. McGraw-Hill Book Company New York, NY, USA.

Krige, D. G. (1951). “A Statistical Approach to Some Basic Mine Valuation Problems on the Witwatersrand”. In: *Journal of the Chemical, Metallurgical and Mining Society of South Africa* 52.6, pp. 119–139. DOI: 10.2307/3006914.

Kögel-Knabner, I. (Feb. 2002). “The macromolecular organic composition of plant and microbial residues as inputs to soil organic matter”. In: *Soil Biology and Biochemistry* 34.2, pp. 139–162. DOI: 10.1016/S0038-0717(01)00158-4.

Lambin, E. F., H. J. Geist, and E. Lepers (2003). “Dynamics of land-use and land-cover change in tropical regions”. In: *Annu. Rev. Environ. Resourc.* 28.1, pp. 205–241. DOI: 10.1146/annurev.energy.28.050302.105459.

Lambrechts, C., B. Woodley, A. Hemp, C. Hemp, and P. Nnyiti (2002). *Aerial survey of the threats to Mt. Kilimanjaro forests*. Dar es Salaam; Tanzania.

Li, J. and A. D. Heap (2011). “A review of comparative studies of spatial interpolation methods in environmental sciences: Performance and impact factors”. In: *Ecological Informatics* 6.3-4, pp. 228–241. DOI: 10.1016/j.ecoinf.2010.12.003.

Lobell, D. B. and G. P. Asner (2002). “Moisture Effects on Soil Reflectance”. In: *Soil Science Society of America Journal* 66.3, pp. 722–727. DOI: 10.2136/sssaj2002.7220.

Matheron, G. (1963). “Principles of geostatistics”. In: *Economic Geology* 58.8, pp. 1246–1266. DOI: 10.2113/gsecongeo.58.8.1246.

Mbonile, M. J. (2005). “Migration and intensification of water conflicts in the Pangani Basin, Tanzania”. In: *Habitat International* 29.1, pp. 41–67. DOI: 10.1016/S0197-3975(03)00061-4.

Miller, C. E. (2001). “Near-infrared technology in the agricultural and food industries”. In: ed. by W. P. and N. K. Vol. 2. Minnesota, USA: American Association of Cereal Chemists. Chap. Chemical principles of near infrared technology, pp. 19–37.

Minasny, B. and A. E. Hartemink (2011). “Predicting soil properties in the tropics”. In: *Earth-Science Reviews* 106.1-2, pp. 52–62. DOI: 10.1016/j.earscirev.2011.01.005.

Misana, S. B., A. E. Majule, H. V. Lyaruu, and L. U. Change (2003). *Linkages between changes in land use, biodiversity and land degradation on the slopes of Mount Kilimanjaro, Tanzania*. LUCID Project, International Livestock Research Institute.

Mishra, U., R. Lal, D. Liu, and M. van Meirvenne (2010). “Predicting the Spatial Variation of the Soil Organic Carbon Pool at a Regional Scale”. In: *Soil Science Society of America Journal* 74.3, pp. 906–914. DOI: 10.2136/sssaj2009.0158.

Mishra, U., M. S. Torn, E. Masanet, and S. M. Ogle (2012). “Improving regional soil carbon inventories: Combining the IPCC carbon inventory method with regression kriging”. In: *Geoderma* 189–190.0, pp. 288–295. DOI: 10.1016/j.geoderma.2012.06.022.

Morgan, C. L., T. H. Waiser, D. J. Brown, and C. T. Hallmark (2009). “Simulated in situ characterization of soil organic and inorganic carbon with visible near-infrared diffuse reflectance spectroscopy”. In: *Geoderma* 151.3-4, pp. 249–256. DOI: 10.1016/j.geoderma.2009.04.010.

Nocita, M., L. Kooistra, M. Bachmann, A. Müller, M. Powell, and S. Weel (Nov. 2011). “Predictions of soil surface and topsoil organic carbon content through the use of laboratory and field spectroscopy in the Albany Thicket Biome of Eastern Cape Province of South Africa”. In: *Geoderma* 167-168.0, pp. 295–302. DOI: 10.1016/j.geoderma.2011.09.018.

Nocita, M., A. Stevens, C. Noon, and B. van Wesemael (2013). “Prediction of soil organic carbon for different levels of soil moisture using Vis-NIR spectroscopy”. In: *Geoderma* 199.0, pp. 37–42. DOI: 10.1016/j.geoderma.2012.07.020.

Nonnotte, P., H. Guillou, B. L. Gall, M. Benoit, J. Cotten, and S. Scaillet (2008). “New K – Ar age determinations of Kilimanjaro volcano in the North Tanzanian diverging rift, East Africa”. In: *Journal of Volcanology and Geothermal Research* 173.1–2, pp. 99–112. DOI: 10.1016/j.jvolgeores.2007.12.042.

Nunan, N., K. Wu, I. M. Young, J. W. Crawford, and K. Ritz (May 2003). "Spatial distribution of bacterial communities and their relationships with the micro-architecture of soil". In: *FEMS Microbiology Ecology* 44.2, pp. 203–215. DOI: 10.1016/S0168-6496(03)00027-8.

Ridolfi, L., P. D'Odorico, A. Porporato, and I. Rodriguez-Iturbe (Mar. 2003). "Stochastic soil moisture dynamics along a hillslope". In: *Journal of Hydrology* 272.1-4, pp. 264–275. DOI: 10.1016/S0022-1694(02)00270-6.

Rockström, J., M. Falkenmark, L. Karlberg, H. Hoff, S. Rost, and D. Gerten (2009). "Future water availability for global food production: The potential of green water for increasing resilience to global change". In: *Water Resources Research* 45.7. DOI: 10.1029/2007WR006767.

Rodionov, A., S. Pätzold, G. Welp, R. C. Pallares, L. Damerow, and W. Amelung (Apr. 2014). "Sensing of Soil Organic Carbon Using Visible and Near-Infrared Spectroscopy at Variable Moisture and Surface Roughness". In: *Soil Science Society of America Journal* 0. DOI: 10.2136/sssaj2013.07.0264.

Rohr, P. and A Killingtveit (2003). "Rainfall distribution on the slopes of Mt Kilimanjaro". English. In: *Hydrological Sciences Journal –Journal des sciences hydrologiques* 48.1, 65–77. DOI: {10.1623/hysj.48.1.65.43483}.

Sankey, J. B., D. J. Brown, M. L. Bernard, and R. L. Lawrence (2008). "Comparing local vs. global visible and near-infrared (VisNIR) diffuse reflectance spectroscopy (DRS) calibrations for the prediction of soil clay, organic C and inorganic C". In: *Geoderma* 148.2, pp. 149–158. DOI: 10.1016/j.geoderma.2008.09.019.

Shepherd, K. D. and M. G. Walsh (2007). "Infrared spectroscopy - enabling an evidence-based diagnostic surveillance appro to agricultural and environmental management in developing countries". In: *Journal of Near Infrared Spectroscopy* 15, pp. 1–19. DOI: 10.1255/jnirs.716.

Shepherd, K. D. and M. G. Walsh (2002). "Development of Reflectance Spectral Libraries for Characterization of Soil Properties". In: *Soil Science Society of America Journal* 66.3, pp. 988–998. DOI: 10.2136/sssaj2002.9880.

Stenberg, B. (2010). "Effects of soil sample pretreatments and standardised rewetting as interacted with sand classes on Vis-NIR predictions of clay and soil organic carbon". In: *Geoderma* 158.1-2. Diffuse reflectance spectroscopy in soil science and land resource assessment, pp. 15–22. DOI: DOI:10.1016/j.geoderma.2010.04.008.

Stenberg, B. and R. V. Rossel (2010). "Diffuse reflectance spectroscopy for high-resolution soil sensing". In: *Proximal Soil Sensing*. Springer, pp. 29–47.

Stevens, A., M. Nocita, G. Tóth, L. Montanarella, and B. van Wesemael (2013). "Prediction of soil organic carbon at the European scale by visible and near infrared reflectance spectroscopy". In: *PLoS ONE* 8.6. DOI: 10.1371/journal.pone.0066409.

Stoyan, H., H. De-Polli, S. Böhm, G. Robertson, and E. Paul (2000). "Spatial heterogeneity of soil respiration and related properties at the plant scale". In: 222.1-2, pp. 203–214–. DOI: 10.1023/A%3A1004757405147.

Stuart, B. H. (2004). *Infrared spectroscopy: fundamentals and applications*. Wiley. com.

Torgo, L., R. P. Ribeiro, B. Pfahringer, and P. Branco (2013). “SMOTE for Regression”. In: *Progress in Artificial Intelligence*. Springer, pp. 378–389.

Viscarra Rossel, R., D. Walvoort, A. McBratney, L. Janik, and J. Skjemstad (2006). “Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties”. In: *Geoderma* 131.1-2, pp. 59–75. DOI: 10.1016/j.geoderma.2005.03.007.

Viscarra Rossel, R., S. Cattle, A. Ortega, and Y. Fouad (2009). “In situ measurements of soil colour, mineral composition and clay content by vis-NIR spectroscopy”. In: *Geoderma* 150.3-4, pp. 253–266. DOI: 10.1016/j.geoderma.2009.01.025.

Viscarra Rossel, R. and T. Behrens (Aug. 2010). “Using data mining to model and interpret soil diffuse reflectance spectra”. In: *Geoderma* 158.1-2, pp. 46–54. DOI: 10.1016/j.geoderma.2009.12.025.

Vågen, T.-G., K. D. Shepherd, and M. G. Walsh (2006). “Sensing landscape level change in soil fertility following deforestation and conversion in the highlands of Madagascar using Vis-NIR spectroscopy”. In: *Geoderma* 133.3-4, pp. 281–294. DOI: 10.1016/j.geoderma.2005.07.014.

Wardle, D. A., R. D. Bardgett, J. N. Klironomos, H. Setälä, W. H. van der Putten, and D. H. Wall (2004). “Ecological Linkages Between Aboveground and Belowground Biota”. In: *Science* 304.5677, pp. 1629–1633. DOI: 10.1126/science.1094875. eprint: http://www.sciencemag.org/content/304/5677/1629.full.pdf.

Wetterlind, J. and B. Stenberg (2010). “Near-infrared spectroscopy for within-field soil characterization: small local calibrations compared with national libraries spiked with local samples”. In: *European Journal of Soil Science* 61.6, pp. 823–843. DOI: 10.1111/j.1365-2389.2010.01283.x.

Wetzel, D. L. (1983). “Near-infrared reflectance analysis”. In: *Analytical Chemistry* 55.12, 1165A–1176A. DOI: 10.1021/ac00262a001.

Young, I. M. and J. W. Crawford (2004). “Interactions and Self-Organization in the Soil-Microbe Complex”. In: *Science* 304.5677, pp. 1634–1637. DOI: 10.1126/science.1097394. eprint: http://www.sciencemag.org/content/304/5677/1634.full.pdf.

Zech, W., N. Senesi, G. Guggenberger, K. Kaiser, J. Lehmann, T. M. Miano, A. Miltner, and G. Schroth (1997). “Factors controlling humification and mineralization of soil organic matter in the tropics”. In: *Geoderma* 79.1-4, pp. 117–161. DOI: 10.1016/S0016-7061(97)00040-2.

Šimůnek, J., N. J. Jarvis, M. van Genuchten, and A. Gärdenäs (Mar. 2003). “Review and comparison of models for describing non-equilibrium and preferential flow and transport in the vadose zone”. In: *Journal of Hydrology* 272.1-4, pp. 14–35. DOI: 10.1016/S0022-1694(02)00252-4.

# Visualizing small scale variability of soil chemical properties on Mt. Kilimanjaro by VIS-NIR spectroscopy

ANNA KÜHNEL[1], CHRISTINA BOGNER[2], HOLGER PABST[3] AND BERND HUWE[1]

[1]Soil Physics Group, BayCEER, University of Bayreuth, Germany
[2]Ecological Modelling, BayCEER, University of Bayreuth, Germany
[3]Department of Soil Science of Temperate Ecosystems, University of Göttingen, Germany

published in *Bornimer Agrartechnische Berichte*, Oktober 2013, Heft 82, pp. 123-128, ISBN: 0947-7314 available at: http://opus.kobv.de/slbp/volltexte/2013/5069

corresponding author: Anna Kühnel (anna.kuehnel@uni-bayreuth.de)

**Abstract**

We investigated the feasibility of VIS-NIR reflectance spectra to predict clay content for different land-use systems in situ. We used partial least square regression on an independent validation dataset and root mean squared error and the Akaike information criterion to evaluate our model. The model was then used to predict clay content in four soil profiles on a 3 x 3 cm scale. Models performed well for spectra taken in the laboratory (RPD > 2; $R^2$ > 0.76). The accuracy for in situ predictions however varies between the land-use systems and predictions are preliminary.

*Keywords:* agricultural soils; diffuse reflectance spectroscopy; spatial variation, clay

## 1.1 Introduction

The conversion of natural or semi-natural ecosystems to anthropogenic land-use forms often results in changes of ecosystem functions like a decrease in water and carbon storage and erosion control. To infer the implications of these changes, fast and accurate predictions are required. This is especially important for the sub-Saharan ecosystems where information on soil properties is still rather scarce. Visible (VIS) and near-infrared (NIR) spectroscopy is a fast method to predict various soil properties simultaneously at comparatively low costs (Minasny et al., 2011b) and has been widely used under laboratory conditions (Viscarra Rossel et al., 2006a; Chang et al., 2001; Awiti et al., 2008). Using VIS-NIR spectroscopy directly in the field is not yet as reliable (Morgan et al., 2009; Nocita et al., 2011). However, it provides a direct and non-destructive method, if stable models can be developed.

The goal of this study is to visualize the small scale variability of clay content, in situ. We use VIS-NIR reflectance spectra of soil to build a model based on partial least square regression (PLSR) to predict clay content for different land-use forms.

## 1.2 Materials and methods

### 1.2.1 Study site

The study was conducted on the southern slopes of Mt. Kilimanjaro, Tanzania (3°4'33"S, 37°21'12"E). The natural ecosystems of the lowlands around Mt. Kilimanjaro (up to 1100 m a.s.l) is savannah that developed on superficial deposits from

the volcano (GeologicalMap, 1955). The mean annual rainfall fluctuates between 400–900 mm (Soini, 2005), the main soil type of the plains is Vertisol and *Balanitis aegyptiaca* and different Acacias species dominate. The savannah ecosystem is threatened by the transformation into fields, as the increasing population needs arable land, where maize and sunflowers are grown. The submontane zone, an area between 1100–1800 m a.s.l. on the southern slopes of Mt. Kilimanjaro is covered mainly by homegarden ecosystems, with mean annual rainfall between 1200 and 2000 mm (Soini, 2005). The main soil types of the higher elevations are Andosols, more weathered soils develop into Vertisols. The homegarden is a traditional agroforestry system, where banana (*Musa* spp.) and coffee (*Coffea arabica*) trees are grown together with a variety of smaller crops (Fernandes et al., 1985). Big trees, often remaining from the former natural rain forest, provide shade and protection against erosion. In this zone, besides the traditional homegarden, extensive coffee plantations were established, as soil and climate provide optimal conditions for coffee.

In the lowlands and in the submontane zone of Mt. Kilimanjaro we have selected two typical ecosystems each, namely natural savannah and maize field of the lowlands and traditional homegarden and coffee plantation of the submontane zone. Soil under coffee plantation was described as Haplic Vertisol, soil under homegarden and savannah as Sodic Vertisols and soil under maize field as Thephric Cambisol (IUSS Working Group WRB, 2007).

### 1.2.2 Soil sampling and laboratory analysis

In each of the four selected ecosystems a soil pit was dug to a depth of at least 100 cm or until continuous rock was reached. One profile wall was carefully cleaned of roots and debris and a frame of $0.5 \times 1$ m with $3 \times 3$ cm segments was put on the wall of the soil profile. Each segment was then scanned with the contact probe attached to an Agrispec portable spectrometer (ASD, Boulder Colorado) in the spectral range of 350–2500 nm in 1 nm intervals. Small soil core samples (diameter 2.5 cm) were taken for validation.

For the calibration of the models, soil samples were collected from 25 different sites (maize fields, savannah, coffee plantations, homegardens and grasslands) on the southern slopes of Mt. Kilimanjaro. The samples were collected with a soil auger and different soil horizons were separated resulting in 146 samples.

The samples were oven-dried at 45° for 24h, sieved $<2$ mm and an aliquot was taken for further analysis. The sand fraction was determined by wet sieving with Na-hexa-meta-phosphate as dispersion agent, after destroying the organic substances. Silt and clay fractions were then measured using a Master Sizer S particle size analyzer (Malvern Instruments).

### 1.2.3 Spectral measurements

For spectral measurements a well-mixed aliquot of the dried sample was placed in a small cup and the surface was smoothed with a spatula. Then it was scanned with the same device as used in the field. The instrument was calibrated with a Spectralon® white tile prior to measurements. For each sample as well as for the calibration with the white reference 30 reflectance spectra were averaged to reduce the noise. In order to validate the field predictions, the small soil core samples from the profiles were air dried and scanned in the laboratory as described above.

### 1.2.4 Model calibration

Each spectrum was corrected for the ASD offset between the three detectors (VNIR, SWIR1 and SWIR2) with the additive method (Becvar et al., (2006 - 2008)). Then, a wavelet transformation was performed using the Daubechies least asymmetric wavelet number 10 and the spectra were transformed into absorbance (log(1/reflectance)) values. Afterwards, noisy portions of the spectra were removed and only the range from 500 nm to 2400 nm was kept. For the calibration of PLSR models the dataset was split into a calibration and a validation datasets by randomly choosing 3/4 of the dataset for calibration. The number of components for the optimal PLSR models was chosen based on the leave-one-out cross validation. The root mean squared error of prediction (RMSE)

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - y_i)^2} \tag{1.1}$$

the Akaike Information Criterion (AIC)

$$\mathrm{AIC} = N\log(\mathrm{RMSE}) + 2m \tag{1.2}$$

the coefficient of determination ($R^2$)

$$\mathrm{R}^2 = 1 - \frac{\sum_{i=1}^{N}(x_i - y_i)^2}{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \tag{1.3}$$

and the residual prediction deviation (RPD)

$$\mathrm{RPD} = \frac{\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}}{RMSE} \tag{1.4}$$

Table 1.1: Parameter of the PLSR models

| Clay (%) | Mean | Range | $R^2$ | RMSE | RPD |
|---|---|---|---|---|---|
| Calibration | 51.2 | 20.0 - 85.0 | 0.84 | 8.0 | 2.04 |
| Validation | 50.8 | 21.0 - 76.0 | 0.80 | 7.1 | 2.26 |

were calculated, where $N$ is the number of samples, m is the number of model components, $x_i$ is the predicted value, $y_i$ is the observed value and $\hat{y}_i$ is the mean of the observed values. The model with the lowest AIC was chosen, as it helps to select a model that represents the variability in the data without causing it to overfit (Viscarra Rossel, 2008). Furthermore, for model validation with field data, the squared regression coefficient $R^2$ between the observed and the modelled values was determined. All analyses were performed in R (R Development Core Team, 2011).

## 1.3 Results and Discussion

### 1.3.1 Model calibration and validation

We have chosen a model with 10 components. The mean values and the range of clay content in calibration and validation datasets were similar, with the validation dataset covering the whole range of measurements (Table 1.1). Other studies predicting the clay content showed slightly better $R^2$ and RMSE values. Stenberg (2010), for example, analysed the effect of different pre-treatments of the samples on clay content and found $R^2 > 0.86$. Considering the classification of RPD values by Viscarra Rossel et al. (2006a), our model performed well and quantitative predictions are very good (RPD > 2.0).

### 1.3.2 Model validation for field samples

Correlations between a) the spectra taken in the laboratory and b) the spectra taken in the field for the same samples and measured values were calculated (Figure 1.1). When looking at all plots, the model for the prediction of clay content with the air dried spectra performed quite well ($R^2 = 0.75$; RPD = 2.02), whereas predicting clay with field spectra resulted only in $R^2$ of 0.27 and a RPD value of 1.18 (Table 1.2). There are, however, large differences between the individual plots. The clay content of the homegarden profile could be predicted quite accurately from air dried as well as from field spectra. In contrast, the clay content in the maize

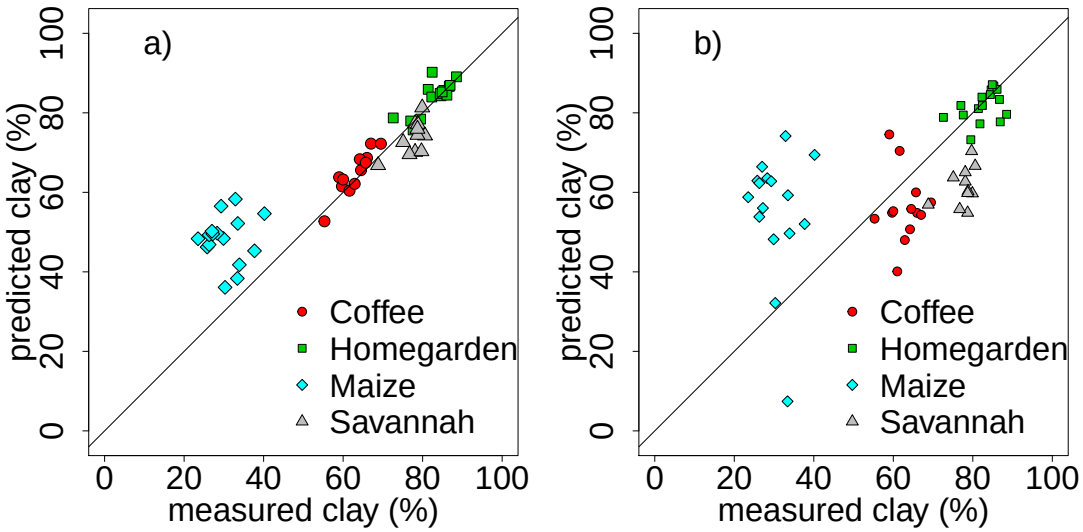



Figure 1.1: Predicted versus measured clay content for a) spectra taken under laboratory conditions and b) field spectra

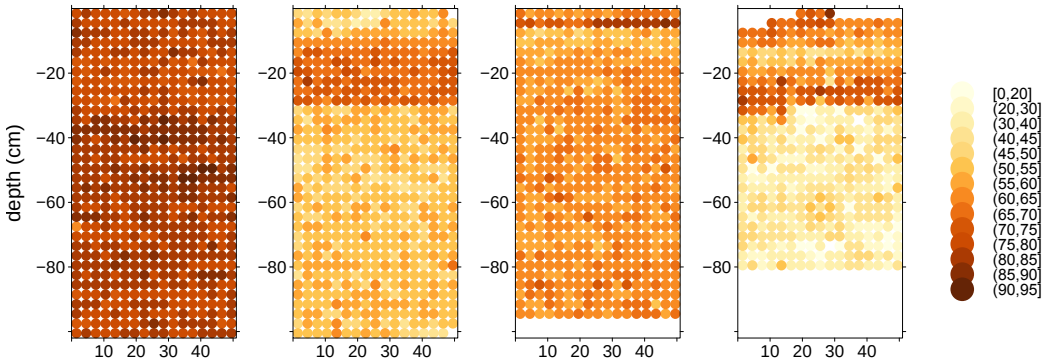
field was poorly predicted with both approaches. Due to the fact, that scanning was conducted during the rainy season, field spectra were probably affected by the moisture content of the soil. As soil moisture has a strong influence on the reflectance spectra (Lobell et al., 2002), spectra taken under field conditions are often not reliable. The poor prediction for the air-dried samples for the maize profile could be due to a high amount of volcanic material in the soil. In our study we had only few samples of this material, so that it is probably under-represented in our validation dataset. Predictions for clay content of the coffee profile with air-dried spectra were very good in contrast to the prediction with field spectra. Apart from soil moisture, other factors are probably influencing the predictions in the field, like size and shape of the particles and the voids between them (Chang et al., 2001) or smearing of the clay during surface preparation.

Table 1.2: Validation parameters for laboratory and field predictions of clay content

| | Air-dried spectra | | | Field spectra | | |
|---|---|---|---|---|---|---|
| Plot | RMSE | RPD | $R^2$ | RMSE | RPD | $R^2$ |
| Homegarden | 3.0 | 1.48 | 0.51 | 4.4 | 0.96 | 0.15 |
| Coffee plantation | 3.0 | 1.33 | 0.38 | 11.6 | 0.33 | -8.8 |
| Savannah | 5.1 | 0.65 | -1.64 | 16.7 | 0.20 | -27.6 |
| Maize field | 19.3 | 0.24 | -17.8 | 29.7 | 0.15 | -43.6 |
| All Plots | 11.0 | 2.02 | 0.75 | 18.5 | 1.18 | 0.27 |

### 1.3.3 Small scale variability in the field

For each $3 \times 3$ cm segment of the profiles clay content was predicted from field spectra with the respective model (Figure 1.2). Differences between the ecosystems are clearly visible. Soil in the homegarden ecosystem showed high clay content throughout the profile. In the maize profile the starting of the Cv-horizon at about $-30$ cm was clearly visible. The accuracy of these predictions however is not yet satisfying and statements regarding differences between the ecosystems are preliminary.




(a) Homegarden (b) Coffee plantation (c) Savannah (d) Maize field

Figure 1.2: Small scale variability of clay content (%) in the different ecosystems.

## 1.4 Conclusions

VIS-NIR spectroscopy is a fast and promising tool, but not yet applicable for detecting small scale differences in the field. Moisture content in the field and the different structure of the soil in situ compared to sieved samples needs to be taken into account. Possible solutions are preprocessing the calibration dataset with external parameter orthogonalisation, as suggested by Minasny et al. (2011b) or to consider the difference between field and laboratory spectra. Whenever predictions are more accurate, VIS-NIR spectroscopy can be used to asses the spatial organisation of soils rapidly and helps to understand the functioning of the soil within the ecosystem.

## Acknowledgements

This study was funded by the German Research Foundation (DFG) within the Research-Unit 1246 (KiLi) and supported by the Tanzanian Commission for Science

and Technology (COSTECH), the Tanzania Wildlife Research Institute (TAWIRI) and the Mount Kilimanjaro National Park (KINAPA). Additionally we want to thank Johannes Hepp for assistance in the field and laboratory.

## References

Awiti, A. O., M. G. Walsh, K. D. Shepherd, and J. Kinyamario (Jan. 2008). "Soil condition classification using infrared spectroscopy: A proposition for assessment of soil condition along a tropical forest-cropland chronosequence". In: *Geoderma* 143.1-2, pp. 73–84. DOI: 10.1016/j.geoderma.2007.08.021.

Becvar M.and Hirner, A. and U. Heiden ((2006 - 2008)). DLR Spectral Archive. URL: http://cocoon.caf.dlr.de/intro_en.html.

Chang, C.-W., D. A. Laird, M. J. Mausbach, and C. R. Hurburgh (2001). "Near-Infrared Reflectance Spectroscopy-Principal Components Regression Analyses of Soil Properties". In: *Soil Science Society of America Journal* 65.2, pp. 480–490. DOI: 10.2136/sssaj2001.652480x.

Fernandes, E. C. M., A. Oktingati, and J. Maghembe (1985). "The Chagga homegardens: a multistoried agroforestry cropping system on Mt. Kilimanjaro (Northern Tanzania)". In: *Agroforestry Systems* 2 (2), pp. 73–86. DOI: 10.1007/BF00131267.

IUSS Working Group WRB (2007). *World reference base for soil resources 2006, first update 2007. World Soil Resources Reports No. 103*. http://www.fao.org/fileadmin/templates/nr/images/resources/pdf_documents/wrb2007_red.pdf. [Accessed on 2014-04-08].

Lobell, D. B. and G. P. Asner (2002). "Moisture Effects on Soil Reflectance". In: *Soil Science Society of America Journal* 66.3, pp. 722–727. DOI: 10.2136/sssaj2002.7220.

Minasny, B., A. B. McBratney, V. Bellon-Maurel, J.-M. Roger, A. Gobrecht, L. Ferrand, and S. Joalland (2011b). "Removing the effect of soil moisture from NIR diffuse reflectance spectra for the prediction of soil organic carbon". In: *Geoderma* 167-168.0, pp. 118–124. DOI: 10.1016/j.geoderma.2011.09.008.

Morgan, C. L., T. H. Waiser, D. J. Brown, and C. T. Hallmark (2009). "Simulated in situ characterization of soil organic and inorganic carbon with visible near-infrared diffuse reflectance spectroscopy". In: *Geoderma* 151.3-4, pp. 249–256. DOI: 10.1016/j.geoderma.2009.04.010.

Nocita, M., L. Kooistra, M. Bachmann, A. Müller, M. Powell, and S. Weel (Nov. 2011). "Predictions of soil surface and topsoil organic carbon content through the use of laboratory and field spectroscopy in the Albany Thicket Biome of Eastern Cape Province of South Africa". In: *Geoderma* 167-168.0, pp. 295–302. DOI: 10.1016/j.geoderma.2011.09.018.

R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*. ISBN 3-900051-07-0. R Foundation for Statistical Computing. Vienna, Austria.

Soini, E. (2005). “Changing livelihoods on the slopes of Mt. Kilimanjaro, Tanzania: Challenges and opportunities in the Chagga homegarden system”. In: *Agroforestry Systems* 64.2, pp. 157–167. DOI: 10.1007/s10457-004-1023-y.

Stenberg, B. (2010). “Effects of soil sample pretreatments and standardised rewetting as interacted with sand classes on Vis-NIR predictions of clay and soil organic carbon”. In: *Geoderma* 158.1-2. Diffuse reflectance spectroscopy in soil science and land resource assessment, pp. 15–22. DOI: DOI:10.1016/j.geoderma.2010.04.008.

Viscarra Rossel, R., R. McGlynn, and A. McBratney (Dec. 2006a). “Determining the composition of mineral-organic mixes using UV–vis–NIR diffuse reflectance spectroscopy”. In: *Geoderma* 137.1-2, pp. 70–82. DOI: 10.1016/j.geoderma.2006.07.004.

Viscarra Rossel, R. A. (Jan. 2008). “ParLeS: Software for chemometric analysis of spectroscopic data”. In: *Chemometrics and Intelligent Laboratory Systems* 90.1, pp. 72–83. DOI: 10.1016/j.chemolab.2007.06.006.

Wold, S., M. Sjöström, and L. Eriksson (2001). “PLS-regression: a basic tool of chemometrics”. In: *Chemometrics and intelligent laboratory systems* 58.2, pp. 109–130. DOI: 10.1016/S0169-7439(01)00155-1.

# Predicting with limited data – Increasing the accuracy in VIS-NIR diffuse reflectance spectroscopy by SMOTE

CHRISTINA BOGNER[1], ANNA KÜHNEL[2], AND BERND HUWE[2]

[1]Ecological Modelling, BayCEER, University of Bayreuth, Germany
[2]Soil Physics Group, BayCEER, University of Bayreuth, Germany

published in

*Proceedings of the 6th Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing*, Whispers 2014, Lausanne, Switzerland

corresponding author: Christina Bogner (christina.bogner@uni-bayreuth.de)

**Abstract**

Diffuse reflectance spectroscopy is a powerful technique to predict soil properties. It can be used *in situ* to provide data inexpensively and rapidly compared to the standard laboratory measurements. Because most spectral data bases contain air-dried samples scanned in the laboratory, field spectra acquired *in situ* are either absent or rare in calibration data sets. However, when models are calibrated on air-dried spectra, prediction using field spectra are often inaccurate. We propose a framework to calibrate partial least squares models when field spectra are rare using synthetic minority oversampling technique (SMOTE). We calibrated a model to predict soil organic carbon content using air-dried spectra spiked with synthetic field spectra. The root mean-squared error of prediction decreased from 6.18 to 2.12 mg g$^{-1}$ and $R^2$ increased from $-0.53$ to 0.82 compared to the model calibrated on air-dried spectra only.

*Keywords:* diffuse reflectance spectroscopy, soil, partial least squares, calibration, SMOTE

## 2.1 Introduction

Diffuse reflectance spectroscopy in the visible and near-infrared range (VIS-NIR DRS) has proved to be useful to assess various soil properties (Stenberg et al., 2010). It can be employed to provide more data rapidly and inexpensively compared to classical laboratory analysis. Therefore, DRS is increasingly used for vast soil surveys in agriculture and environmental research (Shepherd et al., 2007; Vågen et al., 2006). Recently, several studies have shown the applicability of VIS-NIR DRS *in situ* as a proximal soil sensing technique (Viscarra Rossel et al., 2009; Waiser et al., 2007).

To predict soil properties from soil spectra, a model is calibrated, often using partial least squares (PLS) regression. However, when calibration is based on air-dried spectra collected under laboratory conditions, predictions of soil properties from field spectra tend to be less accurate (Viscarra Rossel et al., 2009). Usually, this decrease in accuracy is attributed to varying moisture between air-dried calibration samples and field spectra recorded with a variable moisture content. Different remediation techniques have been proposed, ranging from advanced preprocessing of the spectra (Minasny et al., 2011b) to "spiking" the calibration set with field spectra (Viscarra Rossel et al., 2009).

In our study, we adopt a slightly different view on the calibration problem. It does not only apply to the varying moisture conditions between the calibration data set

and the field spectra. Indeed, it is also valid if we want to predict soil properties in a range where calibration samples were rare. Mining with rarity or learning from imbalanced data is an ongoing research topic in Machine Learning (Weiss, 2004). Because there are not enough rare samples compared to frequent ones, the model will be better at predicting frequent cases than rare ones. Two different approaches exist to take care of this data imbalance: we can either adjust the model or "balance" the data. The latter approach has the advantage that we can use the usual modelling framework. Synthetic minority oversampling technique (SMOTE) is one way to balance the data. It was first proposed for classification (Chawla et al., 2002) and recently for regression (Torgo et al., 2013). SMOTE oversamples the rare data by generating synthetic points and thus helps to balance the data.

In this study, we propose a strategy to increase the prediction accuracy of soil properties from field spectra when they are rare in calibration. The goal of this study is to build a calibration model to predict soil organic carbon content (SOC) from field spectra by air-dried samples spiked with synthetic field spectra.

## 2.2 Material and methods

### 2.2.1 Data acquisition

The studied soil was sampled at the southern slopes of Mt. Kilimanjaro, Tanzania ($3^\circ\ 4'\ 33''$ S, $37^\circ\ 21'\ 12''$ E) in coffee plantations. Due to favourable soil and climate in this region, extensive coffee plantations constitute a frequent form of land use. We took 31 samples for calibration at 4 different study sites. For validation, we scanned 12 field spectra at a wall of a soil pit and sampled soil material for chemical analysis at the scanned spots. We call these validation field spectra F.

After collection, the calibration samples were dried in an oven at 45°C and sieved $< 2$ mm. Subsequently, they were scanned with an AgriSpec portable spectrophotometer equipped with a Contact Probe (Analytical Spectral Devices, Boulder, Colorado) in the range 350–2500 nm with 1 nm intervals. The same spectrometer was used in the field. The instrument was calibrated with a Spectralon white tile before scanning the soil samples. For the measurement, a thoroughly mixed aliquot of the sample was placed in a small cup and the surface was smoothed with a spatula. Each sample was scanned 30 times and the signal averaged to reduce the noise. In the following, we call this calibration data set L.

SOC was measured in a CNS-Analyser by high temperature combustion with conductivity detectors.

### 2.2.2 Generating data by synthetic minority oversampling

To generate new data to spike the calibration data set L, we used SMOTE (Chawla et al., 2002) and its extension for regression (Torgo et al., 2013). This algorithm consists of generating new synthetic data using existing data and is summarized below. In our case, we generated new spectra and the related SOC using the field spectra F. The new spectra are created by calculating the difference between a field spectrum and one of its nearest neighbours and adding this difference (weighted by a random number between 0 and 1) to the field spectrum. The SOC of the synthetic spectrum is then a weighted average between the SOC of the field spectrum and the used nearest neighbour.

SMOTE has two parameters, namely $N$, the number of points to generate for each existing point (given in percent) and $k$, the number of nearest neighbours. To study the influence of these parameters we generated six different synthetic data sets S1 through S6, varying $N = 100, 200, 300$ and $k = 3, 5$.

### 2.2.3 Data pretreatment and explorative analysis

We corrected each spectrum (calibration, validation and synthetic) for the offset at 1000 and 1830 nm and kept only parts with a high signal-to-noise ratio (450–2400 nm). Then, we transformed the spectra to absorbance $(\log_{10}(1/\text{reflectance}))$ and smoothed them using the Singular Spectrum Analysis (SSA). SSA is a non-parametric technique to decompose a signal into additive components that can be identified as the signal itself or as noise (Golyandina et al., 2013). Finally, we divided each spectrum by its maximum and calculated the first derivative.

In order to assess similarities between the calibration, validation and synthetic data sets, we calculated the Principal Component Analysis (PCA) of the (uncorrected original) spectra L and F and projected the synthetic data into the space spanned by the principal components.

### 2.2.4 Partial least squares regression

We calibrated seven different PLS models. For model I we used the data set L, the spectra scanned under laboratory conditions. Model II through VII were calibrated on L spiked with synthetic spectra S1 through S6. To find the best model I through VII, we varied the number of PLS components between 1 and 15. Based on the predictions in the leave-one-out cross-validation (LOOCV) we calculated the corrected Akaike Information Criterion (Sugiura, 1978) $\text{AIC}_\text{c} = n\ln(RMSE^2) + 2p + \frac{2p(p+1)}{n-p-1}$, where $n$ is the number of calibration samples, $p$ the number of PLS

**Algorithm:** SMOTE

**Input**: $T$ original samples to be SMOTEd
Amount of SMOTE $N$%
Number of nearest neighbours $k$

**Output**: $(N/100) \times T$ synthetic samples with their target values (i.e. concentrations)

**if** $N < 100$ **then**
  Randomize the $T$ original samples:
  $T = (N/100) \times T$
  $N = 100$
**end**

$orig.s[i]$: original sample $i, i = 1, \ldots, T$
$orig.t[i]$: target value of original sample $i$
$new.s[j]$: synthetic sample $j, j = 1, \ldots, (N/100) \times T$
$new.t[j]$: target values of synthetic sample $j$
$ng \leftarrow N/100$: number of synthetic samples to compute for each original sample

Generate synthetic samples:

**for** $i$ *in* $1$ *to* $T$ **do**
  $nns \leftarrow$ compute $k$ nearest neighbours for $orig.s[i]$
  **for** $\ell$ *in* $1$ *to* $ng$ **do**
    randomly choose $x \in nns$
    $diff = orig.s[i] - x$
    $new.s[(i-1) \times ng + \ell] = orig.s[i] + \mathrm{RANDOM}(0,1) \times diff$
    $d_1 = \mathrm{DIST}(new.s, orig.s[i])$
    $d_2 = \mathrm{DIST}(new.s, x)$ $target = \frac{d_2 \times orig.t(orig.s) + d_1 \times orig.t(x)}{d_1 + d_2}$
  **end**
**end**

**return** $new.t \cup new.s$

components and $RMSE$ the root mean-squared error. The latter is defined as $\mathrm{RMSE} = \sqrt{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}$, where $\hat{y}_i$ are the predicted and $y_i$ the observed values. We selected the model with the smallest $\mathrm{AIC_c}$ as the most plausible.

To assess the model quality, we used the $RMSE$, the mean error $ME = \frac{1}{n} \sum_{i=1}^{n} \hat{y}_i - y_i$ and the coefficient of determination $R^2 = 1 - \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 / \sum_{i=1}^{n} (y_i - \bar{y})^2$, where $\bar{y}$ is the mean.

### 2.2.5 Monte Carlo simulations

SMOTE has two random components because it selects spectra randomly (with replacement) among the nearest neighbours and weights the difference between

spectra by a random number (between 0 and 1). To study the influence of these random steps, we generated 100 different datasets S1 through S6. Each data set was then used to spike the calibration data set L, to build a new PLS model and to predict the data set F.

## 2.3 Results and discussion

The first principal components (PCs) explain 85.4% and 11.2% of variance, respectively. We can clearly identify two distinct groups of samples: the calibration data set L and the field spectra F (Figure 2.1). In other words, the data sets L and F differ. The synthetic points that were projected into the space spanned by the PCs resemble the field spectra as expected.

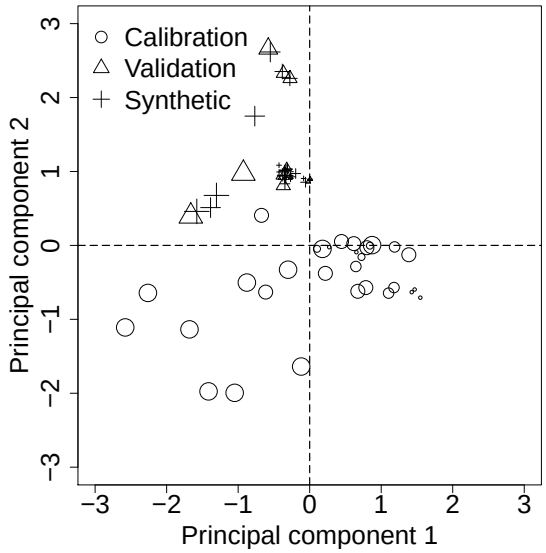



Figure 2.1: Principal component analysis of calibration data set L, validation data set F and one synthetic data set S5. The symbol size was scaled according to the SOC content.

The distinct characteristics of the data sets L and F accord well with the difficulties to predict the data set F by using the laboratory spectra L only (Table 2.1 and Table 2.2). Although the LOOCV of model I yields a moderate $RMSE$ and a large $R^2$, the validation on the data set F fails.

Spiking the calibration data set L with synthetic spectra increases the prediction accuracy of the SOC in the data set F. Actually, the $RMSE$ decreases and $R^2$ increases with increasing number of synthetic points both for the LOOCV and the validation (Table 2.1 and Table 2.2). However, the number of model parameters also increases from 2 to 7.

The Monte Carlo results show only a small variability in the interquartile range. However, some synthetic data sets in model V produced $R^2$ values smaller than

$-0.53$, the value we obtain in model I on air-dried samples only. This might be due to the combination of neighbours during smoting. In general, models with 5 neighbours were more accurate than those with 3 neighbours. However, the number of neighbours had a smaller influence on the prediction accuracy than the number of synthetic points.

It is difficult to decide *a priory* how many synthetic points should be included in the calibration. Indeed, in a classification problem the goal is to approximate an equal distribution of different classes such that the rare class becomes an ordinary one. In regression, however, we do not know which features of the data make them rare. For our data, the range of SOC in the data set L is larger than in the data set F. Therefore, we conclude that concentration is not responsible for the difference between these data sets.

Based on the Monte Carlo results we chose one synthetic data set from model VI, namely the one with the median number of model parameters and the best $R^2$ in the validation. Thus, the calibration data set includes 31 air-dried and 24 synthetic spectra. Compared to model I, spiking the air-dried data set L with these synthetic spectra clearly improves the prediction of the data set F (Figure 2.2).

Table 2.1: Statistics of the PLS calibration. Median values and 25% and 75% quantiles in parenthesis.

| Model | Data set(s) | $N(\%)$ | $k$ | $p$ | $RMSE$ (mg g$^{-1}$) | $R^2$ | $ME$ (mg g$^{-1}$) |
|---|---|---|---|---|---|---|---|
| I | L | – | – | 2 | 6.25 | 0.77 | $-0.20$ |
| II | L and S1 | 100 | 3 | 5 (4; 5) | 5.29 (5.18; 5.47) | 0.80 (0.79; 0.81) | $-0.06$ ($-0.10$; $-0.01$) |
| III | L and S2 | 200 | 3 | 6 (6; 6) | 4.51 (4.47; 4.56) | 0.83 (0.83; 0.84) | 0.07 ( 0.03; 0.11) |
| IV | L and S3 | 300 | 3 | 7 (6; 7) | 4.01 (3.98; 4.06) | 0.85 (0.84; 0.85) | 0.08 ( 0.05; 0.11) |
| V | L and S4 | 100 | 5 | 4 (3; 5) | 5.31 (5.16; 5.55) | 0.80 (0.78; 0.81) | $-0.02$ ($-0.10$; 0.04) |
| VI | L and S5 | 200 | 5 | 6 (6; 6) | 4.51 (4.45; 4.55) | 0.83 (0.83; 0.84) | 0.06 ( 0.01; 0.10) |
| VII | L and S6 | 300 | 5 | 6 (6; 7) | 4.05 (4.02; 4.08) | 0.84 (0.84; 0.85) | 0.07 ( 0.05; 0.09) |

Table 2.2: Statistics of the PLS validation. Median values and 25% and 75% quantiles in parenthesis.

| Model | $RMSE$ (mg g$^{-1}$) | $R^2$ | $ME$ (mg g$^{-1}$) |
|---|---|---|---|
| I | 6.18 | $-0.53$ | $-3.88$ |
| II | 3.09 (2.82; 3.58) | 0.62 (0.49; 0.68) | $-0.03$ ($-0.53$; 0.79) |
| III | 2.00 (1.79; 2.40) | 0.84 (0.77; 0.87) | 0.14 ($-0.01$; 0.36) |
| IV | 1.31 (1.08; 1.58) | 0.93 (0.90; 0.95) | 0.16 ( 0.06; 0.27) |
| V | 3.06 (2.79; 3.56) | 0.62 (0.49; 0.69) | $-0.28$ ($-0.70$; 0.79) |
| VI | 2.12 (1.81; 2.39) | 0.82 (0.77; 0.87) | 0.24 ($-0.04$; 0.48) |
| VII | 1.62 (1.29; 2.07) | 0.89 (0.83; 0.93) | 0.18 ( 0.02; 0.37) |

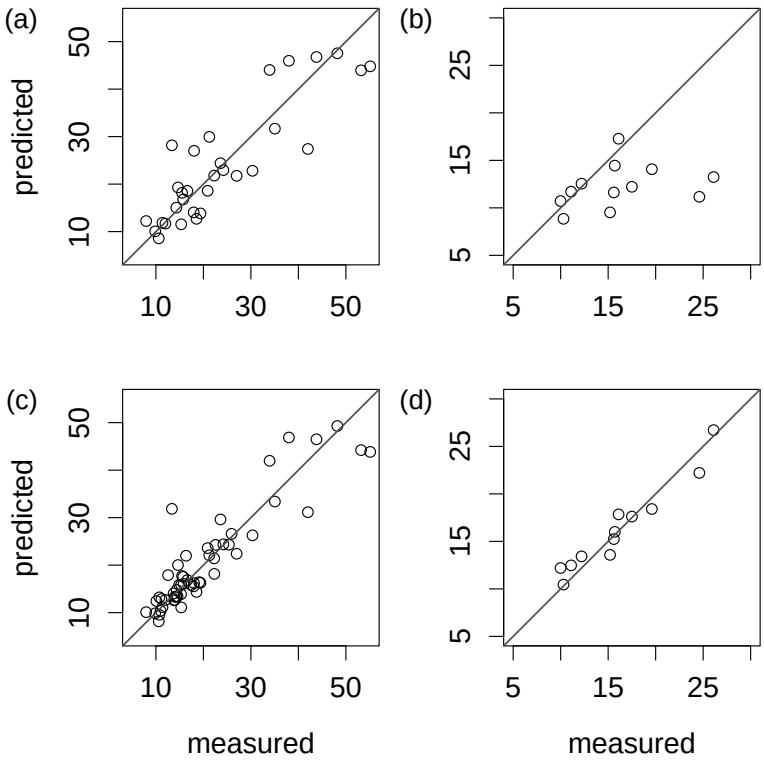



Figure 2.2: Results of (a) leave-one-out cross-validation on data set L (model I), (b) validation on data set F, (c) leave-one-out cross-validation on data set L spiked with a synthetic data set (model VI) and (d) validation on data set F.

## 2.4 Conclusions

We propose a framework to predict soil properties from *in situ* acquired field spectra by spiking air-dried laboratory calibration data by synthetic ones generated from these field spectra. In general, the prediction accuracy increases when a sufficient number of synthetic points is included in the calibration. However, because it is difficult to determine this number *a priori*, we recommend to generate several synthetic data sets to find an appropriate model.

## Acknowledgements

This study is part of the project DFG FOR 1246 "Kilimanjaro ecosystems under global change: Linking biodiversity, biotic interactions and biogeochemical ecosystem processes" and was supported by the Deutsche Forschungsgemeinschaft.

## References

Chawla, N. V., K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer (2002). “SMOTE: Synthetic Minority Over-sampling Technique”. In: *Journal of Artificial Intelligence Research* 16, pp. 321–357.

Golyandina, N. and A. Zhigljavsky (2013). *Singular Spectrum Analysis for time series*. Springer.

Minasny, B., A. B. McBratney, V. Bellon-Maurel, J.-M. Roger, A. Gobrecht, L. Ferrand, and S. Joalland (2011b). “Removing the effect of soil moisture from NIR diffuse reflectance spectra for the prediction of soil organic carbon”. In: *Geoderma* 167-168.0, pp. 118–124. DOI: 10.1016/j.geoderma.2011.09.008.

Shepherd, K. D. and M. G. Walsh (2007). “Infrared spectroscopy - enabling an evidence-based diagnostic surveillance appro to agricultural and environmental management in developing countries”. In: *Journal of Near Infrared Spectroscopy* 15, pp. 1–19. DOI: 10.1255/jnirs.716.

Stenberg, B. and R. V. Rossel (2010). “Diffuse reflectance spectroscopy for high-resolution soil sensing”. In: *Proximal Soil Sensing*. Springer, pp. 29–47.

Sugiura, N. (1978). “Further analysts of the data by akaike’ s information criterion and the finite corrections”. In: *Communications in Statistics - Theory and Methods* 7.1, pp. 13–26. DOI: 10.1080/03610927808827599.

Torgo, L., R. P. Ribeiro, B. Pfahringer, and P. Branco (2013). “SMOTE for Regression”. In: *Progress in Artificial Intelligence*. Springer, pp. 378–389.

Viscarra Rossel, R., S. Cattle, A. Ortega, and Y. Fouad (2009). “In situ measurements of soil colour, mineral composition and clay content by vis-NIR spectroscopy”. In: *Geoderma* 150.3-4, pp. 253–266. DOI: 10.1016/j.geoderma.2009.01.025.

Vågen, T.-G., K. D. Shepherd, and M. G. Walsh (2006). “Sensing landscape level change in soil fertility following deforestation and conversion in the highlands of Madagascar using Vis-NIR spectroscopy”. In: *Geoderma* 133.3-4, pp. 281–294. DOI: 10.1016/j.geoderma.2005.07.014.

Waiser, T. H., C. L. S. Morgan, D. J. Brown, and C. T. Hallmark (Mar. 2007). *In Situ Characterization of Soil Clay Content with Visible Near-Infrared Diffuse Reflectance Spectroscopy*. DOI: 10.2136/sssaj2006.0211.

Weiss, G. M. (June 2004). “Mining with Rarity: A Unifying Framework”. In: *SIGKDD Explorations* 6.1, pp. 7–19. DOI: 10.1145/1007730.1007734.

# In-situ prediction of soil organic carbon by VIS-NIR spectroscopy with limited data

ANNA KÜHNEL[1], CHRISTINA BOGNER[2] AND BERND HUWE[1]

[1]Soil Physics Group, BayCEER, University of Bayreuth, Germany
[2]Ecological Modelling, BayCEER, University of Bayreuth, Germany

submitted to

*European Journal of Soil Science*, (11.09.2014)

corresponding author: Christina Bogner (christina.bogner@uni-bayreuth.de)

**Abstract**

Diffuse reflectance spectroscopy has been widely used to predict soil organic carbon (SOC) in laboratory. Predictions directly from soil spectra measured in-situ under field conditions, however, remain challenging. This study addresses the issue of efficiently incorporating in-situ reflectance spectra into calibration data when only few field measurements are available. We applied the synthetic minority oversampling technique to generate new data using in-situ reflectance spectra from soil profiles. Subsequently, we combined existing spectral libraries with these new synthetic data and compared regional and local partial least squares models. The root mean squared error ($RMSE$) of the regional model varied between 16.45 and 40.83 mg $g^{-1}$ SOC. In contrast, local models in combination with synthetic data outperformed the regional model and yielded an $RMSE$ varying between 0.72 and 3.19 mg $g^{-1}$ SOC. We used the models to predict the distribution of SOC in soil profiles in five different land use zones at Mt. Kilimanjaro (Tanzania). Based on our results, we propose a framework for prediction of SOC with a limited number of in-situ spectra.

*Keywords:* diffuse reflectance spectroscopy, soil organic carbon, in-situ proximal soil sensing, synthetic minority oversampling technique

## 3.1 Introduction

Visible and near-infrared diffuse reflectance spectroscopy (Vis-NIR DRS) is a rapid, inexpensive and easy to use tool. Multiple studies confirm the ability of this technique to predict soil organic carbon (SOC) content for laboratory conditions (i.e. for dry, sieved and possibly ground soil samples) and various other soil parameters (Stevens et al., 2013; Vågen et al., 2006; Shepherd et al., 2002; Viscarra Rossel et al., 2006b). Vis-NIR DRS uses the principle that molecules can only absorb radiation with a certain amount of enery (i.e. at certain wavelengths). Although the fundamental absorption features of SOC occur in the mid- and far-infrared regions, the overtones and combinations of the fundamental absorptions can be seen in the visible and near-infrared ranges (Hunt, 1977; Vågen et al., 2006).

Depending on the SOC content the amount of absorbed radiation changes. SOC is a mixture of many different substances like acids, lipids, carbohydrates, proteins, lignin and cellulose (e.g. Kögel-Knabner, 2002; Baldock, 2007). The goal of Vis-NIR DRS is to get information about SOC by scanning a soil sample which itself consists of a mixture of mineral and organic materials, air and water. Therefore, no simple absorption patterns exist, and a variety of modelling techniques to predict SOC

have been developed. Partial least squares regression (PLSR) proved to be one of the best methods (Wold et al., 2001; Vasques et al., 2008). A comparison of different modelling approaches and pre-processing transformations of soil spectra can be found in Vasques et al. (2008).

Nowadays, a certain number of spectral libraries (or databases) containing spectra taken under standard conditions (air-dried and sieved to $< 2$ mm) and covering large geographic areas is available (Shepherd et al., 2002; Brown, 2007; Viscarra Rossel et al., 2012). One commonly used technique is to predict new data using an existing library. If the library, however, lacks samples with the same or similar spectral characteristics as the new data, a calibration model based on the library samples might fail. Site specific calibration models based on spectral information from only a few, geographically close locations, have been shown to lead to better predictions compared to a global library (Wetterlind et al., 2010a; Sankey et al., 2008).

Creating a new comprehensive local database, however, reduces the effectiveness of Vis-NIR DRS as an inexpensive method. Therefore, including a few new samples from the target area in an existing library has been introduced as an alternative method to improve predictions (Brown, 2007; Sankey et al., 2008; Viscarra Rossel et al., 2009; Wetterlind et al., 2010a; Guerrero et al., 2010).

For practical use and as a further reduction of sampling time and costs, the prediction of SOC from in-situ spectra collected directly in the field with portable scanners would be a great advantage (Reeves III, 2010). Yet, several problems with in-situ spectra have been identified. Indeed, they differ from those collected on air-dried and sieved soil samples by size and shape of soil particles, bulk density, pore size distribution, surface properties and soil moisture (Nocita et al., 2013; Morgan et al., 2009; Chang et al., 2001; Vågen et al., 2006; Wetzel, 1983; Gras et al., 2014). Nocita et al. (2013), for example, showed that when models calibrated on air-dried and sieved samples were used to predict SOC of field-moist samples, the model error increased considerably.

Several different methods like removing the water peak of the spectrum (Wu et al., 2009), external parameter orthogonalisation (Minasny et al., 2011b), classification of samples according to their moisture content and using different models for each moisture content (Nocita et al., 2013; Rodionov et al., 2014) have been tested to overcome this problem. However, by removing the water peak, spectral ranges that are important for prediction of soil properties might also be removed, reducing the prediction accuracy. External parameter orthogonalisation as well as classification of soil depending on its moisture content both require intensive sample preparation.

A simpler solution to deal with in-situ soil spectra might be to augment an already existing library with field spectra. Viscarra Rossel et al. (2009), for example, showed

that predictions of clay content from field spectra improved when adding in-situ soil spectra to the calibration database. The creation of a data set with in-situ soil spectra that is extensive enough to influence the calibration model, however, is also time demanding. Therefore, we need a method to generate a data set from a limited number of in-situ spectra.

The synthetic minority oversampling technique (SMOTE) has been proposed to deal with limited (or rare) data in classification and regression (Chawla et al., 2002; Torgo et al., 2013). Applied to Vis-NIR DRS, it generates additional synthetic spectra from collected in-situ spectra that can subsequently be used to extend the calibration data set. These synthetic data help to balance the existing spectral database in favour of the local area for which predictions are made and in favour of particular soil properties like distinct moisture content of in situ-spectra, for example. Hence SMOTE deals simultaneously with the problem of too few spectral information from a new target area as well as with varying conditions in the field. Bogner et al. (2014) demonstrated in a case study the great potential of SMOTE for improving spectral predictions.

In this study we develop a framework for prediction of SOC content directly from in-situ soil spectra. Specifically we use a regional spectral library of soils from Mt. Kilimanjaro and in-situ spectra from soil profiles in five different land use zones. We apply SMOTE to generate data sets from the in-situ soil spectra to be included in the regional library and local subsets from this library. Our objectives are to compare prediction of i) a regional model ii) local models of the different land use zones, and iii) models from local and regional libraries augmented with synthetic spectra.

## 3.2 Materials and methods

### 3.2.1 Study site

The study was conducted in the colline and the submontane zones on the southern slopes of Mt. Kilimanjaro, Tanzania (3°4'33"S, 37°21'12"E). The colline zone comprises an area between 700 m and 1000 m a.s.l. (Misana et al., 2003) and receives a mean annual precipitation of 400–900 mm (Soini, 2005). On the small and steep volcanic craters in the East of Mt. Kilimanjaro, the main soil type is Leptosol, whereas in the plains Acrisols, Ferralsols, Lixisols, Nitisols and Vertisols dominate (Zech et al., 2014).

The natural ecosystem of the colline zone is savannah that developed on superficial deposits from Mt. Kilimanjaro (GeologicalMap, 1955). *Balanitis aegyptiaca* and

different Acacias (*Acacia tortilis*, *Acacia senegal*, *Acacia nilotica*) constitute the main tree species, with various different grass species underneath. As the population in the Kilimanjaro area is continuously growing, arable land is urgently needed (Mbonile, 2003). Therefore, the savannah ecosystem is increasingly transformed into agricultural fields where maize and occasionally sunflowers are grown.

The submontane zone, an area between 1000–1800 m a.s.l. on the southern slopes of Mt. Kilimanjaro averages a mean annual precipitation between 1200 and 2000 mm (Soini, 2005). The main soil types of the higher elevations are Andosols, more weathered soils develop into Vertisols and Umbrisols.

During the centuries, the Chagga tribe at Mt. Kilimanjaro has preserved a traditional agroforestry system on the southern and eastern slopes of the mountain, the so called homegarden that covers an area of about 1200 $km^2$ (Fernandes et al., 1985). A variety of different crops is grown in these multi-storey systems, usually some cash crops like coffee and different types of banana, along with food crops like sweet potato, taro and beans (Fernandes et al., 1985). Big trees often remaining from the former natural rain forest provide shade and protection against soil erosion.

Because soil and climate provide optimal conditions for coffee in this area, extensive plantations were established where *Coffea arabica* is grown by big companies as a cash crop. Shading trees, remnants of the former forest, still exist on some plantations. However, they are often replaced by exotic trees, for example *Grevilla robosta*, as new trees can fit better in line with coffee rows.

In the submontane zone, grasslands developed out of the former natural forest as the growing population needed building material and firewood and cut the forest trees. Nowadays the grasslands often provide fodder for the livestock.

In our study, we focused on five different land use zones, namely the traditional Homegarden, Coffee plantations, Grasslands, Maize fields and Savannah (Figure 3.1).

### 3.2.2 Data collection

In the five above-mentioned land use zones, 191 soil samples at 26 sites were collected with a soil auger. We call these data the *regional* data set in the following. Additionally, one plot (in the savannah two plots) per land use was selected to collect in-situ spectra from a soil profile. We refer to these plots as Hom, Cof, Gra, Mai, Sav1 and Sav2, respectively. The soil was classified as Sodic Vertisol, Haplic Vertisol, Haplic Andosol, Thephric Cambisol, Sodic Vertisol and Rendzic Leptosol in Hom, Cof, Gra, Mai, Sav1 and Sav2, respectively (IUSS Working Group WRB, 2007).

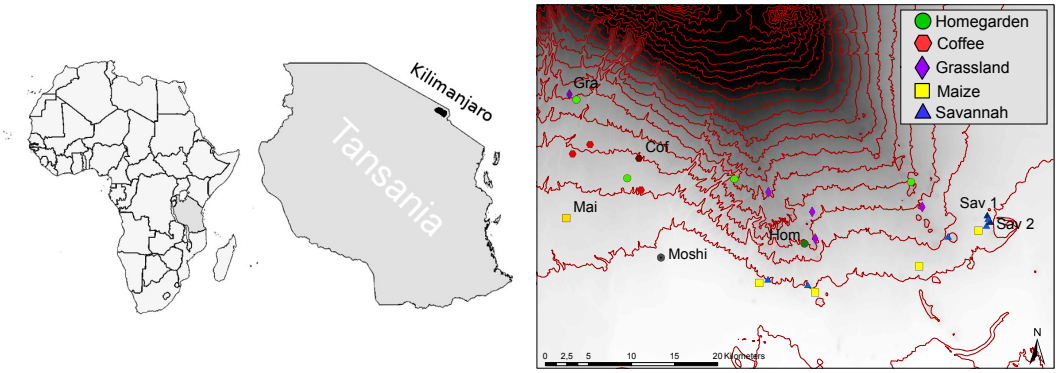



Figure 3.1: Study area and research plots. Points designated by Hom, Cof, Gra, Mai, Sav1 and Sav2 refer to plots where soil profiles were analysed in detail.

A soil pit was dug to a depth of approximately 1 m or until continuous bedrock was reached. Subsequently, a profile wall was carefully cleaned of roots and debris and a frame of 0.5 m $\times$ 1 m with 3 $\times$ 3 cm segments was put on the wall of the soil profile. Each segment was then scanned with a contact probe attached to an Agrispec portable spectrometer (ASD, Boulder Colorado, spectral range 350–2500 nm). The instrument was calibrated with a Spectralon white tile prior to measurements. For each sample as well as for the calibration with the white reference 30 reflectance spectra were averaged to reduce the noise. We refer to these six scanned profiles as the *raster* data sets. In randomly selected segments, between eight and thirteen small soil core samples (diameter 2.5 cm) were taken for SOC analysis (*profile* data set).

The collected soil samples (*regional* and *profile*) were oven-dried at 45 °C for 24 h, sieved $< 2$ mm. Then, an aliquot was used to determine the SOC content using a CNS-Analyser with conductivity detectors by high temperature combustion.

For spectral measurements of all *regional* soil samples a well-mixed aliquot of the dried and sieved sample was placed in a small cup and the surface was smoothed with a spatula. It was then scanned with the same device and same settings as used in the field. To assess the influence of water content on predictions of SOC exemplary, we collected moist spectra on a part of the samples from Homegarden by re-moistening them and taking several scans during the drying process (*moist* data set). The Table C.2 in the online Supplementary Material shows an overview of analysed data sets.

### 3.2.3 Generating synthetic data

The synthetic minority oversampling technique (SMOTE) has been developed by Chawla et al. (2002), in order to increase the number of rare samples in classification tasks. Torgo et al. (2013) demonstrated that this method also worked

for regression. For every point $i$, SMOTE inserts synthetic points along a line that connects this point to one of its $k$ nearest neighbours. To generate the new synthetic point, it calculates the difference between the point $i$ and the chosen nearest neighbour, weights this difference by a random number between 0 and 1 and adds the weighted difference to the point $i$. Several $k$ nearest neighbours can be chosen randomly and several points can be generated along one connecting line, depending on the oversampling rate $N$ (given in percent of input data). For $N = 100$, for example, one new point is generated for every input point. Figure 3.2 illustrates this principle in two dimensions. Our data can be seen as points in a 2151-dimensional real space because we use spectra of 2151 wavelengths for smoting.

To generate synthetic spectra, we only used the *profile* data sets (i.e. soil spectra acquired in-situ). As the soil profiles come from different land use zones and were scanned on different days under different environmental conditions, every profile was smoted separately. We chose $N = 300$ and set the number of nearest neighbours to five, according to the study by Bogner et al. (2014). In other words, for every point in the *profile* data set three new spectra and their SOC contents were generated using randomly one of the five nearest neighbours. The SOC content of the synthetic point was calculated as the weighted average of the SOC content of point $i$ and the used nearest neighbour (c.f. SMOTE algorithm in the online Supplementary Material).

Bogner et al. (2014) showed that, in general, prediction accuracy increased considerably when available data sets were combined with synthetic spectra. However, it varied between different synthetic data sets and depends probably on the combination of nearest neighbours that were included. Therefore, we generated 100 different *synthetic* data sets from every *profile* data set (i.e. six collections of 100 data sets) to compare the prediction performances.

### 3.2.4 Pre-processing of soil spectra

The following pre-processing steps were applied to all collected spectra (spectra from the *regional* data set, all in-situ *raster* spectra, *profile* spectra, *moist* spectra and all *synthetic* spectra). Every spectrum was corrected for the detector offsets with the additive method (Dorigo et al., 2006) and was cut at the edges, so that only wavelengths with a high signal-to-noise ratio were kept (450–2400 nm). The remaining spectrum was then smoothed by singular spectrum analysis (SSA) with a window length of five. SSA is a method for time series analysis and can be used for smoothing, as it decomposes a series into trend, periodicities and noise (Broomhead et al., 1986; Golyandina et al., 2013). When a small window length $L$

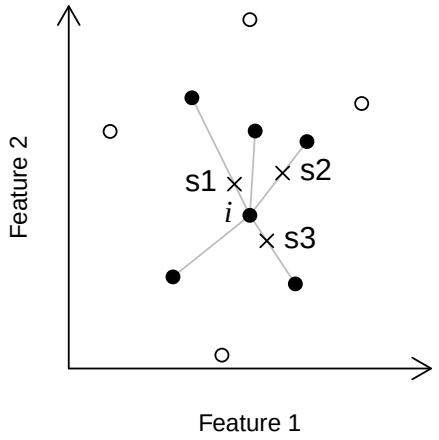



Figure 3.2: Illustration of the synthetic minority oversampling technique (SMOTE) in two dimensions. The $k$ nearest neighbours (black dots) are chosen for an existing point $i$ to generate synthetic points (crosses denoted by $s1$ through $s3$) along the connection lines between $i$ and its nearest neighbours. In this case, $k = 5$ and $N = 300$, i.e. three of the five nearest neighbours are selected randomly to generate three new points. Circles show samples that are not the $k$ nearest neighbours of $i$.

is used, SSA is similar to a weighted moving average procedure. That means, the points included in smoothing are weighted by their distance; the weights thus create a nearly triangular shape around a data point (Golyandina et al., 2013). Compared to moving average SSA uses a different algorithm and can for example also smooth the ending points of a spectrum.

Different standard pre-treatments were tested for reflectance as well as for absorbance ($A = \log(1/\text{reflectance})$) values, namely z-transformation, 1st derivative of z-transformed data, normalization by the maximum value and 1st derivative of data normalized by the maximum. The best combination of pre-processing steps for our data was to use the 1st derivative of absorbance values that were normalized by the maximum ($RMSE$ and $R^2$ as decision criteria). This procedure was consequently used in all analyses.

### 3.2.5 Principal component analysis

We calculated the principal components (PCs) of the raw spectra (cut to 450–2400 nm) of all *raster* samples combined with the *regional* data set and individually for each profile combined with the respective *local* set. By visualising the first two PCs, we want to inspect similarities between the *raster* samples and respectively the *regional* and *local* data sets. Furthermore, we projected the spectra of the *profile*, the *synthetic* and the *moist* data sets (for Homegarden only) into the space spanned by the first two principal components. This shows how spectroscopically close these data are to the *raster* and the *local* sets.

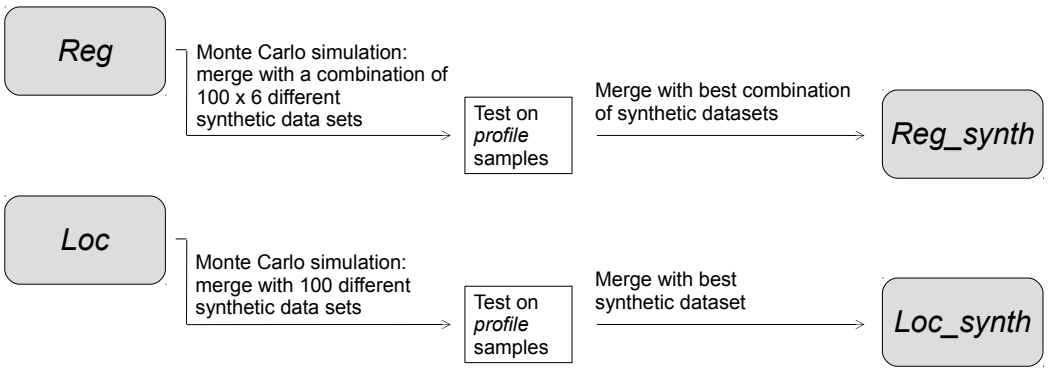



Figure 3.3: Work flow to create different models: *Reg* = regional model, *Reg_synth* = regional model augmented with best combination of synthetic data sets, *Loc* = local models and *Loc_synth* = local models augmented with one synthetic data set from the respective profile.

### 3.2.6 Modelling framework

We compared four different modelling strategies to predict the SOC content in soil profiles at Hom, Cof, Gra, Mai, Sav1 and Sav2 (Figure 3.3). First, we calibrated a PLSR model on the *regional* data set (i.e. the dried soil spectra) containing spectra and SOC values from all land use zones (*Reg* model). Then we divided the *regional* data into five *local* sets based on the different land use zones Homegarden, Coffee plantations, Grasslands, Maize fields and Savannah. For each *local* set, we calibrated an individual PLSR model (i.e. in total five different *Loc* models). Subsequently, we augmented each *local* data set with one synthetic data set from the corresponding profile and recalibrated the model. Each *local* data set was only combined with synthetic data from the same profile. For the Savannah set, this step was done twice (individually for the Sav1 and Sav2 profiles that were smoted separately). We repeated this procedure for the 100 synthetic data sets and obtained 100 different models per profile (Monte Carlo simulation).

All PLSR models were built with leave-one-out cross validation (LOOCV) and the number of model parameters was chosen based on the corrected Akaike Information Criterion ($AIC_c$) (Sugiura, 1978).

$$AIC_c = n\log(RMSE^2) + 2m + \frac{2m(m+1)}{n-m-1} \tag{3.1}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \tag{3.2}$$

where $n$ is the number of samples, $m$ the number of model parameters, $\hat{y}_i$ the predicted and $y_i$ the measured SOC content, respectively.

The error of LOOCV approximates the prediction error for data that are similar to those used to calibrate the models (James et al., 2013). Because our goal is to predict the SOC content from in-situ *raster* spectra, we additionally tested the models on the in-situ *profile* data set from the same profile. This not an independent validation in a strict sense because the profile data were used to generate the synthetic spectra. However, the profile data were not used to calibrate the models and thus constitute new (similar) data. We adopted this test procedure because of a limited number of in-situ samples with corresponding SOC values.

Among the 100 models per ecosystem from the Monte Carlo simulation, we chose the one with a median number of model parameters that produced the maximum $R^2$ of predictions on the *profile* test data set. We denote these six models *Loc_synth*.

In order to chose a good *Reg_synth* model, we run a Monte Carlo analysis on a combination of 100 $\times$ 6 synthetic data sets (one for each profile) and augmented the *regional* data set with the best combination (medium number of model parameters and maximum $R^2$).

To assess the model quality we calculated the mean prediction error ($MPE$) and the coefficient of determination $R^2$

$$MPE = \frac{1}{n}\sum_{i=1}^{n} \bar{y} - y_i \tag{3.3}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{3.4}$$

where $\bar{y}$ is the mean of the measured SOC contents.

Additionally, we created a moist model by adding *moist* soil spectra to the *local* data set (*Loc_moist*) and tested the prediction performance for the Homegarden.

All analyses were performed in `R` (R Development Core Team, 2011) using the packages `ChemometricsWithR` (Wehrens, 2011) and `Rssa` (Korobeynikov, 2010).

## 3.3 Results

### 3.3.1 Summary statistics

We found the highest maximum and the largest range of SOC content in Grassland (Table 3.1). However, in this land use zone the median value was quite low. The highest median was observed in Homegarden, the lowest one in Maize. The lowest

Table 3.1: Summary statistics of the SOC content (mg $g^{-1}$) of the *regional* and the different *local* data sets.

| | Min | $Q_1$ [a] | $Q_2$ [b] | Mean | $Q_3$ [c] | Max |
|---|---|---|---|---|---|---|
| Regional | 1.4 | 11.0 | 18.7 | 28.1 | 35.3 | 148.6 |
| Homegarden | 6.4 | 17.0 | 43.4 | 43.2 | 72.3 | 96.5 |
| Coffee | 8.0 | 15.0 | 19.4 | 24.2 | 32.1 | 55.1 |
| Grassland | 2.3 | 10.9 | 16.8 | 34.6 | 57.0 | 148.6 |
| Maize | 3.7 | 7.7 | 10.0 | 10.1 | 11.1 | 21.9 |
| Savannah | 1.4 | 13.9 | 21.7 | 22.0 | 28.9 | 53.6 |

[a] first quartile [b] median [c] third quartile

SOC content was measured in Savannah. Yet, median and maximum values in this ecosystem were similar to Coffee.

The SOC content of all measured soil samples from the *profile* data set lay inside the range of values from the respective *local* data set, except at the Gra plot (Table 3.2). Actually, in the Gra profile the highest measured value was about 10 mg $g^{-1}$ higher than the maximum of the Grassland *local* set. The ranges of SOC values of Mai and Sav2 profiles were quite low, with 7.2 and 10.8 mg $g^{-1}$, respectively. Hom and Sav1 *profile* data sets expressed a similar range of SOC values of about 26 mg $g^{-1}$. In contrast, the difference between maximum and minimum at the Gra plot was considerably large with about 100 mg $g^{-1}$ SOC.

Table 3.2: Summary statistics of the SOC content (mg $g^{-1}$) of the different *profile* data sets.

| | Min | $Q_1$ [a] | $Q_2$ [b] | Mean | $Q_3$ [c] | Max |
|---|---|---|---|---|---|---|
| Hom | 8.3 | 9.0 | 16.0 | 19.2 | 29.1 | 35.0 |
| Cof | 10.0 | 11.9 | 15.7 | 16.2 | 18.0 | 26.1 |
| Gra | 65.2 | 71.5 | 84.9 | 91.2 | 96.1 | 159.0 |
| Mai | 8.4 | 8.8 | 11.6 | 11.5 | 13.5 | 15.6 |
| Sav1 | 6.9 | 8.0 | 9.2 | 14.7 | 17.9 | 33.3 |
| Sav2 | 19.5 | 23.8 | 25.3 | 25.2 | 26.7 | 30.3 |

[a] first quartile [b] median [c] third quartile

### 3.3.2 Principal component analysis

The two first PCs of the *regional* and *raster* data sets explained 91.2 % and 5.84 % of the total variance in the data, respectively. The in-situ *raster* spectra and the *regional* spectra built two rather distinct clusters with little overlap (Figure 3.4). The separation was particularly pronounced along the second PC. In contrast, the first

PC lacked a simple explanation. We scaled the data points from the *regional* set according to their SOC content. Yet, no obvious pattern based on the varying SOC content appeared. The *synthetic* data that were projected into the space spanned by the first two PCs overlapped well with the in-situ spectra. Some *raster* spectra, however, appeared to be different and lay further away from the main cluster.

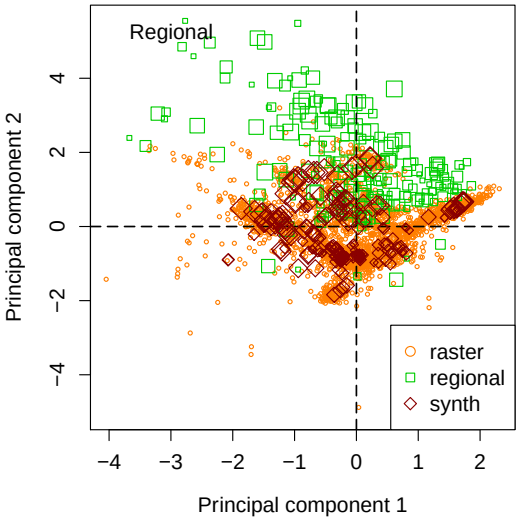



Figure 3.4: Score plot of the first two principal components of the in-situ *raster* spectra (circles) and the *regional* set (squares). Additionally, the projection of the *synthetic* spectra is shown (diamonds). The symbol size for *regional* and *synthetic* points is scaled according to their SOC content.

Score plots for the individual *local* data sets are displayed in Figure 3.5. The first and the second PCs of Homgarden explained 89.0% and 9.4% of the total variance, respectively. The *raster* spectra were clearly separated from the *local* set. Some of the projected *moist* spectra lay near the in-situ spectra. Others, however, created a separate cluster situated even further away than the dry *local* spectra. The *synthetic* spectra and *profile* validation spectra overlapped quite well. Additionally, the principal component analysis (PCA) revealed that some spectral variation of the *raster* spectra was not covered by the *profile* samples.

In the PCA of Coffee the first PC explained 83.8% and the second one 12.6% of the total variance. The *local* spectra were remarkably different from the in-situ *raster* spectra. Two main clusters of *raster* spectra can be observed, which both overlapped quite well with the *synthetic* data set. A few *raster* spectra outside the main clusters lacked corresponding *synthetic* data. Some of these *raster* spectra were more similar to the spectra from the *local* data set.

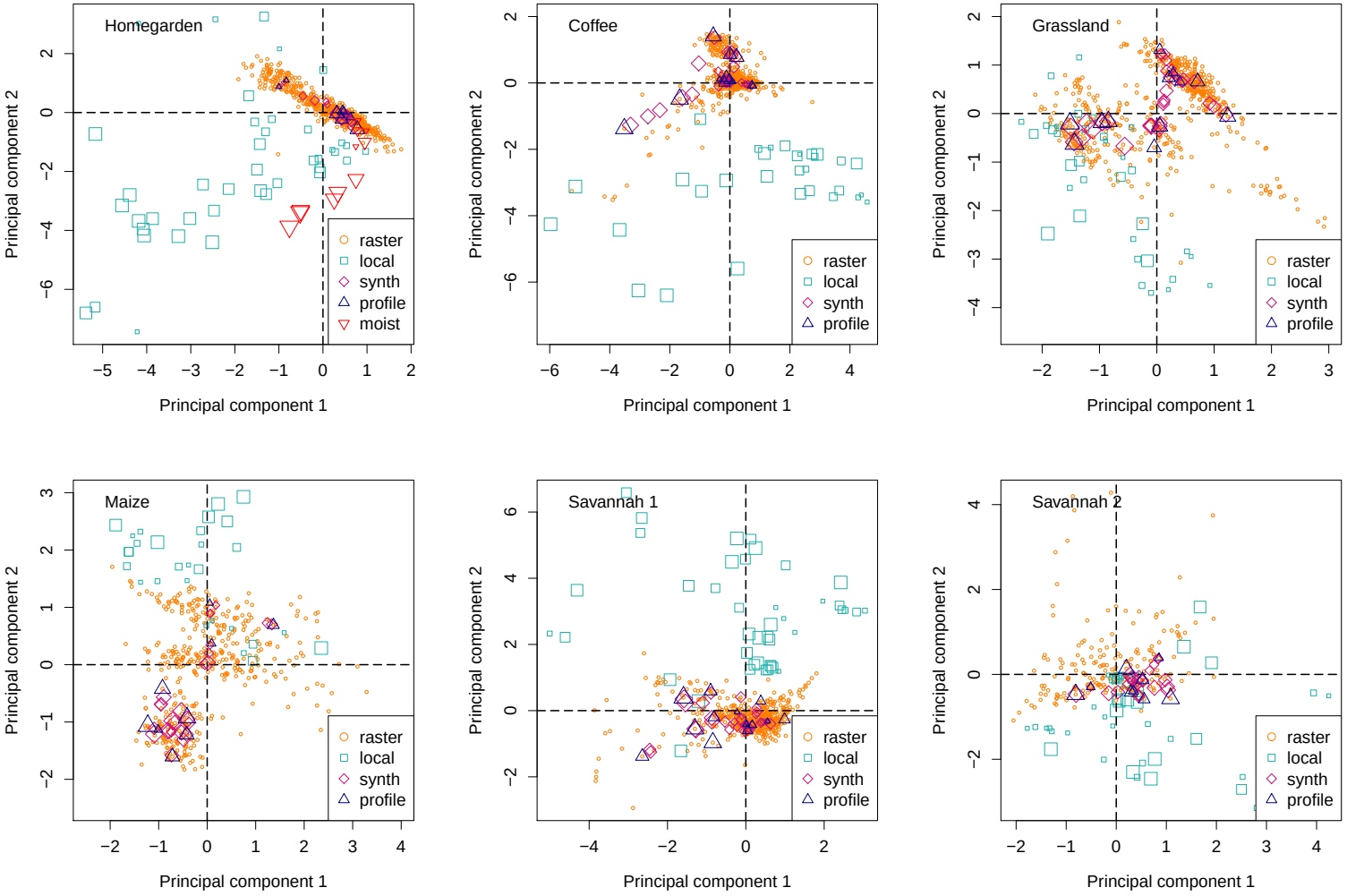



Figure 3.5: Principal component analysis of the spectral data for the six studied soil profiles. The data sets are abbreviated as: *raster* data set (in-situ spectra); profile: *profile* test data set (in-situ spectra with measured SOC content); synth: *synthetic* data set (smoted); loc: *local* data set; moist: *moist* data set (re-moisted spectra from Hom). PCA is based only on the *local* and *raster* datasets, all other points are projected. For detailed description of the data sets see Sections Data collection and Generating synthetic data. The symbol size for *local*, *synthetic* and *profile* points is scaled according to their SOC content.

Score plots for Grassland and Maize revealed two to three clusters in the in-situ *raster* spectra. For Grassland the first two PCs explained 90.6% and 7.5% and for Maize 91.5% and 5.4% of the total variation, respectively. Most *local* spectra were again very different from the *raster* data. Part of the *local* samples, however, were spectroscopically close to one of the *raster* clusters. The *synthetic* spectra lay over the most of the in-situ *raster* spectra, without covering the extreme data points.

For Savannah 1, 88.3% and 8.3% of the total variance was explained by the first two PCs, for Savannah 2, 88.0% and 9.3%, respectively. The *local* samples of the savannah set were substantially different from the *raster* spectra of Savannah 1 and covered only a small part of the *raster* spectra of Savannah 2. The *synthetic* data of Savannah 1 resembled most of the *raster* samples and even some extreme data points. For Savannah 2 only about half of the spectra was overlain by the *synthetic* spectra and many *raster* samples were neither close to the *local* spectra nor to the *synthetic* ones.

In summary *raster* samples of most profiles were spectroscopically far from the *local* samples. The *synthetic* spectra covered well the variation of the *profile* spectra. Yet, some *raster* spectra lacked corresponding *synthetic* spectra, if the former were very different compared to the *profile* spectra.

### 3.3.3 Model calibration and test

Augmenting the *Reg* model with a combination of *synthetic* spectra from all land use zones improved the results of the LOOCV slightly (Table 3.3) and the predictions for the in-situ *profile* test samples substantially (Figure 3.6c and d and Table 3.4). Indeed, the $RMSE$ on the joined *profile* data set decreased with the addition of syntethic spectra by about 70% compared to *Reg*. However, it was still quite large (10.21 mg g$^{-1}$ SOC). Similarly, $R^2$ of the *Reg* model increased from $-0.05$ to 0.90 due to synthetic spectra. Yet, this increase was mainly caused by the Gra profile for it shows a large range of SOC values (Table 3.2). The test results on single profiles were still rather poor (Table 3.4). Actually, because $R^2$ increases with an increasing standard deviation of measured values, $RMSE$ provides a better assessment of model quality.

Compared to the *Reg* and *Reg_synth* models, $RMSE$ from LOOCV for the different *Loc* and *Loc_synth* models decreased, except for the *Loc* model at Gra (Figure 3.7 and Table 3.5). Yet, the test on the *profile* samples with the *Loc* models still led to poor results with large $RMSE$ mostly exceeding 10 mg g$^{-1}$ SOC (Figure 3.7c).

The calibration results of the *Loc_moist* model at Homegarden were similar to those of *Loc* and *Loc_synth* models (Table 3.2). Although the $RMSE$ decreased compared to *Loc*, it still exceeded about four times the error in *Loc_synth*.

Table 3.3: Calibration parameters from leave-one-out cross validation of the regional model (*Reg*) and the regional model augmented with synthetic data (*Reg_synth*).

| model | $m$[a] | $n$[b] | $RMSE$[c] | $R^2$ | $MPE$[d] |
|---|---|---|---|---|---|
| *Reg* | 12 | 191 | 11.32 | 0.80 | 0.21 |
| *Reg_synth* | 15 | 416 | 10.37 | 0.86 | 0.07 |

[a] number of model parameters [b] number of data points [c] root mean squared error [d] mean prediction error

Table 3.4: Parameters of the model tests on the *profile* data sets of the regional model (*Reg*) and the regional model augmented with synthetic data (*Reg_synth*).

| Plot | Model | $n$[a] | $RMSE$[b] | $R^2$ | $MPE$[c] |
|---|---|---|---|---|---|
| Hom | *Reg* | 8 | 16.45 | -1.48 | 4.70 |
| | *Reg_synth* | 8 | 9.63 | 0.15 | -0.74 |
| Cof | *Reg* | 12 | 34.51 | -46.83 | 15.40 |
| | *Reg_synth* | 12 | 7.45 | -1.23 | 2.91 |
| Gra | *Reg* | 12 | 40.83 | -1.21 | -35.41 |
| | *Reg_synth* | 12 | 10.35 | 0.86 | -3.59 |
| Mai | *Reg* | 9 | 36.14 | -207.08 | 23.18 |
| | *Reg_synth* | 9 | 9.70 | -14.00 | 3.05 |
| Sav1 | *Reg* | 13 | 38.03 | -14.28 | 8.90 |
| | *Reg_synth* | 13 | 10.22 | -0.10 | -1.21 |
| Sav2 | *Reg* | 9 | 17.20 | -25.95 | 3.97 |
| | *Reg_synth* | 9 | 12.05 | -12.22 | 1.45 |
| all plots | *Reg* | 63 | 33.26 | -0.05 | 2.50 |
| | *Reg_synth* | 63 | 10.21 | 0.90 | 1.38 |

[a] number of data points [b] root mean squared error [c] mean prediction error

In contrast, $RMSE$ of the different *Loc_synth* models dropped substantially and varied between 0.72 and 3.19 mg g$^{-1}$ SOC. It always decreased compared to the *Loc* model, for most models even by more than 80% (Figure 3.7c and d). Additionally, while the test $R^2$ of the *Loc* models was very low or even negative, it increased in *Loc_synth* models. In fact, at Hom, Cof and Gra profiles $R^2$ exceeded 0.95. In three of the six *Loc_synth* models (Cof, Gra, Mai) the number of model parameters $m$ increased. In contrast, $m$ decreased for Sav1 and remained the same for Sav2 and Hom.

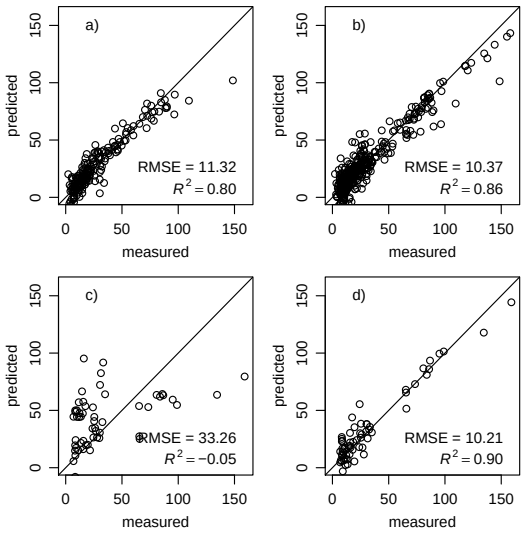



Figure 3.6: Predicted versus measured SOC content (mg $g^{-1}$): a) calibration of the regional model (*Reg*) b) calibration of the regional model augmented with synthetic data (*Reg_synth*) c) test of *Reg* d) test of *Reg_synth*. The calibration is based on a leave-one-out cross validation and the test on the joined *profile* data set from all studied land use zones.

Table 3.5: Statistics of the local model (*Loc*) and the local model augmented with synthetic data (*Loc_synth*). Calibration is based on a leave-one-out cross validation. The models were tested on the *profile* data sets.

| Plot | Model | | Calibration | | | | Test | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | $m$ | $n$ | $RMSE$ | $R^2$ | $MPE$ | $n$ | $RMSE$ | $R^2$ | $MPE$ |
| Hom | *Loc* | 4 | 41 | 9.67 | 0.88 | 0.09 | 8 | 12.67 | -0.47 | 11.74 |
| | *Loc_moist* | 4 | 52 | 8.69 | 0.91 | 0.05 | 8 | 7.94 | 0.42 | 5.83 |
| | *Loc_synth* | 4 | 65 | 7.56 | 0.91 | -0.04 | 8 | 2.15 | 0.96 | 0.59 |
| Cof | *Loc* | 2 | 31 | 6.25 | 0.77 | -0.20 | 12 | 6.18 | -0.53 | -3.88 |
| | *Loc_synth* | 6 | 67 | 3.96 | 0.85 | 0.07 | 12 | 0.72 | 0.98 | 0.19 |
| Gra | *Loc* | 4 | 39 | 14.53 | 0.83 | -0.18 | 12 | 26.30 | 0.08 | -10.39 |
| | *Loc_synth* | 12 | 75 | 8.92 | 0.95 | 0.52 | 12 | 3.19 | 0.99 | 0.72 |
| Mai | *Loc* | 3 | 29 | 3.19 | 0.38 | 0.13 | 9 | 4.27 | -1.91 | -0.13 |
| | *Loc_synth* | 6 | 56 | 2.46 | 0.44 | -0.06 | 9 | 1.45 | 0.66 | -0.25 |
| Sav1 | *Loc* | 8 | 43 | 5.11 | 0.81 | -0.02 | 13 | 19.64 | -3.08 | 7.51 |
| | *Loc_synth* | 5 | 82 | 5.03 | 0.80 | -0.01 | 13 | 2.77 | 0.92 | 0.73 |
| Sav2 | *Loc* | 8 | 43 | 5.11 | 0.81 | -0.02 | 9 | 15.31 | -20.36 | 13.77 |
| | *Loc_synth* | 8 | 70 | 4.64 | 0.76 | -0.19 | 9 | 2.69 | 0.34 | -0.05 |

Hom = Homegarden, Cof = Coffee, Gra = Grassland, Mai = Maize, Sav1 and Sav2 = Savannah1 and Savannah2

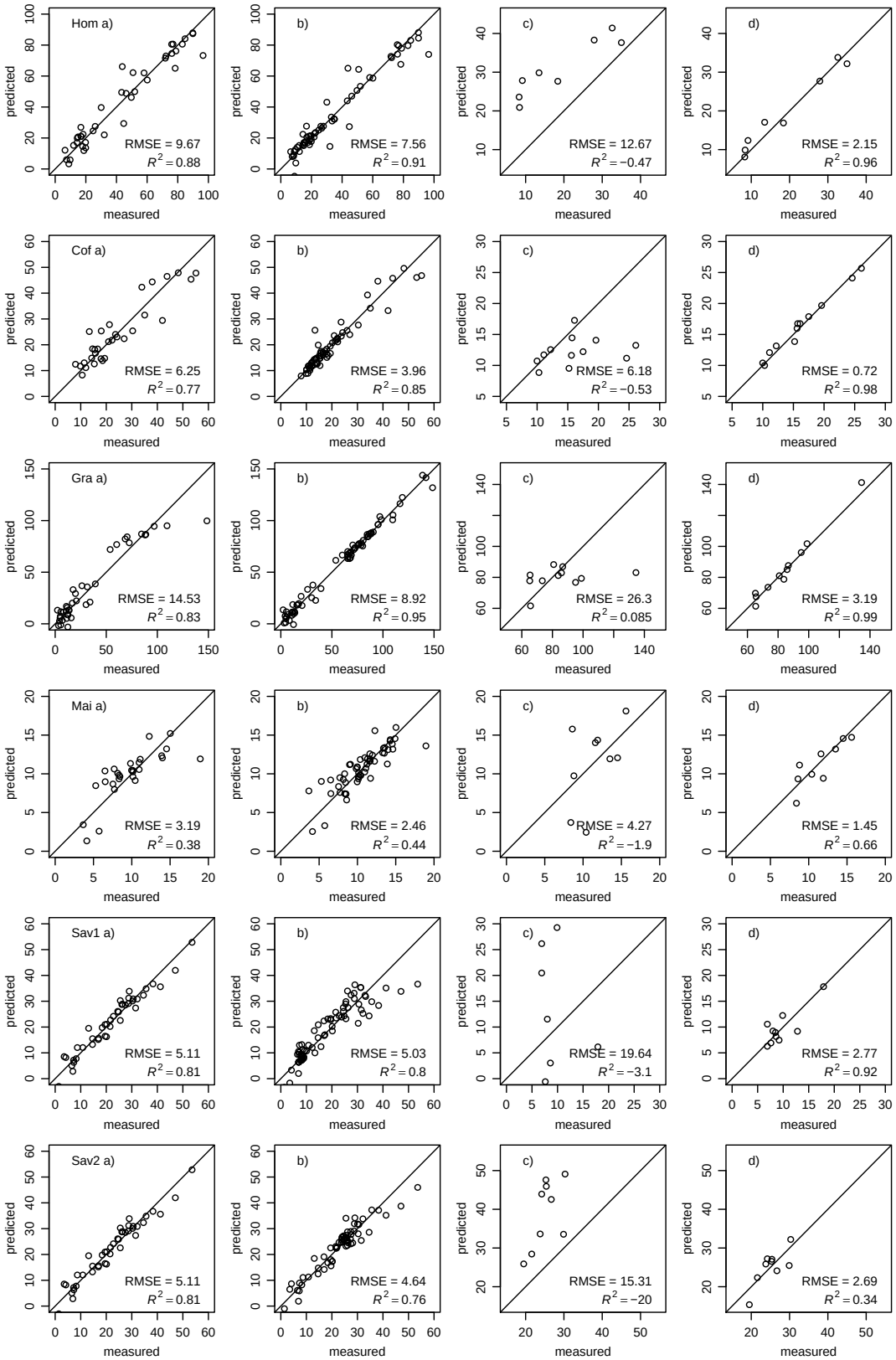



Figure 3.7: Predicted versus measured SOC content (mg $g^{-1}$): Comparison of calibration and model test for the different local models; a) calibration of the local model b) calibration of the local model augmented with synthetic data c) test of the local model d) test of the local model augmented with synthetic data. Results for the local model merged with moist data are displayed in the online Supplementary Material (B.2).

### 3.3.4 Distribution of SOC within the soil profiles

We obtained four sets of predicted SOC contents per profile using the *raster* data sets, namely predictions with the *Reg*, *Reg_synth*, *Loc* and *Loc_synth* models. Figure 3.8 summarizes the variation of SOC with depth as predicted by these models. The predictions with the *Reg* model were generally far from the measured SOC contents. In contrast, *Loc_synth* predictions mostly showed good agreement with the observed values. Note that the measured SOC contents (i.e. the *profile* data sets) constitute point measurements (one or two points per depth). Thus, given the natural variability of SOC content within the profiles, we consider that a model performs well when the *profile* data sets are contained in the interquartile range (between 25% and 75% quantiles) of the predicted values.

Despite this general pattern, in some profiles deviations occurred at different depths. At the Hom profile the interquartile ranges of *Loc*, *Loc_moist* and *Loc_synth* predictions were much smaller than those of *Reg* and *Reg_synth*, which might indicate that these predictions were more stable. The *Loc_synth* model predicted well for an area between $-20$ and $-40$ cm and from $-70$ to $-100$ cm depth. In between, however, predicted SOC contents differed from measurements. All three local models showed a distinctive jump around $-70$ cm, which did not correspond to the measured values. *Loc_moist* predictions lay between those of *Loc_synth* and *Loc*.

This jump in predictions of all local models at the Hom plot is probably due to calibration issues with the white reference. As the detectors of the spectrometer change while they are running and heating up, it is necessary to re-calibrate the instrument with a white reference regularly. This problem probably occurred during the measurements between $-40$ and $-70$ cm of that soil profile.

The interquartile ranges of predictions with the *Reg* and *Reg_synth* models at Cof were very large and many negative predictions occured. On the contrary, ranges of *Loc* and *Loc_synth* were much smaller and their predictions were similar. Compared to the measured values, both models predicted too low SOC contents in the upper part of the profile.

At Gra the *Reg* predictions were substantially inferior to the measurements within the whole profile. While the *Loc* and the *Reg_synth* predicted well in the lower part of the profile, they underestimated the SOC content in the upper soil. Predictions with *Loc_synth* were much closer to the measured values than any other model. They only deviated in the uppermost part of the profile (between 0 and $-10$ cm) where no measured values existed, and in an area around $-70$ cm.

The *Reg* model predicted very high SOC contents in the upper 50 cm and very low and even negative ones for the rest of the profile at Mai. *Reg_synth* predictions

showed the complete opposite, low values in the upper part and high values at the bottom of the profile. Although in general, *Loc* and *Loc_synth* predictions agreed better with measured values, the *Loc* model led to negative predictions in the lower part.

At Sav1 the predictions with all models but *Loc_synth* had a very high interquartile range and were negative for most parts of the profile. Additionally, *Reg* and *Loc* predicted unreasonably high values for the first 5 cm of the profile.

In the upper part of the profile at Sav2, *Reg_synth* and *Loc_synth* predicted low or negative SOC contents. Although the *Reg* and *Loc* models did not show these extremes, their predictions were in general too high. Because we lacked measurements at the top of the profile, the assessment of agreement between data and models is impossible. However, for the part of the profile for which measured values exist, *Loc_synth* and *Reg_synth* matched them the best.

In summary, the *Loc_synth* models outperformed the other ones. However, predictions for the upper segments of the soil profiles disagreed with the measured data. Especially for the upper two segments in the Sav2 profile the SOC content was underestimated. This might partly be associated with the lack of in-situ samples from this area and thus of similar *synthetic* spectra in the calibration data set. The SOC content often changed substantially in the upper parts of soil. Thus, omitting to include these samples might lead to inaccurate predictions for all samples with similar spectral characteristics. Furthermore, as mentioned in Section Spectral characteristics of the data, measurements in the upper part of a profile might have suffered from light entering the contact probe.

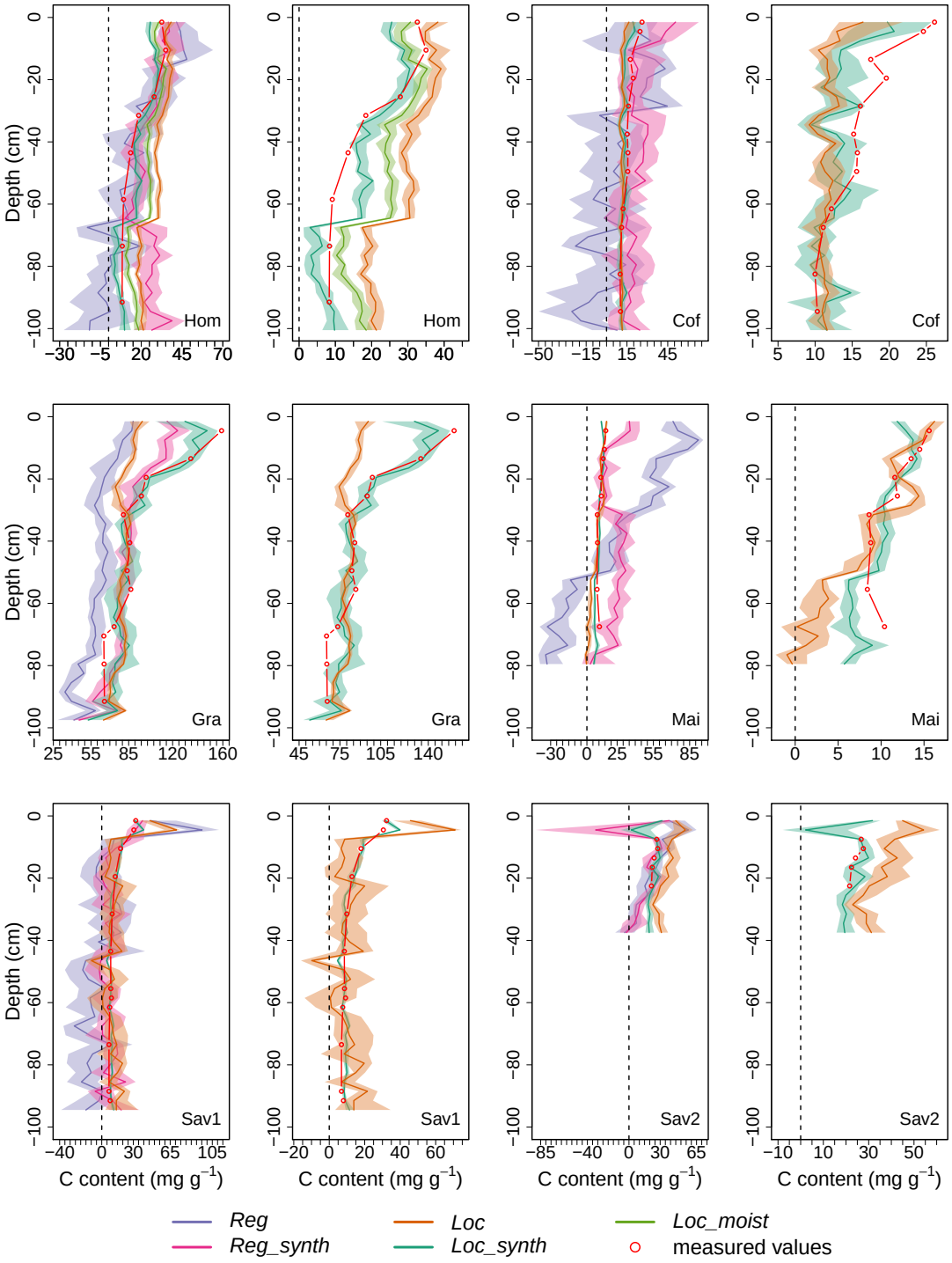



Figure 3.8: Depth profiles of SOC content predicted by the different models (mg g$^{-1}$). The coloured areas indicate the interquartile range (25%–75%) and the plain line the median. The models are designated as *Reg* = regional model, *Reg_synth* = regional model merged with six synthetic data sets (one for each profile), *Loc* = local model, *Loc_synth* = local model augmented with the respective synthetic data set. The second subplot shows details of predictions with local models for each land use zone. Hom = Homegarden, Cof = Coffee, Gra = Grassland, Mai = Maize, Sav1 and Sav2 = Savannah.

## 3.4 Discussion

### 3.4.1 Spectral characteristics of the data

Our results showed that the in-situ *raster* and air-dried *local* spectra differed substantially from each other in all analysed land use zones. This is in agreement with other studies and is due to differences in moisture content and sample preparation. Indeed, reflectance generally decreases with increasing moisture content of the sample (e.g. Nocita et al., 2013; Lobell et al., 2002). Additionally, smearing of soil can lead to higher reflectance compared to sieved soil samples (Morgan et al., 2009). The bulk density probably affects the spectra both directly and indirectly. On the one hand, reflectance intensity increases in compacted soils (i.e. with a larger bulk density) compared to non compacted soils (Dematte et al., 2010). On the other hand, in undisturbed soils bulk density is negatively correlated with the SOC content (Federer et al., 1993) and thus with the spectral response.

The dimension reduction of spectra by PCA revealed distinct clusters among the in-situ *raster* spectra in some profiles. This could be an indication of abrupt changes of soil properties between horizons. At the Mai profile, for example, spectra in the A horizon differed markedly from those in other parts of the profile. They build a distinct cluster in the lower left quadrant in Figure 3.5, probably due to a higher SOC content. Actually, we could clearly distinguish the horizons by their colour in the field. In contrast, at the Hom profile the soil was very homogeneous and no horizon boundaries were identified. This corresponds well with the spectral information, as no clusters formed and soil properties probably change continuously with depth. The spectral variation with depth in all six profiles is displayed in Figure B.1 in the online Supplementary Material.

*Raster* spectra outside the main clusters could be either due to extreme sample characteristics (i.e. very high or low SOC content) or to different surface conditions which occurred while scanning the profile. Even though the profile wall was carefully prepared prior to scanning, surface roughness in some segments might have been different. Especially in the upper part of the profiles, the soil was partly crumbly and external light might have entered the contact probe sensor. Furthermore environmental conditions might have changed as the scanning of a whole profile took about two hours.

The *synthetic* samples projected into the space spanned by the first two PCs overlap very well with the *profile* samples and most of the *raster* samples. That means that the *synthetic* samples contain the possible spectral changes due to varying bulk density, moisture content and surface roughness. In other words, they cover the spectral variability as it appears in the field regardless of its source. In

contrast, the *moist* spectra in Homegarden were spectroscopically distinct from the in-situ spectra indicating that moisture alone does not explain the difference between spectra of dry and sieved samples and in-situ scanned soils.

### 3.4.2 Modelling accuracy

#### Regional and local models

In general, the *Loc* models outperformed the *Reg* model. Note that in our study the *Loc* models correspond to reduced regional libraries and not to samples from one study site, which are usually referred to as local samples. The inferior accuracy of *Reg* was probably caused by a larger variability of soil properties unrelated to SOC content in the *regional* data set.

Despite better results in LOOCV, the *Loc* models often failed to predict the SOC content of the *raster* samples. Indeed, the PCA revealed sometimes substantial spectral differences between the dry *local* samples and the in-situ *raster* samples. Generally reflectance spectra of soil decreases with increasing SOC content, the same, however, also applies for spectra with increasing soil moisture content (Nocita et al., 2013; Lobell et al., 2002) or different degrees of surface roughness (Stevens et al., 2008). A spectral library that only includes dry and sieved soil spectra is thus likely to fail for in-situ predictions.

#### Augmented models

In our study the *synthetic* spectra and their SOC contents were generated from in-situ *profile* spectra. That means that their spectral characteristics reflect the corresponding SOC content and the environmental conditions like soil moisture, bulk density or surface roughness. By using the *synthetic* data set to extend the existing libraries, all spectral characteristics which are covered by the *profile* validation samples were incorporated in the model.

In other words, the model learned the relationship between in-situ spectra and SOC and was therefore able to predict SOC on similar spectra. This is the basic principle of supervised learning called the smoothness assumption: "If two points $x_1$, $x_2$ are close, then so should be the corresponding outputs $y_1$,$y_2$" (Chapelle et al., 2006). Because the dried and in-situ spectra (i.e. inputs) are dissimilar (i.e. not close), the *Reg* and *Loc* models trained on dry spectra failed to predict the SOC contents of in-situ spectra.

Therefore, augmenting the calibration data sets with *synthetic* spectra increased the validation accuracy on the *profile* samples. The effect was more pronounced in the

*Loc* models than the *Reg* model. This is probably due to two aspects. First the *Loc* models already performed better. Second, the proportion of *synthetic* spectra from the relevant profile was larger in the *Loc* than in the *Reg* model. The *Loc_synth* models did not only outperform the *Loc* models considering the prediction of the *profile* samples, but also led to better calibration results from LOOCV.

A number of studies confirm that including new data in a database improves predictions for a small target area. Shepherd et al. (2002), for instance, mentioned in their work that when analysing a new area that lacks representative soil samples in the calibration database, some soil samples should be analysed and added to the calibration. Furthermore, Sankey et al. (2008) and Brown (2007) found that combining a global spectral database with local samples often provided better predictions than either models based on the global database or the local database alone. In our case the *synthetic* samples correspond to such local samples added to the calibration database. The main advantage of including new data in general is that models comprise the spectral information of the target area, but are more stable and less prone to overfitting than models based on few, local samples alone (Brown, 2007).

Combining the *Loc* model in the Homegarden ecosystem with moist spectra improved predictions only slightly. It shows that in-field spectra differ from those scanned in the laboratory by more than the mere moisture content. Therefore, a simple adjustment for varying moisture content in calibration data might be insufficient. In contrast, including synthetic samples into the calibration data sets improved predictions for *raster* spectra taken under the same field conditions substantially. However, as some extreme points lacked appropriate synthetic spectra, prediction quality for these points remains unknown. This shows that the choice of samples for reference chemical analysis is crucial in order to generate an appropriate synthetic spectral data set.

We are aware that our test procedure differs from a classical validation approach where the data are split (randomly) into a calibration and a validation data set (or hold-out data set). Our *profile* samples were used to create *synthetic* data and to test our models. However, the *synthetic* spectra constitute new data and we avoid reusing the same in-situ spectra both for calibration and testing. The results from LOOCV already confirm the improvement of SOC predictions of the augmented models.

Actually, in LOOCV every data point is predicted once based on all other points. The cross validation error (the mean error of all predictions) is an approximately unbiased estimate for the test error (i.e. the error one obtains when testing on new data) (James et al., 2013). Compared to the validation on a hold-out data set, LOOCV has less potential to overestimate the test error and will always yield

the same result. Indeed, there is no randomness in LOOCV due to data splitting because every point is predicted (James et al., 2013). Testing our models on the in-situ spectra yields an additional information, namely how well the models predict the SOC content on spectra scanned in the field. This performance is relevant for the prediction based on in-situ *raster* data sets.

Additionally, the interquartile ranges of the predicted values in the soil profiles serve as an independent validation. Because the reflectance spectra of the *profile* and the *raster* datasets were acquired in the same profiles, we suppose that the SOC contents are comparable. Thus, if the predictions based on the *raster* spectra match the measured SOC contents of the *profile* data, the model performs satisfactorily.

Augmenting the *Loc_synth* models with synthetic data led to better models (smaller cross validation errors) which can accurately predict SOC from all given spectral information (i.e. dry and in-situ spectra). Additionally, our aim was precisely to show how modelling with limited in-situ data might work. The line of further research is to evaluate the potential of including synthetic spectra in the calibration by validating on a representative hold-out data set.

### 3.4.3 Suggested framework for modelling with in-situ spectra

Based on our main findings we propose the following approach for the prediction of SOC content with an existing library and limited in-situ spectra.

**Selection of available calibration data** If possible only a subset of the library that is geographically close to the target area or comprises similar soil properties should be used, as also suggested by Guerrero et al. (2010).

**Measuring in-situ spectra** While measuring spectra in-situ all samples should be prepared in the same way (i.e. removing of grass, smoothing of surface, no preparation at all etc.) and environmental conditions should be stable.

**Selection of in-situ calibration data** The spectral variability of the in-situ spectra should be assessed directly in the field. One possible approach is a PCA. However, different methods to evaluate the spectral similarity exist (Ramirez-Lopez et al., 2013). In-situ calibration samples can then be chosen accordingly to cover the spectral variation in the field.

**Generation of synthetic data** In case that the in-situ calibration spectra are limited either in absolute number or relative to the existing spectral library, SMOTE can be used to generate new synthetic data. A Monte-Carlo analysis should be run because not all synthetic data sets improve predictions equally well (Bogner et al., 2014).

**Deriving the final model** The final calibration data set is created by combining the spectral library with the chosen synthetic data set. The usual modelling framework can be used including data preprocessing and model validation.

## 3.5 Conclusions

We have developed a modelling framework for prediction of SOC content from in-situ soil spectra by efficiently utilising rare field data. The synthetic minority oversampling technique (SMOTE) was applied to generate synthetic data sets with spectral characteristics resembling those of the in-situ soil samples. These synthetic data were then successfully incorporated in an existing spectral library. This approach is especially promising if in-situ calibration data are limited (i.e. due to sampling time and/or available amount of money). By combining the available calibration data with synthetic spectra, we simultaneously incorporate spectral characteristics from a new site and in-situ features directly into the model. Models combined with moist spectra alone appear to lack spectral information that reflects the field conditions appropriately.

Modelling with in-situ spectra especially in a new area is challenging because the choice of both the field calibration samples and samples from the spectral library is crucial. When synthetic spectra are generated from only a few samples, they should be representative for the studied site. Hence, further studies could focus on optimizing the selection procedure in the field and from spectral libraries.

## Acknowledgements

This study was funded by the German Research Foundation (DFG) within the Research Unit 1246 (KiLi) and supported by the Tanzanian Commission for Science and Technology (COSTECH) and the Tanzania Wildlife Research Institute (TAWIRI). We thank Holger Pabst, Johannes Hepp, Hannes Thomasch, David Kienle, Johannes Kühnel, Jumanne Mwinyi and James Ndimion for assistance during the field campaigns and in the laboratory.

## References


Awiti, A. O., M. G. Walsh, K. D. Shepherd, and J. Kinyamario (Jan. 2008). "Soil condition classification using infrared spectroscopy: A proposition for assessment of soil condition along a tropical forest-cropland chronosequence". In: *Geoderma* 143.1-2, pp. 73–84. DOI: 10.1016/j.geoderma.2007.08.021.

Becvar M.and Hirner, A. and U. Heiden ((2006 - 2008)). DLR Spectral Archive. URL: http://cocoon.caf.dlr.de/intro_en.html.

Brown, D. J. (2007). "Using a global VNIR soil-spectral library for local soil characterization and landscape modeling in a 2nd-order Uganda watershed". In: *Geoderma* 140.4, pp. 444–453. DOI: 10.1016/j.geoderma.2007.04.021.

Chang, C.-W., D. A. Laird, M. J. Mausbach, and C. R. Hurburgh (2001). "Near-Infrared Reflectance Spectroscopy-Principal Components Regression Analyses of Soil Properties". In: *Soil Science Society of America Journal* 65.2, pp. 480–490. DOI: 10.2136/sssaj2001.652480x.

Chawla, N. V., K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer (2002). "SMOTE: Synthetic Minority Over-sampling Technique". In: *Journal of Artificial Intelligence Research* 16, pp. 321–357.

Fernandes, E. C. M., A. Oktingati, and J. Maghembe (1985). "The Chagga homegardens: a multistoried agroforestry cropping system on Mt. Kilimanjaro (Northern Tanzania)". In: *Agroforestry Systems* 2 (2), pp. 73–86. DOI: 10.1007/BF00131267.

GeologicalMap (1955). *Geological Survey of Tanganyika [1:125,000] quarter degree sheets*. Mineral Resources Division and Bundesanstalt für Bodenforschung and V.O. Technoexport. Dodoma, Tanzania.

Guerrero, C., R. Zornoza, I. Gómez, and J. Mataix-Beneyto (2010). "Spiking of NIR regional models using samples from target sites: Effect of model size on prediction accuracy". In: *Geoderma* 158.1-2, pp. 66–77. DOI: 10.1016/j.geoderma.2009.12.021.

Haubrock, S. N., S. Chabrillat, C. Lemmnitz, and H. Kaufmann (Jan. 2008). "Surface soil moisture quantification models from reflectance data under field conditions". In: *International Journal of Remote Sensing* 29.1, pp. 3–29. DOI: 10.1080/01431160701294695.

Hunt, G. (1977). "Spectral Signatures of Particulate Minerals in the Visible and Near Infrared". In: *Geophysics* 42.3, pp. 501–513. DOI: 10.1190/1.1440721.

IUSS Working Group WRB (2007). *World reference base for soil resources 2006, first update 2007. World Soil Resources Reports No. 103*. http://www.fao.org/fileadmin/templates/nr/images/resources/pdf_documents/wrb2007_red.pdf. [Accessed on 2014-04-08].

Korobeynikov, A. (2010). "Computation- and Space-Efficient Implementation of SSA". In: *Statistics and Its Interface* 3.3, pp. 357–368.

Lobell, D. B. and G. P. Asner (2002). "Moisture Effects on Soil Reflectance". In: *Soil Science Society of America Journal* 66.3, pp. 722–727. DOI: 10.2136/sssaj2002.7220.

Minasny, B., A. B. McBratney, V. Bellon-Maurel, J.-M. Roger, A. Gobrecht, L. Ferrand, and S. Joalland (2011b). "Removing the effect of soil moisture from NIR diffuse reflectance spectra for the prediction of soil organic carbon". In: *Geoderma* 167-168.0, pp. 118–124. DOI: 10.1016/j.geoderma.2011.09.008.

Misana, S. B., A. E. Majule, H. V. Lyaruu, and L. U. Change (2003). *Linkages between changes in land use, biodiversity and land degradation on the slopes of Mount Kilimanjaro, Tanzania*. LUCID Project, International Livestock Research Institute.

Morgan, C. L., T. H. Waiser, D. J. Brown, and C. T. Hallmark (2009). "Simulated in situ characterization of soil organic and inorganic carbon with visible near-infrared diffuse reflectance spectroscopy". In: *Geoderma* 151.3-4, pp. 249–256. DOI: 10.1016/j.geoderma.2009.04.010.

Nocita, M., L. Kooistra, M. Bachmann, A. Müller, M. Powell, and S. Weel (Nov. 2011). "Predictions of soil surface and topsoil organic carbon content through the use of laboratory and field spectroscopy in the Albany Thicket Biome of Eastern Cape Province of South Africa". In: *Geoderma* 167-168.0, pp. 295–302. DOI: 10.1016/j.geoderma.2011.09.018.

Nocita, M., A. Stevens, C. Noon, and B. van Wesemael (2013). "Prediction of soil organic carbon for different levels of soil moisture using Vis-NIR spectroscopy". In: *Geoderma* 199.0, pp. 37–42. DOI: 10.1016/j.geoderma.2012.07.020.

R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*. ISBN 3-900051-07-0. R Foundation for Statistical Computing. Vienna, Austria.

Reeves III, J. B. (2010). "Near- versus mid-infrared diffuse reflectance spectroscopy for soil analysis emphasizing carbon and laboratory versus on-site analysis: Where are we and what needs to be done?" In: *Geoderma* 158.1-2, pp. 3–14. DOI: 10.1016/j.geoderma.2009.04.005.

Sankey, J. B., D. J. Brown, M. L. Bernard, and R. L. Lawrence (2008). "Comparing local vs. global visible and near-infrared (VisNIR) diffuse reflectance spectroscopy (DRS) calibrations for the prediction of soil clay, organic C and inorganic C". In: *Geoderma* 148.2, pp. 149–158. DOI: 10.1016/j.geoderma.2008.09.019.

Shepherd, K. D. and M. G. Walsh (2002). "Development of Reflectance Spectral Libraries for Characterization of Soil Properties". In: *Soil Science Society of America Journal* 66.3, pp. 988–998. DOI: 10.2136/sssaj2002.9880.

Soini, E. (2005). "Changing livelihoods on the slopes of Mt. Kilimanjaro, Tanzania: Challenges and opportunities in the Chagga homegarden system". In: *Agroforestry Systems* 64.2, pp. 157–167. DOI: 10.1007/s10457-004-1023-y.

Stenberg, B. (2010). "Effects of soil sample pretreatments and standardised rewetting as interacted with sand classes on Vis-NIR predictions of clay and soil organic carbon". In: *Geoderma* 158.1-2. Diffuse reflectance spectroscopy in soil science and land resource assessment, pp. 15–22. DOI: DOI:10.1016/j.geoderma.2010.04.008.

Sugiura, N. (1978). "Further analysts of the data by akaike' s information criterion and the finite corrections". In: *Communications in Statistics - Theory and Methods* 7.1, pp. 13–26. DOI: 10.1080/03610927808827599.

Torgo, L., R. P. Ribeiro, B. Pfahringer, and P. Branco (2013). “SMOTE for Regression”. In: *Progress in Artificial Intelligence*. Springer, pp. 378–389.

Vasques, G., S. Grunwald, and J. Sickman (2008). “Comparison of multivariate methods for inferential modeling of soil carbon using visible/near-infrared spectra”. In: *Geoderma* 146.1-2, pp. 14–25. DOI: 10.1016/j.geoderma.2008.04.007.

Viscarra Rossel, R. A. and R. Webster (2012). “Predicting soil properties from the Australian soil visible-near infrared spectroscopic database”. In: *European Journal of Soil Science* 63.6, pp. 848–860. DOI: 10.1111/j.1365-2389.2012.01495.x.

Viscarra Rossel, R., R. McGlynn, and A. McBratney (Dec. 2006a). “Determining the composition of mineral-organic mixes using UV–vis–NIR diffuse reflectance spectroscopy”. In: *Geoderma* 137.1-2, pp. 70–82. DOI: 10.1016/j.geoderma.2006.07.004.

Viscarra Rossel, R., D. Walvoort, A. McBratney, L. Janik, and J. Skjemstad (2006b). “Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties”. In: *Geoderma* 131.1-2, pp. 59–75. DOI: 10.1016/j.geoderma.2005.03.007.

Viscarra Rossel, R., S. Cattle, A. Ortega, and Y. Fouad (2009). “In situ measurements of soil colour, mineral composition and clay content by vis-NIR spectroscopy”. In: *Geoderma* 150.3-4, pp. 253–266. DOI: 10.1016/j.geoderma.2009.01.025.

Viscarra Rossel, R. A. (Jan. 2008). “ParLeS: Software for chemometric analysis of spectroscopic data”. In: *Chemometrics and Intelligent Laboratory Systems* 90.1, pp. 72–83. DOI: 10.1016/j.chemolab.2007.06.006.

Vågen, T.-G., K. D. Shepherd, and M. G. Walsh (2006). “Sensing landscape level change in soil fertility following deforestation and conversion in the highlands of Madagascar using Vis-NIR spectroscopy”. In: *Geoderma* 133.3-4, pp. 281–294. DOI: 10.1016/j.geoderma.2005.07.014.

Wehrens, R. (2011). *Chemometrics with R: multivariate data analysis in the natural sciences and life sciences*. Springerverlag Berlin Heidelberg.

Wetterlind, J. and B. Stenberg (2010a). “Near-infrared spectroscopy for within-field soil characterization: small local calibrations compared with national libraries spiked with local samples”. In: *European Journal of Soil Science* 61.6, pp. 823–843. DOI: 10.1111/j.1365-2389.2010.01283.x.

Wold, S., M. Sjöström, and L. Eriksson (2001). “PLS-regression: a basic tool of chemometrics”. In: *Chemometrics and intelligent laboratory systems* 58.2, pp. 109–130. DOI: 10.1016/S0169-7439(01)00155-1.

Zech, M., C. Hörold, K. Leiber-Sauheitl, A. Kühnel, A. Hemp, and W. Zech (2014). “Buried black soils on the slopes of Mt. Kilimanjaro as a regional carbon storage hotspot”. In: *CATENA* 112.0, pp. 125–130. DOI: 10.1016/j.catena.2013.05.015.

# Small scale spatial variability of soil hydraulic properties in different land uses at Mt. Kilimanjaro

ANNA KÜHNEL[1], CHRISTINA BOGNER[2], AND BERND HUWE[1]

[1]Soil Physics Group, BayCEER, University of Bayreuth, Germany
[2]Ecological Modelling, BayCEER, University of Bayreuth, Germany

in preparation for SOIL

corresponding author: Anna Kühnel (anna.kuehnel@uni-bayreuth.de)

**Abstract**

Little is known about the spatial heterogeneity of soil hydraulic properties at small scales and how it is affected by different land uses. As measuring these properties is laborious, they are often predicted with pedotransfer functions. These functions require information on basic soil parameters, which itself is often lacking. This is especially the case for tropical ecosystems, where information on the basic soil parameters such as soil organic carbon, nitrogen and particle size distribution is often scarce. In this study we combined visible and near infrared spectroscopy, which can be used to derive various basic soil properties at the same time, with a pedotransfer function to predict soil physical and hydraulic properties. Subsequently we visualized these physical and hydraulic properties in-situ, i.e. on soil profiles of different land uses at Mt. Kilimanjaro, namely homegarden, a traditional agroforestry system, coffee plantation, maize field and natural savannah and studied the spatial patterns at the centimeter scale. Furthermore we derived carbon stocks and total available water. Root mean squared error for saturated water content and residual water content were as low as 2.7 and 3.6%, respectively. Distribution of saturated and residual water content and thus of the available water content was quite homogenic, although a vertical trend could be observed for most soil profiles. Field air capacity and porosity showed more heterogenic patterns, with higher semivariances already at small scales. Soil carbon stocks were highest in the homegarden and lowest in a maize field, whereas savannah contained most available water. We conclude that combining soil spectroscopy with pedotransfer functions seems a promising way to obtain information on soil hydraulic properties.

*Keywords:* diffuse reflectance spectroscopy; random forest; pedotransfer function; soil carbon stocks; available water capacity, air capacity; hydraulic conductivity; tropical soils; Kilimanjaro

## 4.1 Introduction

Soils are the most diverse ecosystems of the world (Young et al., 2004) and their heterogeneity is the key to biological processes. Carbon (C) storage and available water capacity of soils are important ecosystem services (Power, 2010). The spatial variability of the latter depends on soil hydraulic properties at the centimeter scale (Ritsema et al., 1998), which in turn are associated with other physical, chemical and biological soil parameters. As the habitat space of soil biota is not distributed equally in the soil matrix, spatial heterogeneity of soils is more and more appreciated as an intrinsic property (Young et al., 2001; Nunan et al., 2002).

The turnover and storage of organic matter involves processes at different scales and thus different levels of structural and functional complexity (Christensen, 2001). Studies of the intact soil at levels from centimeter to meter, as this is the level at which C storage changes matter and direct and immediate effects of physical disturbance occur (Christensen, 2001), are of great importance. However, detailed information on the spatial variability of soil physical parameters on the intact soil at that scale are lacking.

Especially in sub-Saharan Africa where soils are susceptible to changes, sustainable management of soil resources is needed for future welfare (Awiti et al., 2008). Compared to many lowland african soils, the Kilimanjaro area in the northern part of Tanzania provides fertile soils and favourable climatic conditions for agriculture. Mt. Kilimanjaro with its rain forests maintains a stable water supply for many people living at the southern slopes of the mountain (Rohr et al., 2003). However, during the last century the population has continuously grown to about 1.4 million in 2002 (National Bureau of Statistics, 2006). Consequently, former forested areas in the submontane zone of Mt. Kilimanjaro were converted into cultivated land like agroforestry systems or coffee plantations. Similarly large areas of savannah bushland at the base of the mountain has been turned into agricultural fields (Misana et al., 2003). Little information is available on how these land use changes affect essential ecosystem services like carbon and water storage.

In recent years visible and near infra-red diffuse reflectance spectroscopy (Vis-NIR DRS) has become popular as a fast and inexpensive analytical tool for many soil parameters like organic C and clay content (Shepherd et al., 2002; Vågen et al., 2006; Viscarra Rossel et al., 2006b; Stevens et al., 2013). This technique is based on the relationship between the reflectance of a surface at different wavelengths (350–2500 nm) and its physical and chemical properties (Hunt, 1977). If robust portable instruments are available and infrastructural problems can be overcome, this method has a huge potential especially for developing countries (Shepherd et al., 2002).

Vis-NIR DRS has been successfully applied to derive soil information from spectra of dried and sieved soil samples and global and regional spectral databases have been established (Shepherd et al., 2002; Brown, 2007). However, the application of Vis-NIR DRS in-situ remains challenging. Indeed, because soil moisture, surface roughness and bulk density vary largely in the field, models calibrated on available spectral databases often fail to predict from in-situ spectra. In a recent study, Bogner et al. (2014) used the synthetic minority oversampling technique (SMOTE) (Chawla et al., 2002; Torgo et al., 2013) and proposed a methods to predict soil organic C directly from in-situ soil spectra.

While Vis-NIR DRS is useful to assess soil chemical and physical properties, hydraulic properties might be difficult to determine. Janik et al. (2007) developed calibration models for several hydraulic properties from air-dried sieved soil mid infrared spectra with low cross validation error. Predictions on an independent test set were however much higher. Santra et al. (2009) mentioned that the performance of calibration models to predict the van-Genuchten parameters from Vis-NIR spectra was not comparable to pedotransfer functions (PTFs) with basic soil properties. Indeed, Minasny et al. (2011a) stated that such pore-space characteristics should not be predicted by spectroscopy, because the modelled relationships lack a physical basis. Additionally, measurements of hydraulic properties that are necessary to calibrate a model are more laborious than those of other soil properties like C and N content. Therefore, indirect prediction with PTFs might be more appropriate.

PTFs have been developed in order to derive properties that are difficult to measure, like soil hydraulic properties, from easily measurable soil properties (Rawls et al., 1991). Many different PTFs have been developed, mainly concentrating on particle size distribution, bulk density and organic matter content as input variables (Schaap et al., 2001). Most studies have focused on soils of the temperate regions (Minasny et al., 2011a) and only few PTFs explicitly deal with tropical soils (Tomasella et al., 2004). In particular, PTFs that can handle characteristics from volcanic soils are lacking.

To create a PTF, machine learning techniques like random forest (RF) (Breiman, 2001) can be used. RF can handle a large amount of input variables compared to the number of samples and still does not overfit the model. Furthermore no assumptions about the relationship between the predictors and the target variable have to be made, so that non linear relationships can be equally well incorporated in the model (Prasad et al., 2006). Only recently RF models have been used as pedotransfer functions to predict slope stability (Ließ et al., 2011), soil texture (Ließ et al., 2012), soil parent material (Heung et al., 2014), the strength of preferential transport (Koestel et al., 2014) or bulk density (Sequeira et al., 2014). However, to our knowledge, RF has not yet been used to predict soil hydraulic properties.

Given the natural variability of soil properties even at small scales (Rossi et al., 2009), direct physical measurements are time consuming and hard to obtain without the use of prediction techniques. By combining spectral models with RF models, soil hydraulic properties can be obtained for a large amount of samples with minimal effort in-situ on the intact soil. These techniques can be used to derive information from the centimeter to the kilometer scale at a high spatial resolution.

The aim of this study is to characterise the soils of four different typical land uses at the southern slopes of Mt. Kilimanjaro regarding soil organic matter, soil physical and soil hydraulic properties and assess their spatial variability at the profile scale.

The specific methodological objectives are i) to test the applicability of including SMOTE in spectroscopy for the prediction of various soil parameters from in-situ soil spectra and ii) to assess the performance of RF models for the prediction of different soil hydraulic properties.

## 4.2 Methods

### 4.2.1 Study area

The study was conducted in the colline and the submontane zones on the southern slopes of Mt. Kilimanjaro, Tanzania (3°4'33"S, 37°21'12"E). This large shield volcano is the highest mountain in Africa with elevations ranging from 800 m up to 5895 m a.s.l.. Its formation began 2.5 million years ago, with different phases of volcanic activity, the last about 200,000 thousand years ago (Nonnotte et al., 2008). The volcanic soils that consequently formed on the superficial deposits are relatively young and quite different from the deeply weathered soils in the surrounding savannah plains. In the plains Acrisols, Ferralsols, Lixisols, Nitisols and Vertisols dominate, whereas the main soil type of the higher zones of Mt. Kilimanjaro is Andosol (Zech et al., 2014), with Vertisols and Umbrisols in more weathered areas. In the south east of the mountain various small strombolian-type parasitic cones formed during the final stage of volcanic activity (Nonnotte et al., 2008). The main soil type of theses small volcanoes is Leptosol.

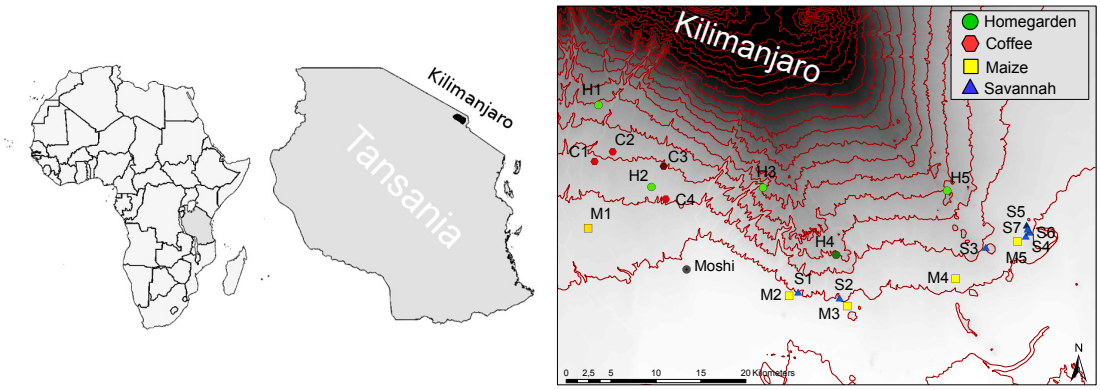



Figure 4.1: Study area and research plots. C= coffee, H = homegarden, M = maize, S = savannah, C3, H4, M1 and S5 were chosen for the detailed soil profile study.

In our study, we focused on four different land uses, namely savannah and maize fields in the colline zone at the base of the mountain and coffee plantations and traditional homgarden in the submontane zone (Figure 4.1).

The colline zone extends around the mountain base, between 700 m and 1000 m a.s.l. (Misana et al., 2003) and comprises part of the adjacent plains and the small

volcanoes. It receives a mean annual precipitation of 400–900 mm (Soini, 2005). The rainfall pattern in this zone, as on the whole mountain, is bimodal, with an extended rainy season from March to May and a short rainy season in November (Rohr et al., 2003). The natural ecosystem of the colline zone is savannah with *Balanitis aegyptiaca* and different Acacias (*Acacia tortilis*, *Acacia senegal*, *Acacia nilotica*) as the main tree species and various different grass species underneath. Nowadays natural savannah is restricted mostly to the small volcanoes, as the arable land in the plains is increasingly transformed into agricultural fields, where maize and occasionally sunflowers are grown (Mbonile, 2003).

The submontane zone, reaching from 1000–1800 m a.s.l. on the southern slopes of Mt. Kilimanjaro, receives a mean annual precipitation of 1200 – 2000 mm (Soini, 2005). A traditional agroforestry system, the so called homegarden, preserved by the Chagga tribe during the last centuries, covers an area of about 1200 km$^2$ in this zone (Fernandes et al., 1985). The homegarden is a multi-storey agricultural system, with up to four vegetation layers. Different vegetables, like sweet potato, taro and beans (Fernandes et al., 1985) are usually grown under the coffee and banana layers. Big trees provide shade and protection against soil erosion and can be used as timber. Apart from the homegardens, the submontane zone comprises large areas, where extensive coffee plantations were established. On these plantations *Coffea arabica* is grown as a cash crop, as soil and climate provide optimal conditions for coffee cultivation.

Soil types of the studied coffee and homegarden land uses within the submontane zone were diverse with Vertisols, Umbrisol, Nitisols and Andosols. Despite the fact, that some soils were not classified as Andosols, the volcanic origin was still visible in soils from more than half of the plots. Soils of the savannah plots on the small volcanic hills were all classified as Leptosol, those in the plains as Vertisols. Main soil type under the maize was Nitisol.

### 4.2.2 Field sampling and laboratory analysis

On 21 plots, comprising the four different land uses, soil samples were collected with a soil auger down to a depth of 1 m or until continuous bedrock was reached. Soil auger samples were separated by horizons and collected separately. All soil samples were sieved $< 2$ mm and oven-dried at 45 ℃ for 24 h. The sand fraction was determined by wet sieving, silt and clay content were measured with a Master Sizer S particle size analyzer. C and N contents were determined using a CNS-Analyser with conductivity detectors by high temperature combustion. Furthermore, in order to classify the soil of each plot, pH, cation exchange capacity (CEC), base saturation (BS) and soil color were analysed in samples from one soil auger per

plot. Andic properties were tested by placing an aliquot of the sample on a filter paper, previously soaked in a mixture of phenolphthalein and 1 M sodium fluoride (NaF) (Jahn et al., 2006).

Spectral measurements were conducted on a well-mixed aliquot of each dried soil sample. Therefore it was placed in a small cup and the soil surface was carefully smoothed before scanning it with an Agrispec portable spectrometer (ASD, Boulder Colorado, spectral range 350–2500 nm). The instrument was calibrated with a white reference prior to analyses and at regular intervals. To reduce the noise 30 reflectance spectra were averaged. For each soil sample the spectral information and the corresponding soil parameters are stored separated by land use as the so called $local_{\text{landuse}}$ spectral databases.

### 4.2.3 Soil profile study

One plot of each land use was selected for a detailed soil profile study (Figure 4.1). On these plots, a soil pit was dug to a depth of approximately 1 m or until continuous bedrock was reached. One profile wall was carefully cleaned of debris and roots and a frame of 0.5 m $\times$ 1 m with 3 $\times$ 3 cm segments was placed on the wall. Subsequently each segment was scanned with the same spectrometer and calibration details as used for the spectral database. After the scanning process, 24 soil core samples (diameter 2.5 cm) were taken as reference in randomly selected segments. About half of the samples were used for soil texture analyses, in the remaining samples C and N content were determined.

Furthermore the soil profile was separated into horizons and bulk samples were taken from each horizon. Soil texture, C and N content and the particle density ($\rho_p$) were determined for each horizon. Additionally five undisturbed samples per horizon for the analysis of soil water retention characteristics and bulk density ($\rho_b$) were extracted using 100 $cm^{-3}$ soil cores.

### 4.2.4 Spectral model development

As the soil spectra, that were scanned directly in the field have different spectral characteristics compared to the *local* datasets, we developed an independent model for each profile and parameter (C, N, clay, silt and sand content). Therefore the in-situ spectra of the segments with reference samples were used to generate synthetic spectra with the synthetic minority oversampling technique (SMOTE) and its extension for regression (Chawla et al., 2002; Torgo et al., 2013). Individually

for every profile and parameter, three new spectra were generated for every in-situ spectra with the following equation:

$$X_s = X_o + \delta(X_n - X_o) \tag{4.1}$$

where $X_s$ is the synthetic spectra, $X_o$ is the original spectra, $X_n$ is a randomly chosen neighbour of $X_o$ and $\delta$ is a random number between 0 and 1. The values of the respective parameter were generated in the same way.

Following the modeling approach of Bogner et al. (2014), we augmented the respective *local* spectral database with the new synthetic spectra. Subsequently partial least squares regression (PLSR) models were developed and validated by leave-one-out-cross validation (LOOCV). To assess the performances of the PLSR, we calculated the $RMSE$, the mean prediction error $MPE$ and the coefficient of determination $R^2$ as measures of the agreement between observed and predicted values:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \tag{4.2}$$

$$MPE = \frac{1}{n}\sum_{i=1}^{n}\bar{y} - y_i \tag{4.3}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{4.4}$$

where $n$ is the number of samples, $\hat{y}_i$ are the predicted values with the respective model, $y_i$ are the observed values and $\bar{y}$ is the mean of the observed values.

The PLSR models were additionally tested on the small soil core reference samples, individually for each profile. They were then used to predict C, N, clay, silt and sand content at every segment of the soil profile directly from in-situ spectra.

We are aware that this test procedure differs from a classical validation approach where the data are split into a calibration and a validation data set. The small soil core reference samples were used to create synthetic data and to test the models. However, the small soil core reference spectra were not used for calibration. The error from LOOCV already is an approximately unbiased estimate for the test error, as every data point is predicted once based on all other points (James et al., 2013).

Additionally, the bulk horizon samples from the individual profiles can serve as an independent validation, as they were not used to generate new spectra. If

predictions on the in-situ spectra are comparable to bulk horizon samples, we conclude that models perform satisfactorily.

A detailed description of the SMOTE algorithm as well as specifications about the pre-treatments of soil spectra can be found in the Online Supplementary. All analyses were performed in `R` (R Development Core Team, 2011) using the packages `ChemometricsWithR` (Wehrens, 2011) and `Rssa` (Korobeynikov, 2010).

### 4.2.5 Estimation of soil physical and hydraulic properties

#### Water retention curve

The retention characteristics were determined on the 19 undisturbed soil cores from the soil horizons of the four profiles. Ceramic plates were used, starting with water-saturated samples the water content was measured at subsequent successive dehydration at suctions of $10^{0.5}$, $10^{1}$, $10^{1.5}$ and $10^{2}$ hPa. The water content at the potential of $10^{4.2}$ was determined with a pressure chamber. A water retention curve was fitted for each horizon of the four soil profiles with the van Genuchten equation (Equation 4.5, (Van Genuchten, 1980)):

$$\theta_h = \theta_r + \frac{\theta_s - \theta_r}{(1 + (\alpha h)^n)^m} \tag{4.5}$$

where $\theta_r$ is the residual water content, $\theta_s$ is the saturated water content, $h$ is the matric potential, $\alpha$ and $n$ are fitted parameters and $m = 1 - 1/n$. The relative unsaturated hydraulic conductivity ($K_r$) could then be determined with the following equation (Van Genuchten, 1980; Mualem, 1976):

$$K_r(h) = \frac{(1 - (\alpha h)^{(n-1)}(1 + (\alpha h)^n)^{-m})^2}{(1 + (\alpha h)^n)^{m/2}} \tag{4.6}$$

Consequently we determined the water content and the relative hydraulic conductivity at a matric potential of $10^{1.8}$ hPa ($\theta_{1.8}$ or field capacity and $K_{r(1.8)}$), the available water capacity ($AWC = \theta_{1.8} - \theta_{4.2}$) and the field air capacity ($FAC = \phi - \theta_{1.8}$) for all soil horizons. These parameters were now available for every soil horizon of the four profiles (in total 19 samples), additionally to the measured values C, N, clay, silt and sand content, $\theta_s$, $\theta_{4.2}$ and $\rho_b$.

**Pedotransfer functions**

In order to predict the hydraulic properties for every segment of the soil profiles build a pedotransfer function with only those components, that were available at the segments, namely C, N, clay, silt and sand content. We chose random forest (RF), as no a priori assumptions about the relationship between input and response variable have to be made (Breiman, 2001; Prasad et al., 2006). A RF is a combination of tree predictors and can be used in classification and in regression. The advantage of RF over single regression tree analysis is, that it does not overfit, as a large number of individually trees are grown and the output is averaged (Breiman, 2001).

For the creation of a tree in a RF about one third of the input samples are left out (the so called out of bag samples). A single tree is now build by randomly selecting an input variable and splitting the dataset in two, so that the homogeneity (of the response variable) of the two resulting groups is maximized (Prasad et al., 2006). This procedure is then repeated and a tree is grown to the maximum size. At each tree node RF estimates the importance of the single variables. This permutation accuracy gives an estimate about how much the prediction error would increase, if that variable is left out. The absolute number is however less important than the proportions of the variables. $RSQ$, a measure similar to $R^2$ is calculated by

$$RSQ = \frac{\sum_{j=1}^{ntree}(1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2})}{ntree} \tag{4.7}$$

where $ntree$ is the number of grown trees, $n$ is the number of samples $\hat{y}_i$ are the predicted values, $y_i$ are the observed values and $\bar{y}$ is the mean of the observed values. It is thus the mean $R^2$ of all trees.

From the database of the 19 undisturbed soil cores from the soil horizons of the four profiles, we build individual RF models for the saturated water content ($\theta_s$), the field capacity at a potential of $10^{1.8}$ hPa ($\theta_{1.8}$), the residual water content at a potential of $10^{4.2}$ hPa ($\theta_{4.2}$), the available water content (*AWC*), porosity ($\phi = (1 - \rho_b/\rho_s) \times 100$), field air capacity (*FAC* =$\phi$ - $\theta_{1.8}$), unsaturated hydraulic conductivity ($K_{r(1.8)}$) and $\rho_b$. These parameters were consequently predicted for all segments of the soil profiles.

As input parameter we chose C, as C and N content were highly correlated and only two out of the soil texture parameters (clay and sand; silt is yielded directly out of those two, as they should add up to 100%). The number of trees to grow was set to 5000 and the number of input variables that were randomly selected at each node was set to 1. Variable importance was calculated and models evaluated by the $RSQ$. That means, that for every model, C, clay and sand content were tested,

but only those parameters, that led to the best model were chosen. Even though individual models were build for *AWC* and *FAC*, we displayed and computed the spatial pattern on the calculated values with *AWC* = $\theta_{1.8}$ - $\theta_{4.2}$ and *FAC* =$\phi$ - $\theta_{1.8}$.

Unfortunately no independent test set was available to asses the performance of the RF models on the in-situ spectra. The median predictions within the profiles, however, can provide an estimate of how well the RF models predict soil hydraulic properties of the four profiles. The package `randomForest` (Breiman, 2001) was used for analysis.

### 4.2.6 Variability of soil parameters within the profile

A spatial picture (0.5 m $\times$ 1 m or until continuous rock was reached) with values at every 3 cm is now available for C, N, clay, silt and sand content from the direct prediction out of the spectral PLSR models and for the soil hydraulic properties from the RF models. From these values C stocks as well as available water were calculated for all available depth steps, up to maximum soil depth or 102 cm respectively, as continuous rock was reached at – 85 cm in the maize profile and at –98 cm in the savannah. Furthermore the spatial autocorrelation of these values was consequently analysed by calculating the semivariances $Y(h)$ dependent on the distance to each other (Matheron, 1963):

$$Y(h) = \frac{1}{2} \cdot \frac{1}{N(h)} \sum_{i=1}^{N(h)} (O(s_i) - O(s_i + h))^2 \tag{4.8}$$

where $N(h)$ is the number of compared point pairs per distance $h$, $O(s_i)$ is the value at the location $s_i$ and $O(s_i + h)$ is the value at the distance $h$ from the location $s_i$. Subsequently, the semivariances are averaged by distance intervals and displayed in the so called empirical variograms. The semivariance can be calculated for different directions by only considering point pairs in the specific direction for each variogram. Empirical variograms enable the analyses of autocorrelation with distance without the need of variogram models. A model would only be required if the goal is to predict at unknown locations. Empirical variograms were calculated for two directions, vertically (180) and horizontally (90), as a strong vertical trend was suspected.

Analyses were performed with `gstat` (Pebesma, 2004).

## 4.3 Results

### 4.3.1 Soil properties

The soil properties of all 21 plots, which constitute the individual *local* databases, differ between the studied land uses (Figure 4.2). Median C and N content of $local_{\mathrm{Mai}}$ were much less than those of the others, with 10.0 and 0.83 mg $g^{-1}$ for C and N, respectively. The range of C and N content was highest in the $local_{\mathrm{Hom}}$, with values between 6 and 97 mg $g^{-1}$ for C. The range of clay content of $local_{\mathrm{Cof}}$ was much smaller than that of $local_{\mathrm{Hom}}$, however, the median was higher than thatof $local_{\mathrm{Hom}}$ with 61% compared to 40% clay content. Clay content of of $local_{\mathrm{Mai}}$ was comparable to that of $local_{\mathrm{Cof}}$. Median clay content of $local_{\mathrm{Sav}}$ was very small, with values of only 27%, whereas high values of up to 81% were also observed in that land use. Sand content on the other hand, was much higher in $local_{\mathrm{Sav}}$, with median of 38% compared to 8–14% sand of the other land uses.

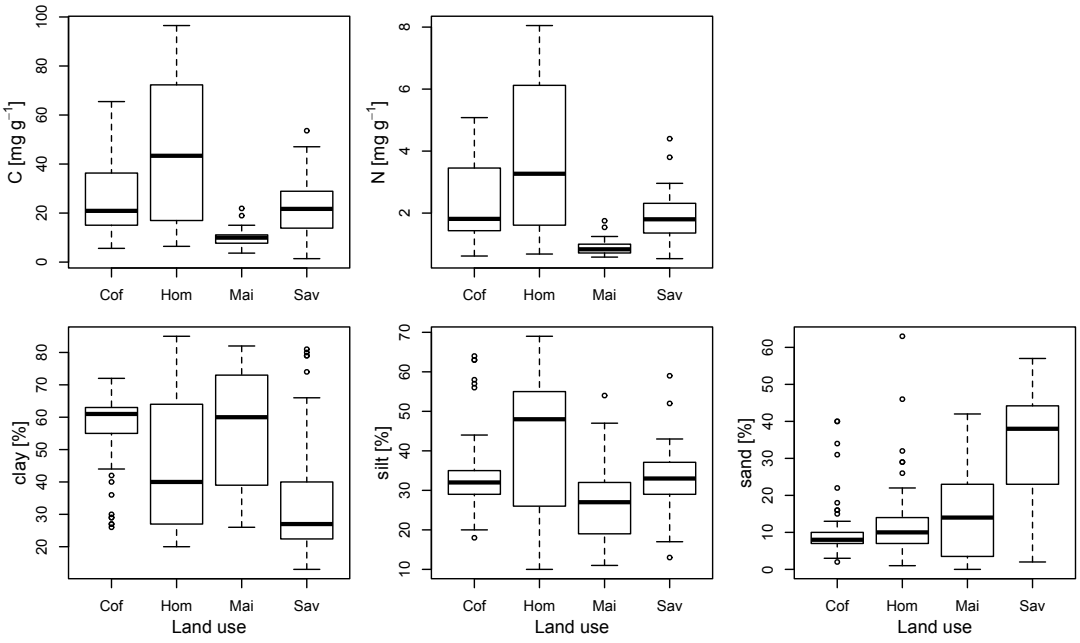



Figure 4.2: Summary statistics of the different *local* datasets, that comprise all horizons from the 21 plots, separated by land use; Cof = coffee, Hom= homegarden, Mai=maize, Sav = savannah.

### 4.3.2 Spectral models

Table 4.1: Calibration accuracy of the of the partial least squares models for the prediction of different soil parameters with visible and near infrared spectroscopy. Calibration is based on a leave-one-out cross validation, separately for each *local* dataset. The test set is constituted of the small soil core reference samples from the individual profile for each land use.

| Parameter | Plot | | Calibration | | | | Test | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | $m$ | $n$ | $RMSE$ | $R^2$ | $MPE$ | $n$ | $RMSE$ | $R^2$ | $MPE$ |
| Clay | Hom | 9 | 129 | 8.34 | 0.88 | 0.43 | 16 | 3.23 | 0.38 | 0.06 |
| | Cof | 6 | 136 | 6.92 | 0.38 | -0.11 | 13 | 1.85 | 0.75 | 0.51 |
| | Mai | 6 | 87 | 8.68 | 0.78 | 0.24 | 16 | 5.18 | -0.35 | -0.45 |
| | Sav | 5 | 95 | 7.33 | 0.92 | -0.07 | 11 | 4.50 | -1.07 | 0.20 |
| Silt | Hom | 7 | 129 | 7.84 | 0.85 | -0.20 | 16 | 2.33 | 0.46 | -0.05 |
| | Cof | 7 | 136 | 6.12 | 0.29 | 0.00 | 13 | 1.97 | 0.36 | -0.62 |
| | Mai | 8 | 87 | 4.47 | 0.54 | -0.13 | 16 | 2.76 | 0.51 | -0.01 |
| | Sav | 8 | 95 | 4.81 | 0.72 | -0.04 | 11 | 1.25 | 0.82 | 0.35 |
| Sand | Hom | 2 | 129 | 7.08 | 0.21 | -0.07 | 16 | 2.69 | 0.22 | 0.67 |
| | Cof | 1 | 136 | 5.53 | -0.02 | -0.01 | 13 | 1.43 | 0.26 | -0.48 |
| | Mai | 5 | 87 | 7.13 | 0.80 | -0.05 | 16 | 3.52 | 0.76 | 0.22 |
| | Sav | 5 | 95 | 9.64 | 0.76 | 0.06 | 11 | 4.77 | -46.10 | -0.28 |
| N | Hom | 4 | 65 | 0.73 | 0.88 | -0.01 | 8 | 0.28 | 0.93 | 0.07 |
| | Cof | 7 | 67 | 0.37 | 0.83 | 0.01 | 12 | 0.09 | 0.98 | -0.02 |
| | Mai | 1 | 44 | 0.24 | 0.14 | 0.00 | 5 | 0.17 | 0.06 | 0.00 |
| | Sav | 11 | 82 | 0.31 | 0.88 | 0.00 | 13 | 0.12 | 0.98 | -0.01 |
| C | Hom | 4 | 65 | 7.56 | 0.91 | -0.04 | 8 | 2.15 | 0.96 | 0.59 |
| | Cof | 6 | 67 | 3.96 | 0.85 | 0.07 | 12 | 0.72 | 0.98 | 0.19 |
| | Mai | 6 | 56 | 2.46 | 0.44 | -0.06 | 9 | 1.45 | 0.66 | -0.25 |
| | Sav | 5 | 82 | 5.03 | 0.80 | -0.01 | 13 | 2.77 | 0.92 | 0.73 |

$m$ = number of model parameters, $n$ = number of data points, $RMSE$ = root mean squared error (% for clay, silt and sand and g kg$^{-1}$ for C and N), $R^2$ = coefficient of determination, $MPE$ = mean prediction error, Hom = homegarden, Cof = coffee, Mai = maize, Sav =savannah

The performance of the PLSR calibration models depended on the studied soil parameter and the land use for which the model was developed (Table 4.1). C and N models were generally good, with $R^2$ up to 0.91 and $RMSE$ between 2.5 and 7.6 mg g$^{-1}$ for C and between 0.24 and 0.73 mg g$^{-1}$ for N, respectively. Even though the $R^2$ of the $local_{\mathrm{Mai}}$ model was inferior to the other models, $RMSE$ was still very low. $RMSE$ of the clay, silt and sand models were between 4.5 and 9.6%, with

generally higher $R^2$ in the clay models, but again with differences within the studied land use. Models for the prediction of sand content were less reliable. For most land uses the problem was the lack of enough samples with a high sand content, which were consequently predicted inaccurately.

$RMSE$ of prediction for the test sets of all parameters and land uses were lower than those of the calibration (Table 4.1). Prediction of C and N was very good, with the exception of N for the Mai profile. Here only a few measured values from the profile were available, as N content was often lower than the detection limit of the CN-analyser.

Although some $R^2$ of calibration from LOOCV were low, $RMSE$ of prediction for some models were still small, i.e. sand, silt and clay for Cof, with $RMSE$ between 1.43 and 1.85%. $RMSE$ was never higher than 5.18% in any land use. $R^2$ of prediction, however, differed between the systems and parameters. $R^2$ for the clay model for Mai and Sav was very low, whereas clay of Cof and Hom was predicted with higher accuracy. Prediction for silt and sand were all acceptable, except for the sand content in the Sav system, which had a very low $R^2$ and a relatively high $RMSE$. As the sand model for Sav failed, instead of using predictions for every segment, we assumed a constant sand content of 2%, as this was the measured content in all references samples.

### 4.3.3 Water retention curves

The water retention for the different horizons of the selected profiles and the water retention curves of the van-Genuchten equation (Equation 4.5) are presented in Figure 4.3. All soils had a high saturated water content ($\theta_s$). The high residual water content ($\theta_{4.2}$) of the homegarden and savannah profile was typical for clayey soils, whereas $\theta_{4.2}$ of B1 to B3 of the coffee profile was unusually high and did not match the measured soil texture. $\theta_{4.2}$ of the maize profile was lower than those of the other profiles, due to the lower clay content.

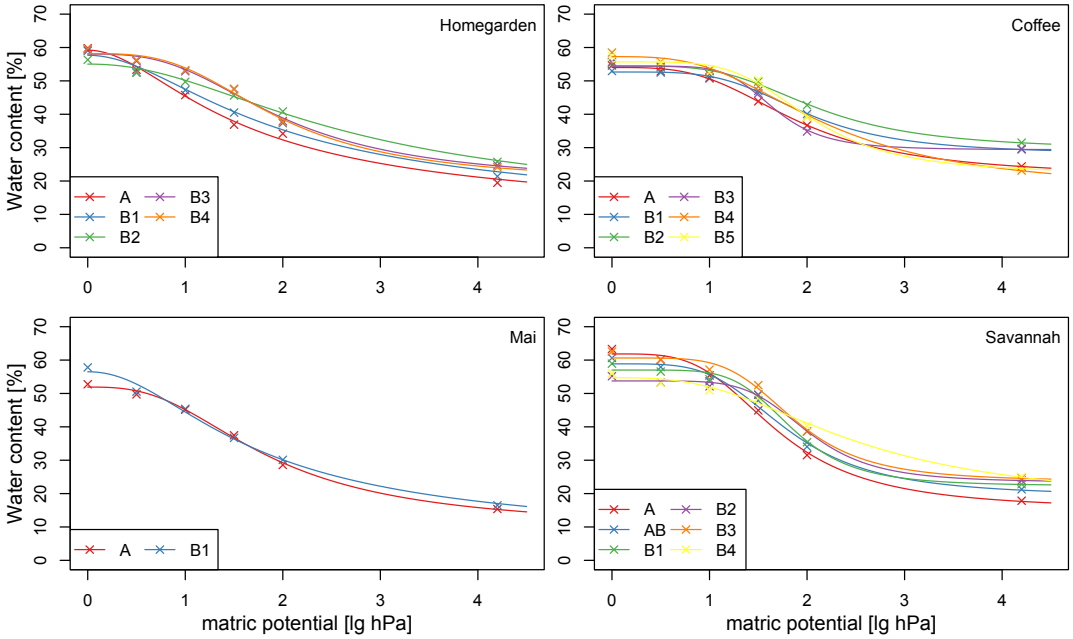



Figure 4.3: Measured water content at different matric potentials and fitted water retention curves for each horizon; each point is the mean of five measurements and the curve was fitted with the van-Genuchten model.

### 4.3.4 Pedotransfer functions

$RSQ$ for the prediction of $\theta_{1.8}$ and $\theta_{4.2}$ were 0.51 and 0.24, respectively (Table 4.2). Despite the low $RSQ$, $RMSE$ was still low with values of about 2.7 and 3.6 % water content. Predictions accuracies for $\phi$ and *FAC* were slightly lower with $RMSE$ of 4.7 and 5.7 %, respectively.

For the prediction of water content at different potentials all three selected parameters were important, with C being the least important parameter and clay being more important than sand for $\theta_s$ and $\theta_{4.2}$. In the RF models for the estimation of *AWC* and $K_r$ clay content was not considered and C content was the most important parameter. The *FAC* was best predicted with only C and clay as input parameters with C content again being most important. For the prediction of $\rho_b$ all three tested parameters were equally important.

Table 4.2: Quality criteria of the random forest models for the prediction of hydraulic properties with clay, sand and carbon content as input parameters.

| Parameter | $\theta_{1.8}$ | $\theta_{4.2}$ | *AWC* | $\phi$ | *FAC* | $K_r(1.8)$ | $\rho_b$ |
|---|---|---|---|---|---|---|---|
| | % | % | % | % | % | cm $d^{-1}$ | g $cm^{-3}$ |
| *RMSE* | 2.69 | 3.59 | 2.07 | 4.70 | 5.70 | 0.05 | 0.12 |
| *RSQ* | 0.51 | 0.24 | 0.41 | 0.44 | 0.34 | 0.20 | 0.36 |
| *cor* | 0.80 | 0.56 | 0.59 | 0.70 | 0.60 | 0.45 | 0.63 |
| *MPE* | 0.38 | 0.58 | -0.10 | -0.08 | -0.72 | 0.00 | 0.01 |
| input variables | sa, cl, C | cl, sa, C | C, sa | sa, cl, C | C, cl | C, sand | C, cl, sa |

$\theta$ = water content, *AWC* = available water capacity, $\phi$ = porosity, *FAC* = field air capacity $K_r$ = relative hydraulic conductivity, $\rho_b$ = bulk density, cor = Pearson's correlation coefficient; input variables are in order of importance, cl = clay content, sa = sand content, C =carbon content

### 4.3.5 Variability of soil properties within the profile

#### Carbon and nitrogen

The decrease of C with depth is clearly visible in all profiles (Figure 4.4 and Figure 4.5). The sharp border, that is visible in the homegarden profile seems to be an error, that occurred during the scanning process. In the coffee profile C content remained stable after a depth of about –10 cm, with higher variations within the same depth. Independent of depth, some segments expressed a much higher C content than those around it, especially in the homegarden and in the savannah profile. Highest C content values were found for some segments of the homegarden profile, which also expressed the highest C contents down to a depth of more than –60 cm. C content of the upper horizons in the savannah profile was also very high with some values more than 50 mg $g^{-1}$, after a rapid decrease, however, C content remains low at about 10 mg $g^{-1}$. The maize profile expressed the lowest C values of all analysed profiles, with only some segments showing C contents of about 20 mg $g^{-1}$.

Pictures of N content show a similar pattern, except for the savannah profile (Figure 4.4 and Figure 4.5). Here, the model for the prediction of N was probably not suitable, as the empty segments correspond to negative predictions that were excluded before drawing the picture. Again the homegarden profile expressed the highest and the maize profile the lowest N values.

Prediction of C content in the homegarden profile match measured values only in an area between –20 and –40 cm depth and again down from about –70 cm (Figure 4.5). In the coffee profile predicted values were lower than measured

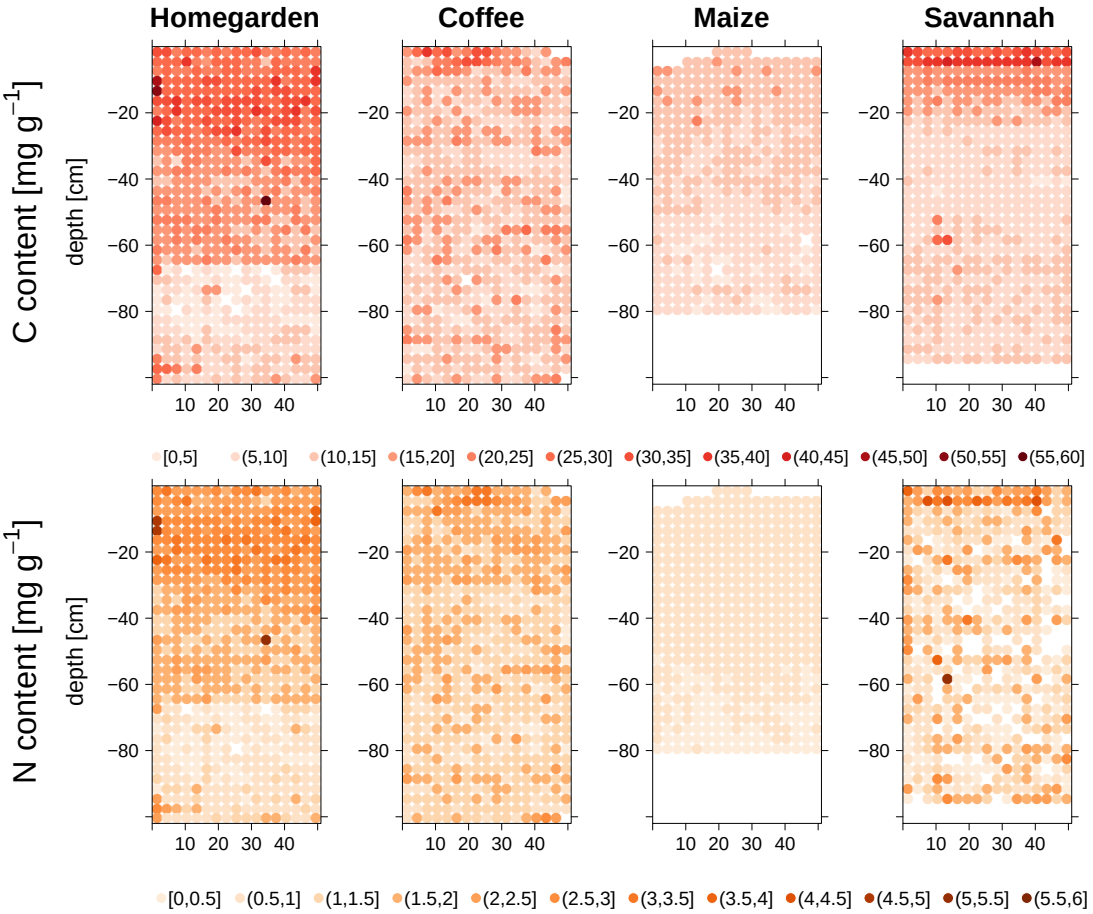



Figure 4.4: Distribution of carbon (C) and nitrogen (N) content (mg g$^{-1}$) in the different profiles.

values in the top of the profile and agreed quite well from –40 to –100 cm depth. Predictions of C content in the maize profile matched the small soil core samples well between –5 and –25 cm; for the remaining parts predictions differed from measurements. Predictions of the savannah profile agreed very well for all parts of the profile, both with measured samples from horizons and small soil cores. Measured values from the small soil cores and the bulk horizon samples for C were usually similar, except for the A-horizon of the maize profile. Those for N differed more, probably because of difficulties in N measurements.

Predictions of N content in the maize profile did not agree with measured samples in any part of the profile, probably because of a poor N model for the maize profile due to a low amount of samples. Median predictions for N in the savannah profile agreed quite well to measured value. The interquartile range, however, was quite high and the median expresses big differences within short depths, which could be an indication of poor predictions.

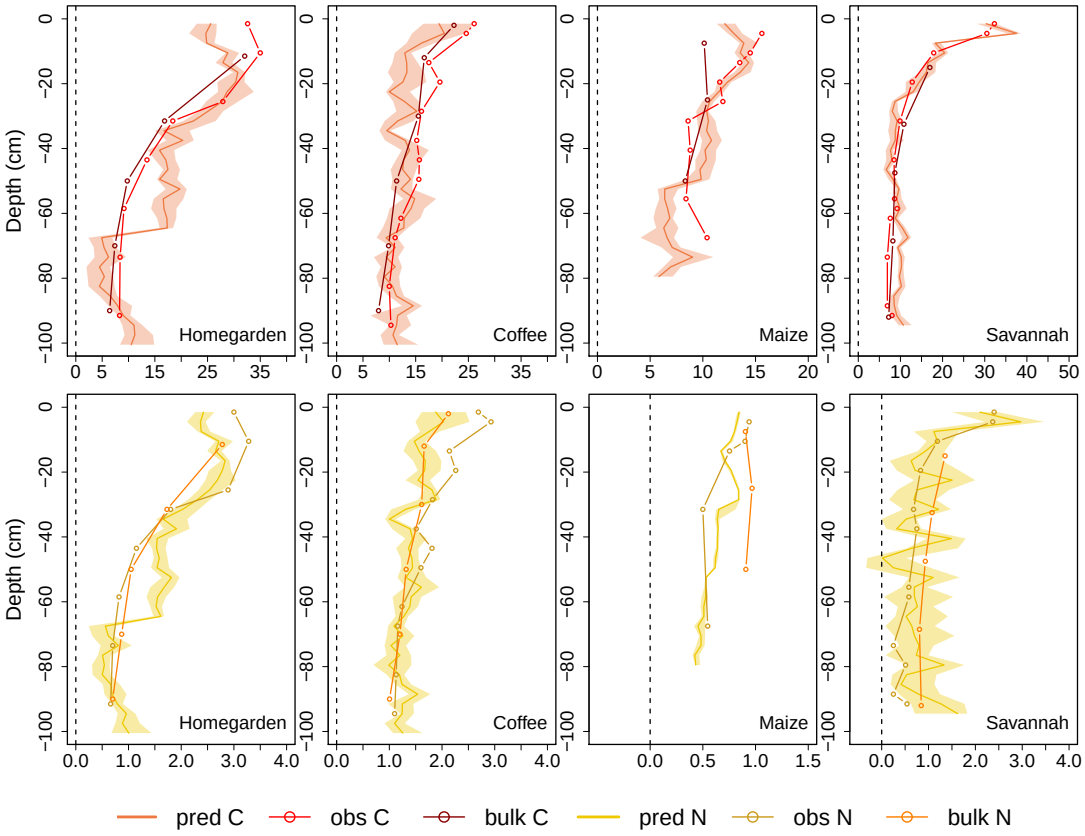



Figure 4.5: Carbon (C) and nitrogen (N) content (mg $g^{-1}$) in the different profiles; pred = predicted, obs = observed, bulk = bulk horizon samples; obs values are from the small soil core reference samples, that were taken directly after scanning, coloured areas indicate the interquartile range (25%–75%) of predictions and the plain line the median.

### Soil physical properties

Median predictions of particle size distribution with depth are displayed in Figure 4.6. Although some of the models did exhibit low $R^2$ values, overall predictions of the soil textural classes (clay, silt and sand) agreed very well to the measured contents. Only at locations where measured values were lacking (first two segments of the savannah profile and lower segments of the maize profile), predictions were off. Measured values of the bulk soil samples per horizon are well in line with values of the small segments for coffee and savannah, in the homegarden and maize profiles differences could be observed.

For some segments of the homegarden profile clay content of more than 100% was predicted, these segments were thus excluded (Figure 4.7). Consequently some of the high predictions for clay between 90 and 100% might be prediction errors as well. Nevertheless, the homegarden profile was characterised by an increase of clay with depth and a generally high clay content of about 80%. Furthermore, the

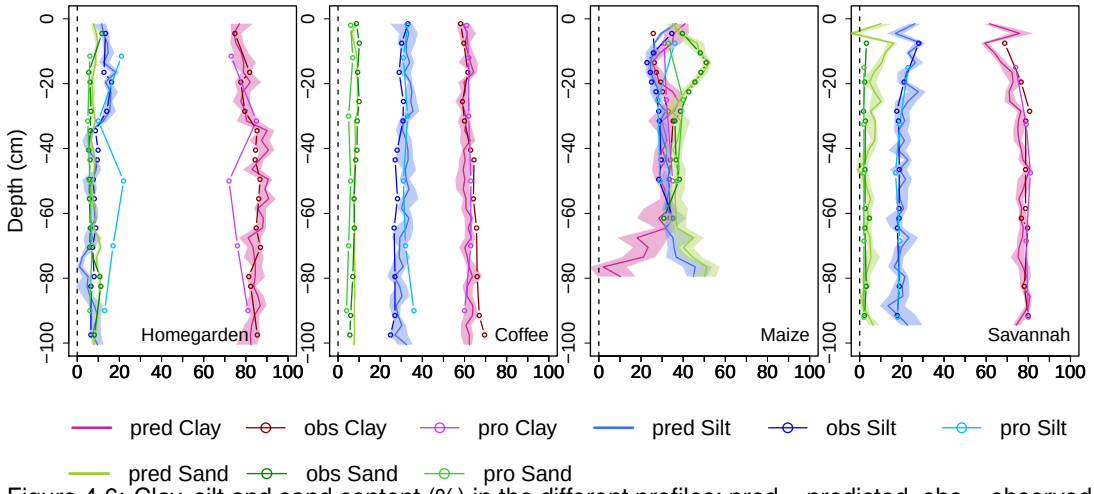



Figure 4.6: Clay, silt and sand content (%) in the different profiles; pred = predicted, obs = observed, pro = observed horizon samples, coloured areas indicate the interquartile range (25%–75%) and the plain line the median.

semivariances were quite high in this profile and increased with vertical distance. Silt content of the homegarden decreases accordingly, as sand content remained constantly low throughout the whole profile. Here some values were excluded, as silt content values below 0% were predicted.

In the coffee profile, clay, silt and sand content remained constant with depth. This even distribution within the whole profile was also confirmed by the semivariograms, which did not increase with distance.

Sand content of the maize profile was generally high throughout the profile with some segments showing a sand content of up to 65%. Sand content of the maize profile first increased to a depth of about –15 cm up to 50%, decreased again down to –30 cm and then remained stable with about 40%. The second increase below –65 cm could not be validated, as small soil core samples were missing. This increase seems still reasonable, as a second Cv horizon adjoined at –65 cm.

Some unreasonable values were predicted for sand content of the savannah, as the model for the prediction of sand failed for that profile. The unusual peak of high clay and low sand and silt values in the savannah profile at –6 cm depth (Figure 4.6) is probably a measurement artefact and no property of the soil profile. From 9 cm on, clay content was first increasing with depth and then remained stable around 80% clay content.

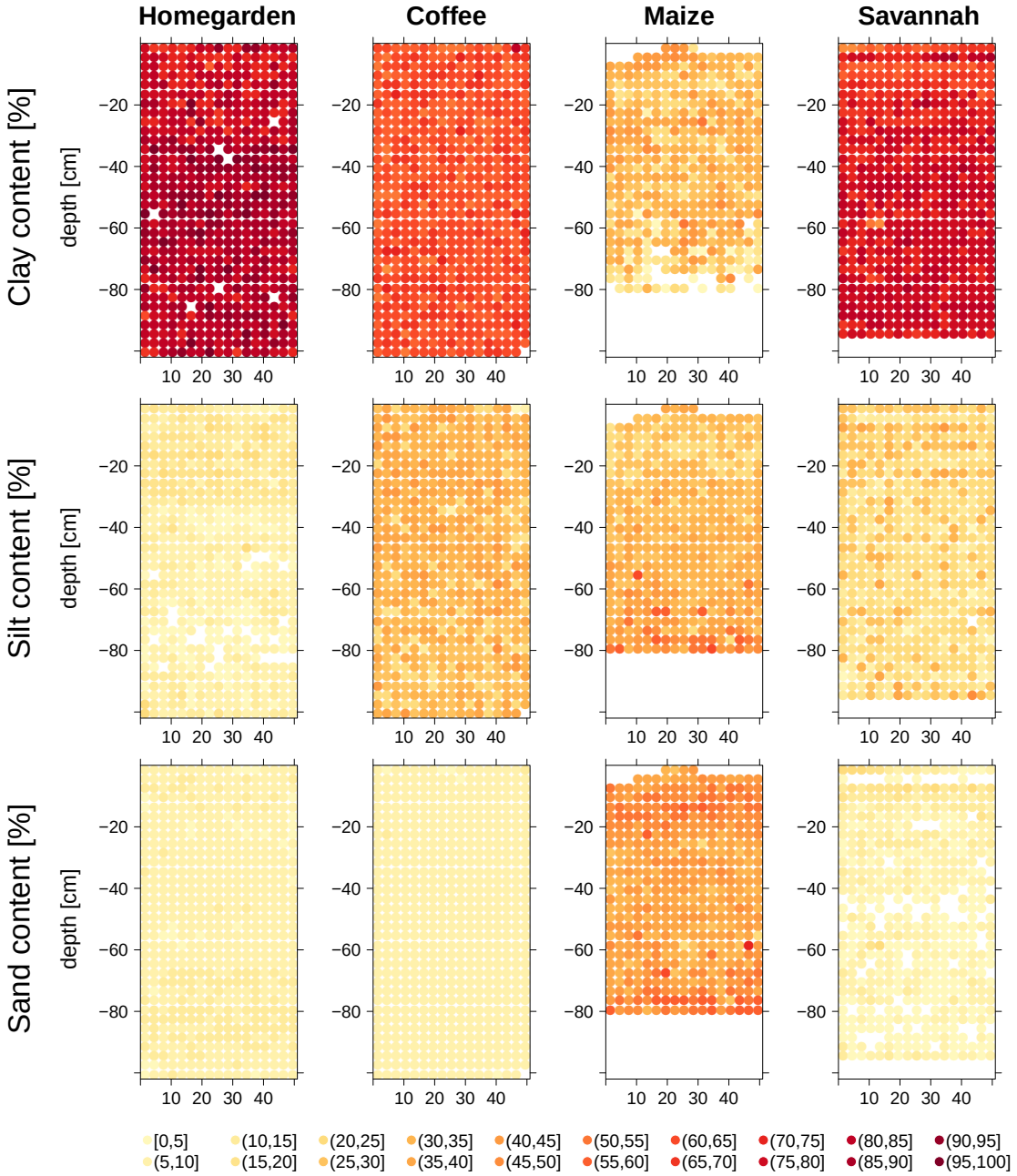



Figure 4.7: Distribution of clay, silt and sand content (%) in the different profiles; Hom = homgarden, Cof = coffee, Mai = maize, Sav = savannah; empty spaces within the measured soil depth mark points outside the prediction the range of soil texture (0 - 100 %).

$\rho_b$ of the homegarden was much lower than those of the other profiles, with the maize profile expressing highest $\rho_b$ values throughout (Figure 4.8). $\rho_b$ increased with depth for homegarden and savannah, whereas it was quite constant in the maize and and coffee profile. This is also visible in the semivariograms. The semivariance remained constant for coffee and maize and and increased for homegarden and savannah in the vertical direction (Appendix C). Semivariances in the savannah profile are higher than those of the other profiles.

The distributions of field air capacity *FAC* and $\phi$ were quite similar, with highest values in the upper part of the homegarden profile and in the first centimetres of the savannah profile (Figure 4.9). In the homegarden it decreased slowly with depth,

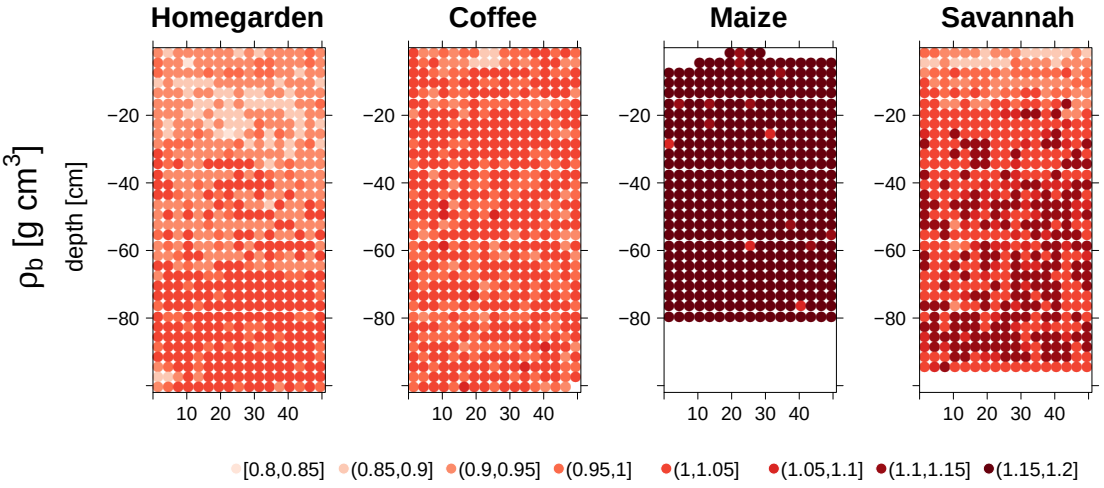



Figure 4.8: Distribution of bulk density ($\rho_b$) in the different profiles.

whereas the decrease in the savannah profile was more pronounced. *FAC* and $\phi$ are both lowest in the maize profile with values around only 45% $\phi$ and 10% *FAC*. Some segments showed distinctly higher values, especially in the upper part of the maize profile. The overall semivariance in the maize profile is however negligible, whereas semivariances of homegarden, coffee and savannah were relatively high with values between 2 and 4 $(\%)^2$ (Appendix C).

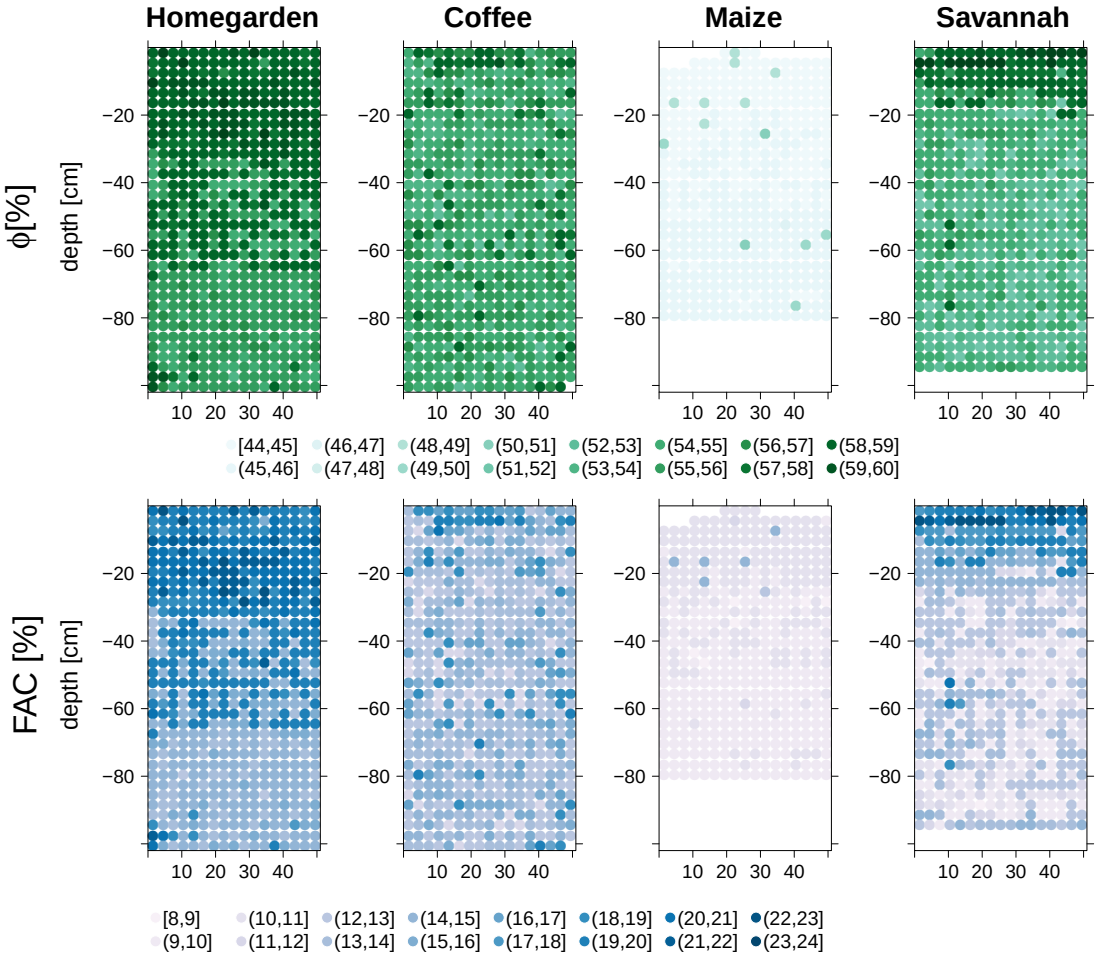



Figure 4.9: Porosity ($\phi$) and field air capacity (*FAC* = $\phi$ - $\theta_{1.8}$) in the different profiles.

### Soil hydraulic properties

$\theta_{1.8}$ and $\theta_{4.2}$ of homegarden, coffee and savannah was higher than at the maize profile, with a slight increase with depth at the homegarden and savannah profile (Figure 4.10). *AWC* of the subsoil in the savannah and maize profile was much higher than in the upper parts of the profiles. At the coffee and homegarden *AWC* was distributed evenly with depth, with some lower values in the middle part of the homegarden profile. Generally semivariances of $\theta_{1.8}$, $\theta_{4.2}$ and *AWC* were small with absolute values between 0.3 and 2.5% water content.

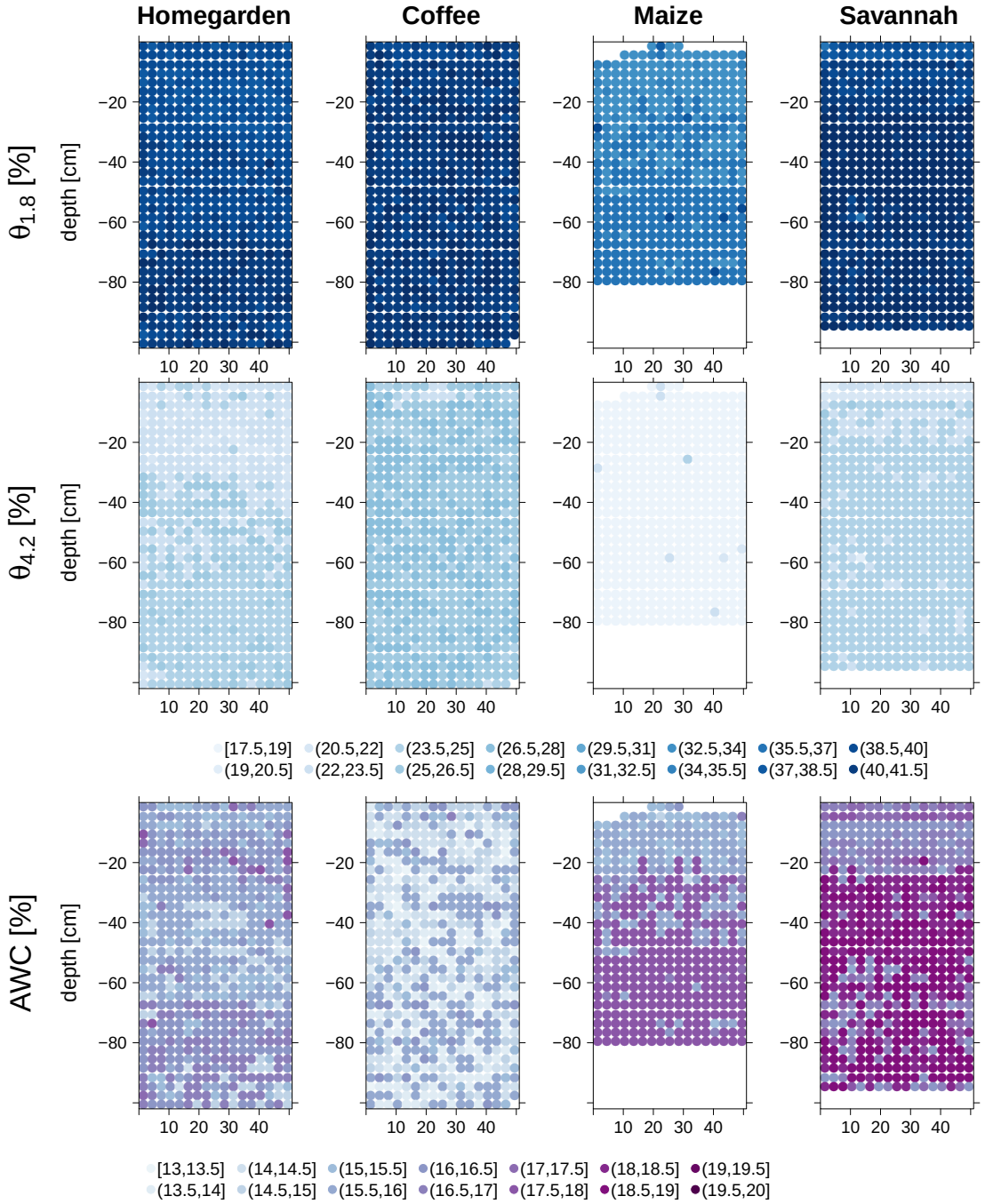



Figure 4.10: Distribution of water content at matric potentials of $10^{1.8}$ and $10^{4.2}$ hPa ($\theta_{1.8}$ and $\theta_{4.2}$) and available water capacity (*AWC* = $\theta_{1.8}$ - $\theta_{4.2}$) in the different profiles.

Predicted and observed values of *AWC* and $\theta_{4.2}$ agreed well in the homegarden profile, especially in the lower parts of the profiles (Figure 4.11). At the coffee profile predictions of $\theta_{1.8}$ and $\theta_{4.2}$ were lower than observed values in the upper part of the profile. *AWC* was predicted quite well for that part. Here the observed value of the B3 horizon (–50 cm) was probably measured incorrectly. At the maize profile predicted values of $\theta_{4.2}$ were only slightly higher than the observed values, and the predicted *AWC* agreed well with observed values. However, measured values were only available for the first two horizons and no test value was available for the lower part of the profile. Predictions of all values were good in the lower parts of the savannah profile, only at the upper part predictions of $\theta_{4.2}$ were higher than observed values and consequently predictions of *AWC* were lower.

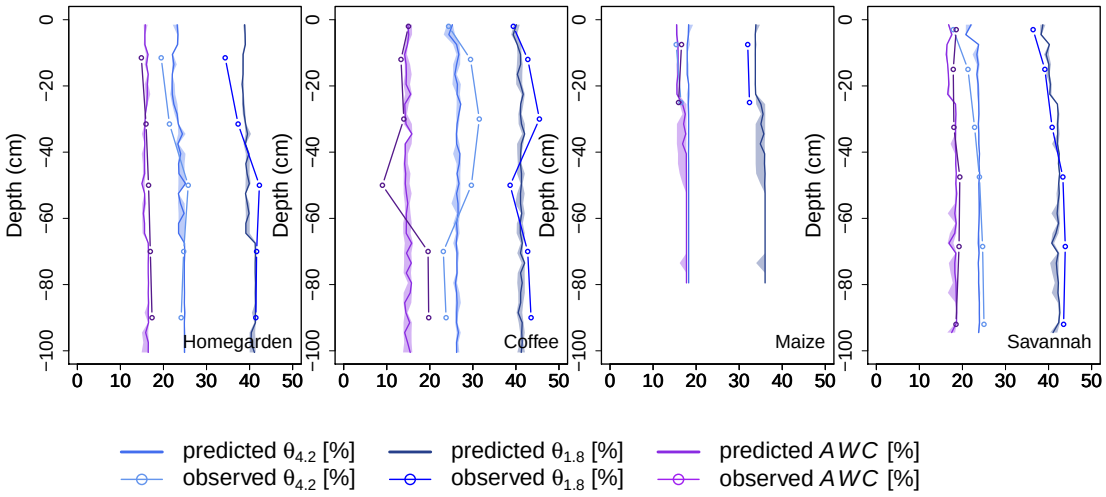



Figure 4.11: Water content at a potential of $10^{1.8}$ and $10^{4.2}$ hPa ($\theta_{1.8}$ and $\theta_{4.2}$ (%)) and available water capacity (*AWC* = $\theta_{1.8}$ - $\theta_{4.2}$) in the different profiles; coloured areas indicate the interquartile range (25%–75%) and the plain line the median.

The spatial pictures of $K_r(1.8)$ in the homegarden profile was somewhat similar to that of $\rho_b$, with lower values in the top and higher values at the bottom of the soil profile (Figure 4.12). In the coffee profile $K_r(1.8)$ varied between 0.049 and 0.111 with no differentiation with depth. $K_r(1.8)$ in the maize and savannah profiles was much lower in the upper part than at the bottom. The savannah profile expressed highest values in the middle part of the profile with maximal $K_r(1.8)$ of 0.172 cm $d^{-1}$. Semivariances of $K_r(1.8)$ in the savannah were much higher than those of the other profiles.

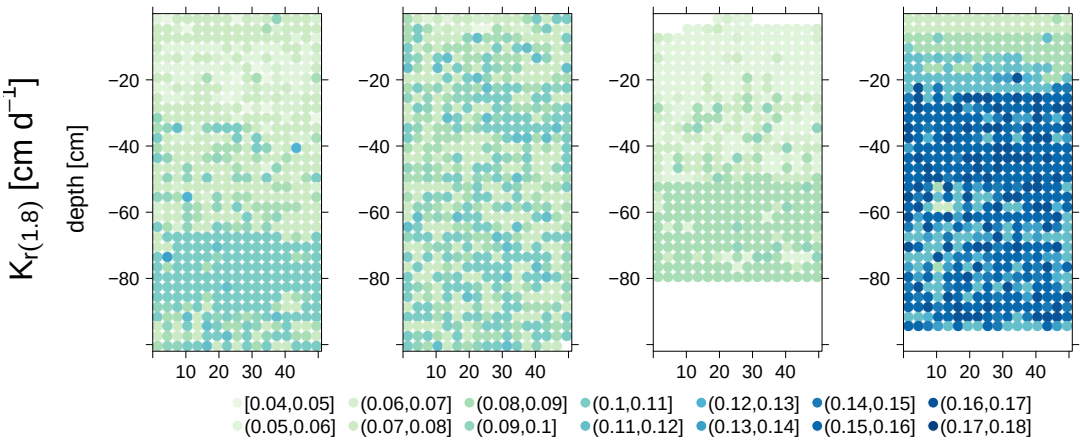



Figure 4.12: Distribution of the unsaturated hydraulic condictivity at a potential of $10^{1.8}$ hPa ($K_r(1.8)$).

### 4.3.6 Carbon and water stocks

As expected C stocks of the homegarden and coffee profile were higher than those of maize and savannah (Figure 4.13). C stocks of the homegarden were highest with values of 16.8 kg $m^{-2}$. A high amount of available water was present in this profile ($\sim$160mm). C stocks of the coffee profile on the other hand were much lower than those of the homegarden, whereas available water was similar at around 150 mm. The savannah profile contained about 1.3 times more C than the maize profile. *AWC* in the savannah profile was really high with values of up to 20%, resulting in overall highest amount of available water. Low C content and shallow soil depth of only about –85 cm resulted in lowest C stocks in the maize profile compared to all studied land uses. Even though *AWC* in the bottom of the maize profile was quite high, total available water was lower compared to all other land uses, because of the shallow soil.

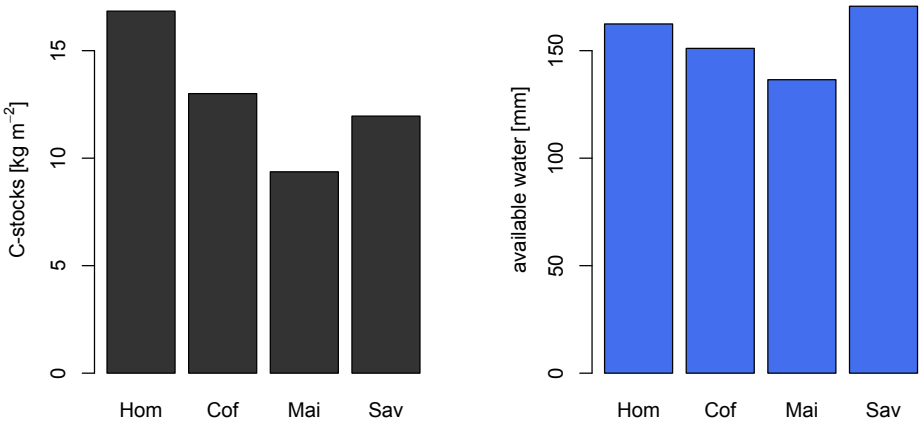



Figure 4.13: Carbon (C) stocks (kg $m^{-2}$) and available water (mm) of maximum soil depth/102cm respectively; Cof = coffee, Hom = homegarden, Mai = maize, Sav =savannah.

## 4.4 Discussion

### 4.4.1 Prediction accuracies

We developed PLSR models for the prediction of various soil parameters from in-situ soil spectra. Prediction accuracies, however, varied substantially between the parameters and land uses. In detail, the RMSE values of the individual *local* models for clay content were comparable to other studies, that were working on air-dried spectra (Sankey et al., 2008; Wetterlind et al., 2010b) and better than models with in-situ scans (Waiser et al., 2007; Viscarra Rossel et al., 2009). To our knowledge, no models from in-situ soil spectra for the prediction of silt and sand have been tested so far, and our models for silt are comparable to those from air-dried spectra (Shepherd et al., 2002; Viscarra Rossel et al., 2006b; Wetterlind et al., 2010a). Bogner et al. (2014) already demonstrated the potential of this approach for the in-situ prediction of C content; in this study we could show that it is extendible to the soil parameters clay, silt and N content. The prediction accuracies for sand content were less satisfying. There were great variations between the four *local* models. Models that predominantly contained samples with a low amount of sand (between 0 and 20%) performed worse.

Models for the prediction of C and N content generally performed well. The inferior $R^2$ of calibration of the $local_{\mathrm{Mai}}$ models for the prediction of C and N can be explained by the low range of values within that database, which affects the $R^2$, without giving information on the absolute accuracy of predictions. When comparing models that were developed with different databases for the same parameter, the $RMSE$ might thus be more applicable.

In spite of the low $R^2$ of the test set of some models, median predictions of all segments for most parameters and sites were very close to the measured values. Thus, spectra that are generated by SMOTE and then incorporated into the database seem to represent the in-situ conditions for clay, silt, C and N content quite well.

The RF models for the prediction of additional soil physical and hydraulic parameters performed well. Our models had lower $RMSE$ values than those of other authors, who developed more general functions (Tomasella et al., 2004; Minasny et al., 1999; Minasny et al., 2011a). As some of the hydraulic properties did not vary much in the calibration dataset, some low $RSQ$ were observed. The Pearson's correlation coefficient between observed and predicted variable on the other hand is much higher than the very conservative measure of $RSQ$.

Although only a limited amount of samples was available, satisfactory predictions for most points in the profiles were achieved. The differing values between observed and predicted $\theta$ at the coffee profile could be a measurement error of $\theta$ and not a prediction failure, as unusually high values were measured for the AB, B2 and

B3 horizon of the coffee profile for $\theta_{4.2}$. Soils with the particle size distribution as observed in the coffee profile (around 60% clay content) can usually hold less water than measured in this study (Hodnett et al., 2002; Tomasella et al., 2004; Minasny et al., 2011a).

The relation between the basic soil parameters clay, sand and C content and soil hydraulic properties has been shown by many authors and is used in most pedotransfer functions (Berg et al., 1997; Minasny et al., 2011a). As expected most RF included all of these three variables. Clay was not included in the prediction of *AWC*, which is probably due to the small dataset and not representative for the soils.

Considering this modelling approach, where soil hydraulic properties are predicted in two steps, the agreement between measured and predicted values in the soil profiles is surprisingly high. Even though no independent test set was available, we could show that median predictions of all segments in the profiles agreed well with measured values for different water contents. The direct approach to predict soil hydraulic properties from spectra, as suggested by Santra et al. (2009) would have been difficult in our case. Firstly because the in-situ spectra and the air-dried calibration spectra differed substantially. Secondly, a prediction of hydraulic properties from soil spectra is indirect (Janik et al., 2007) and there would be too much unknown variation in the spectra. Thirdly, it is more laborious to create a sufficiently large spectral database for soil hydraulic properties than for the basic soil parameters. RF models on the other hand performed well with just a minimum amount of samples.

### 4.4.2 Comparison of land uses

#### Soil physical properties

Soil physical properties varied between the land uses. Three of the four analysed soil profiles were classified as Vertisol. These soil types usually have a high internal turnover of soil and thus more or less uniform morphological characteristics throughout the profile (IUSS Working Group WRB, 2007). With the exception of the A-horizon, horizon boundaries were indeed hard to differentiate in most profiles. In the soil of the coffee profile all soil physical properties are distributed evenly, which is also supported by the low semivariances both horizontally and vertically. On Mt. Kilimanjaro, when establishing a new rotation of coffee trees, soil is tilled mechanically to provide better conditions for root growth. On the investigated field site, drip irrigation was installed, which provided water supply within the coffee rows, but not in between. During the dry season, vertical cracks typical for Vertisols, might form and contribute to the uniform structure of the soil.

Soils of the homegarden and the savannah show more structured features. The lower $\rho_b$ of the upper horizons can be attributed to higher amounts of C. Furthermore there is probably less soil compaction, as these two sites are not passed over with machines. Nevertheless, $\rho_b$ values were comparatively small in all three Vertisols. Other studies on Vertisols reported $\rho_b$ values between 1.25 and 1.54 for B-horizons (Hati et al., 2007; Dinka et al., 2013), which is much higher than at all our investigated sites. As the soils of Mt. Kilimanjaro developed on volcanic material, volcanic minerals, such as allophanes are still present in the soil and might be responsible to these reduced bulk densities. This is especially visible in the soil of the homegarden. *FAC* and $\phi$ were negatively correlated to $\rho_b$ at all sites and thus variable and high in the homegarden profile and in the upper part of savannah. If a soil is better structured, with different characteristics at small scales, thus expresses a high heterogeneity, it might provide more and diverse habitats for microbial communities (Ettema et al., 2002; Young et al., 2004). Indeed, microbial biomass is much higher in the homegarden and savannah soil compared to maize field and coffee plantation (Pabst et al., 2013).

The soil of the maize profile is characterised by a high sand content and a high $\rho_b$ and classified as Vitric Cambisol. Its vitric material (primary volcanic minerals) dominates the sand fraction of the soil. As the physical behaviour of vitric material in the soil depends on its type and the degree of weathering (Maeda et al., 1977; Arnalds, 2011), high $\rho_b$ values are reasonable. Together with the low soil depth (Cv-horizon starting at –35 cm) this might be a restraint for plant growth, depending on the capability and requirements of the roots.

#### Soil hydraulic properties

At the measured centimeter scale water content at different pF levels as well as the *AWC* did not change considerably regarding the horizontal direction and only little with depth. To our knowledge, no information is available how these properties vary at the profile scale, as the scales analysed in other studies were much greater. Mallants et al. (1996) analysed the spatial variability of soil hydraulic properties in a soil transect of 31 m, with smallest sampling intervals of 0.1 m. They found semivariances about 10 times higher at a distance of 0.1 m than we did, indicating at a more heterogeneous site.

The differences between the studied land uses on the other hand were more pronounced. Field capacity of savannah and coffee were similar. However, coffee contained a high amount of residual water, which resulted in relatively low *AWC*. Given the particle size distribution, less water should be bound at that potential.

The residual water content in the topsoil of the homegarden was relatively low, despite of the high clay content. Thus, a relatively high amount of water is available

for the plants. In addition to the particle size distribution, C content and $\rho_b$ are often used in pedotransfer functions to estimate *AWC* or the respective water contents for its calculation (Berg et al., 1997; Tomasella et al., 2004; Børgesen et al., 2005). These two parameters could have a positive influence on *AWC* in the topsoil. However, there is no simple correlation in our case and as *AWC* is also high in the subsoil other factors are probably additionally responsible for the high *AWC*.

*AWC* of maize and savannah were high, especially in the subsoil. Due to its shallow depth, total available water of the maize is less than that of savannah. Furthermore, the soil of the maize contained much less water at both potentials than the soil of the savannah, due to higher drainage. Even though the calculated *AWC* is similar in these two profiles, implications might still differ. The drainage characteristics of the maize lead to faster de-watering of that profile, which might get problematic if the rainfall pattern is irregular during the growing season. In the savannah soil de-watering is reduced, at a potential of $10^{4.2}$ hPa water is still present in the profile, even though it is bound more strongly to the particles.

The concept of available water content is rather arbitrary. It is assumed, that plants can not use the water that is stored in the soil at a potential of $10^{4.2}$ hPa, as this is assumed to correspond to the soil water content at which sunflower leaves wilt irreversibly (Batjes, 1996; *Soil survey manual*). This has not been proven for all plants and savannah species might be able to use part of this water. Furthermore, the drainage behaviour of a soil horizon is, in addition to its own soil properties, also dependent on the underlying soil properties and/or properties of the bedrock (Hillel, 1998).

Comparisons to other studies are difficult, as not all studies use the same potentials for the calculation of *AWC*. We calculated with $10^{1.8}$ hPa, but potential of $10^{2}$ hPa or $10^{2.5}$ hPa are also sometimes used (Berg et al., 1997). Batjes (1996) estimated *AWC* and the total available water for different soil types of the world. Vertisols in his study contained 130 mm available water and coarse textured Cambisols 115 mm, which is less than the amount contained in our soils. However, as the *AWC* was defined as *AWC* = $\theta_{2.5}$ - $\theta_{4.2}$ in his study, it is less than for our soils per definition.

#### Carbon and water stocks

Soil C stocks of the plots in the colline zone were lower than those in the submontane zone. This can partly be explained by the higher elevation and thus lower turnover of the organic matter in the submontane zone (Zech et al., 2011). Furthermore, the soils of the submontane zone contain more short-range minerals, such as allophanes, which occur in soils on volcanic substrates (Zech et al., 2014).

The amount of above ground biomass and thus C stocks of the savannah and the maize plots was similar. However, most biomass of the maize plot is concentrated in the herbaceous layer, that is harvested every year (Ensslin et al., 2014). This C loss is clearly reflected in the below ground C stocks. The reduced C stocks of the maize plots is thus probably a consequence of the agricultural intensification.

Comparing the two plots of the submontane zone, below ground C stocks of the coffee plantation were considerably reduced to those of the homegarden. This is in line with other studies that compared agroforestry systems to monocultures (Vagen et al., 2005; Luedeling et al., 2011; Hergoualc'h et al., 2012). Ensslin et al. (2014), who analysed the aboveground biomass on these plots found C stocks of about 40 g $cm^{-2}$ in homegarden compared to about 15 g $cm^{-2}$ on the coffee site. They attributed this difference largely to a reduction of trees. The higher abundance of trees, together with higher biomass in the shrub layer, might be responsible for the conversion and/or higher input of the below-ground biomass. In the homegarden of our study, the shrub layer is mostly composed of banana plants (Ensslin et al., 2014). Whenever the offshoots are planted, small soil pits are excavated and old banana litter is added to provide fast available nutrients for the new banana plant. This helps to explains the occurrence of the deep A-horizon of more than 20 cm, which could not solely be explained by manual tillage. The large horizontal semivariance in C content could also be explained by this cultivation procedure, although spatial analyses on larger scales would be necessary to support this concept.

## 4.5 Conclusion

In this study we provide an approach to gain information about many soil parameters on the intact soil. The characteristics of the soils under the studied land uses differ and the detailed spatial analysis revealed interesting pattern. In the homegarden and savannah, spatial semivariance of most parameters and thus the soil heterogeneity was high. The coffee and the maize soil were homogeneous with low semivariances and no visible trend for most parameters in the coffee profile and small semivariances in the maize profile and thus an overall lower heterogeneity. As spatial heterogeneity can promote soil biodiversity (Ettema et al., 2002), the physical and chemical spatial structure of soils is of great importance. To differentiate better between pedogenetic and land use effects on soil physical and soil hydraulic properties, however, the sampling design ought to be adjusted.

In order to understand the soil as it is, i.e. the undisturbed soil in the field and to asses its specific characteristics, in-situ measurements are essential. Combining Vis-NIR-DRS and pedotransfer functions permitted the in-situ estimation of soil physical and hydraulic properties with a high spatial resolution.

The results presented in this study demonstrate the potential of the combination of Vis-NIR DRS with pedotransfer functions. The fast assessment of various soil parameters at the same time with just one in-situ spectra can facilitate data acquirement. These data can then be used in digital soil mapping and modelling or in landscape simulations. Further research is needed if this approach can be employed at larger scales or combined with hyper-spectral images. Following the approach by ICRAF-ISRIC (2010), who made the spectral database publicly available, databases on measured water contents and the corresponding basic soil parameters should be published also. Regression equations alone can not provide the same service, as new data can not be incorporated. Some measured additional input points might be enough to improve pedotransfer functions for a new area. The user himself could then decide which modelling approach fits best to his needs and could also use non-linear functions, such as random forests.

## Acknowledgements

This study was funded by the German Research Foundation (DFG) within the Research Unit 1246 (KiLi) and supported by the Tanzanian Commission for Science and Technology (COSTECH) and the Tanzania Wildlife Research Institute (TAWIRI). We thank especially thank Johannes Hepp for discussion about and assistance with field design and measurements. Furthermore we thank Holger Pabst, Hannes Thomasch, David Kienle, Johannes Kühnel and Jumanne Mwinyi for help during the field campaigns and in the laboratory and additionally Holger Pabst and Yakov Kuzyakov for background data analysis.

## References

Arnalds, O. (2011). *Andosols*. Tech. rep. Springer-Verlag Berlin Heidelberg, 2008: Earth Sciences Series. Encyclopedia of Soil Science: SpringerReference.

Awiti, A. O., M. G. Walsh, K. D. Shepherd, and J. Kinyamario (Jan. 2008). “Soil condition classification using infrared spectroscopy: A proposition for assessment of soil condition along a tropical forest-cropland chronosequence”. In: *Geoderma* 143.1-2, pp. 73–84. DOI: 10.1016/j.geoderma.2007.08.021.

Batjes, N. (May 1996). “Development of a world data set of soil water retention properties using pedotransfer rules”. In: *Geoderma* 71.1-2, pp. 31–52. DOI: 10.1016/0016-7061(95)00089-5.

Berg, M. van den, E. Klamt, L. P. van Reeuwijk, and W. G. Sombroek (1997). “Pedotransfer functions for the estimation of moisture retention characteristics of Ferralsols and related soils”. In: *Geoderma* 78.3-4, pp. 161–180. DOI: 10.1016/S0016-7061(97)00045-1.

Bogner, C., A. Kühnel, and B. Huwe (2014). "Predicting with limited data – Increasing the accuracy in VIS-NIR diffuse reflectance spectroscopy by SMOTE". In: *Proceedings of the 6th Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing*. accepted. Lausanne, Switzerland.

Breiman, L. (2001). "Random Forests". English. In: *Machine Learning* 45.1, pp. 5–32. DOI: 10.1023/A:1010933404324.

Brown, D. J. (2007). "Using a global VNIR soil-spectral library for local soil characterization and landscape modeling in a 2nd-order Uganda watershed". In: *Geoderma* 140.4, pp. 444–453. DOI: 10.1016/j.geoderma.2007.04.021.

Børgesen, C. D. and M. G. Schaap (2005). "Point and parameter pedotransfer functions for water retention predictions for Danish soils". In: *Geoderma* 127.1-2, pp. 154–167. DOI: 10.1016/j.geoderma.2004.11.025.

Chawla, N. V., K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer (2002). "SMOTE: Synthetic Minority Over-sampling Technique". In: *Journal of Artificial Intelligence Research* 16, pp. 321–357.

Christensen, B. T. (2001). "Physical fractionation of soil and structural and functional complexity in organic matter turnover". In: *European Journal of Soil Science* 52.3, pp. 345–353. DOI: 10.1046/j.1365-2389.2001.00417.x.

Dinka, T. M., C. L. Morgan, K. J. McInnes, A. S. Kishné, and R. Daren Harmel (Jan. 2013). "Shrink-swell behavior of soil across a Vertisol catena". In: *Journal of Hydrology* 476.0, pp. 352–359. DOI: 10.1016/j.jhydrol.2012.11.002.

Ensslin, A., G. Rutten, U. Pommer, R. Zimmermann, A. Hemp, and M. Fischer (2014). "Biomass of trees, shrubs and herbs for important natural and human influenced habitat types at Mount Kilimanjaro". In: submitted.

Fernandes, E. C. M., A. Oktingati, and J. Maghembe (1985). "The Chagga homegardens: a multistoried agroforestry cropping system on Mt. Kilimanjaro (Northern Tanzania)". In: *Agroforestry Systems* 2 (2), pp. 73–86. DOI: 10.1007/BF00131267.

Hati, K. M., A. Swarup, A. Dwivedi, A. Misra, and K. Bandyopadhyay (Feb. 2007). "Changes in soil physical properties and organic carbon status at the topsoil horizon of a vertisol of central India after 28 years of continuous cropping, fertilization and manuring". In: *Agriculture, Ecosystems & Environment* 119.1-2, pp. 127–134. DOI: 10.1016/j.agee.2006.06.017.

Hergoualc'h, K., E. Blanchart, U. Skiba, C. Hénault, and J.-M. Harmand (Feb. 2012). "Changes in carbon stock and greenhouse gas balance in a coffee (*Coffea arabica*) monoculture versus an agroforestry system with Inga densiflora, in Costa Rica". In: *Agriculture, Ecosystems & Environment* 148.0, pp. 102–110. DOI: 10.1016/j.agee.2011.11.018.

Heung, B., C. E. Bulmer, and M. G. Schmidt (2014). "Predictive soil parent material mapping at a regional-scale: A Random Forest approach". In: *Geoderma* 214-215.0, pp. 141–154. DOI: 10.1016/j.geoderma.2013.09.016.

Hillel, D. (1998). *Environmental soil physics: Fundamentals, applications, and environmental considerations*. Academic press.

Hodnett, M. and J. Tomasella (Aug. 2002). "Marked differences between van Genuchten soil water-retention parameters for temperate and tropical soils: a new water-retention pedo-transfer functions developed for tropical soils". In: *Geoderma* 108.3-4, pp. 155–180. DOI: 10.1016/S0016-7061(02)00105-2.

Hunt, G. (1977). "Spectral Signatures of Particulate Minerals in the Visible and Near Infrared". In: *Geophysics* 42.3, pp. 501–513. DOI: 10.1190/1.1440721.

ICRAF-ISRIC (2010). *A Globally Distributed Soil Spectral Library: Visible Near Infrared Diffuse Reflectance Spectra*. World Agroforestry Centre (ICRAF) and ISRIC - World Soil Information http://africasoils.net/data/spectral-libraries.

IUSS Working Group WRB (2007). *World reference base for soil resources 2006, first update 2007. World Soil Resources Reports No. 103*. http://www.fao.org/fileadmin/templates/nr/images/resources/pdf_documents/wrb2007_red.pdf. [Accessed on 2014-04-08].

Jahn, R., H. Blume, V. Asio, O. Spaargaren, and P. Schad (2006). *Guidelines for soil description, 4th edition*. Rome: Food and Agriculture Organization of the United Nations.

James, G., D. Witten, T. Hastie, and R. Tibshirani (2013). *An introduction to statistical learning*. Springer.

Janik, L. J., R. H. Merry, S. T. Forrester, D. M. Lanyon, and A. Rawson (2007). "Rapid Prediction Of Soil Water Retention Using Mid Infrared Spectroscopy". In: *Soil Science Society of America Journal* 71.2, pp. 507–514. DOI: 10.2136/sssaj2005.0391.

Koestel, J. and H. Jorda (2014). "What determines the strength of preferential transport in undisturbed soil under steady-state flow?" In: *Geoderma* 217-218.0, pp. 144–160. DOI: 10.1016/j.geoderma.2013.11.009.

Korobeynikov, A. (2010). "Computation- and Space-Efficient Implementation of SSA". In: *Statistics and Its Interface* 3.3, pp. 357–368.

Ließ, M., B. Glaser, and B. Huwe (2011). "Functional soil-landscape modelling to estimate slope stability in a steep Andean mountain forest region". In: *Geomorphology* 132.3-4, pp. 287–299. DOI: 10.1016/j.geomorph.2011.05.015.

– (2012). "Uncertainty in the spatial prediction of soil texture: Comparison of regression tree and Random Forest models". In: *Geoderma* 170, pp. 70–79. DOI: 10.1016/j.geoderma.2011.10.010.

Luedeling, E., G. Sileshi, T. Beedy, and J. Dietz (2011). "Carbon Sequestration Potential of Agroforestry Systems in Africa". In: *Advances in Agroforestry*. Ed. by B. M. Kumar and P. K. R. Nair. Vol. 8. Springer Netherlands, pp. 61–83–. DOI: 10.1007/978-94-007-1630-8_4.

Maeda, T, H Takenaka, and B. Warkentin (1977). "Physical properties of allophane soils". In: *Advances in Agronomy*. Vol. 29. Academic Press, pp. 229–264.

Mallants, D., B. P. Mohanty, D. Jacques, and J. Feyen (1996). "Spatial Variability of Hydraulic Properties in a Multi-Layered Soil Profile". In: *Soil Sci.* 161.3, pp. 167–181.

Matheron, G. (1963). “Principles of geostatistics”. In: *Economic Geology* 58.8, pp. 1246–1266. DOI: 10.2113/gsecongeo.58.8.1246.

Minasny, B., A. B. McBratney, and K. L. Bristow (1999). “Comparison of different approaches to the development of pedotransfer functions for water-retention curves”. In: *Geoderma* 93.3-4, pp. 225–253. DOI: 10.1016/S0016-7061(99)00061-0.

Minasny, B. and A. E. Hartemink (2011a). “Predicting soil properties in the tropics”. In: *Earth-Science Reviews* 106.1-2, pp. 52–62. DOI: 10.1016/j.earscirev.2011.01.005.

Misana, S. B., A. E. Majule, H. V. Lyaruu, and L. U. Change (2003). *Linkages between changes in land use, biodiversity and land degradation on the slopes of Mount Kilimanjaro, Tanzania*. LUCID Project, International Livestock Research Institute.

Mualem, Y. (1976). “A new model for predicting the hydraulic conductivity of unsaturated porous media”. In: *Water Resources Research* 12.3, pp. 513–522. DOI: 10.1029/WR012i003p00513.

National Bureau of Statistics (2006). *Population and Housing Census 2002*. Analytical Report. Ministry of Planning, Economy and Empowerment. Dar es Salaam, Tanzania.

Nonnotte, P., H. Guillou, B. L. Gall, M. Benoit, J. Cotten, and S. Scaillet (2008). “New K – Ar age determinations of Kilimanjaro volcano in the North Tanzanian diverging rift, East Africa”. In: *Journal of Volcanology and Geothermal Research* 173.1–2, pp. 99–112. DOI: 10.1016/j.jvolgeores.2007.12.042.

Nunan, N., K. Wu, I. Young, J. Crawford, and K. Ritz (2002). “In Situ Spatial Patterns of Soil Bacterial Populations, Mapped at Multiple Scales, in an Arable Soil”. English. In: *Microbial Ecology* 44.4, pp. 296–305. DOI: 10.1007/s00248-002-2021-0.

Pabst, H., A. Kühnel, and Y. Kuzyakov (May 2013). “Effect of land-use and elevation on microbial biomass and water extractable carbon in soils of Mt. Kilimanjaro ecosystems”. In: *Applied Soil Ecology* 67.0, pp. 10–19. DOI: 10.1016/j.apsoil.2013.02.006.

Pebesma, E. J. (Aug. 2004). “Multivariable geostatistics in S: the gstat package”. In: *Computers & Geosciences* 30.7, pp. 683–691. DOI: 10.1016/j.cageo.2004.03.012.

Power, A. G. (2010). “Ecosystem services and agriculture: tradeoffs and synergies”. In: *Philosophical transactions of the royal society B: biological sciences* 365.1554, pp. 2959–2971. DOI: 10.1098/rstb.2010.0143.

Prasad, A., L. Iverson, and A. Liaw (2006). “Newer Classification and Regression Tree Techniques: Bagging and Random Forests for Ecological Prediction”. English. In: *Ecosystems* 9.2, pp. 181–199. DOI: 10.1007/s10021-005-0054-1.

R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*. ISBN 3-900051-07-0. R Foundation for Statistical Computing. Vienna, Austria.

Rawls, W., T. Gish, and D. Brakensiek (1991). “Estimating Soil Water Retention from Soil Physical Properties and Characteristics”. In: *Advances in Soil Science*. Ed. by B. Stewart. Vol. 16. Springer New York, pp. 213–234. DOI: 10.1007/978-1-4612-3144-8_5.

Ritsema, C. J. and L. W. Dekker (1998). "Three-dimensional patterns of moisture, water repellency, bromide and pH in a sandy soil". In: *Journal of Contaminant Hydrology* 31.3-4, pp. 295–313. DOI: 10.1016/S0169-7722(97)00067-3.

Rohr, P. and A Killingtveit (2003). "Rainfall distribution on the slopes of Mt Kilimanjaro". English. In: *Hydrological Sciences Journal –Journal des sciences hydrologiques* 48.1, 65–77. DOI: {10.1623/hysj.48.1.65.43483}.

Rossi, J., A. Govaerts, B. D. Vos, B. Verbist, A. Vervoort, J. Poesen, B. Muys, and J. Deckers (2009). "Spatial structures of soil organic carbon in tropical forests–A case study of Southeastern Tanzania". In: *CATENA* 77.1, pp. 19–27. DOI: DOI:10.1016/j.catena.2008.12.003.

Sankey, J. B., D. J. Brown, M. L. Bernard, and R. L. Lawrence (2008). "Comparing local vs. global visible and near-infrared (VisNIR) diffuse reflectance spectroscopy (DRS) calibrations for the prediction of soil clay, organic C and inorganic C". In: *Geoderma* 148.2, pp. 149–158. DOI: 10.1016/j.geoderma.2008.09.019.

Santra, P., R. N. Sahoo, B. S. Das, R. N. Samal, A. K. Pattanaik, and V. K. Gupta (2009). "Estimation of soil hydraulic properties using proximal spectral reflectance in visible, near-infrared, and shortwave-infrared (VIS-NIR-SWIR) region". In: *Geoderma* 152.3-4, pp. 338–349. DOI: 10.1016/j.geoderma.2009.07.001.

Schaap, M. G., F. J. Leij, and M. T. van Genuchten (2001). "ROSETTA: a computer program for estimating soil hydraulic parameters with hierarchical pedotransfer functions". In: *Journal of Hydrology* 251.3-4, pp. 163–176. DOI: 10.1016/S0022-1694(01)00466-8.

Sequeira, C. H., S. A. Wills, C. A. Seybold, and L. T. West (2014). "Predicting soil bulk density for incomplete databases". In: *Geoderma* 213.0, pp. 64–73. DOI: 10.1016/j.geoderma.2013.07.013.

Shepherd, K. D. and M. G. Walsh (2002). "Development of Reflectance Spectral Libraries for Characterization of Soil Properties". In: *Soil Science Society of America Journal* 66.3, pp. 988–998. DOI: 10.2136/sssaj2002.9880.

Soil Conservation Service. *Soil survey manual*. 18th ed. Soil Survey Division Staff. 1993. U.S. Department of Agriculture Handbook.

Soini, E. (2005). "Changing livelihoods on the slopes of Mt. Kilimanjaro, Tanzania: Challenges and opportunities in the Chagga homegarden system". In: *Agroforestry Systems* 64.2, pp. 157–167. DOI: 10.1007/s10457-004-1023-y.

Stevens, A., M. Nocita, G. Tóth, L. Montanarella, and B. van Wesemael (2013). "Prediction of soil organic carbon at the European scale by visible and near infrared reflectance spectroscopy". In: *PLoS ONE* 8.6. DOI: 10.1371/journal.pone.0066409.

Tomasella, J and M Hodnett (2004). "Pedotransfer functions for tropical soils". In: *Development of pedotransfer functions in soil hydrology* 30, pp. 415–435. DOI: 10.1016/S0166-2481(04)30021-8.

Torgo, L., R. P. Ribeiro, B. Pfahringer, and P. Branco (2013). "SMOTE for Regression". In: *Progress in Artificial Intelligence*. Springer, pp. 378–389.

Vagen, T.-G., R. Lal, and B. R. Singh (2005). “Soil carbon sequestration in sub-Saharan Africa: a review”. In: *Land Degrad. Dev.* 16.1, pp. 53–71. DOI: 10.1002/ldr.644.

Van Genuchten, M. T. (1980). “A closed-form equation for predicting the hydraulic conductivity of unsaturated soils”. In: *Soil science society of America journal* 44.5, pp. 892–898. DOI: 10.2136/sssaj1980.03615995004400050002x.

Viscarra Rossel, R., D. Walvoort, A. McBratney, L. Janik, and J. Skjemstad (2006b). “Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties”. In: *Geoderma* 131.1-2, pp. 59–75. DOI: 10.1016/j.geoderma.2005.03.007.

Viscarra Rossel, R., S. Cattle, A. Ortega, and Y. Fouad (2009). “In situ measurements of soil colour, mineral composition and clay content by vis-NIR spectroscopy”. In: *Geoderma* 150.3-4, pp. 253–266. DOI: 10.1016/j.geoderma.2009.01.025.

Vågen, T.-G., K. D. Shepherd, and M. G. Walsh (2006). “Sensing landscape level change in soil fertility following deforestation and conversion in the highlands of Madagascar using Vis-NIR spectroscopy”. In: *Geoderma* 133.3-4, pp. 281–294. DOI: 10.1016/j.geoderma.2005.07.014.

Waiser, T. H., C. L. S. Morgan, D. J. Brown, and C. T. Hallmark (Mar. 2007). *In Situ Characterization of Soil Clay Content with Visible Near-Infrared Diffuse Reflectance Spectroscopy*. DOI: 10.2136/sssaj2006.0211.

Wehrens, R. (2011). *Chemometrics with R: multivariate data analysis in the natural sciences and life sciences*. Springerverlag Berlin Heidelberg.

Wetterlind, J. and B. Stenberg (2010a). “Near-infrared spectroscopy for within-field soil characterization: small local calibrations compared with national libraries spiked with local samples”. In: *European Journal of Soil Science* 61.6, pp. 823–843. DOI: 10.1111/j.1365-2389.2010.01283.x.

Wetterlind, J., B. Stenberg, and M. Söderström (2010b). “Increased sample point density in farm soil mapping by local calibration of visible and near infrared prediction models”. In: *Geoderma* 156.3-4, pp. 152–160. DOI: 10.1016/j.geoderma.2010.02.012.

Young, I. M., J. W. Crawford, and C. Rappoldt (2001). “New methods and models for characterising structural heterogeneity of soil”. In: *Soil and Tillage Research* 61.1-2, pp. 33–45. DOI: 10.1016/S0167-1987(01)00188-X.

Young, I. M. and J. W. Crawford (2004). “Interactions and Self-Organization in the Soil-Microbe Complex”. In: *Science* 304.5677, pp. 1634–1637. DOI: 10.1126/science.1097394. eprint: http://www.sciencemag.org/content/304/5677/1634.full.pdf.

Zech, M., K. Leiber, W. Zech, T. Poetsch, and A. Hemp (Oct. 2011). “Late Quaternary soil genesis and vegetation history on the northern slopes of Mt. Kilimanjaro, East Africa”. In: *Quaternary International* 243.2, pp. 327–336. DOI: 10.1016/j.quaint.2011.05.020.

Zech, M., C. Hörold, K. Leiber-Sauheitl, A. Kühnel, A. Hemp, and W. Zech (2014). “Buried black soils on the slopes of Mt. Kilimanjaro as a regional carbon storage hotspot”. In: *CATENA* 112.0, pp. 125–130. DOI: 10.1016/j.catena.2013.05.015.

# Spatial patterns of microbial biomass and fauna activity in savannah soils at Mt. Kilimanjaro

ANNA KÜHNEL[1], HOLGER PABST[2], JULIANE RÖDER[3], CHRISTINA BOGNER[4], YAKOV KUZYAKOV[2,5] AND BERND HUWE[1]

[1]Soil Physics Group, BayCEER, University of Bayreuth, Germany
[2]Department of Soil Science of Temperate Ecosystems, University of Göttingen, Germany
[3]Animal Ecology Group, University of Marburg
[4]Ecological Modelling, BayCEER, University of Bayreuth, Germany
[5] Department of Agricultural Soil Science, University of Göttingen, Germany

submitted to

*European Journal of Soil Science*, (17.07.2014)

corresponding author: Anna Kühnel (anna.kuehnel@uni-bayreuth.de)

**Abstract**

The knowledge of spatial distribution of soil microbial biomass and the soil fauna activity (SFA) is important in order to understand the functioning of an ecosystem. For East African savannah ecosystems this detailed spatial knowledge is still lacking. This study compares the precision and results of the four spatial prediction techniques (i) ordinary kriging, (ii) multiple linear regression, (iii) regression kriging, and (iv) geographically weighted regression. These techniques were used to predict microbial biomass carbon ($C_{mic}$), microbial biomass nitrogen ($N_{mic}$) and SFA. The predictions were used to create detailed maps of two savannah sites. As predictor variables we used parameters that are a) easy to measure, and b) correlated to the target variables – such as $C_{org}$, nitrogen, clay content and pH value. Regression kriging, multiple linear regression and geographically weighted regression performed better than ordinary kriging with $R^2$ for the prediction of $C_{mic}$ and $N_{mic}$ up to 0.71 and 0.76, respectively. SFA did not show any spatial dependencies or relations to the chosen predictor variables. $C_{mic}$ and $N_{mic}$ showed diverse spatial relations to the predictors and contrasting degrees of heterogeneity on the two plots. Concluding, multivariate methods are advantageous for the estimation of microbial parameters, producing maps and ecological interpretations of the local relationships in soil.

## 5.1 Introduction

Tropical ecosystems with their high biodiversity play an important role in global climate and biogeochemical cycles, especially in carbon (C) turnover and sequestration (Detwiler et al., 1988; Malhi et al., 2004). This is also true for savannah ecosystems, which cover nearly 20% of the earth's land surface and are affected by global change (Sankaran et al., 2013; D'Odorico et al., 2013). Therefore savannahs have been the subject of several studies focusing on soil microbial biomass, C turnover, fire effects and greenhouse-gas exchange (Jones, 1973; Singh et al., 1989; Hagos et al., 2005).

A basic concept in biology is that there is a positive relationship between environmental heterogeneity and species diversity (Tamme et al., 2010). However, the concept of environmental heterogeneity is not clearly defined and the relations are scale dependent (Wilson, 2000). The animal species diversity in savannah ecosystem, for example, is closely linked to the occurrence of large trees, as they function as food resource, shelter or nesting site (Tews et al., 2004). Canopy structure of savannah trees on the other hand influences throughfall, which affects soil moisture, soil fertility (Vetaas, 1992) and tree growth (Plath et al., 2011). But also the other way around, biological parameters can influence physical and chemical properties of soils (Wiens, 2000). Therefore detailed knowledge of the spatial distribution of

different parameters in a system and the dependencies between them are of great concern.

Spatial interpolation techniques are a common tool for the estimation of C and nutrient pools (Kumar et al., 2012; Kuzyakova et al., 2001; Mishra et al., 2012). Spatial analyses focus on the variability of a given parameter over space and thus the heterogeneity of this parameter in the system at the studied scale. However, spatial analysis of savannah soils was used in only few studies, mostly in South and West Africa (Hagos et al., 2005; Wang et al., 2009). In addition and since each physical and chemical analysis implies further costs, soil sample collection is often sparsely distributed over space. As measuring biological soil parameters like microbial C and N ($C_{mic}$, $N_{mic}$) or soil fauna activity (SFA) are very laborious in the field as well as in the lab, the application of prediction techniques might make better use of the data and might lead to precise estimations based on only few data points.

Over the last century, soil-landscape modeling has shifted from qualitative methods, e.g. soil classification and soil survey maps towards quantitative methods like fuzzy sets and multivariate geospatial models (Grunwald, 2005). Geospatial models can be used to estimate a soil property at an unknown location and modeling is considered more detailed and less error-prone than for example soil survey maps (Thompson et al., 2005). One to several predictor variables which are available in a high resolution within the study area are used to estimate the variability of the sparsely sampled target variable (Thompson et al., 2005; Mishra et al., 2010). However, there is no universal single best prediction method for all target variables and therefore caution is advised when selecting the most suitable method for a certain variable (Li et al., 2011). The traditional technique of ordinary kriging (OK) uses data of the target variable, available at the observation points, to predict its value at new locations (Cressie, 1988). OK is widely used and its computation is easy compared to the more advanced multivariate methods, like regression kriging (RK). RK uses the available data of the dependent variable and, in addition, information from auxiliary or co-variables (e.g. topography, data that can be derived from satellite images, variables that are more easy to measure than the dependent variable, etc.) (Hengl, 2009). In addition to the commonly used geostatistical methods, Brunsdon et al. (1996) introduced the multivariate approach of the geographically weighted regression (GWR). Compared to OK and RK, GWR has the advantage that it considers the possibility of varying relationships between the model variables over space (Brunsdon et al., 1996).

In order to gain information on soil parameters correlated to microbial biomass (i.e. $C_{org}$, soil texture), visible to near-infrared diffuse reflectance spectroscopy (Vis-NIR-DRS) can be used (Awiti et al., 2008; Viscarra Rossel et al., 2010; Chang et al.,

2001). It is an established method to predict several soil physical and chemical properties. Spectral measurements are mostly non-destructive, faster and less expensive compared to classical physical and chemical soil analyses. One of the main reasons for the speed and cost-efficiency of Vis-NIR-DRS is that several different soil properties can be derived from a single spectroscopic measurement, like $C_{org}$, nitrogen (N) and clay content (Viscarra Rossel et al., 2006b).

The objectives of this study were therefore a) to use easily measurable soil properties as co-variables in multiple linear regression (MLR), RK and GWR to predict $C_{mic}$, $N_{mic}$ as well as SFA and b) to test if these methods can increase the prediction quality compared to OK. Our goal is to characterize and compare microbial biomass and SFA on a plain and a sloping site in an East African savannah.

## 5.2 Materials and methods

### 5.2.1 Study site

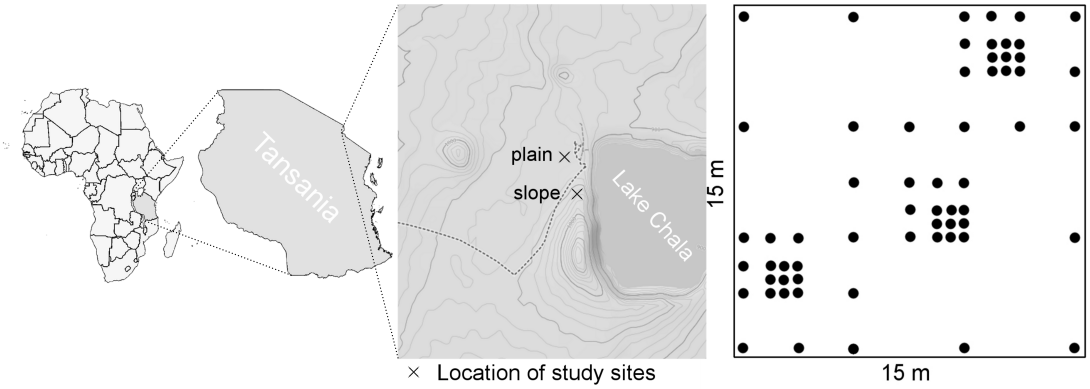



Figure 5.1: Study area with the location of the study plots (x) and study design. Source: commons.wikimedia and OpenStreetMap.

The study was conducted in a semi-arid natural savannah near Lake Chala in the East of Mt. Kilimanjaro, Tanzania (3°18'39" S, 37°41'8" E, Figure 5.1). The mean annual rainfall equals about 565 mm, with a long rainy season from March to May and a short rainy season between October and December (Moernaut et al., 2010). Soils developed on superficial deposits from Kibo and Mawenzi peaks of Mt. Kilimanjaro and from the various small and steep craters in the east of the mountain complex (Nonnotte et al., 2008). On the slopes of these small volcanoes the main soil type is Leptosol, whereas Vertisols are dominant in the plains.

We worked on two different study sites. One site ($P_{slope}$) is situated on the outer foot slope of the crater rim of the Lake Chala caldera at an elevation of 960 m

(Figure 5.1). It is north exposed with an inclination of about 10°. The soil is very shallow with a maximum depth of 25 cm, with bedrock appearing directly at the surface in some parts of the area. It was classified as Rendzic Leptosol (Calcaric, Tephric, Sodic, Eutric, Skeletic) according to WRB (IUSS Working Group WRB, 2007). Dominating tree species at this site is Combretum molle, with some Acacias in between. The second study site ($P_{plain}$) is situated 400 m to the north-west, in the plains surrounding Lake Chala at 950 m a.s.l. with no inclination. The soil was classified as Sodic Vertisol (Hypereutric, Chromic). *Balanitis aegyptiaca* and different Acacias (*Acacia tortilis, Acacia senegal, Acacia nilotica*) are the dominating tree species at this site.

### 5.2.2 Study design and field sampling

A detailed sampling campaign was carried out in October 2012. The sampling was designed in a hierarchically nested grid on two 15 m x 15 m plots, consisting of 61 sampling points each (Figure 5.1). Soil fauna activity was estimated with the bait-lamina method (Toerne, 1990). The bait lamina consisted of plastic strips with 16 biconical holes of 2 mm diameter and a distance of 5 mm. The holes were filled with a bait of cellulose (70%), wheat bran (27%) and active coal (3%) (Kratz, 1998). Five bait laminas were inserted vertically into the soil at each grid point until all bait units were covered with soil or the maximum soil depth was reached, respectively. The bait lamina were placed about 10 cm apart and within 20 cm from the actual sampling/grid point. After two weeks, the bait laminas were collected. The number of perforated bait units was then evaluated on a light table after cleaning the strips from attached soil and the percentage of eaten units was calculated.

Directly after the retrieval of the bait lamina, a mixed soil sample of the upper 5 cm was collected at each grid point. The samples were sieved through a 2 mm mesh screen and about 10-15 g of each of the thoroughly mixed samples was oven-dried at 45°C for spectral analysis. The remaining material was stored under field moist conditions at 4-6°C until analysis for microbial biomass. On most of the 61 observation points per plot a soil sample could be taken. However, some positions are missing because of surface adjacent bedrock.

### 5.2.3 Laboratory measurements

#### Soil microbial biomass

$C_{mic}$ and $N_{mic}$ were analyzed by the fumigation-extraction method. Summarily, 7-8 g of field moist soil were fumigated in an exsiccator with ethanol-free $CHCl_3$.

Afterwards, soluble C and N from fumigated and non-fumigated samples was extracted with 60 ml of 0.5 M $K_2SO_4$. C and N in the solution were measured with a C-N-analyser (multi N/C 2100S, analytikjena, Jena, Germany). Since not all of the C and N can be extracted, a $k_{EC}$ factor of 0.45 (Joergensen, 1996) and a $k_{EN}$ factor of 0.54 (Joergensen et al., 1996) was used to convert microbial C and N flush into $C_{mic}$ and $N_{mic}$, respectively.

#### Visible to near infrared diffuse reflectance spectroscopy

Each soil sample was scanned with an AgriSpec portable spectrophotometer equipped with a contact Probe (Analytical Spectral Devices, Boulder, Colorado) in the range of 350 – 2500 nm with 1 nm intervals. The instrument was calibrated with a white reference prior to measurements. Each sample was scanned 30 times and the spectra were averaged to reduce the noise. Partial least square regression with leave-one-out cross validation (LOOCV) (Davis, 1987; James et al., 2013) was used to predict $C_{org}$, N and clay content, as it is a common method to predict soil properties from spectral data (Wold et al., 2001). Clay content was predicted using a model that was based on samples from different soils in the Mt. Kilimanjaro area. Details of the modeling process and validation results can be found in Study 1 and Appendix E

For the acquisition of $C_{org}$ and N we chose 16 of the 61 sampling points that were equally distributed over the sampling area with a distance of 5 m each. $C_{org}$ and N were measured with a C-N-analyzer in the selected samples. The corresponding spectra were subsequently used to generate synthetic spectra with the synthetic minority oversampling technique (SMOTE) and its extension for regression (Chawla et al., 2002; Torgo et al., 2013). For every soil spectra with known $C_{org}$ and N content, three new spectra were generated with the following equation:

$$X_s = X_o + \delta(X_n - X_o) \tag{5.1}$$

where $X_s$ is the synthetic spectra, $X_o$ is the original spectra, $X_n$ is a randomly chosen neighbour of $X_o$ and $\delta$ is a random number between 0 and 1.

Following the modeling approach of Bogner et al., 2014, we augmented a spectral database consisting of savannah soils from a larger area in the East of Mt. Kilimanjaro with the new synthetic spectra, separately for each plot and parameter. The resulting partial least squares regression models were validated by LOOCV. An additional validation was conducted with a separate dataset containing the 16 samples with known $C_{org}$ and N content . All models performed well with $R^2$ of the LOOCV (see Equation 5.9 for definition) between 0.70 and 0.84.

The detailed SMOTE algorithm, a graphical illustration of SMOTE as well as error parameters of the different regression models can be found in the Appendix D.

### 5.2.4 Spatial methods

#### Ordinary Kriging

Ordinary kriging uses the differences of values, depending on the distance to each other, to estimate the spatial autocorrelation structure (Krige, 1951; Matheron, 1963; Hengl, 2009; Goovaerts, 1997). The semivariances $Y(h)$ of these differences are calculated by:

$$Y(h) = \frac{1}{2}E[(O(s_i) - O(s_i + h))^2] \tag{5.2}$$

where $O(s_i)$ is the target variable at the location $s_i$ and $O(s_i + h)$ is the value at a distance $h$ from the location $s_i$. Subsequently, the semivariances are summarized by their separation distance $h$ (called lag) and a variogram model is fitted (Figure 5.3).

Variograms are interpreted via three main values: the range,the sill and the nugget. The range is the distance at which the variances between points are more or less equal to the variance of all observed values of the dataset and the sill is the total variance at the range distance. The nugget is the semivariance at zero distance or in other words, it is the variance of sampling points within distances smaller than the smallest sampling interval, including unknown measurement errors.

Finally, predictions are made with the formula

$$P(s_0) = \sum_{i=1}^{N} (\Omega_i(s_0) \cdot O(s_i)) \tag{5.3}$$

where $P(s_0)$ stands for the predicted value at location $s_0$, $\Omega$ is the spatial weighting function based on the semivariogram, $O(s_i)$ is the observation at location $s_i$ and $N$ is the number of observations. In other words, in OK the value at a location is calculated as a weighted linear combination of measured values at locations $s_i$ ($i$ = 1, 2, ..., $N$).

#### Multiple Linear Regression

The general regression model used in MLR is

$$P = \alpha_0 + \sum_{i=1}^{l} \alpha_k \cdot V_{i,k} + \varepsilon_i \tag{5.4}$$

where $P$ is the $i$-th predicted value, $\alpha_0$ ist the intercept, $\alpha_k$ is the regression coefficient of the $k$-th predictor variable, $l$ is the number of predictor variables, $V_{i,k}$ is the $i$-th observation of the $k$-th predictor variable and $\varepsilon$ is an independent normally

distributed error. MLR is not itself a spatial prediction method, but can be used as such, if the predictor variables are known at every location.

The best predictive model was selected for each target variable based on the corrected Akaike information criterion ($AIC_c$) (Akaike, 1973; Webster, 1989):

$$AIC_c = N\log(RMSE^2) + 2m + \frac{2m\cdot(m+1)}{N-m-1} \tag{5.5}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(p_i - o_i)^2} \tag{5.6}$$

where $m$ is the number of model parameters, $N$ is the number of observations, $p_i$ are the predicted values and $o_i$ the observed values. All possible combinations of explanatory variables were considered. None of the chosen models showed signs of multicollinearity, since all of the predictor variables showed variance inflation factors $< 4$ (Rogerson, 2001).

#### Regression Kriging

Regression kriging combines MLR with a variogram analysis of the model residuals (Hengl et al., 2004; Bourennane et al., 2000). First, a MLR between the dependent and independent variables is calculated. Then the residuals $r$ of this model are predicted at all locations of interest using their respective variogram parameters. The RK model resembles the MLR model, but instead of using one constant error term, the residual variance is calculated as the sum of the predicted residuals, weighted by distance:

$$\begin{aligned} P(s_0) &= \sum_{k=0}^{l}\alpha_k \cdot V_k(s_0) + \sum_{i=1}^{N} w_i(s_0)\cdot r(s_0) \\ V_0(s_0) &= 1 \end{aligned} \tag{5.7}$$

where $\alpha_k$ are the estimated linear model coefficients, $V_k$ is the $k$-th predictor variable and $w_i$ are weights based on the variogram analysis of the residuals.

#### Geographically Weighted Regression

GWR uses a linear regression model similar to MLR (Equation Equation 5.4). However, instead of estimating only one set of regression parameters for all sampling positions combined, it allows the estimation of local parameters (Fotheringham et al., 2002):

$$P(s_0) = \alpha_0(s_0) + \sum_{k=1}^{l} \alpha_k(s_0) \cdot V_{i,k} + \varepsilon_i \tag{5.8}$$

where $s_0$ stands for the location of the $i$-th point. Based on the proximity of an observation to a point $i$, the correlation of this observation to the point is estimated by weighted least squares regression. This approach allows local variations in relationships between response and explanatory variables (Brunsdon et al., 1996; Fotheringham et al., 2002).

### 5.2.5 Spatial predictions and mapping

We used OK to predict $C_{org}$, N and clay content as well as pH at every point in a regular grid (15 m $\times$ 15 m grid, spacing 0.625 m) where no samples had been taken. This resulted in a grid with information about these parameters at 625 locations. $C_{org}$, N, clay and pH were chosen, because they can either be easily acquired with Vis-NIR-DRS ($C_{org}$, N, clay) or are easy and cheap to measure (pH-value) and are related to the target variables $C_{mic}$ and $N_{mic}$. Using OK to predict the co-variables is a compromise, as the fitted variograms are not always satisfying (see section Predictor variables 5.3.2). If the geostatistical analysis is promising however, Vis-NIR-DRS could be used to get soil data on a very dense grid in a very short time with low costs.

In a second step, depending on the chosen MLR model, $C_{org}$, N, clay and pH were used as co-variables for the prediction of $C_{mic}$ and $N_{mic}$. We used LOOCV to evaluate the different prediction techniques OK, RK, MLR and GWR. A variogram model was estimated by leaving out one data point for each variogram fit and repeating this procedure $N$ times, where $N$ is the number of available observations. That means, that every point was left out once and predicted with the remaining points. We tested three different variogram model types, namely exponential, linear and spherical. The variogram model type for the respective variable was then selected based on the highest $R^2$ from LOOCV:

$$R^2 = 1 - \frac{\sum_{i=1}^{N} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{N} (y_i - \bar{y})^2} \tag{5.9}$$

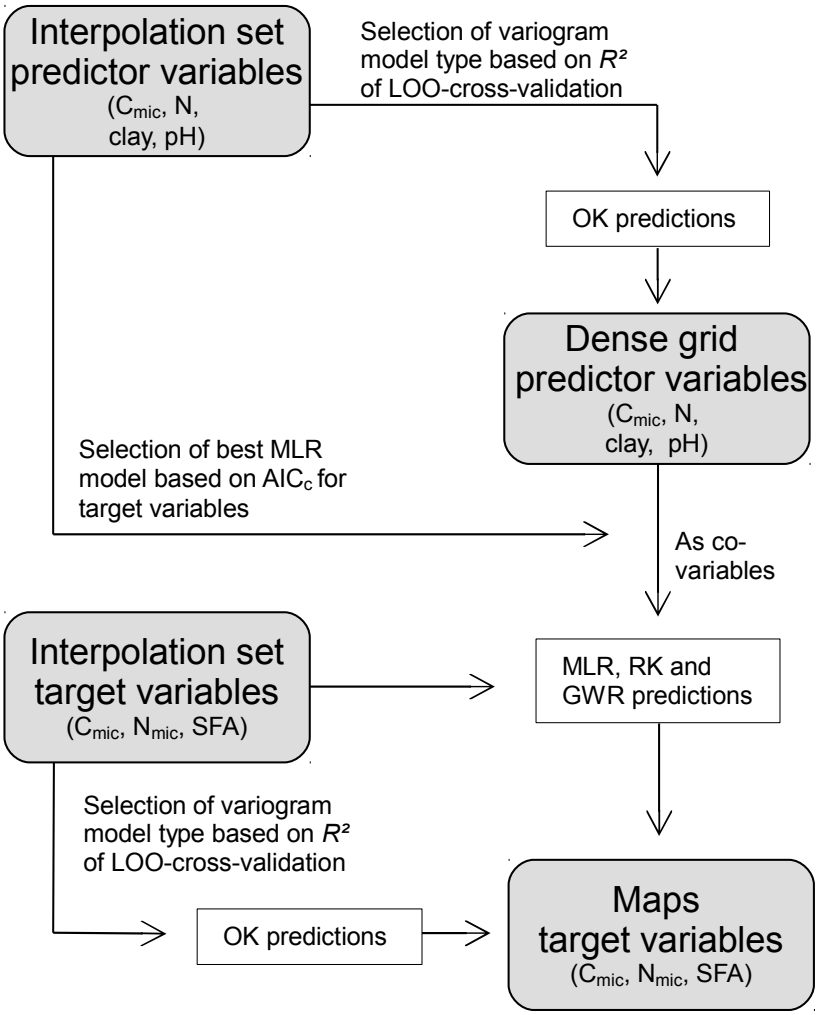



Figure 5.2: Workflow for the production of maps for $C_{mic}$ and $N_{mic}$ on both savannah plots; $C_{mic}$ = microbial biomass carbon, $N_{mic}$ = microbial biomass nitrogen, SFA= soil fauna activity, OK = Ordinary Kriging, MLR = Multiple linear regression, RK = Regression Kriging, GWR = Geographically weighted regression, LOO = Leave-One-Out

where $\hat{y}_i$ are the predicted values, $y_i$ the observed values and $\bar{y}$ is the mean of the observed values. In other words we chose the variogram model type based on the predictive quality and not the best global fit to the data. In order to assess model accuracies of the different variogram model types as well as the different geostatistical methods, we used the $R^2$ and the $RMSE$. Furthermore, to compare prediction accuracies directly, the $RMSE$ (Equation Equation 5.6) was divided by the known standard deviation of the target variable $sd(o_i)$:

$$RMSE_r(\%) = \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(p_i - o_i)^2}}{sd(o_i)} \cdot 100 \tag{5.10}$$

The resulting relative root mean squared error of prediction $RMSE_r$ is scale independent.

The chosen variogram model was consequently used to predict all 625 values of the grid $N$ times and median values of all predictions were calculated. Finally we created maps of $C_{mic}$ and $N_{mic}$, using the median predictions from the best modeling technique to visualize the spatial distribution.

Figure 5.2 shows the workflow of how to obtain spatial predictions and create maps of the target variables.

All statistical analyses were done using R 3.0 (R Development Core Team, 2011) and the packages gstat (Pebesma, 2004), spgwr (Bivand et al., 2013) and automap (Hiemstra et al., 2009).

## 5.3 Results and Discussion

### 5.3.1 Descriptive Statistics

Table 5.1: Descriptive statistics for predictor and target variables

| | | Predictor variables | | | | Target variables | | |
|---|---|---|---|---|---|---|---|---|
| | | $C_{org}$ g kg$^{-1}$ | N g kg$^{-1}$ | clay % | pH - | $C_{mic}$ g kg$^{-1}$ | $N_{mic}$ g kg$^{-1}$ | SFA % |
| $P_{slope}$ | Min | 29.0 | 2.30 | 29.8 | 7.80 | 0.53 | 0.063 | 0.10 |
| | max | 73.6 | 5.87 | 43.1 | 8.65 | 3.01 | 0.348 | 60.00 |
| | mean | 42.0 | 3.46 | 37.8 | 8.29 | 1.40 | 0.157 | 22.86 |
| | sd | 9.4 | 0.68 | 2.9 | 0.19 | 0.70 | 0.079 | 13.33 |
| | median | 38.7 | 3.32 | 38.0 | 8.32 | 1.15 | 0.121 | 21.98 |
| | mad | 9.8 | 0.60 | 2.4 | 0.21 | 0.63 | 0.061 | 12.93 |
| $P_{plain}$ | Min | 18.8 | 1.63 | 45.9 | 6.28 | 0.20 | 0.027 | 8.75 |
| | max | 35.7 | 3.05 | 57.7 | 6.96 | 1.00 | 0.103 | 80.00 |
| | mean | 27.1 | 2.23 | 50.5 | 6.60 | 0.59 | 0.059 | 37.18 |
| | sd | 3.4 | 0.28 | 2.1 | 0.16 | 0.17 | 0.017 | 18.17 |
| | median | 27.0 | 2.22 | 50.5 | 6.58 | 0.58 | 0.058 | 32.50 |
| | mad | 3.2 | 0.21 | 1.9 | 0.16 | 0.15 | 0.018 | 18.53 |

$C_{mic}$ = microbial biomass carbon, $N_{mic}$ = microbial biomass nitrogen, SFA = soil fauna activity; sd = standard deviation, mad = median absolute deviation

Basic descriptive statistics of all variables are summarized in 5.1. The median is more robust than the mean and was used to interpret the data. The within-plot

variation of most of the variables is up to 1.8 times higher on $P_{slope}$ than on $P_{plain}$, most probably due to shallow-to-no soil areas on the $P_{slope}$. $C_{org}$, N, $C_{mic}$ and $N_{mic}$ values on $P_{slope}$ were generally higher than on $P_{plain}$ (Table 5.1). These results are in general agreement with previous studies of savannah ecosystems (Jones, 1973; Wang et al., 2009; Michelsen et al., 2004). However, the maximum $C_{org}$ content of $P_{slope}$ is unusually high with 73.6 g $kg^{-1}$. C/N ratios varied between 10.4 and 13.4 on $P_{slope}$ and 10.7 and 13.3 on $P_{plain}$, with median values of 12.2 for both plots, which is similar to the findings by Jones (1973) and Hernández-Hernández et al. (2002). Distributions of $C_{mic}$ and $N_{mic}$ showed similarities, with much higher and more variable values for $P_{slope}$.

For Eastern Kilimanjaro Pabst et al. (2013) found $C_{org}$ and N contents as well as $C_{mic}$ in the same range as measured on $P_{plain}$ in this study. Michelsen et al. (2004) and other studies (Singh et al., 1989; Hernández-Hernández et al., 2002) reported $C_{mic}$ values up to 0.8 g $kg^{-1}$ for the topsoil of a wooded grassland which are in general lower than our findings. These studies however did analyze savannah soils, but not natural savannah ecosystems. A study of natural savannah in the Serengeti, East Africa, reported $C_{mic}$ content of up to 3.1 g $kg^{-1}$ (Ruess et al., 1987) which corresponds to the very high $C_{mic}$ contents found on $P_{slope}$. Furthermore, as $C_{mic}$ and $N_{mic}$ may be highly seasonal (Michelsen et al., 2004), depending on the study period, high amounts of microbial biomass might not have been covered due to sampling times (Singh et al., 1989).

The typical soil pH in savannah ecosystems is assumed to vary around values between 4 and 6 (Hagos et al., 2005; Hernández-Hernández et al., 2002). In contrast, pH values found in this study are much higher, with values on $P_{slope}$ exceeding pH 7, which is probably due to the parent material.

Soil fauna activity was higher on $P_{plain}$ with values up to 80 % of eaten bait units and median values of 33 % compared to 60 % and 22 % on $P_{slope}$.

### 5.3.2 Spatial data analysis

#### Predictor variables – $C_{org}$, N, clay content & pH value

The best predictive quality for $C_{org}$ content on $P_{slope}$ was obtained with a linear model (Figure 5.3), which was therefore chosen. A spherical model for N content and exponential models for clay content and pH on $P_{slope}$ were selected, respectively. For $C_{org}$ on $P_{plain}$ we chose a linear model and for N, clay content and pH an exponential model.

Both, $C_{org}$ and N contents showed higher spatial ranges on $P_{plain}$ (Figure 5.3) compared to $P_{slope}$, which were similar to ranges found in southern African savannahs

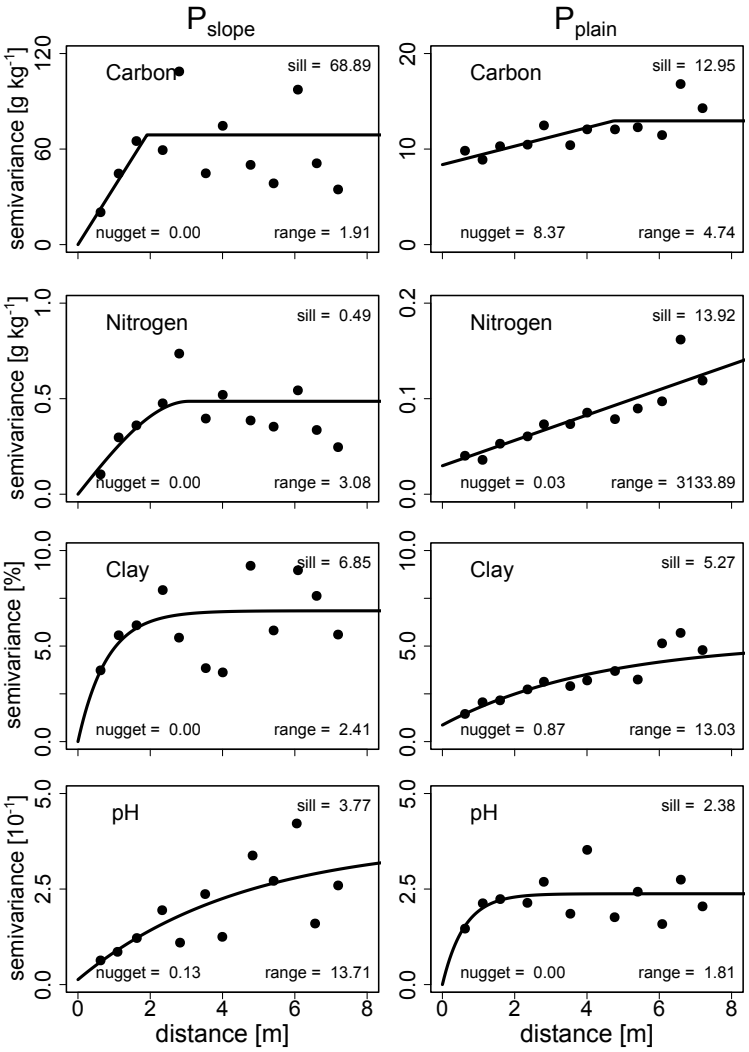



Figure 5.3: Variogram models of the predictor variables for $P_{slope}$ and $P_{plain}$.

(Wang et al., 2009). This indicates lower heterogeneity on $P_{plain}$, since points are still correlated to each other even at large distances. We predicted a very high range value for N on $P_{plain}$ (3.13 km). Typically a variogram is computed up to one third of the maximal distance between points – in our case around 7 m. In greater distances, the number of point pairs and the quality of the experimental variogram decrease rapidly. As we chose the variogram models based on the predictive quality, an exponential model was selected. In contrast, the best global fit to the data was provided by a linear variogram model with a range of 6.09 m and a sill of 0.01 (data not shown). Both, the exponential as well as linear model follow a seemingly linear shape within the distance of 7 m and consequently the differences in predictions between these models are expected to be negligible. Wang et al. (2009) assumed that a loss of woody vegetation and regional drying results in more

heterogeneous soil pools. This acts as a possible explanation for the rather small ranges found for $C_{org}$ and N on $P_{slope}$, since it was characterized by sparse woody vegetation, shallow soil depth and adjacent bedrock.

The spatial ranges of clay contents on the two investigated plots were differing (Figure 5.3). The difference is explained by the relief of the two sites: $P_{slope}$ is prone for soil erosion by wind and water. Clay is easily erodible, and therefore shows high heterogeneity. Soil particles are deposited on the flat relief of $P_{plain}$, leading to deeper soils and a more homogeneous distribution of clay particles.

With the exception of pH, all parameters showed a low range on $P_{slope}$, indicating that values are not dependent on each other already at small distances. This is probably because of the patchy vegetation and the high heterogeneity of soil thickness. A consequence of the clustered sampling design was that the distance classes of the variogram models of $P_{slope}$ often contained points clustered together at locations with either low or high vegetation cover, respectively. This resulted in high variation in the semivariances at higher distances and the high range (Figure 5.3). Including more points in the sampling design would probably strengthen the variogram and also lead to a lower range for pH on $P_{slope}$.

As described in a review by Li et al. (2011), the performance of spatial interpolation methods is significantly influenced by data variation and sampling design and to a lesser extent sampling density. They argued that the effects caused by sampling density are mainly dominated by the data variation in the area (Li et al., 2011). In this study, small sample size and variation within the data resulted in severe problems in finding suitable variogram models (e.g. for pH on $P_{slope}$, N on $P_{plain}$) and low or even negative $R^2$ values (Figure 5.4).

Due to the small sample size it was also not possible to account for anisotropy of the data set and some of the variogram models showed a clear trend in the data, partly explaining the high ranges. The ratio of nugget/sill is an estimate of the spatial dependence within the investigated area. Both low and high spatial dependencies were observed within the smallest sampling interval, whereas it stands out that for $C_{org}$ on $P_{plain}$ 65% of the total variance is explained by the nugget effect. For $C_{org}$ on $P_{slope}$ and N on both plots, the nugget effect was negligible (Figure 5.3). Yet, we can assume that some variability occurs within the smallest sampling interval of 0.625 m and/or measurement errors occurred in the analysis.

The observed spatial structures vary depending on sampling density across the study area (Mishra et al., 2010). Since we used the same sampling density/design in our study on both plots and for all variables, this indicates that the spatial structures of the investigated variables vary not only depending on sampling density but also because of local characteristics of the study sites.

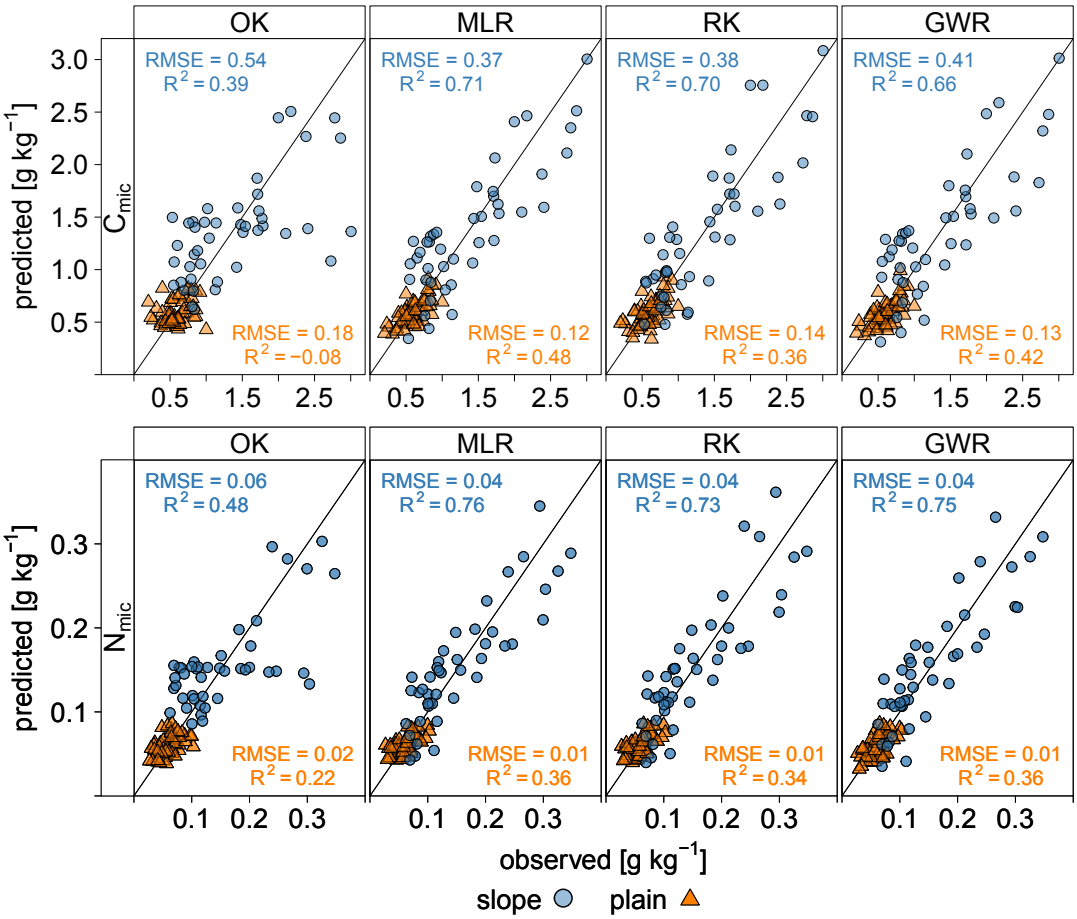



Figure 5.4: Observed versus predicted data of different geostatistical methods on $P_{plain}$ (orange) and $P_{slope}$ (blue): OK = ordinary kriging, RK = regression kriging, GWR = geographically weighted regression, MLR = multiple linear regression, $C_{mic}$ = microbial biomass carbon (top), $N_{mic}$ = microbial biomass nitrogen (bottom).

### Target variables – $C_{mic}$, $N_{mic}$ & SFA

**Variogram parameters** An exponential variogram model type was selected for $C_{mic}$ on $P_{plain}$ and $N_{mic}$ on $P_{slope}$ (Figure 5.5). For $C_{mic}$ on $P_{slope}$ a linear model and for $N_{mic}$ on $P_{plain}$ a spherical model was used, respectively. The linear variogram model type was selected for SFA on both plots (Figure 5.5).

Similar to $C_{org}$ and N contents the sparse vegetation and patchy soil cover on $P_{slope}$ resulted in lower ranges for $C_{mic}$ and $N_{mic}$ compared to $P_{plain}$. On both plots, $N_{mic}$ showed ranges similar to the findings for soil by Wang et al. (2009). The nugget effects were zero on $P_{slope}$, and on $P_{plain}$ they did not exceed 20% of the total variance. Consequently at least 80% of the spatial variation was explained by the chosen variogram model. The variogram models for SFA, however, showed huge nugget-effects and unreasonably high ranges and sills, indicating that there

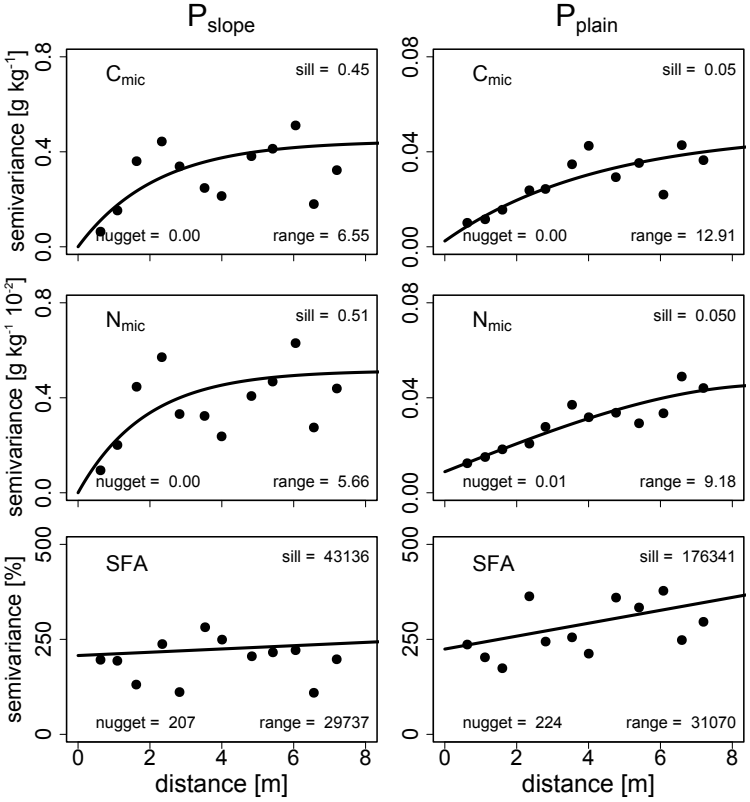



Figure 5.5: Variogram models of the target variables ($C_{mic}$, $N_{mic}$) for $P_{slope}$ and $P_{plain}$.

is a high heterogeneity and no dependency between the individual observations (Figure 5.5).

**Multiple linear regression** The multiple regression models selected for RK and GWR are shown in Table 5.2. No significant model could be established for SFA for either of the two investigated plots, as it was not correlated to any of the explanatory variables and was therefore excluded from further analysis.

Table 5.2: Selected regression models for MLR, RK and GWR prediction methods

| Target variable | Predictors | AICc | F value | | P value | $R^2$ |
|---|---|---|---|---|---|---|
| $P_{slope}$ | | | | | | |
| $C_{mic}$ | $C_{org}$, pH | 76.1 | F(2,40) | 50.1 | $< 0.001^*$ | 0.71 |
| $N_{mic}$ | N, pH | 68.4 | F(2,40) | 63.8 | $< 0.001^*$ | 0.76 |
| $P_{plain}$ | | | | | | |
| $C_{mic}$ | clay, N | 136.7 | F(2,56) | 25.8 | $< 0.001^*$ | 0.48 |
| $N_{mic}$ | N | 148.6 | F(2,58) | 33.1 | $< 0.001^*$ | 0.36 |

The high variation of consumed bait units per grid point seems to be a matter of the method-inherent variability and the low soil moisture content. Kratz (1998) suggested to use a minimum of 4 x 16 bait lamina per study site, which equals four replicates per site and soil depth. Considering that the vertical distribution of feeding activity was not measured in this study, we are in line with these suggestions, with five bait lamina per grid point. Anyhow, the bait-lamina method seems not to be sensitive enough to adequately mirror the small-scale heterogeneity relevant in this study.

$C_{mic}$ on $P_{slope}$ could be well explained with $C_{org}$ content and pH value (Table 5.2). Since $C_{org}$ is the main substrate for soil microbes, there are numerous studies linking $C_{mic}$ to $C_{org}$ (Singh et al., 1989; Michelsen et al., 2004). Similar, $N_{mic}$ showed a good correlation with N and pH as a predictor variable in the regression model (Hernández-Hernández et al., 2002). For $P_{plain}$ the $R^2$ of the multiple regressions were lower compared to $P_{slope}$. $C_{mic}$ on $P_{plain}$ was best explained with a combination of N and clay content; $N_{mic}$ with N alone. Clay content and rainfall explained 50% of the variation of soil organic matter in the Serengeti, Tanzania (Jones, 1973). In addition, clay is known to limit organic matter mineralization in soils (Traoré et al., 2007) and consequently was always negatively correlated to either $C_{mic}$ and $N_{mic}$. As we found higher clay contents on $P_{plain}$, an increased amount of $C_{org}$ could consequently be adsorbed to these particles and therefore be protected from microbial utilization (Traoré et al., 2007). We assume that, on $P_{plain}$, the effect of clay content on $C_{mic}$ overshadows the correlation between $C_{org}$ and $C_{mic}$.

**Regression Kriging** For all variables and both plots, the sills and ranges of the residual variograms were lower than that of the original data (data not shown). Nevertheless, RK provided results more or less similar to MLR since the regression part accounts for most of the spatial interpolation in our study.

**Geographically weighted regression** Although it is possible to only use a subset of the available observations in the regression model, in our case most or all of the observation points were included for both plots and variables. This indicates that larger numbers of sample points would be necessary to sufficiently explain the variations in the observed data.

### 5.3.3 Comparison of geostatistical methods

OK showed high $RMSE$ and low $R^2$ for both plots and variables, for $C_{mic}$ on $P_{plain}$ even $< 0$. $R^2$ was generally higher on $P_{slope}$ (Figure 5.4). The estimates improved

when additional information in form of explanatory variables was used. $RMSE$ values of MLR, RK and GWR were similar and consistently lower than those of OK (Figure 5.4). Accordingly, $R^2$ increased by the use of MLR, RK and GWR (Figure 5.4). The improvement of prediction accuracy by using multivariate approaches – which is also the fact in this study – has been reported in several studies (Mishra et al., 2012; Li et al., 2011; Mishra et al., 2010; Bourennane et al., 2000). For the calculation of $C_{org}$ pools in the Midwestern United States, RK and GWR were seen as the best suited methods (Mishra et al., 2010). However, on $P_{plain}$, $R^2$ remained rather low even with the use of the multivariate prediction techniques since the variability within the observed data of $P_{plain}$ was low. However, the ability of the presented three methods to further improve the predictions compared to the mean, clearly showed the usefulness of including explanatory variables.

Table 5.3: Error parameters for the prediction of $C_{mic}$ and $N_{mic}$ with methods OK, MLR, RK, GWR and the relative improvement ($RI$) by the use of MLR, RK, GWR for $P_{slope}$ and $P_{plain}$.

| | | $P_{slope}$ | | $P_{plain}$ | |
|---|---|---|---|---|---|
| | | $RMSE_r$ | $RI$ | $RMSEr$ | $RI$ |
| | | % | % | % | % |
| $C_{mic}$ | OK | 77.12 | - | 103.28 | - |
| | MLR | 52.79 | 31.55 | 71.55 | 30.72 |
| | RK | 53.99 | 29.99 | 79.62 | 22.90 |
| | GWR | 57.97 | 24.38 | 75.52 | 26.88 |
| $N_{mic}$ | OK | 71.11 | - | 87.60 | - |
| | MLR | 48.27 | 32.12 | 79.14 | 9.65 |
| | RK | 51.43 | 27.68 | 80.74 | 7.83 |
| | GWR | 49.24 | 30.76 | 79.39 | 9.37 |

$C_{mic}$ = microbial biomass carbon, $N_{mic}$ = microbial biomass nitrogen, OK = ordinary kriging, MLR = multiple linear regression, RK = regression kriging, GWR = geographically weighted regression

The $RMSE_r$ indicates the variation of prediction errors within the observed range of the sampled data. Is the variation in the observed data low, methods have to be more accurate to obtain low $RMSE_r$ values. Similarly, high variations in the observed data lead more easily to low $RMSE_r$ values. Compared to OK, the multivariate methods clearly improved the $RMSE_r$ on both plots (Table 5.3). However, the on-plot variation of the observed data was low on $P_{plain}$ and consequently the $RMSE_r$ was still up to 81% for MLR, RK and GWR. For all target variables and both study plots, compared to OK the addition of explanatory variables in the models reduced the global estimation error ($RI$) by 8-32% (Table 5.3, $RI = (RMSE_{OK} - RMSE_N)/RMSE_{OK} \times 100$, where $N$ is the respective new method). However, only

very small to no differences could be observed in the accuracy of prediction between the different multivariate interpolation methods (Table 5.3). This is in contrast to a study of Mishra et al. (2010), where a strong improvement of GWR over MLR for the estimation of soil C pools was found. Also Zhang et al. (2011) evaluated the prediction quality of GWR on $C_{org}$ values in Ireland and found an improvement over OK and MLR. The size of the study area, as well as the allocated explanatory variables were however quite different in both studies. The prediction quality of MLR, RK and GWR clearly depend on available auxiliary information and their correlation to the response variables. As MLR already showed good correlations and this multivariate model is subsequently used as a basis of RK and GWR, there is not much range for improvement for RK and GWR over MLR.

### 5.3.4 Maps

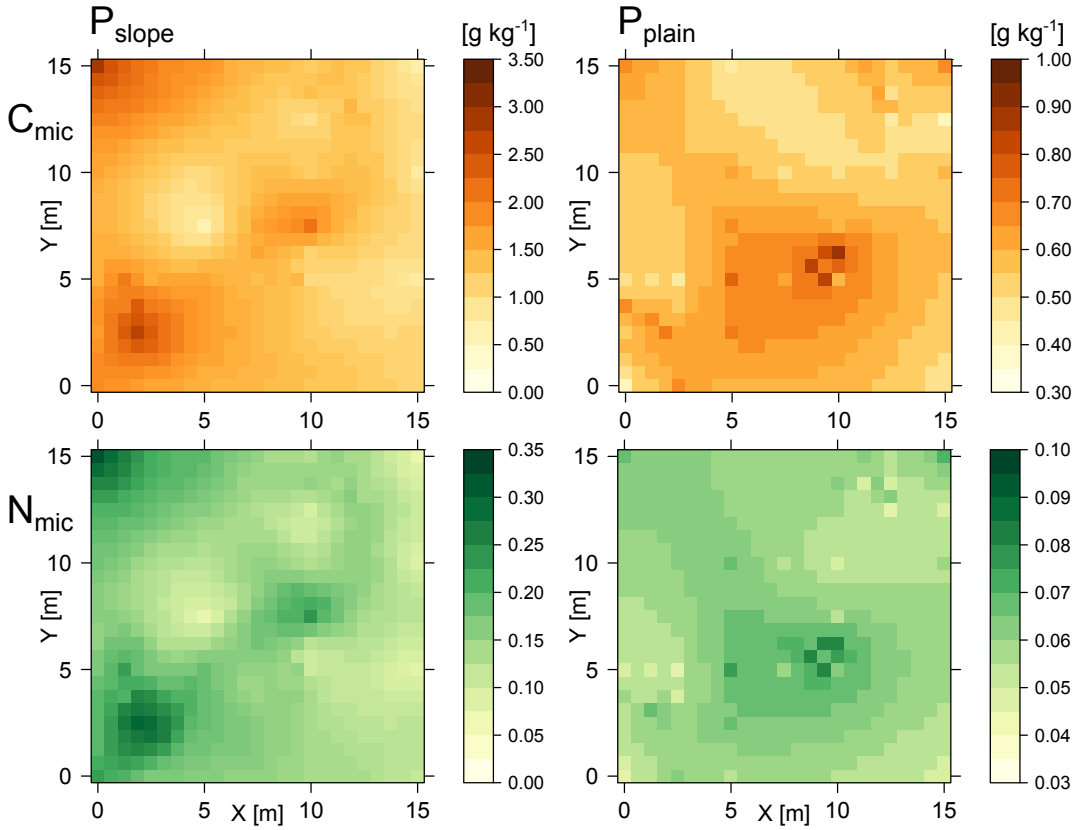



Figure 5.6: Maps of $C_{mic}$ (top) and $N_{mic}$ (bottom) for $P_{slope}$ (left) and $P_{plain}$ (right), produced by the method with the highest $R^2$ (here MLR)

High levels of $C_{mic}$ were observed at the upper-left corner of $P_{slope}$ (Figure 5.6). A possible explanation is the presence of a mature individual of *Sclerocarya birrea*.

Trees and shrubs in savannah ecosystems are known to have a great influence on nutrients and microorganisms (Wang et al., 2009; Vetaas, 1992). In dry savannah ecosystems, the C input through tree litter is restricted to the under canopy areas and trees have a strong influence on the distribution and storage of C in soil (Wang et al., 2009). Compared to the remaining area of $P_{slope}$, different grass species occur within the tree's shading radius and a humus layer developed. Due to the high availability of substrate ($C_{org}$), higher amounts of $C_{mic}$ could be observed. Also the heterogeneous terrain/soil structure of $P_{slope}$ is replicated in the spatial distribution of $C_{mic}$. On this plot, shrubs, grasses and bigger stones act as a protection of erosion and as a result, higher $C_{mic}$ values were observed in these patches, whereas low $C_{mic}$ values were observed close to the areas with adjacent bedrock (Figure 5.6).
$C_{mic}$ values of $P_{plain}$ were generally lower than on $P_{slope}$, the pattern with higher values under trees and bushes however remained the same. In the middle of the plot a cluster of Acacia trees was observed, whereas the area in the upper right was only scarcely covered with grasses. Spatial distribution of $N_{mic}$ showed very similar patterns as $C_{mic}$ on both plots (Figure 5.6).

## 5.4 Conclusion

We have demonstrated that including additional variables ($C_{org}$, N, clay content, pH) improves the spatial prediction of soil microbial parameters such as $C_{mic}$ and $N_{mic}$. The methods MLR, RK and GWR use the specific information provided by the parameters $C_{org}$, N, clay content and soil pH for higher accuracy of local prediction and/or less prediction errors. Because of strong linear correlations between the microbial parameters and the predictor variables, regression kriging, multiple linear regression and geographically weighted regression performed equally and but better than ordinary kriging. In this case the global linear relationships between the predictor variables and $C_{mic}$ and $N_{mic}$ were already quite strong. Regression kriging and geographically weighted regression did thus not provide additional accuracy over MLR. SFA, however, did not show spatial dependencies or relation to the chosen predictor variables on the examined scale and therefore could not be predicted with the multivariate methods. This study suggests multivariate methods for the estimation of soil microbial parameters and ecological interpretation of the local relationships. $C_{org}$, N, clay content and soil pH are seen as suitable variables to predict spatial relations of soil microbial parameters at small scales in natural savannah ecosystems of East Africa. The integration of soil parameters predicted with Vis-NIR DRS proved useful. In combination with geostatistical methods, $C_{mic}$ and $N_{mic}$ can thus be predicted easily and with a minimum amount of laboratory analyses.

## Acknowledgments

This study was funded by the German Research Foundation (DFG) within the Research-Unit 1246 (KiLi) and supported by the Tanzanian Commission for Science and Technology (COSTECH) and the Tanzania Wildlife Research Institute (TAWIRI). For permission to conduct this study in its surrounding area, thanks goes to the Lake Chala Safari Camp. Additionally the authors want to thank T. Leipold, I. Otte, T. Appelhans, A. Ensslin and K. Schmidt for support in the field and in the laboratory.

## References

Akaike, H. (1973). "Information theory and the maximum likelihood principle". In: *2nd International Symposium on Information Theory*. Ed. by B. N. Petrov and F. Csàki. Akademiai Kiado, Budapest, pp. 267–281.

Awiti, A. O., M. G. Walsh, K. D. Shepherd, and J. Kinyamario (Jan. 2008). "Soil condition classification using infrared spectroscopy: A proposition for assessment of soil condition along a tropical forest-cropland chronosequence". In: *Geoderma* 143.1-2, pp. 73–84. DOI: 10.1016/j.geoderma.2007.08.021.

Bivand, R. S. and D. Yu (2013). *spgwr: Geographically weighted regression*.

Bogner, C., A. Kühnel, and B. Huwe (2014). "Predicting with limited data – Increasing the accuracy in VIS-NIR diffuse reflectance spectroscopy by SMOTE". In: *Proceedings of the 6th Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing*. accepted. Lausanne, Switzerland.

Bourennane, H., D. King, and A. Couturier (2000). "Comparison of kriging with external drift and simple linear regression for predicting soil horizon thickness with different sample densities". In: *Geoderma* 97.3-4, pp. 255–271. DOI: 10.1016/S0016-7061(00)00042-2.

Brunsdon, C., A. S. Fotheringham, and M. E. Charlton (1996). "Geographically weighted regression: a method for exploring spatial nonstationarity". In: *Geographical analysis* 28.4, pp. 281–298. DOI: 10.1111/j.1538-4632.1996.tb00936.x.

Chang, C.-W., D. A. Laird, M. J. Mausbach, and C. R. Hurburgh (2001). "Near-Infrared Reflectance Spectroscopy-Principal Components Regression Analyses of Soil Properties". In: *Soil Science Society of America Journal* 65.2, pp. 480–490. DOI: 10.2136/sssaj2001.652480x.

Chawla, N. V., K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer (2002). "SMOTE: Synthetic Minority Over-sampling Technique". In: *Journal of Artificial Intelligence Research* 16, pp. 321–357.

Cressie, N. (1988). "Spatial prediction and ordinary kriging". In: *Mathematical Geology* 20.4, pp. 405–421. DOI: 10.1007/BF00892986.

Davis, B. M. (1987). “Uses and abuses of cross-validation in geostatistics”. In: *Mathematical Geology* 19.3, pp. 241–248. DOI: 10.1007/BF00897749.

Detwiler, R. P. and C. A. S. Hall (1988). “Tropical Forests and the Global Carbon Cycle”. In: *Science* 239.4835, pp. 42–47. DOI: 10.1126/science.239.4835.42.

D’Odorico, P., A. Bhattachan, K. F. Davis, S. Ravi, and C. W. Runyan (2013). “Global desertification: Drivers and feedbacks”. In: *Advances in Water Resources* 51.0, pp. 326–344. DOI: 10.1016/j.advwatres.2012.01.013.

Fotheringham, A. S., C. Brunsdon, and M. Charlton (2002). *Geographically weighted regression: the analysis of spatially varying relationships*. Chichester, UK: John Wiley & Sons, Ltd.

Goovaerts, P. (1997). *Geostatistics for natural resources evaluation*. Oxford: Oxford University Press on Demand.

Grunwald, S. (2005). *Environmental Soil-Landscape Modeling: Geographic Information Technologies and Pedometrics*. Boca Raton, Florida: CRC Press.

Hagos, M. and G. Smit (Apr. 2005). “Soil enrichment by *Acacia mellifera subsp. detinens* on nutrient poor sandy soil in a semi-arid southern African savanna”. In: *Journal of Arid Environments* 61.1, pp. 47–59. DOI: 10.1016/j.jaridenv.2004.08.003.

Hengl, T. (2009). *A Practical Guide to Geostatistical Mapping*. available at lulu.com: 2nd ed., Amsterdam, p. 291.

Hengl, T., G. B. Heuvelink, and A. Stein (2004). “A generic framework for spatial prediction of soil variables based on regression-kriging”. In: *Geoderma* 120.1–2, pp. 75–93. DOI: 10.1016/j.geoderma.2003.08.018.

Hernández-Hernández, R. and D. López-Hernández (Nov. 2002). “Microbial biomass, mineral nitrogen and carbon content in savanna soil aggregates under conventional and no-tillage”. In: *Soil Biology and Biochemistry* 34.11, pp. 1563–1570. DOI: 10.1016/S0038-0717(02)00125-6.

Hiemstra, P. H., E. J. Pebesma, C. J. W. Twenhöfel, and G. B. M. Heuvelink (2009). “Real-time automatic interpolation of ambient gamma dose rates from the Dutch radioactivity monitoring network”. In: *Computers & Geosciences* 35.8, pp. 1711–1721. DOI: 10.1016/j.cageo.2008.10.011.

IUSS Working Group WRB (2007). *World reference base for soil resources 2006, first update 2007. World Soil Resources Reports No. 103*. http://www.fao.org/fileadmin/templates/nr/images/resources/pdf_documents/wrb2007_red.pdf. [Accessed on 2014-04-08].

James, G., D. Witten, T. Hastie, and R. Tibshirani (2013). *An introduction to statistical learning*. Springer.

Joergensen, R. G. (1996). “The fumigation-extraction method to estimate soil microbial biomass: Calibration of the kEC value”. In: *Soil Biology and Biochemistry* 28.1, pp. 25–31. DOI: 10.1016/0038-0717(95)00102-6.

Joergensen, R. G. and T. Mueller (1996). “The fumigation-extraction method to estimate soil microbial biomass: Calibration of the kEN value”. In: *Soil Biology and Biochemistry* 28.1, pp. 33–37. DOI: 10.1016/0038-0717(95)00101-8.

Jones, M. J. (1973). “The organic matter content of the savanna soils of West Africa”. In: *Journal of Soil Science* 24.1, pp. 42–53. DOI: 10.1111/j.1365-2389.1973.tb00740.x.

Kratz, W. (1998). “The bait-lamina test – General aspects, applications and perspectives”. In: *Environmental Science and Pollution Research* 5, pp. 94–96. DOI: 10.1007/BF02986394.

Krige, D. G. (1951). “A Statistical Approach to Some Basic Mine Valuation Problems on the Witwatersrand”. In: *Journal of the Chemical, Metallurgical and Mining Society of South Africa* 52.6, pp. 119–139. DOI: 10.2307/3006914.

Kumar, S., R. Lal, and D. Liu (2012). “A geographically weighted regression kriging approach for mapping soil organic carbon stock”. In: *Geoderma* 189, pp. 627–634. DOI: 10.1016/j.geoderma.2012.05.022.

Kuzyakova, I., V. Romanenkov, and Y. V. Kuzyakov (2001). “Application of geostatistics in processing the results of soil and agrochemical studies”. In: *Eurasian Soil Science c/c Pochvovedenie* 34.11, pp. 1219–1228.

Li, J. and A. D. Heap (2011). “A review of comparative studies of spatial interpolation methods in environmental sciences: Performance and impact factors”. In: *Ecological Informatics* 6.3-4, pp. 228–241. DOI: 10.1016/j.ecoinf.2010.12.003.

Malhi, Y. et al. (2004). “The above-ground coarse wood productivity of 104 Neotropical forest plots”. In: *Global Change Biology* 10.5, pp. 563–591. DOI: 10.1111/j.1529-8817.2003.00778.x.

Matheron, G. (1963). “Principles of geostatistics”. In: *Economic Geology* 58.8, pp. 1246–1266. DOI: 10.2113/gsecongeo.58.8.1246.

Michelsen, A., M. Andersson, M. Jensen, A. Kjøller, and M. Gashew (2004). “Carbon stocks, soil respiration and microbial biomass in fire-prone tropical grassland, woodland and forest ecosystems”. In: *Soil Biology and Biochemistry* 36.11, pp. 1707–1717. DOI: 10.1016/j.soilbio.2004.04.028.

Mishra, U., R. Lal, D. Liu, and M. van Meirvenne (2010). “Predicting the Spatial Variation of the Soil Organic Carbon Pool at a Regional Scale”. In: *Soil Science Society of America Journal* 74.3, pp. 906–914. DOI: 10.2136/sssaj2009.0158.

Mishra, U., M. S. Torn, E. Masanet, and S. M. Ogle (2012). “Improving regional soil carbon inventories: Combining the IPCC carbon inventory method with regression kriging”. In: *Geoderma* 189–190.0, pp. 288–295. DOI: 10.1016/j.geoderma.2012.06.022.

Moernaut, J., D. Verschuren, F. Charlet, I. Kristen, M. Fagot, and M. D. Batist (2010). “The seismic-stratigraphic record of lake-level fluctuations in Lake Challa: Hydrological stability and change in equatorial East Africa over the last 140 kyr”. In: *Earth and Planetary Science Letters* 290.1-2, pp. 214–223. DOI: 10.1016/j.epsl.2009.12.023.

Nonnotte, P., H. Guillou, B. L. Gall, M. Benoit, J. Cotten, and S. Scaillet (2008). “New K – Ar age determinations of Kilimanjaro volcano in the North Tanzanian diverging rift, East Africa”. In: *Journal of Volcanology and Geothermal Research* 173.1–2, pp. 99–112. DOI: 10.1016/j.jvolgeores.2007.12.042.

Pabst, H., A. Kühnel, and Y. Kuzyakov (May 2013). “Effect of land-use and elevation on microbial biomass and water extractable carbon in soils of Mt. Kilimanjaro ecosystems”. In: *Applied Soil Ecology* 67.0, pp. 10–19. DOI: 10.1016/j.apsoil.2013.02.006.

Pebesma, E. J. (Aug. 2004). “Multivariable geostatistics in S: the gstat package”. In: *Computers & Geosciences* 30.7, pp. 683–691. DOI: 10.1016/j.cageo.2004.03.012.

Plath, M., K. Mody, C. Potvin, and S. Dorn (Feb. 2011). “Establishment of native tropical timber trees in monoculture and mixed-species plantations: Small-scale effects on tree performance and insect herbivory”. In: *Forest Ecology and Management* 261.3, pp. 741–750. DOI: 10.1016/j.foreco.2010.12.004.

R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*. ISBN 3-900051-07-0. R Foundation for Statistical Computing. Vienna, Austria.

Rogerson, P. A. (2001). *Statistical methods for geography*. London: SAGE Publications.

Ruess, R. and S. McNaughton (1987). “Grazing and the dynamics of nutrient and energy regulated microbial processes in the Serengeti grasslands”. In: *Oikos*, pp. 101–110.

Sankaran, M. and J. Ratnam (2013). “African and Asian Savannas”. In: *Encyclopedia of Biodiversity (Second Edition)*. Ed. by S. A. Levin. Waltham: Academic Press, pp. 58–74.

Singh, J. S., A. S. Raghubanshi, R. S. Singh, and S. C. Srivastava (Apr. 1989). “Microbial biomass acts as a source of plant nutrients in dry tropical forest and savanna”. In: *Nature* 338.6215, pp. 499–500. DOI: 10.1038/338499a0.

Tamme, R., I. Hiiesalu, L. Laanisto, R. Szava-Kovats, and M. Pärtel (2010). “Environmental heterogeneity, species diversity and co-existence at different spatial scales”. In: *Journal of Vegetation Science* 21.4, pp. 796–801. DOI: 10.1111/j.1654-1103.2010.01185.x.

Tews, J., U. Brose, V. Grimm, K. Tielbörger, M. C. Wichmann, M. Schwager, and F. Jeltsch (2004). “Animal species diversity driven by habitat heterogeneity/diversity: the importance of keystone structures”. In: *Journal of Biogeography* 31.1, pp. 79–92. DOI: 10.1046/j.0305-0270.2003.00994.x.

Thompson, J. A. and R. K. Kolka (July 2005). “Soil Carbon Storage Estimation in a Forested Watershed using Quantitative Soil-Landscape Modeling”. In: *Soil Science Society of America Journal* 69.4, pp. 1086–1093. DOI: 10.2136/sssaj2004.0322.

Toerne, E. von (1990). “Assessing feeding activities of soil-living animals. I: Bait-lamina-tests”. In: *Pedobiologia*. Vol. 34 (2). Urban & Fischer. Chap. Assessing feeding activities of soil-living animals - 1. Bait-lamina-tests, pp. 89–101.

Torgo, L., R. P. Ribeiro, B. Pfahringer, and P. Branco (2013). “SMOTE for Regression”. In: *Progress in Artificial Intelligence*. Springer, pp. 378–389.

Traoré, S., L. Thiombiano, J. R. Millogo, and S. Guinko (Mar. 2007). “Carbon and nitrogen enhancement in Cambisols and Vertisols by Acacia spp. in eastern Burkina Faso: Relation to soil respiration and microbial biomass”. In: *Applied Soil Ecology* 35.3, pp. 660–669. DOI: 10.1016/j.apsoil.2006.09.004.

Vetaas, O. R. (1992). “Micro-site effects of trees and shrubs in dry savannas”. In: *Journal of Vegetation Science* 3.3, pp. 337–344. DOI: 10.2307/3235758.

Viscarra Rossel, R., D. Walvoort, A. McBratney, L. Janik, and J. Skjemstad (2006b). “Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties”. In: *Geoderma* 131.1-2, pp. 59–75. DOI: 10.1016/j.geoderma.2005.03.007.

Viscarra Rossel, R. and T. Behrens (Aug. 2010). “Using data mining to model and interpret soil diffuse reflectance spectra”. In: *Geoderma* 158.1-2, pp. 46–54. DOI: 10.1016/j.geoderma.2009.12.025.

Wang, L., G. S. Okin, K. K. Caylor, and S. A. Macko (Mar. 2009). “Spatial heterogeneity and sources of soil carbon in southern African savannas”. In: *Geoderma* 149.3-4, pp. 402–408. DOI: 10.1016/j.geoderma.2008.12.014.

Webster R. & McBratney, A. B. (1989). “On the Akaike Information Criterion for choosing models for variograms of soil properties”. In: *Journal of Soil Science* 40.3, pp. 493–496. DOI: 10.1111/j.1365-2389.1989.tb01291.x.

Wiens, J. (2000). “Ecological heterogeneity: an ontogeny of concepts and approaches”. In: *The ecological consequences of environmental heterogeneity. The 40th symposium of the British Ecological Society*. Ed. by E. Hutchings M.J.and John and A. Stewart. Univeristy of Sussex: Blackwell Science, Oxford. Chap. Ecological heterogeneity: an ontogeny of concepts and approaches, pp. 9–32.

Wilson, S. (2000). “Heterogeneity, diversity and scale in plant communities”. In: *The ecological consequences of environmental heterogeneity. The 40th symposium of the British Ecological Society*. Ed. by M. Hutchings, E. John, and A. Stewart. Univeristy of Sussex: Blackwell Science, Oxford. Chap. Heterogeneity, diversity and scale in plant communities, pp. 53–69.

Wold, S., M. Sjöström, and L. Eriksson (2001). “PLS-regression: a basic tool of chemometrics”. In: *Chemometrics and intelligent laboratory systems* 58.2, pp. 109–130. DOI: 10.1016/S0169-7439(01)00155-1.

Zhang, C., Y. Tang, X. Xu, and G. Kiely (July 2011). “Towards spatial geochemical modelling: Use of geographically weighted regression for mapping soil organic carbon contents in Ireland”. In: *Applied Geochemistry* 26.7, pp. 1239–1248. DOI: 10.1016/j.apgeochem.2011.04.014.

# Appendix A

## In situ prediction of soil chemical properties with visible and near infrared spectroscopy in an African savannah

ANNA KÜHNEL[1], CHRISTINA BOGNER[2] AND BERND HUWE[1]

[1]Soil Physics Group, BayCEER, University of Bayreuth, Germany
[2]Ecological Modelling, BayCEER, University of Bayreuth, Germany

published in

*GlobalSoilMap: Basis of the global spatial soil information system*, 409-413, CRC Press (2014)

**Abstract**

The performance of a general spectral model, developed for the Kilimanjaro region, was tested in a dry savannah ecosystem east of Mt. Kilimanjaro, Tanzania. Soil organic carbon and nitrogen were predicted with visible to near-infrared spectroscopy on spectra that were taken directly in the field and on standardised samples taken under laboratory conditions. As this general model did not perform well for the field spectra, we used a spiking approach, that could improve the predictions.

## Introduction

Savannah ecosystems cover a huge area of the worlds surface and its dynamics and complexity have been studied by various authors (Marchant, 2010; Kashaigili et al., 2010; Groen et al., 2011). Around Mt. Kilimanjaro they are restricted to small areas in the East and West and are threatened by the conversion into arable land. The conversion of these natural ecosystems often results in degradation of soil quality and altered ecosystem functions like water and carbon storage or erosion control. Fast and accurate measurements of soil parameters are required in order to infer the implications of changes to the soil and the ecosystem functioning. Visible (VIS) and near-infrared (NIR) spectroscopy is a fast and easy method to obtain soil characteristic information at low costs and is especially useful in sub-Saharan ecosystems, where soil information is scarce. It has been widely used under laboratory conditions (Viscarra Rossel et al., 1998; Viscarra Rossel et al., 2006b; Chang et al., 2001; Awiti et al., 2008). Field applications however are not yet as reliable (Morgan et al., 2009; Nocita et al., 2011), probably due to changing soil moisture and surface conditions. In this study we tested i) the prediction quality of a global calibration model, that was developed for soil carbon and nitrogen of the Mt. Kilimanjaro region, directly in the field and ii) if spiking this model with field spectra from the respective site can lead to improved predictions. Furthermore we tested the application of the NSMI (Haubrock et al., 2008) to asses the influence of water content.

## Methods

### Study area

The sampling was conducted on a hierarchically nested grid design on two 15x15m plots, consisting of 61 sampling points each, in a natural Savanna ecosystem east of Mt. Kilimanjaro, Tanzania. One research plot was situated on the outer foot slope of the crater rim of the Lake Chala caldera at an elevation of 990m. The soil is very shallow with a maximum depth of 25cm, with bedrock appearing directly at the surface in some parts of the area. It was classified as Rendzic Leptosol (Calcaric, Tephric, Sodic, Eutric, Skeletic) according to IUSS Working Group WRB (2007). The other research plot was situated 400m apart, in the plains surrounding Lake Chala at an elevation of 950m, with deeply developed soil that was classified as Sodic Vertisol (Hypereutric, Chromic).

### Field campaign and laboratory analysis

On each plot, soil surface reflectance spectra and soil moisture content were recorded for every grid point, at 5 am before sunrise (night) and at 12 am (day). Soil surface reflectance spectra were taken with an ASD field spectrometer with a contact probe attachment using an internal light source. A small plastic ring, that was attached to the contact probe, enabled us to put the probe directly to the surface, while eliminating the influence of external light. At the grid points we took three spectra each, taking care of putting the contact probe directly to the soil surface. Before measurement fallen leaves and dried grass were removed and if there were tufts of grass or big stones, we searched for a point with bare soil in the direct neighbourhood. Every half hour the instrument was calibrated with a white reference spectra and 30 reflectance spectra were averaged at each point. After the spectral measurements, a mixed soil sample of the upper 5 cm was collected at each point, air-dried and sieved to 2 mm. Subsequently a well-mixed aliquot of the sieved and air-dried samples was placed in a small cup and the same measuring procedure as in the field was applied. Carbon and nitrogen content were measured using a CNS-Analyser by high temperature combustion with conductivity detectors.

### Pre-processing of soil spectra

An independent dataset with 150 soil samples from different sites around Mt. Kilimanjaro was used to predict C and N content. The following pre-processing steps were performed:

1. jump correction for the detector offsets
2. cutting the edges of the spectra to wavelengths from 500 to 2400 nm
3. range transformation by dividing the reflectance values through the maximum reflectance
4. taking the 1.derivative
5. splitting into calibration dataset (5/6 of the data) and validation dataset (1/6 of the data)

We used partial least square regression analysis on the calibration dataset for model development and tested different models on the validation dataset. The number of components for the best model, were chosen based on the root mean squared error of prediction ($RMSE$, Equation 1.1) of the validation dataset. To validate the model we also used the $R^2$ as calculated with Equation 1.2 and residual

prediction deviation or ratio of percent deviation $RPD$ (Equation 1.3). The possible values of $R^2$ range from $-\infty$ to 1, with values closer to 1 indicate more accurate models. An $R^2$ of 0 signifies that the model is just as good as the mean of the observed data. $RPD$ is a factor that indicates by how much the prediction accuracy of the model has been increased compared to the mean of the data (Viscarra Rossel et al., 2006a).

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(p_i - o_i)^2} \tag{1.1}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(p_i - o_i)^2}{\sum_{i=1}^{N}(o_i - \hat{o}_i)^2} \tag{1.2}$$

$$RPD = \frac{\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(o_i - \hat{o}_i)^2}}{RMSE} \tag{1.3}$$

$p_i$ = predicted values
$o_i$ = observed values
$\hat{o}_i$ = mean of the observed values
$N$ = number of observations

### Spiking

We therefore tried an approach that was presented by Shepherd et al. (2002) and is now known as spiking (Viscarra Rossel et al., 2009; Guerrero et al., 2010). The spectral library for model development is "spiked" with samples from the target area. The new calibration model should now contain the characteristics from the target spectra and should predict better. The usual practice is to spike the database with spectra that were obtained under the same conditions as the calibration model. In this study we also used field spectra and spiked the calibration model accordingly. We had six different spectral datasets, consisting of 16 samples each, were carbon and nitrogen content is known (lab,night and day for each plot). For every spiking approach we used five of these datasets to supplement the calibration model and predicted the sixths dataset. A relative improvement factor for $R^2$ and $RMSEP$ was calculated, Rel = (new value - old value ) / old value *100% (Table A.3).

Table A.1: Summary statistics of the input parameter for the models

| Carbon (mg $g^{-1}$) | Min | Med | Mean | Max |
|---|---|---|---|---|
| Calibration | 5.0 | 18.2 | 24.8 | 89.9 |
| Validation | 10.6 | 22.6 | 27.6 | 76.0 |
| Nitrogen (mg $g^{-1}$) | Min | Med | Mean | Max |
| Calibration | 0.48 | 1.67 | 2.22 | 7.59 |
| Validation | 0.46 | 1.36 | 1.92 | 6.71 |

## Results and Discussion

### Model calibration

Carbon content for the model calibration ranged from 5 mg $g^{-1}$ to almost 90 mg $g^{-1}$, nitrogen content from 0.48 to 7.59 mg $g^{-1}$, respectively (Table A.1). The values for the validation model are within or very close to this range.

The respective error parameters are shown in Table A.2. Viscarra Rossel et al. (2006a) classified the $RPD$ values, where $RPD$ < 1.0 indicates very poor models and values > 2.5 indicate excellent models. Considering this classification, our models performed well for C an N with $RPD$ values of 2.26 and 2.25, respectively. These models however are only valid for air-dried and sieved soils, the application in the field has to be tested individually, as for example soil moisture and surface properties have an influence on the reflectance spectra (Lobell et al., 2002; Chang et al., 2001).

Table A.2: Parameter of the calibration model and the prediction of the validation dataset; n = number of chosen model components, $RMSE$ = root mean squared error, $R^2$ = model efficiency calculated with Equation 1.2, $RPD$ = residual prediction deviation

| Parameter | | calibration | | | validation | | |
|---|---|---|---|---|---|---|---|
| | n | $RMSE$ | $R^2$ | $RPD$ | $RMSE$ | $R^2$ | $RPD$ |
| C | 6 | 8.7 | 0.80 | 2.26 | 6.8 | 0.84 | 2.8 |
| N | 7 | 0.71 | 0.80 | 2.25 | 0.65 | 0.84 | 2.53 |

### Spectra

The processed spectra for the individual plots and sampling times (in the laboratory, at night and during the day) are displayed in Figure A.1. The spectra taken in the

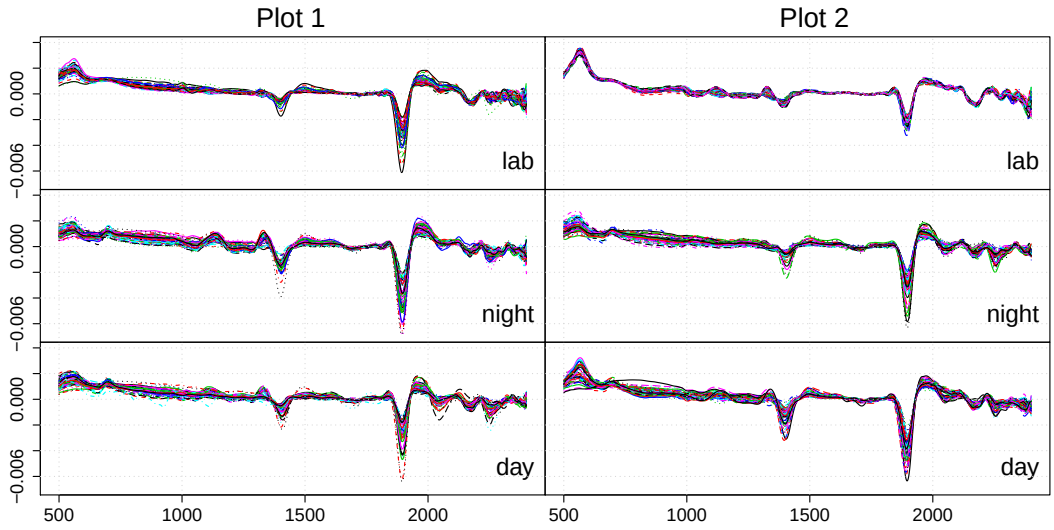



Figure A.1: Processed soil spectra of the dried soil samples (lab), the spectra taken at night (night) and the spectra taken during the day (day).

field show a lot more variation and more pronounced peaks than the spectra in the laboratory. Some individual spectra do not follow well the characteristic curve for soils, indicating that partly stones or dead leaves were sampled instead of bare soil.

### Predictions

Figure A.2 shows the predicted vs. the observed values for carbon for the six different datasets. As expected the predictions from field spectra are unsatisfactory with $R^2 < 0$. Unfortunately the predictions for the laboratory samples are also only good for plot 2. The predictions for plot 1 lead to very high $RMSE$ values.

### Spiking

The predictions with spiked models for the field spectra are still very poor ($RPD < 1$, Table A.4). One reason for this is, that the $RPD$ value is dependent on the range of the target parameter, with small ranges leading to low $RPD$ values. Another problem of our sampling approach is that we took an area of about 10 $cm^2$ around the grid point into account. Possibly there is already quite a high variation of the soil parameters within this area. Furthermore, the soil that was used for the laboratory analysis is a mixed sample of the upper 5 cm, whereas the spectra in the field were only taken on the surface of the soil. We could however show, that spiking has a positive effect on the prediction error, with higher $R^2$ and lower $RMSEP$. The Rel is positive for almost all datasets, except for the dried samples of Plot 2, where the predictions with the general calibration model were already very good.

Table A.3: Comparison of the old models, that were developed for dried soils and the spiked models (new), $Rel_r$= relative improvement factor of $RMSEP$ (%), $Rel_{R^2}$= relative improvement factor of $R^2$ (%)

| | Carbon | | | | | | Nitrogen | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | old | | new | | $Rel_{R^2}$ | $Rel_r$ | old | | new | | $Rel_{R^2}$ | $Rel_r$ |
| | $R^2$ | $RMSEP$ | $R^2$ | $RMSEP$ | | | $R^2$ | $RMSEP$ | $R^2$ | $RMSEP$ | | |
| Lab Plot1 | -1.07 | 16.8 | 0.40 | 9.03 | 138 | 46 | -0.67 | 1.3 | 0.42 | 0.77 | 163 | 41 |
| Lab Plot2 | 0.38 | 3.7 | 0.36 | 3.80 | -5 | -3 | -0.41 | 0.45 | 0.73 | 0.20 | 278 | 56 |
| Day Plot1 | -8.30 | 35.6 | -0.64 | 14.97 | 92 | 58 | -6.41 | 2.76 | -5.94 | 1.28 | 7 | 54 |
| Day Plot2 | -28.50 | 25.8 | -5.53 | 11.98 | 81 | 54 | -32.10 | 2.17 | -9.18 | 1.20 | 71 | 45 |
| Night Plot1 | -3.09 | 23.6 | -0.85 | 15.89 | 72 | 33 | -2.01 | 1.76 | -0.77 | 1.35 | 62 | 23 |
| Night Plot2 | -49.70 | 33.9 | -1.39 | 7.36 | 97 | 78 | -57.30 | 2.88 | -2.34 | 0.69 | 96 | 76 |

Table A.4: $RPD$ values of the predictions and number of components used for the spiked models

| | Carbon | | Nitrogen | |
|---|---|---|---|---|
| | $RPD$ | comp | $RPD$ | comp |
| Lab Plot1 | 1.34 | 4 | 1.36 | 6 |
| Lab Plot2 | 0.82 | 2 | 1.97 | 7 |
| Night Plot1 | 0.76 | 3 | 0.77 | 7 |
| Night Plot2 | 0.67 | 3 | 0.56 | 3 |
| Day Plot1 | 0.77 | 4 | 0.82 | 4 |
| Day Plot2 | 0.41 | 3 | 0.32 | 3 |

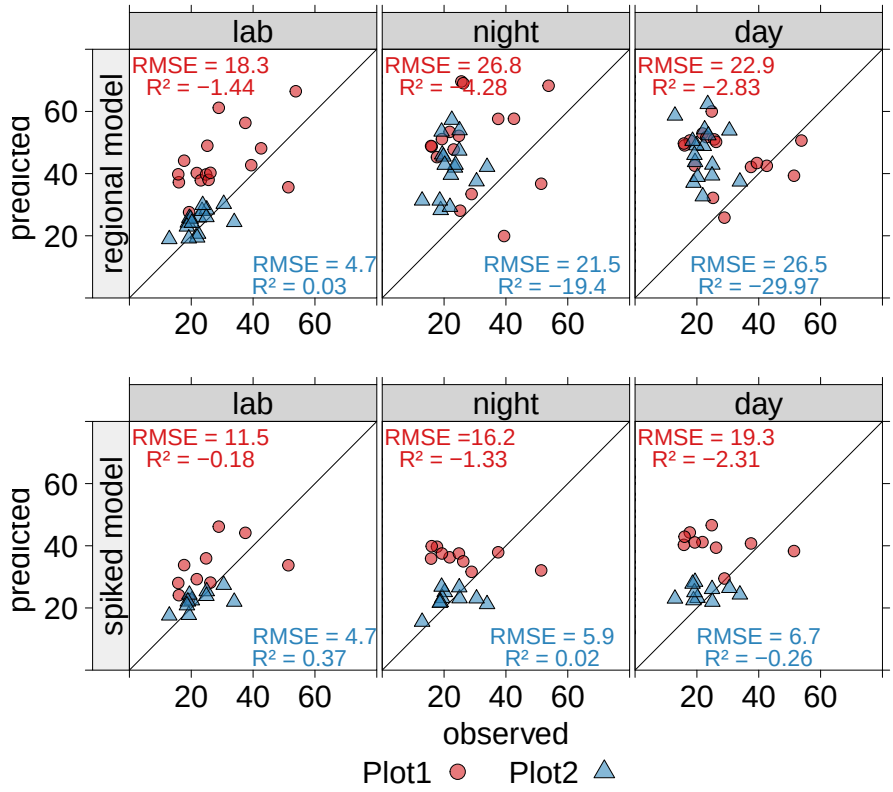



Figure A.2: Observed vs. predicted values of carbon content (mg $g^{-1}$); lab = soil spectra measured on the dry soil, night = soil spectra measured in the field at night, day = soil spectra measured in the field during the day

### NSMI

In order to asses the influence of the water content, it is desirable to get it directly from the spectra without having to measure it simultaneously. Haubrock et al. (2008) developed the normalized soil moisture index (NSMI), that should be insensitive to soil type and should be correlated to the gravimetric soil moisture content. It is a dimensionless parameter based only on the relation between the reflectance values at 1800 and 2119 nm wavelength. We calculated this index for all the field spectra and related it to volumetric soil moisture content, as we didn't measure the gravimetric moisture content (Figure A.3). For our field site however we could not find any correlations between soil moisture and NSMI. The Vertisols and Leptosols from our study site probably do not show the same characteristics as the samples used in the study of Haubrock et al. (2008). They analysed different substrates to show the effect of soil type, their samples however, came from only one field site in Germany.

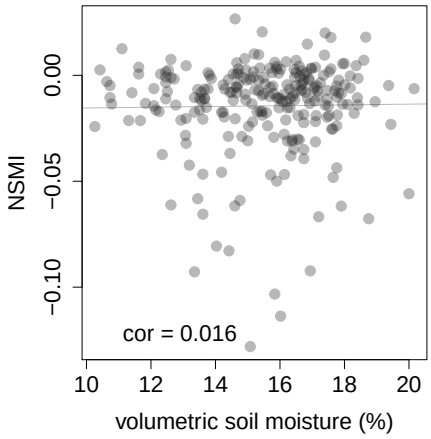



Figure A.3: NSMI vs. volumetric soil moisture for both plots and times

## Conclusions

Using a small amount of soil samples, we developed a global calibration model for soil carbon and nitrogen for the lower Mt. Kilimanjaro area and its adjacent savannahs. This model however, is only applicable for general predictions within this area. Small differences that occur within one individual field site might not be predicted sufficiently precise, as it was the case for our study area. Spiking this global model with local spectra from the target field site improved the predictions.

## Acknowledgements

We thank Juliane Röder and Holger Pabst for support during field work. This study was funded by the German Research Foundation (DFG) within the Research-Unit 1246 (KiLi) and supported by the Tanzanian Commission for Science and Technology (COSTECH) and the Tanzania Wildlife Research Institute (TAWIRI).

## References

Awiti, A. O., M. G. Walsh, K. D. Shepherd, and J. Kinyamario (Jan. 2008). “Soil condition classification using infrared spectroscopy: A proposition for assessment of soil condition along a tropical forest-cropland chronosequence”. In: *Geoderma* 143.1-2, pp. 73–84. DOI: 10.1016/j.geoderma.2007.08.021.

Groen, T., F. van Langevelde, C. van de Vijver, A. de Raad, J. de Leeuw, and H. Prins (Aug. 2011). “A continental analysis of correlations between tree patterns in African savannas and human and environmental variables”. In: *Journal of Arid Environments* 75.8, pp. 724–733. DOI: 10.1016/j.geoderma.2010.02.001.

Haubrock, S. N., S. Chabrillat, C. Lemmnitz, and H. Kaufmann (Jan. 2008). “Surface soil moisture quantification models from reflectance data under field conditions”. In: *International Journal of Remote Sensing* 29.1, pp. 3–29. DOI: 10.1080/01431160701294695.

IUSS Working Group WRB (2007). *World reference base for soil resources 2006, first update 2007. World Soil Resources Reports No. 103*. http://www.fao.org/fileadmin/templates/nr/images/resources/pdf_documents/wrb2007_red.pdf. [Accessed on 2014-04-08].

Kashaigili, J. and A. Majaliwa (2010). “Integrated assessment of land use and cover changes in the Malagarasi river catchment in Tanzania”. In: *Physics and Chemistry of the Earth, Parts A/B/C* 35, pp. 730–741. DOI: 10.1016/j.pce.2010.07.030.

Marchant, R. (2010). “Understanding complexity in savannas: climate, biodiversity and people”. In: *Current opinion in environmental sustainability* 2.1-2, pp. 101–108. DOI: 10.1016/j.cosust.2010.03.001.

Morgan, C. L., T. H. Waiser, D. J. Brown, and C. T. Hallmark (2009). “Simulated in situ characterization of soil organic and inorganic carbon with visible near-infrared diffuse reflectance spectroscopy”. In: *Geoderma* 151.3-4, pp. 249–256. DOI: 10.1016/j.geoderma.2009.04.010.

Nocita, M., L. Kooistra, M. Bachmann, A. Müller, M. Powell, and S. Weel (Nov. 2011). “Predictions of soil surface and topsoil organic carbon content through the use of laboratory and field spectroscopy in the Albany Thicket Biome of Eastern Cape Province of South Africa”. In: *Geoderma* 167-168.0, pp. 295–302. DOI: 10.1016/j.geoderma.2011.09.018.

Viscarra Rossel, R., R. McGlynn, and A. McBratney (Dec. 2006a). “Determining the composition of mineral-organic mixes using UV–vis–NIR diffuse reflectance spectroscopy”. In: *Geoderma* 137.1-2, pp. 70–82. DOI: 10.1016/j.geoderma.2006.07.004.

# Appendix B

## Supplementary Material to Manuscript 3 "In-situ prediction of soil organic carbon by VIS-NIR spectroscopy with limited data"

Table B.1: Overview of analysed data sets.

| Name | Scanned in | SOC content | Land use zone or profile | Number of samples |
|---|---|---|---|---|
| *regional*[a] | laboratory[b] | yes | Homegarden | 41 |
| | | | Coffee | 31 |
| | | | Grasslands | 39 |
| | | | Maize | 29 |
| | | | Savannah | 43 |
| *profile*[c] | in-situ | yes | Hom | 8 |
| | | | Cof | 12 |
| | | | Gra | 12 |
| | | | Mai | 9 |
| | | | Sav1 | 13 |
| | | | Sav2 | 9 |
| *raster* | in-situ | no | Hom | 578 |
| | | | Cof | 577 |
| | | | Gra | 560 |
| | | | Mai | 443 |
| | | | Sav1 | 544 |
| | | | Sav2 | 195 |
| *moist* | laboratory[d] | yes | Homegarden | 11 |

[a] comprised of five *local* data sets [b] oven-dried at 45 °C for 24 h [c] used to generate synthetic data and to test the models
[d] re-moistened and scanned during drying

**Algorithm:** SMOTE

**Input**: $T$ original samples to be smoted
Amount of SMOTE $N$%
Number of nearest neighbours $k$

**Output**: $(N/100) \times T$ synthetic samples with their target values (i.e. concentrations)

**if** $N < 100$ **then**
  Randomize the $T$ original samples:
  $T = (N/100) \times T$
  $N = 100$
**end**

$orig.s[i]$: original sample $i, i = 1, \ldots, T$

$orig.t[i]$: target value of original sample $i$

$new.s[j]$: synthetic sample $j, j = 1, \ldots, (N/100) \times T$

$new.t[j]$: target values of synthetic sample $j$

$ng \leftarrow N/100$: number of synthetic samples to compute for each original sample

Generate synthetic samples:

**for** $i$ *in* $1$ *to* $T$ **do**
  $nns \leftarrow$ compute $k$ nearest neighbours for $orig.s[i]$
  **for** $\ell$ *in* $1$ *to* $ng$ **do**
    randomly choose $x \in nns$
    $diff = orig.s[i] - x$
    $new.s[(i-1) \times ng + \ell] = orig.s[i] + \text{RANDOM}(0,1) \times diff$
    $d_1 = \text{DIST}(new.s, orig.s[i])$
    $d_2 = \text{DIST}(new.s, x)$
    $target = \frac{d_2 \times orig.t(orig.s) + d_1 \times orig.t(x)}{d_1 + d_2}$
  **end**
**end**

**return** $new.t \cup new.s$

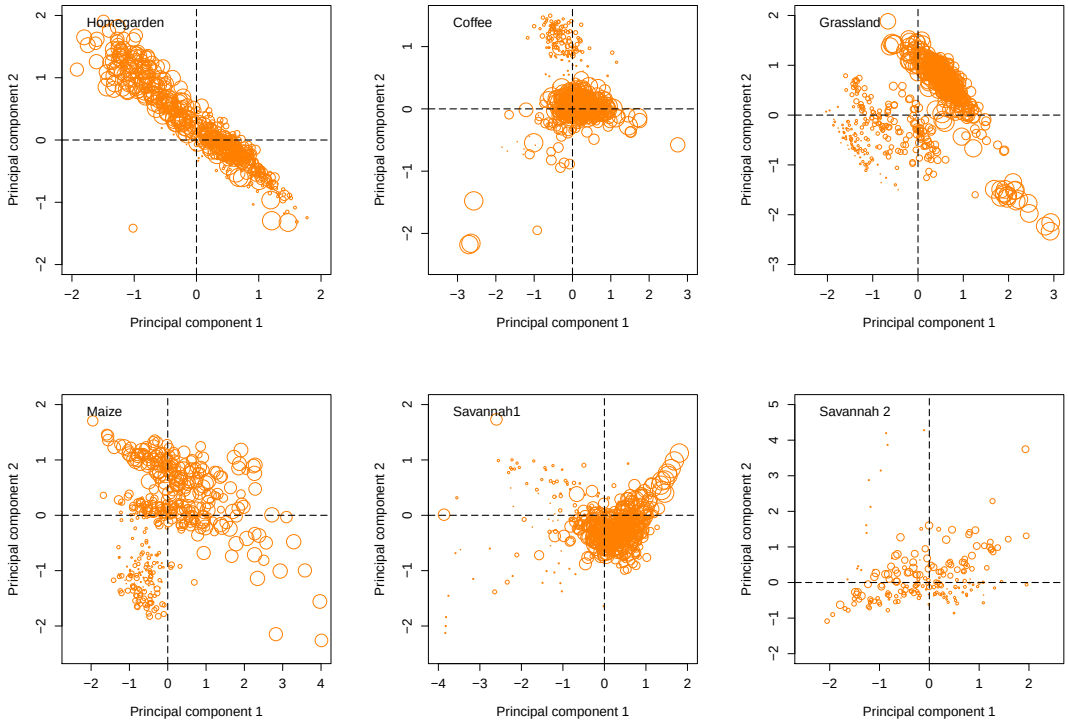



Figure B.1: Principal component analysis of the in-situ *raster* data sets. The symbol size is scaled according to the depth within the profile with smaller points closer to the surface.

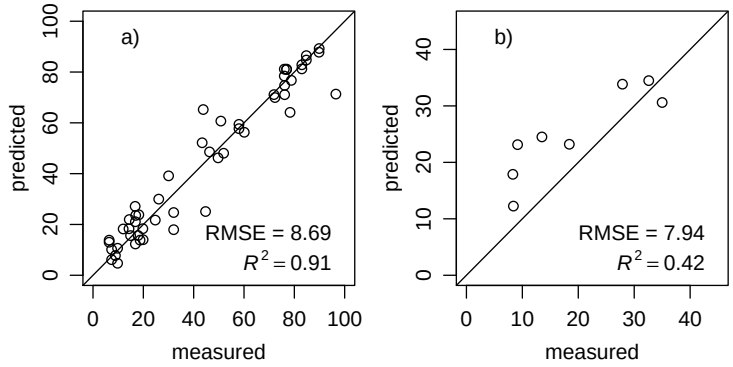



Figure B.2: Predicted versus measured SOC content (mg $g^{-1}$) in the *Loc_moist* model in Homegarden (local model spiked with moist data): a) calibration by leave-one-out cross-validation b) validation on the *profile* data set.

# Appendix C

## Supplementary Material to Manuscript 4 "Small scale spatial variability of soil hydraulic properties in different land uses at Mt. Kilimanjaro"

### Spectral model development

#### Spectral pre-treaments

Every spectrum was corrected for the detector offsets with the additive method (Dorigo et al., 2006) and was cut at the edges, so that only wavelengths with a high signal-to-noise ratio were kept (450–2400 nm). The remaining spectrum was then smoothed by singular spectrum analysis (SSA,(Broomhead et al., 1986; Golyandina et al., 2013)) with a window length of five.

Different standard pre-treatments were tested for reflectance as well as for absorbance ($A = \log(1/\text{reflectance})$) values, namely z-transformation, 1st derivative of z-transformed data, normalization by the maximum value and 1st derivative of data normalized by the maximum. The best combination of pre-processing steps dependent on the model and different best pre-treatments were observed. However, we chose the same pre-treatment steps for all models, as they performed best for the majority of models and $RMSE$ and $R^2$ were only slightly increased/decreased for the remaining models. These steps were as follows: absorbance, normalization by the maximum and the 1st derivative.

## SMOTE

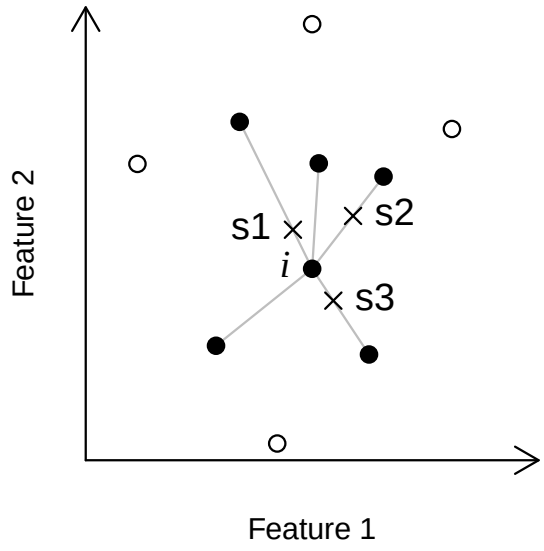



Figure C.1: Illustration of the synthetic minority oversampling technique (SMOTE) in two dimensions. The $k$ nearest neighbours (black dots) are chosen for an existing point $i$ to generate synthetic points (crosses denoted $s1$ through $s3$) along the connection lines between $i$ and its nearest neighbours. Circles show samples that are not the $k$ nearest neighbours of $i$.

**Algorithm:** SMOTE

**Input**: $T$ original samples to be smoted
Amount of SMOTE $N$%
Number of nearest neighbours $k$

**Output**: $(N/100) \times T$ synthetic samples with their target values (i.e. concentrations)

**if** $N < 100$ **then**
  Randomize the $T$ original samples:
  $T = (N/100) \times T$
  $N = 100$
**end**

$orig.s[i]$: original sample $i, i = 1, \ldots, T$

$orig.t[i]$: target value of original sample $i$

$new.s[j]$: synthetic sample $j, j = 1, \ldots, (N/100) \times T$

$new.t[j]$: target values of synthetic sample $j$

$ng \leftarrow N/100$: number of synthetic samples to compute for each original sample

Generate synthetic samples:

**for** $i$ *in* $1$ *to* $T$ **do**
  $nns \leftarrow$ compute $k$ nearest neighbours for $orig.s[i]$
  **for** $\ell$ *in* $1$ *to* $ng$ **do**
    randomly choose $x \in nns$
    $diff = x - orig.s[i]$
    $new.s[(i-1) \times ng + \ell] = orig.s[i] + \mathrm{RANDOM}(0,1) \times diff$
    $d_1 = \mathrm{DIST}(new.s, orig.s[i])$
    $d_2 = \mathrm{DIST}(new.s, x)$
    $target = \frac{d_2 \times orig.t(orig.s) + d_1 \times orig.t(x)}{d_1 + d_2}$
  **end**
**end**

**return** $new.t \cup new.s$

We chose $N = 300$ and set the number of nearest neighbours to five, according to the study of Bogner et al., 2014. That means, that for every spectra three new spectra and their corresponding target values were generated using randomly one of the five nearest neighbours (without replacment). The target value of the synthetic point was then calculated as the weighted average of the target value of point $i$ and the used nearest neighbour.

As the prediction accuracy varies between different synthetic data sets (Bogner et al., 2014), we used a Monte Carlo simulation and created 100 different synthetic datasets. Among the 100 models, we chose the one that produced the minimum root mean squared error ($RMSE$) of predictions on the profile reference data set out of those datasets where the number of model parameters was the median of all

100 datasets.

## Model calibration and testing

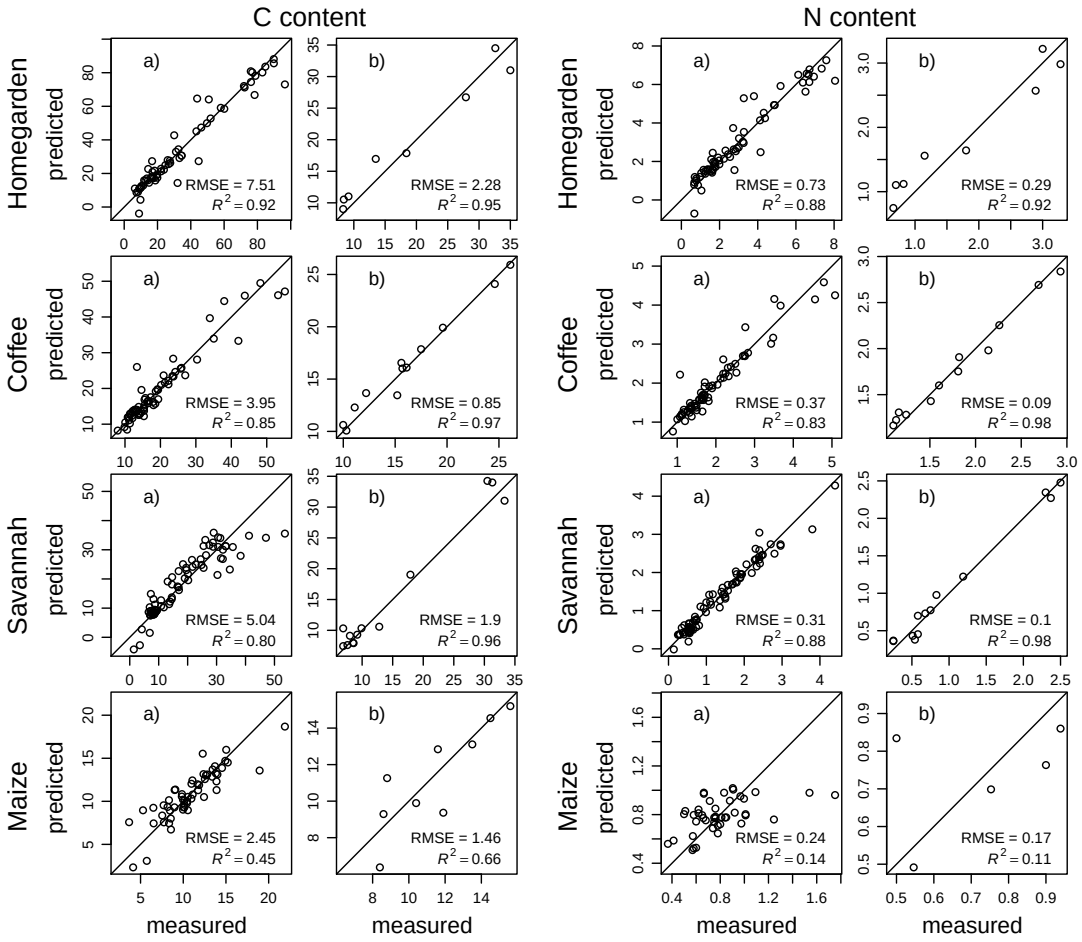



Figure C.2: Predicted versus measured C and N content (mg $g^{-1}$) of partial least squares regression models; a) = model calibration with leave-one-out cross validation, b) = test for the small soil core reference samples.

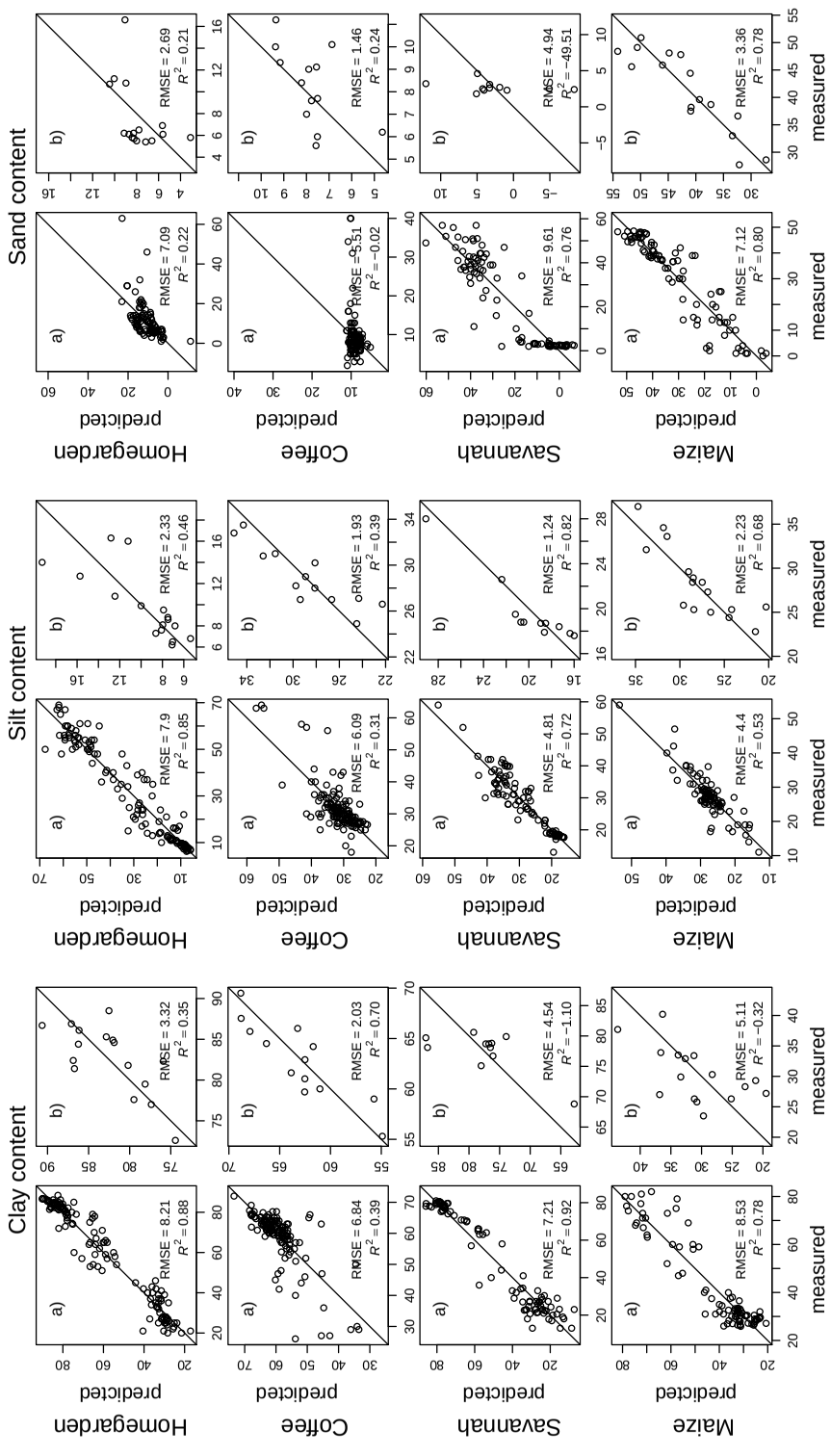



Figure C.3: Prediction versus measured soil texture values (%) of partial least squares regression models ; a) = model calibration with leave-one-out cross validation, b) = test for the small soil core reference samples.

## Pedotransfer function

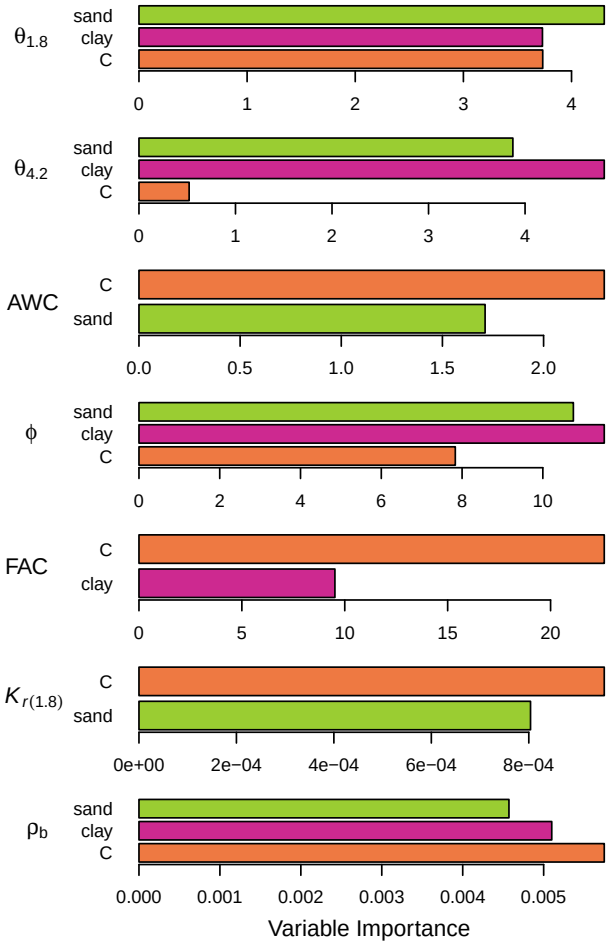



Figure C.4: Importance of the input parameter in the random forest models; $\theta$ = water content at the respective pF value, *AWC* = available field capacity, $\rho_b$ = bulk density, $\phi$ = porosity, $K_r$ = relative hydraulic conductivity, *FAC* = field air capacity.

Table C.1: Data table for estimation of pedotransfer functions with random forest; means of five soil cores, C= carbon conten, N = nitrogen content, $\theta$ = water content at different matric potentials, $\rho_b$ = bulk density, $\phi$ = porosity, *AWC* = available water capacity, $K_{r(1.8)}$ = relative hydraulic conductivity at a matric potential of $10^{1.8}$ hPa, *FAC* = field air capacity

| Plot | horizon | depth | C | N | clay | silt | sand | $\theta_0$ | $\theta_{0.5}$ | $\theta_1$ | $\theta_{1.5}$ | $\theta_2$ | $\theta_{4.2}$ | $\rho_b$ | $\phi$ | $\theta_{1.8}$ | *AWC* | $K_{r(1.8)}$ | *FAC* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Cof | A | 4 | 22.27 | 2.12 | 61 | 33 | 6 | 54.6 | 52.9 | 50.7 | 43.9 | 36.7 | 24.4 | 0.99 | 55.4 | 39.4 | 15.0 | 0.085 | 16.0 |
| | B1 | 20 | 16.63 | 1.66 | 62 | 31 | 7 | 52.9 | 52.5 | 51.0 | 47.0 | 40.1 | 29.5 | 1.08 | 52.0 | 42.8 | 13.3 | 0.141 | 9.2 |
| | B2 | 40 | 15.51 | 1.61 | 62 | 33 | 5 | 55.4 | 54.0 | 52.2 | 49.8 | 42.7 | 31.5 | 1.13 | 47.9 | 45.5 | 14.0 | 0.168 | 2.5 |
| | B3 | 60 | 11.38 | 1.32 | 63 | 31 | 6 | 55.7 | 54.3 | 52.0 | 46.3 | 34.8 | 29.7 | 0.96 | 56.8 | 38.7 | 9.0 | 0.045 | 18.1 |
| | B4 | 80 | 9.86 | 1.2 | 63 | 32 | 5 | 58.5 | 55.7 | 53.3 | 48.3 | 39.2 | 23.1 | 0.92 | 57.2 | 42.7 | 19.6 | 0.129 | 14.5 |
| | B5 | 100 | 7.95 | 1 | 60 | 36 | 4 | 57.4 | 55.3 | 52.5 | 50.3 | 39.2 | 23.8 | 0.90 | 58.9 | 43.5 | 19.7 | 0.188 | 15.4 |
| Hom | A | 23 | 32.05 | 2.78 | 73 | 21 | 6 | 59.8 | 53.3 | 45.7 | 36.9 | 34.2 | 19.5 | 0.77 | 61.7 | 34.3 | 14.9 | 0.015 | 27.4 |
| | B1 | 40 | 16.84 | 1.73 | 85 | 10 | 5 | 59.0 | 52.6 | 47.2 | 40.5 | 37.3 | 21.4 | 0.83 | 61.0 | 37.3 | 15.9 | 0.029 | 23.7 |
| | B2 | 60 | 9.77 | 1.05 | 72 | 22 | 6 | 56.3 | 52.4 | 49.7 | 45.6 | 40.8 | 25.8 | 1.00 | 56.3 | 42.3 | 16.5 | 0.104 | 14.0 |
| | B3 | 80 | 7.4 | 0.87 | 76 | 17 | 7 | 59.4 | 55.9 | 52.9 | 47.5 | 38.0 | 24.7 | 0.94 | 58.2 | 41.6 | 16.9 | 0.082 | 16.6 |
| | B4 | 100 | 6.42 | 0.7 | 81 | 13 | 6 | 59.6 | 56.3 | 53.2 | 47.6 | 37.6 | 24.1 | 0.96 | 58.1 | 41.5 | 17.3 | 0.085 | 16.6 |
| Mai | A | 15 | 14.54 | 1.25 | 31 | 36 | 33 | 52.8 | 49.7 | 45.2 | 37.5 | 28.6 | 15.4 | 1.21 | 42.9 | 32.0 | 16.6 | 0.053 | 10.9 |
| | B1 | 35 | 10.96 | 0.98 | 32 | 28 | 40 | 57.8 | 50.6 | 45.4 | 36.7 | 30.1 | 16.5 | 1.25 | 41.0 | 32.4 | 15.9 | 0.023 | 8.6 |
| Sav | A | 6 | 34.57 | 2.42 | 66 | 32 | 2 | 63.3 | 60.1 | 55.4 | 44.9 | 31.5 | 17.9 | 0.78 | 63.6 | 36.4 | 18.5 | 0.049 | 27.1 |
| | AB | 24 | 16.99 | 1.35 | 74 | 23 | 2 | 60.6 | 57.9 | 53.7 | 47.9 | 34.0 | 21.3 | 0.90 | 60.9 | 39.1 | 17.9 | 0.077 | 21.7 |
| | B1 | 41 | 10.83 | 1.08 | 79 | 19 | 2 | 58.9 | 56.6 | 54.3 | 49.7 | 35.4 | 22.8 | 1.20 | 48.5 | 40.8 | 18.0 | 0.125 | 7.7 |
| | B2 | 54 | 8.68 | 0.93 | 81 | 17 | 2 | 55.2 | 53.3 | 52.1 | 49.6 | 38.6 | 23.9 | 1.07 | 54.7 | 43.3 | 19.4 | 0.238 | 11.3 |
| | B3 | 83 | 8.21 | 0.81 | 79 | 19 | 2 | 62.5 | 60.1 | 57.1 | 52.5 | 38.6 | 24.7 | 1.14 | 52.1 | 43.9 | 19.2 | 0.123 | 8.2 |
| | B4 | 101 | 7.21 | 0.84 | 80 | 18 | 2 | 56.0 | 53.2 | 50.9 | 48.6 | 40.3 | 25.0 | 1.29 | 46.3 | 43.5 | 18.5 | 0.171 | 2.8 |

## Analysed datasets

Table C.2: Overview over analysed data sets; amount of samples, for which soil parameters and spectra were available, C = carbon content, N = nitrogen content, texture = clay, silt and sand content.

| Name | Scanned in | measured | Land use zone or profile | C &N | texture |
|---|---|---|---|---|---|
| *Local* | laboratory[a] | yes | savannah | 43 | 62 |
| | | | maize | 29 | 39 |
| | | | coffee | 31 | 97 |
| | | | homegarden | 41 | 81 |
| *profile* | in-situ | yes | Sav | 13 | 11 |
| | | | Mai | 9 | 16 |
| | | | Cof | 12 | 13 |
| | | | Hom | 8 | 16 |
| *raster* | in-situ | no | Sav | 544 | 544 |
| | | | Mai | 443 | 443 |
| | | | Cof | 577 | 577 |
| | | | Hom | 578 | 578 |

[a] oven-dried at 45°C for 24 h

## Soil properties

Table C.3: Soil parameters used for soil type classification

| Plot | horizon | uhb | lhb | Clay | Silt | Sand | C | N | pH | CEC | BS | $\rho_b$ | Color | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | cm | cm | % | % | % | mg $g^{-1}$ | mg $g^{-1}$ | | cmol $kg^{-1}$ | % | g $cm^{-3}$ | H | V | C |
| C1 | A | 0 | 17 | 63 | 30 | 6 | 42.00 | 3.43 | 4.01 | 11.3 | 29.5 | 0.8 | 5YR | 3 | 2 |
| | Bw1 | 17 | 57 | 62 | 32 | 5 | 19.38 | 1.81 | 4.24 | 9.1 | 51.4 | NA | 5YR | 2 | 3 |
| | Bw2 | 57 | 98 | 61 | 35 | 4 | 14.81 | 1.44 | 4.84 | 10.1 | 91.7 | NA | 5YR | 3 | 3 |
| C2 | A | 0 | 17 | 30 | 63 | 7 | 55.12 | 4.56 | 5.39 | 12.8 | 96.8 | 0.7 | 5YR | 3 | 2 |
| | Bw1 | 17 | 50 | 29 | 63 | 8 | 43.81 | 3.67 | 4.98 | 7.0 | 91.7 | NA | 5YR | 2 | 2 |
| | Bw2 | 50 | 84 | 29 | 64 | 7 | 33.94 | 2.76 | 5.34 | 10.4 | 98.1 | NA | 5YR | 3 | 1 |
| | Cv | 84 | 96 | 39 | 52 | 9 | 18.04 | 1.65 | 5.33 | 8.5 | 98.2 | NA | 5YR | 3 | 2 |
| C3 | A | 0 | 4 | 61 | 33 | 6 | 22.27 | 2.12 | 4.24 | 9.8 | 49.0 | 1.0 | 2.5YR | 2 | 2 |
| | Bw1 | 4 | 20 | 62 | 31 | 7 | 16.63 | 1.66 | 4.31 | 8.4 | 51.8 | 1.1 | 2.5YR | 2 | 2 |
| | Bw2 | 20 | 40 | 62 | 33 | 5 | 15.51 | 1.61 | 4.38 | 9.2 | 69.8 | 1.1 | 2.5YR | 2 | 2 |
| | Bi1 | 40 | 80 | 63 | 31 | 6 | 10.62 | 1.26 | 4.77 | 8.5 | 95.9 | 0.9 | 5YR | 2 | 3 |
| | Bi2 | 80 | 100 | 60 | 36 | 4 | 7.95 | 1.00 | 4.94 | 8.8 | 98.5 | 0.9 | 5YR | 2 | 4 |
| C4 | A | 0 | 30 | 61 | 33 | 7 | 30.34 | 2.39 | 5.64 | 21.9 | 97.9 | 1.0 | 7.5YR | 3 | 2 |
| | Bw | 30 | 81 | 62 | 33 | 5 | 20.90 | 1.71 | 6.15 | 21.1 | 98.9 | NA | 7.5YR | 3 | 3 |
| | Bi | 81 | 93 | 64 | 31 | 5 | 13.37 | 1.07 | 6.21 | 17.3 | 99.5 | NA | 7.5YR | 3 | 2 |
| H1 | Ah | 0 | 25 | 24 | 69 | 8 | 84.76 | 6.71 | 4.76 | 2.8 | 28.2 | 0.5 | 2.5YR | 2 | 3 |
| | Bw1 | 25 | 47 | 26 | 67 | 7 | 76.46 | 6.34 | 5.00 | 1.8 | 47.1 | 0.6 | 2.5YR | 2 | 3 |
| | Bw2 | 47 | 84 | 26 | 67 | 6 | 79.56 | 6.52 | 4.93 | 2.3 | 59.9 | 0.6 | 2.5YR | 2 | 3 |
| | Bw3 | 84 | 100 | 33 | 61 | 6 | 58.00 | 4.85 | 5.03 | 2.1 | 70.4 | 0.7 | 2.5YR | 2 | 3 |
| H2 | Ap | 0 | 32 | 54 | 37 | 9 | 26.11 | 2.31 | 5.45 | 17.9 | 96.4 | 0.9 | 7.5YR | 3 | 2 |
| | Bw1 | 32 | 72 | 57 | 35 | 8 | 16.96 | 1.67 | 5.40 | 13.5 | 97.4 | NA | 7.5YR | 3 | 4 |
| | Bw2 | 72 | 78 | 59 | 34 | 7 | 11.98 | 1.25 | 5.33 | 12.3 | 96.9 | NA | 5YR | 3 | 2 |

| Plot | horizon | uhb | lhb | Clay | Silt | Sand | C | N | pH | CEC | BS | $\rho_b$ | Color | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | cm | cm | % | % | % | mg $g^{-1}$ | mg $g^{-1}$ | | cmol $kg^{-1}$ | % | g $cm^{-3}$ | H | V | C |
| H3 | A | 0 | 36 | 58 | 40 | 3 | 71.90 | 5.20 | 4.45 | 10.4 | 17.1 | 0.5 | 5YR | 2 | 3 |
| | Bw1 | 36 | 53 | 60 | 36 | 3 | 43.80 | 3.27 | 4.53 | 7.9 | 20.8 | NA | 5YR | 2 | 4 |
| | Bw2 | 53 | 100 | 45 | 54 | 1 | 8.90 | 0.68 | 4.42 | 7.2 | 14.7 | NA | 5YR | 3 | 2 |
| H4 | A | 0 | 18 | 71 | 24 | 5 | 44.72 | 4.15 | 5.04 | 22.7 | 97.5 | 0.8 | 5YR | 3 | 4 |
| | Bw | 18 | 60 | 64 | 30 | 5 | 24.74 | 2.44 | 5.08 | 18.0 | 98.1 | NA | 5YR | 3 | 6 |
| | Bt | 60 | 100 | 77 | 19 | 4 | 14.92 | 1.57 | 4.83 | 16.2 | 96.7 | NA | 2.5YR | 3 | 4 |
| H5 | Ap | 0 | 12 | 62 | 26 | 12 | 51.90 | 4.32 | 5.93 | 29.3 | 98.3 | 0.7 | 2.5YR | 3 | 2 |
| | Bw1 | 12 | 60 | 65 | 22 | 13 | 30.03 | 2.71 | 5.61 | 17.9 | 98.3 | NA | 2.5YR | 3 | 3 |
| | Bw2 | 60 | 103 | 66 | 23 | 11 | 15.58 | 1.60 | 5.39 | 8.7 | 98.3 | NA | 2.5YR | 3 | 3 |
| M1 | A | 0 | 15 | 31 | 36 | 33 | 14.54 | 1.25 | 4.56 | 7.0 | 88.6 | 1.2 | 7.5YR | 3 | 2 |
| | Bw | 15 | 35 | 32 | 28 | 40 | 10.96 | 0.98 | 4.38 | 5.6 | 73.4 | 1.3 | 7.5YR | 3 | 4 |
| | Cv | 35 | 65 | 35 | 29 | 37 | 7.54 | 0.74 | 4.32 | 5.0 | 49.2 | NA | 7.5YR | 4 | 4 |
| M2 | Ap | 0 | 29 | 74 | 23 | 3 | 18.94 | 1.54 | 5.65 | 24.0 | 98.5 | 1.1 | 5YR | 3 | 3 |
| | Bw1 | 29 | 70 | 81 | 18 | 1 | 7.73 | 0.83 | 4.94 | 15.0 | 97.8 | NA | 5YR | 3 | 3 |
| | Bw2 | 70 | 85 | 80 | 19 | 1 | 5.29 | 0.66 | 5.04 | 15.5 | 98.2 | NA | 5YR | 3 | 4 |
| M3 | Ap | 0 | 19 | 52 | 29 | 19 | 21.90 | 1.75 | 6.34 | 35.5 | 99.6 | 1.1 | 5YR | 2 | 3 |
| | Bw1 | 19 | 70 | 60 | 27 | 13 | 15.03 | 1.09 | 6.54 | 36.5 | 99.9 | NA | 5YR | 2 | 4 |
| | Bw2 | 70 | 84 | 59 | 28 | 13 | 12.28 | 0.66 | 7.12 | 50.2 | 100.0 | NA | 5YR | 2 | 4 |
| M4 | Ap | 0 | 28 | 69 | 16 | 15 | 10.13 | 0.90 | 6.22 | 18.2 | 99.8 | 1.1 | 7.5YR | 3 | 4 |
| | Bw | 28 | 61 | 71 | 18 | 10 | 10.44 | 0.97 | 5.40 | 13.7 | 99.4 | NA | 7.5YR | 3 | 4 |
| | Bt | 61 | 80 | 80 | 11 | 8 | 8.33 | 0.91 | 5.39 | 11.6 | 99.4 | NA | 7.5YR | 3 | 4 |

| Plot | horizon | uhb | lhb | Clay | Silt | Sand | C | N | pH | CEC | BS | $\rho_b$ | Color | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | cm | cm | % | % | % | mg $g^{-1}$ | mg $g^{-1}$ | | cmol $kg^{-1}$ | % | g $cm^{-3}$ | H | V | C |
| M5 | Ap | 0 | 32 | 64 | 34 | 2 | 10.13 | 1.00 | 5.11 | 16.4 | 97.3 | 1.1 | 5YR | 3 | 3 |
| | Bt | 32 | 71 | 73 | 26 | 1 | 8.34 | 0.88 | 5.73 | 15.7 | 99.4 | NA | 5YR | 3 | 4 |
| | Bw | 71 | 85 | 77 | 22 | 0 | 6.53 | 0.71 | 5.74 | 16.9 | 99.8 | NA | 5YR | 3 | 4 |
| S1 | A | 0 | 29 | 58 | 40 | 2 | 20.19 | 1.57 | 6.00 | 26.6 | 99.6 | 1.1 | 5YR | 3 | 3 |
| | Cv | 29 | 53 | 44 | 33 | 23 | 14.60 | 1.36 | 6.06 | 22.1 | 99.5 | NA | 5YR | 3 | 4 |
| S2 | A | 0 | 31 | 27 | 42 | 33 | 21.56 | 1.36 | 7.08 | 78.0 | 99.9 | 1.1 | 7.5YR | 3 | 4 |
| | Bw1 | 31 | 69 | 25 | 52 | 24 | NA | 0.68 | 7.32 | 105.9 | 99.9 | NA | 7.5YR | 3 | 4 |
| | Bw2 | 69 | 81 | 29 | 59 | 11 | NA | 1.39 | 7.30 | 110.7 | 99.9 | NA | 5YR | 3 | 3 |
| S3 | A | 0 | 32 | 64 | 26 | 10 | 36.37 | 2.33 | 6.29 | 32.2 | 99.3 | 1.0 | 5YR | 2 | 4 |
| | B | 32 | 54 | 57 | 25 | 17 | 32.01 | 1.97 | 6.48 | 31.8 | 99.4 | NA | 5YR | 3 | 3 |
| S4 | A | 0 | 29 | 34 | 33 | 33 | 47.56 | 1.42 | 7.35 | 85.2 | 100.0 | 1.2 | 7.5YR | 3 | 4 |
| | Cv | 29 | 57 | 41 | 43 | 16 | NA | 1.11 | 7.41 | 77.0 | 100.0 | NA | 7.5YR | 3 | 4 |
| S5 | A | 0 | 6 | 66 | 32 | 2 | 34.57 | 2.42 | 5.62 | 31.8 | 99.7 | 0.8 | 7.5YR | 2 | 3 |
| | AB | 6 | 24 | 74 | 23 | 2 | 16.99 | 1.35 | 5.02 | 21.9 | 99.5 | 0.9 | 5YR | 2 | 3 |
| | Bw | 24 | 54 | 80 | 18 | 2 | 9.75 | 1.00 | 5.03 | 20.4 | 99.6 | 1.1 | 5YR | 2 | 4 |
| | Bi | 54 | 101 | 80 | 19 | 2 | 7.71 | 0.83 | 5.31 | 21.8 | 99.8 | 1.2 | 7.5YR | 3 | 4 |
| S6 | A | 0 | 5 | NA | NA | NA | 48.81 | 2.96 | 7.07 | 56.3 | 100.0 | 1.2 | NA | NA | NA |
| | Bw1 | 5 | 10 | NA | NA | NA | 40.18 | 2.28 | 7.29 | 72.0 | 100.0 | 1.2 | NA | NA | NA |

C1-C4 = Coffee, H1-H5 = Homegarden, M1-M5 = Maize, S1-S6 = Savannah

uhb = upper horizon boundary, lhb = lower horizon boundary, C = Carbon, N = Nitrogen, CEC = Cation exchange capacity, BS = Base saturation, $\rho_b$ = bulk density, HVC = Hue Value Chroma

Table C.4: Particle size fractionation of additional soil auger samples.

| Plot | auger | horizon | uhb | lhb | clay | silt | sand |
|---|---|---|---|---|---|---|---|
| | | | cm | cm | % | % | % |
| C1 | 1 | Ah | 0 | 34 | 60 | 32 | 8 |
| | 1 | B1 | 34 | 80 | 65 | 27 | 8 |
| | 1 | B2 | 80 | 100 | 66 | 25 | 9 |
| C2 | 1 | A | 0 | 7 | 26 | 56 | 18 |
| | 1 | B1 | 7 | 36 | 27 | 58 | 16 |
| | 1 | B2 | 36 | 72 | 27 | 57 | 16 |
| | 1 | C | 72 | 80 | 36 | 30 | 34 |
| C3 | 1 | H1 | 0 | 20 | 57 | 36 | 7 |
| | 2 | H1 | 0 | 20 | 46 | 44 | 10 |
| | 3 | H1 | 0 | 20 | 53 | 39 | 8 |
| | 4 | H1 | 0 | 20 | 60 | 34 | 6 |
| | 5 | H1 | 0 | 20 | 58 | 35 | 7 |
| | 5 | H5 | 80 | 100 | 64 | 29 | 7 |
| | 6 | H1 | 0 | 20 | 52 | 40 | 8 |
| | 7 | H1 | 0 | 20 | 59 | 32 | 8 |
| | 7 | H2 | 20 | 40 | 59 | 32 | 9 |
| | 7 | H3 | 40 | 60 | 64 | 29 | 7 |
| | 7 | H4 | 60 | 80 | 61 | 31 | 8 |
| | 7 | H5 | 80 | 100 | 63 | 30 | 7 |
| | 8 | H1 | 0 | 20 | 62 | 30 | 7 |
| | 8 | H2 | 20 | 40 | 61 | 29 | 10 |
| | 8 | H3 | 40 | 60 | 65 | 28 | 7 |
| | 8 | H4 | 60 | 80 | 63 | 27 | 10 |
| | 8 | H5 | 80 | 100 | 42 | 18 | 40 |
| | 9 | H1 | 0 | 20 | 59 | 32 | 10 |
| | 9 | H2 | 20 | 40 | 60 | 32 | 9 |
| | 9 | H3 | 40 | 60 | 65 | 28 | 7 |
| | 9 | H4 | 60 | 80 | 64 | 29 | 8 |
| | 9 | H5 | 80 | 100 | 67 | 27 | 6 |
| | 10 | H1 | 0 | 20 | 58 | 31 | 11 |
| | 10 | H2 | 20 | 40 | 62 | 29 | 9 |
| | 10 | H3 | 40 | 60 | 57 | 35 | 8 |
| | 10 | H4 | 60 | 80 | 64 | 29 | 7 |
| | 10 | H5 | 80 | 100 | 67 | 27 | 6 |
| | 11 | H1 | 0 | 20 | 58 | 31 | 11 |
| | 11 | H2 | 20 | 40 | 59 | 34 | 7 |
| | 11 | H3 | 40 | 60 | 61 | 30 | 9 |
| | 11 | H4 | 60 | 80 | 62 | 30 | 8 |
| | 11 | H5 | 80 | 100 | 66 | 28 | 6 |
| | 12 | H1 | 0 | 20 | 56 | 31 | 13 |
| | 12 | H2 | 20 | 40 | 60 | 32 | 7 |
| | 12 | H3 | 40 | 60 | 52 | 26 | 22 |

| Plot | auger | horizon | uhb | lhb | clay | silt | sand |
|---|---|---|---|---|---|---|---|
| | | | cm | cm | % | % | % |
| C3 | 12 | H4 | 60 | 80 | 61 | 31 | 8 |
| | 12 | H5 | 80 | 100 | 60 | 29 | 10 |
| | 13 | H1 | 0 | 20 | 44 | 25 | 31 |
| | 13 | H2 | 20 | 40 | 55 | 36 | 9 |
| | 13 | H3 | 40 | 60 | 60 | 33 | 7 |
| | 13 | H4 | 60 | 80 | 62 | 30 | 8 |
| | 13 | H5 | 80 | 100 | 62 | 32 | 7 |
| | 14 | H1 | 0 | 20 | 57 | 33 | 10 |
| | 14 | H2 | 20 | 40 | 54 | 35 | 11 |
| | 14 | H3 | 40 | 60 | 60 | 32 | 8 |
| | 14 | H4 | 60 | 80 | 62 | 27 | 11 |
| | 14 | H5 | 80 | 100 | 62 | 31 | 7 |
| | 15 | H1 | 0 | 20 | 55 | 30 | 15 |
| | 15 | H2 | 20 | 40 | 59 | 33 | 8 |
| | 15 | H3 | 40 | 60 | 40 | 20 | 40 |
| | 15 | H4 | 60 | 80 | 55 | 32 | 13 |
| | 15 | H5 | 80 | 100 | 63 | 31 | 7 |
| | 16 | H1 | 0 | 20 | 47 | 40 | 13 |
| | 16 | H2 | 20 | 40 | 45 | 43 | 12 |
| | 16 | H3 | 40 | 60 | 50 | 39 | 11 |
| | 16 | H4 | 60 | 80 | 48 | 43 | 9 |
| | 16 | H5 | 80 | 100 | 62 | 32 | 5 |
| | 17 | H1 | 0 | 20 | 57 | 35 | 8 |
| | 17 | H2 | 20 | 40 | 54 | 37 | 9 |
| | 17 | H3 | 40 | 60 | 47 | 42 | 10 |
| | 17 | H4 | 60 | 80 | 59 | 34 | 7 |
| | 17 | H5 | 80 | 100 | 51 | 36 | 13 |
| C4 | 1 | A | 0 | 35 | 63 | 29 | 8 |
| | 1 | B1 | 35 | 58 | 62 | 29 | 9 |
| | 1 | B2 | 58 | 66 | 63 | 29 | 9 |
| | 1 | B3 | 66 | 77 | 64 | 29 | 7 |
| | 1 | B4 | 77 | 95 | 64 | 28 | 8 |
| C5 | 1 | Ap | 0 | 37 | 63 | 29 | 8 |
| | 1 | B1 | 37 | 70 | 67 | 23 | 11 |
| | 1 | B2 | 70 | 97 | 72 | 25 | 3 |
| H2 | 1 | Ap | 0 | 32 | 51 | 39 | 10 |
| | 1 | B1 | 32 | 63 | 53 | 37 | 10 |
| | 1 | B2 | 63 | 69 | 54 | 35 | 10 |
| H3 | 1 | A | 0 | 23 | 53 | 43 | 4 |
| | 1 | B1 | 23 | 60 | 57 | 40 | 2 |
| | 1 | B2 | 60 | 80 | 60 | 37 | 3 |
| | 1 | B3 | 80 | 105 | 59 | 40 | 1 |

| Plot | auger | horizon | uhb | lhb | clay | silt | sand |
|---|---|---|---|---|---|---|---|
| | | | cm | cm | % | % | % |
| H4 | 1 | A | 0 | 19 | 79 | 15 | 6 |
| | 1 | B1 | 19 | 66 | 79 | 14 | 7 |
| | 1 | B2 | 66 | 102 | 80 | 13 | 7 |
| | 2 | A | 0 | 7 | 70 | 27 | 3 |
| | 2 | B1 | 7 | 37 | 75 | 18 | 7 |
| | 2 | B2 | 37 | 83 | 78 | 17 | 5 |
| M1 | 1 | A | 0 | 17 | 31 | 29 | 39 |
| | 1 | B1 | 17 | 38 | 27 | 35 | 39 |
| | 1 | Cv | 38 | 54 | 38 | 42 | 20 |
| | 1 | C | 54 | 60 | 31 | 47 | 22 |
| M1 | 1 | A | 0 | 25 | 26 | 32 | 42 |
| | 1 | B1 | 25 | 55 | 30 | 40 | 30 |
| | 1 | Cv | 55 | 75 | 32 | 54 | 14 |
| M2 | 1 | Ap | 0 | 11 | 73 | 23 | 4 |
| | 1 | B1 | 11 | 52 | 75 | 23 | 3 |
| | 1 | B2 | 52 | 65 | 79 | 18 | 2 |
| M3 | 1 | Ah | 0 | 18 | 47 | 28 | 24 |
| | 1 | B1 | 18 | 37 | 48 | 32 | 20 |
| | 1 | B2 | 37 | 75 | 59 | 26 | 14 |
| M3 | 1 | Ap | 0 | 13 | 40 | 36 | 25 |
| | 1 | Ah | 13 | 35 | 41 | 34 | 25 |
| | 1 | B1 | 35 | 74 | 60 | 25 | 15 |
| | 1 | B2 | 74 | 85 | 58 | 30 | 12 |
| M4 | 1 | Ap | 0 | 19 | 69 | 17 | 14 |
| | 1 | B1 | 19 | 38 | 67 | 19 | 15 |
| | 1 | B2 | 38 | 53 | 68 | 19 | 13 |
| | 1 | B3 | 53 | 75 | 76 | 14 | 10 |
| M5 | 1 | Ap | 0 | 25 | 63 | 32 | 4 |
| | 1 | B1 | 25 | 86 | 71 | 27 | 2 |
| | 1 | B2 | 86 | 100 | 82 | 17 | 1 |
| S1 | 1 | AB | 0 | 20 | 29 | 23 | 48 |
| | 1 | Cv | 20 | 51 | 27 | 23 | 50 |
| | 1 | A | 51 | 55 | 23 | 28 | 49 |
| S2 | 1 | A | 0 | 21 | 18 | 29 | 52 |
| | 1 | B1 | 21 | 63 | 17 | 42 | 41 |
| | 1 | B2 | 63 | 81 | 13 | 30 | 57 |
| | 2 | A | 0 | 18 | 13 | 30 | 57 |
| | 2 | B | 18 | 29 | 13 | 31 | 56 |
| S3 | 1 | A | 0 | 23 | 40 | 13 | 47 |
| S4 | 1 | A | 0 | 16 | 20 | 42 | 38 |

| Plot | auger | horizon | uhb | lhb | clay | silt | sand |
|---|---|---|---|---|---|---|---|
| | | | cm | cm | % | % | % |
| S6 | 1 | A | 0 | 5 | 19 | 34 | 47 |
| | 2 | A | 0 | 5 | 18 | 31 | 51 |
| | 3 | A | 0 | 5 | 21 | 41 | 38 |
| | 4 | A | 0 | 5 | 27 | 31 | 41 |
| S7 | 1 | A | 0 | 5 | 63 | 33 | 5 |
| | 2 | A | 0 | 5 | 66 | 29 | 6 |
| | 3 | A | 0 | 5 | 65 | 32 | 4 |
| | 4 | A | 0 | 5 | 63 | 32 | 4 |

C1-C4 = Coffee, H1-H5 = Homegarden, M1-M5 = Maize, S1-S7 = Savannah, uhb = upper horizon boundary, lhb = lower horizon boundary

Table C.5: Carbon and nitrogen content of additional soil auger samples.

| Plot | auger | horizon | uhb | lhb | C | N |
|---|---|---|---|---|---|---|
| | | | cm | cm | mg $g^{-1}$ | mg $g^{-1}$ |
| C1 | 1 | A | 0 | 35 | 23.6 | 2.19 |
| | 1 | B1 | 35 | 77 | 12 | 1.07 |
| | 1 | B2 | 77 | 90 | 10.6 | 0.89 |
| C2 | 1 | A | 0 | 25 | 53.2 | 5.08 |
| | 1 | B1 | 25 | 46 | 48.2 | 4.79 |
| | 1 | B2 | 46 | 85 | 38 | 3.51 |
| | 1 | B3 | 85 | 91 | 21.25 | 1.91 |
| C3 | 1 | A | 0 | 29 | 35.05 | 3.48 |
| | 1 | B1 | 29 | 63 | 24.2 | 2.53 |
| | 1 | B2 | 63 | 92 | 15.75 | 1.8 |
| | 1 | B3 | 92 | 97 | 14.65 | 1.71 |
| C4 | 1 | A | 0 | 27 | 27 | 2.33 |
| | 1 | B1 | 27 | 90 | 18.55 | 1.49 |
| | 1 | B2 | 90 | 102 | 17.95 | 1.33 |
| C5 | 1 | Ap | 0 | 38 | 46.1 | 4.26 |
| | 1 | B1 | 38 | 68 | 37.6 | 3.61 |
| | 1 | B2 | 68 | 95 | 28.15 | 2.67 |
| | 1 | B3 | 95 | 100 | 18.7 | 1.8 |
| H1 | 1 | A | 0 | 30 | 89.8 | 7.59 |
| | 1 | B1 | 30 | 53 | 78.8 | 6.94 |
| | 1 | B2 | 53 | 84 | 50.8 | 3.8 |
| | 1 | B3 | 84 | 94 | 60.1 | 4.89 |
| H2 | 1 | Ap | 0 | 24 | 46.3 | 4.12 |
| | 1 | B1 | 24 | 57 | 18.2 | 1.76 |
| | 1 | B2 | 57 | 63 | 14.3 | 1.4 |

| Plot | auger | horizon | uhb | lhb | C | N |
|---|---|---|---|---|---|---|
| | | | cm | cm | mg $g^{-1}$ | mg $g^{-1}$ |
| H3 | 1 | A | 0 | 23 | 96.5 | 8.05 |
| | 1 | B1 | 23 | 42 | 78.3 | 6.5 |
| | 1 | B2 | 42 | 66 | 49.7 | 4.38 |
| | 1 | B3 | 66 | 100 | 19.9 | 1.5 |
| H4 | 3 | A | 0 | 18 | 44.72 | 4.148 |
| | 3 | B1 | 18 | 60 | 24.74 | 2.442 |
| | 3 | B2 | 60 | 100 | 14.92 | 1.568 |
| H5 | 1 | Ap | 0 | 25 | 72.3 | 6.7 |
| | 1 | B1 | 25 | 75 | 19.8 | 1.79 |
| | 1 | B2 | 75 | 92 | 17.9 | 1.61 |
| | 1 | B3 | 92 | 100 | 18.7 | 1.61 |
| M1 | 1 | A | 0 | 19 | 9.83 | 0.75 |
| | 1 | B1 | 19 | 38 | 5.73 | NA |
| | 1 | Cv | 38 | 58 | 4.13 | NA |
| | 1 | C | 58 | 73 | 3.66 | NA |
| M2 | 1 | Ap | 0 | 29 | 13.9 | 1.01 |
| | 1 | B1 | 29 | 58 | 7.7 | 0.58 |
| | 1 | B2 | 58 | 66 | 6.5 | NA |
| M4 | 1 | Ap | 0 | 19 | 11.1 | 0.75 |
| | 1 | B1 | 19 | 48 | 10.95 | 0.76 |
| | 1 | B2 | 48 | 68 | 9.97 | 0.68 |
| | 1 | B3 | 68 | 83 | 8.16 | 0.6 |
| M5 | 1 | Ap | 0 | 33 | 14 | 1.01 |
| | 1 | B1 | 33 | 59 | 10.15 | 0.76 |
| | 1 | B2 | 59 | 72 | 8.45 | 0.66 |
| S1 | 1 | Ah | 72 | 15 | 25.5 | 1.78 |
| | 1 | Cv | 15 | 31 | 18.4 | 1.44 |
| S2 | 1 | A | 31 | 32 | 16.9 | 1.5 |
| | 1 | B1 | 32 | 55 | 4.16 | NA |
| | 1 | B2 | 55 | 74 | 1.4 | NA |
| S3 | 1 | A | 74 | 12 | 47.05 | 2.96 |
| | 1 | Cv | 12 | 20 | 30.55 | 2.06 |
| S4 | 1 | A | 20 | 10 | 20.1 | 1.5 |
| | 1 | B | 10 | 28 | 7.29 | NA |
| S6 | 1 | A | 0 | 5 | 29 | 2.4 |
| | 2 | A | 0 | 5 | 26.1 | 2.4 |
| | 3 | A | 0 | 5 | 28.7 | 2.7 |
| | 4 | A | 0 | 5 | 41.2 | 3.8 |
| | 5 | A | 0 | 5 | 27.5 | 2.3 |
| | 6 | A | 0 | 5 | 53.6 | 4.4 |

| Plot | auger | horizon | uhb | lhb | C | N |
|---|---|---|---|---|---|---|
| | | | cm | cm | mg $g^{-1}$ | mg $g^{-1}$ |
| S7 | 1 | A | 0 | 5 | 25.5 | 1.9 |
| | 2 | A | 0 | 5 | 24.5 | 1.8 |
| | 3 | A | 0 | 5 | 24.8 | 2.2 |
| | 4 | A | 0 | 5 | 22.8 | 1.9 |
| | 5 | A | 0 | 5 | 21.5 | 1.8 |
| | 6 | A | 0 | 5 | 21.7 | 1.7 |
| | 7 | A | 0 | 5 | 31.5 | 2.8 |
| | 8 | A | 0 | 5 | 19.5 | 1.4 |

C1-C4 = Coffee, H1-H5 = Homegarden, M1-M5 = Maize, S1-S7 = Savannah, uhb = upper horizon boundary, lhb = lower horizon boundary

Table C.6: Soil types of the study plots

| Land use | plot | Soil type | | Elevation (m a.s.l.) |
|---|---|---|---|---|
| Homegarden | H1 | Melanic | Andosol | 1647 |
| | H2 | Andic | Umbrisol | 1170 |
| | H3 | Aluandic | Andosol | 1837 |
| | H4 | Sodic | Vertisol | 1276 |
| | H5 | Mollic | Nitisol | 1560 |
| Coffee | C1 | Andic | Umbrisol | 1306 |
| | C2 | Mollic | Andosol | 1345 |
| | C3 | Haplic | Vertisol | 1305 |
| | C4 | Haplic | Vertisol | 1125 |
| Maize | M1 | Vitric | Cambisol | 1009 |
| | M2 | Haplic | Nitisol | 860 |
| | M3 | Haplic | Luvisol | 886 |
| | M4 | Lixic | Nitisol | 960 |
| | M5 | Lixic | Nitisol | 960 |
| Savannah | S1 | Haplic | Leptosol | 899 |
| | S2 | Rendzic | Leptosol | 906 |
| | S3 | Haplic | Leptosol | 1148 |
| | S4 | Rendzic | Leptosol | 993 |
| | S5 | Sodic | Vertisol | 951 |
| | S6 | Rendzic | Leptosol | 960 |
| | S7 | Sodic | Vertisol[a] | 952 |

[a] S7 is situated very close to S5, soil type was not identified separately

Table C.7: Measured soil properties of the different horizons for Coffee (C3), Homegarden (H4), Maize (M1) and Savannah (S5) profiles

| plot | horizon | uhb | lhb | C | N | clay | silt | sand | $\rho_p$ | $\rho_b$ | $\phi$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | cm | cm | mg $g^{-1}$ | mg $g^{-1}$ | % | % | % | g $cm^{-3}$ | g $cm^{-3}$ | % |
| C3 | A | 0 | 4 | 22.27 | 2.12 | 61 | 33 | 6 | 2.22 | 0.99 | 55.4 |
| C3 | B1 | 4 | 20 | 16.63 | 1.66 | 62 | 31 | 7 | 2.24 | 1.08 | 52.0 |
| C3 | B2 | 20 | 40 | 15.51 | 1.61 | 62 | 33 | 5 | 2.17 | 1.13 | 47.9 |
| C3 | B3 | 40 | 60 | 11.38 | 1.32 | 63 | 31 | 6 | 2.22 | 0.96 | 56.8 |
| C3 | B4 | 60 | 80 | 9.86 | 1.20 | 63 | 32 | 5 | 2.15 | 0.92 | 57.2 |
| C3 | B5 | 80 | 100 | 7.95 | 1.00 | 60 | 36 | 4 | 2.19 | 0.90 | 58.9 |
| H4 | A | 0 | 23 | 32.05 | 2.78 | 73 | 21 | 6 | 2.01 | 0.77 | 61.7 |
| H4 | B1 | 23 | 40 | 16.84 | 1.73 | 85 | 10 | 5 | 2.13 | 0.83 | 61.0 |
| H4 | B2 | 40 | 60 | 9.77 | 1.05 | 72 | 22 | 6 | 2.29 | 1.00 | 56.3 |
| H4 | B3 | 60 | 80 | 7.40 | 0.87 | 76 | 17 | 7 | 2.25 | 0.94 | 58.2 |
| H4 | B4 | 80 | 100 | 6.42 | 0.70 | 81 | 13 | 6 | 2.29 | 0.96 | 58.1 |
| M1 | A | 0 | 15 | 14.54 | 1.25 | 31 | 36 | 33 | 2.12 | 1.21 | 42.9 |
| M1 | B | 15 | 35 | 10.96 | 0.98 | 32 | 28 | 40 | 2.12 | 1.25 | 41.0 |
| M1 | Cv | 35 | 65 | 8.33 | 0.91 | 35 | 29 | 37 | 2.11 | NA | NA |
| S5 | A | 0 | 6 | 34.57 | 2.42 | 66 | 32 | 2 | 2.14 | 0.78 | 63.6 |
| S5 | B1 | 6 | 24 | 16.99 | 1.35 | 74 | 23 | 2 | 2.30 | 0.90 | 60.9 |
| S5 | B2 | 24 | 41 | 10.83 | 1.08 | 79 | 19 | 2 | 2.33 | 1.20 | 48.5 |
| S5 | B3 | 41 | 54 | 8.68 | 0.93 | 81 | 17 | 2 | 2.36 | 1.07 | 54.7 |
| S5 | B4 | 54 | 83 | 8.21 | 0.81 | 79 | 19 | 2 | 2.38 | 1.14 | 52.1 |
| S5 | B5 | 83 | 98 | 7.21 | 0.84 | 80 | 18 | 2 | 2.40 | 1.29 | 46.3 |

uhb = upper horizon boundary, lhb = lower horizon boundary, $\rho_p$ = particle density, $\rho_b$ = bulk density, $\phi$ = porosity = $(1-(\rho_b/\rho_p))\cdot 100$

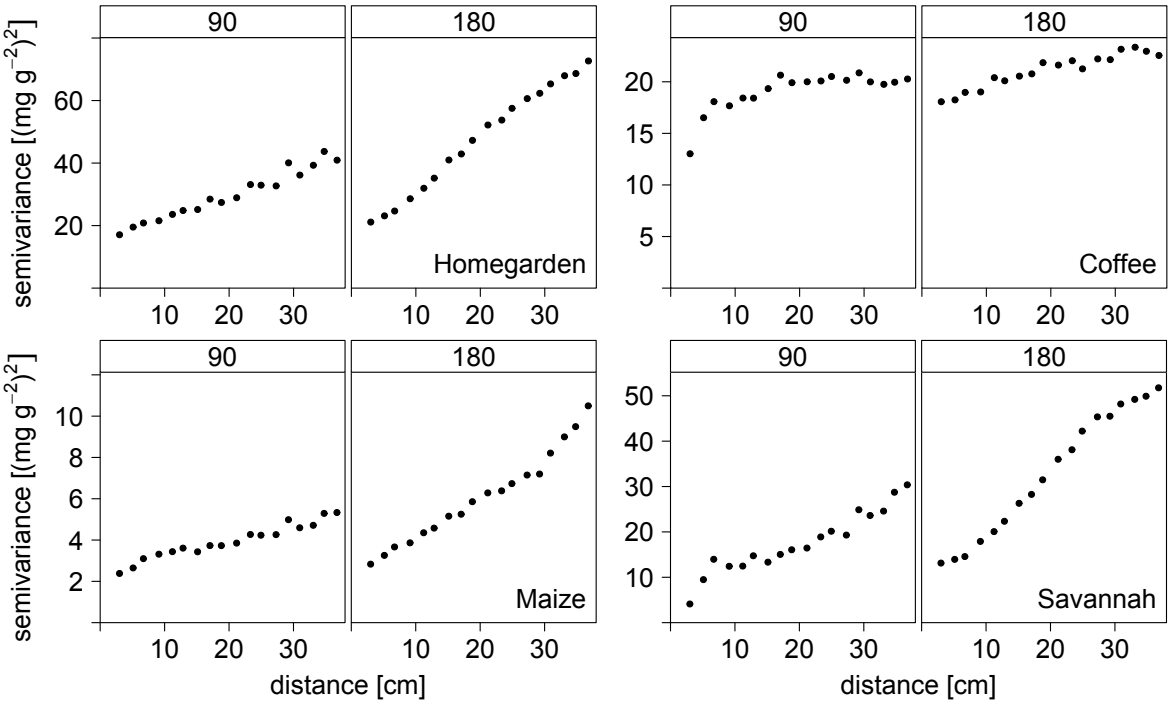



Figure C.5: Anisotropic variograms of carbon content: 90 represents the horizontal direction, 180 the vertical direction within the soil profiles.

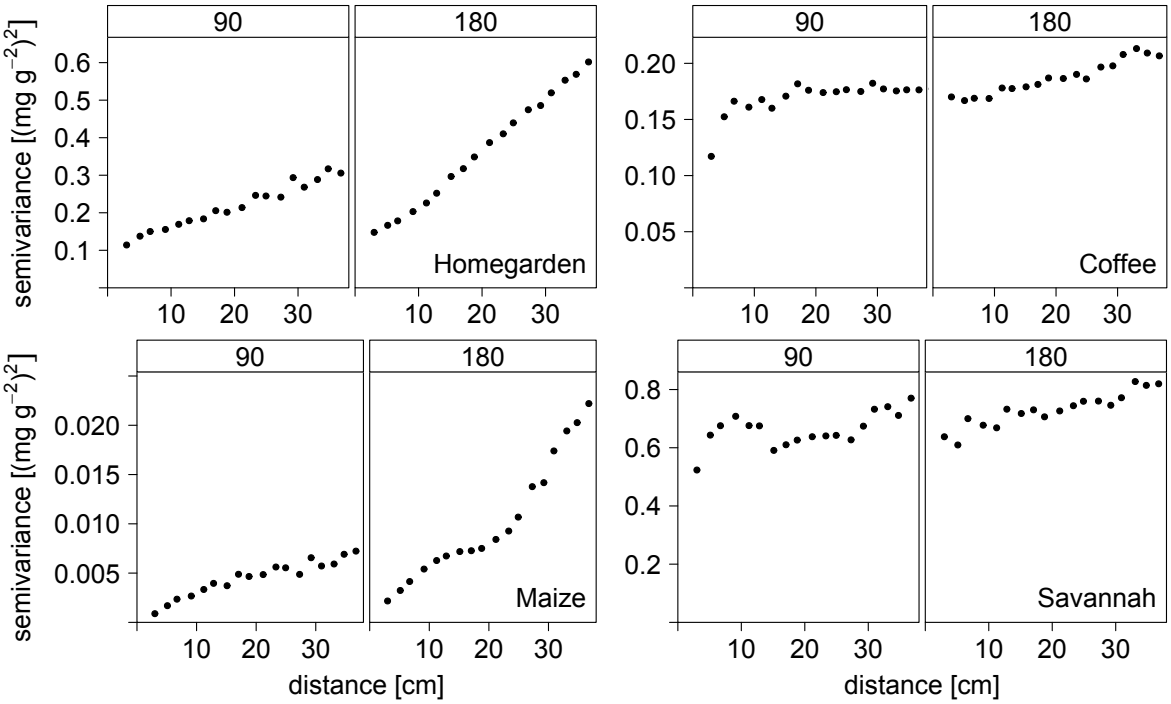



Figure C.6: Anisotropic variograms of nitrogen content: 90 represents the horizontal direction, 180 the vertical direction within the soil profiles.

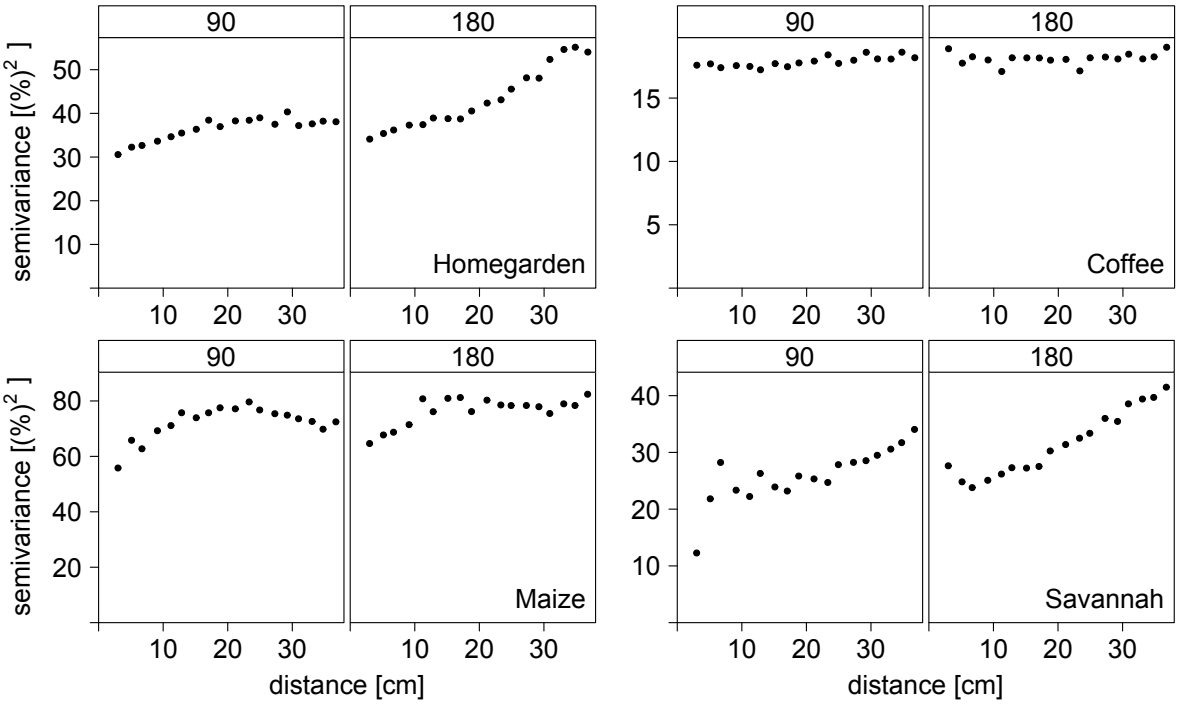



Figure C.7: Anisotropic variograms of clay content: 90 represents the horizontal direction, 180 the vertical direction within the soil profiles.

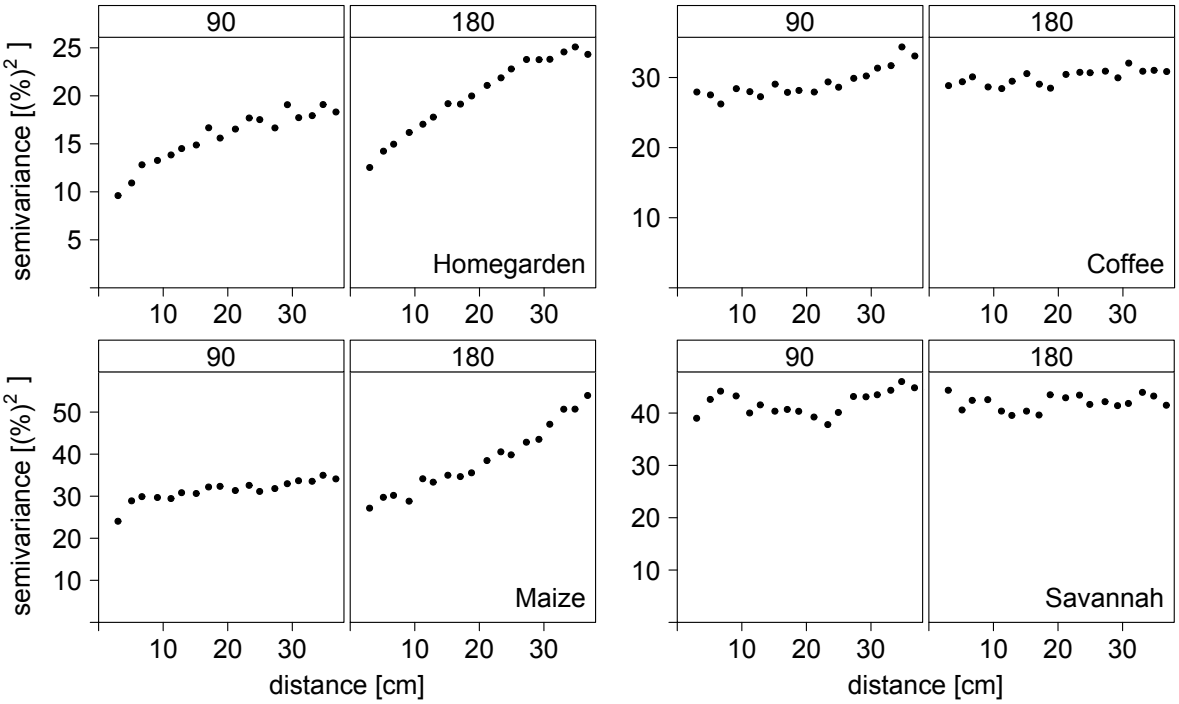



Figure C.8: Anisotropic variograms of silt content: 90 represents the horizontal direction, 180 the vertical direction within the soil profiles

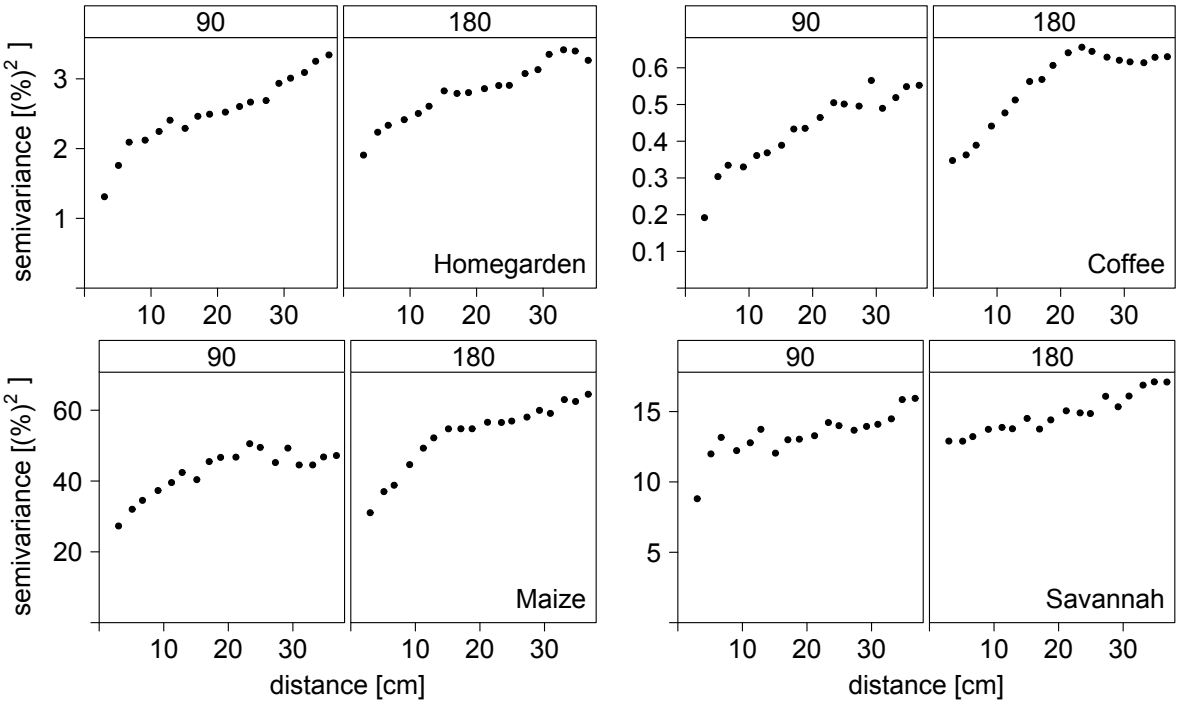



Figure C.9: Anisotropic variograms of sand content: 90 represents the horizontal direction, 180 the vertical direction within the soil profiles.

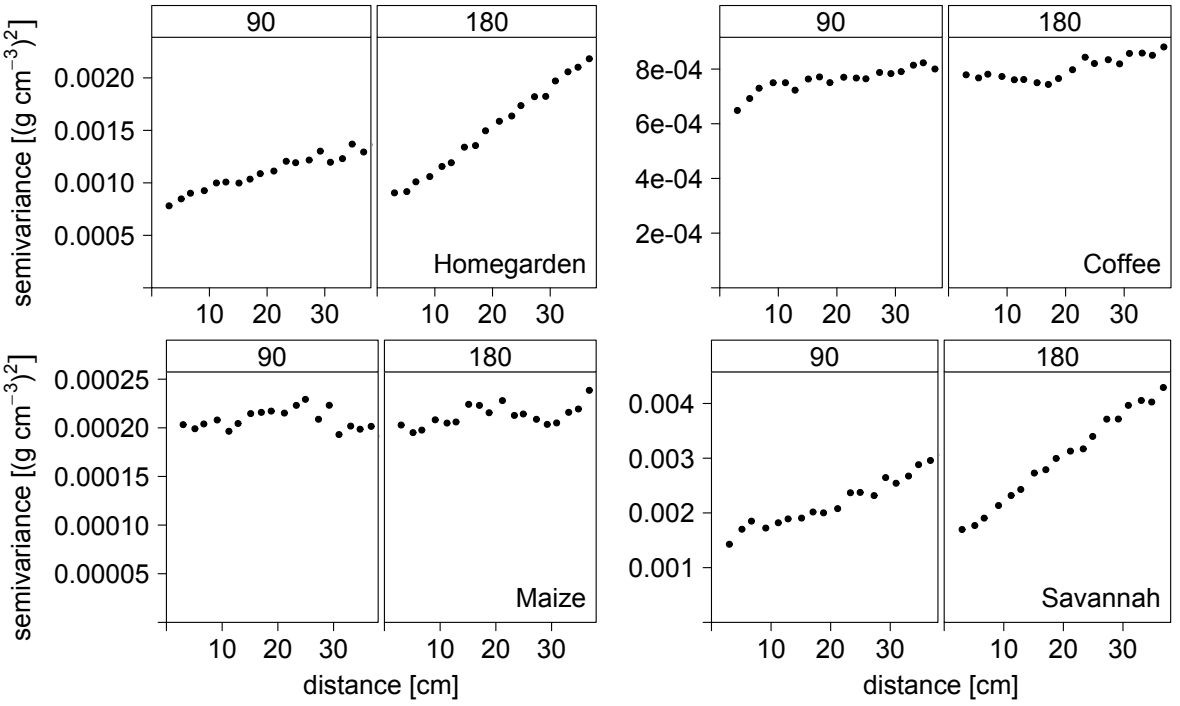



Figure C.10: Anisotropic variograms of $\rho_b$ as predicted with the random forest model: 90 represents the horizontal direction, 180 the vertical direction within the soil profiles.

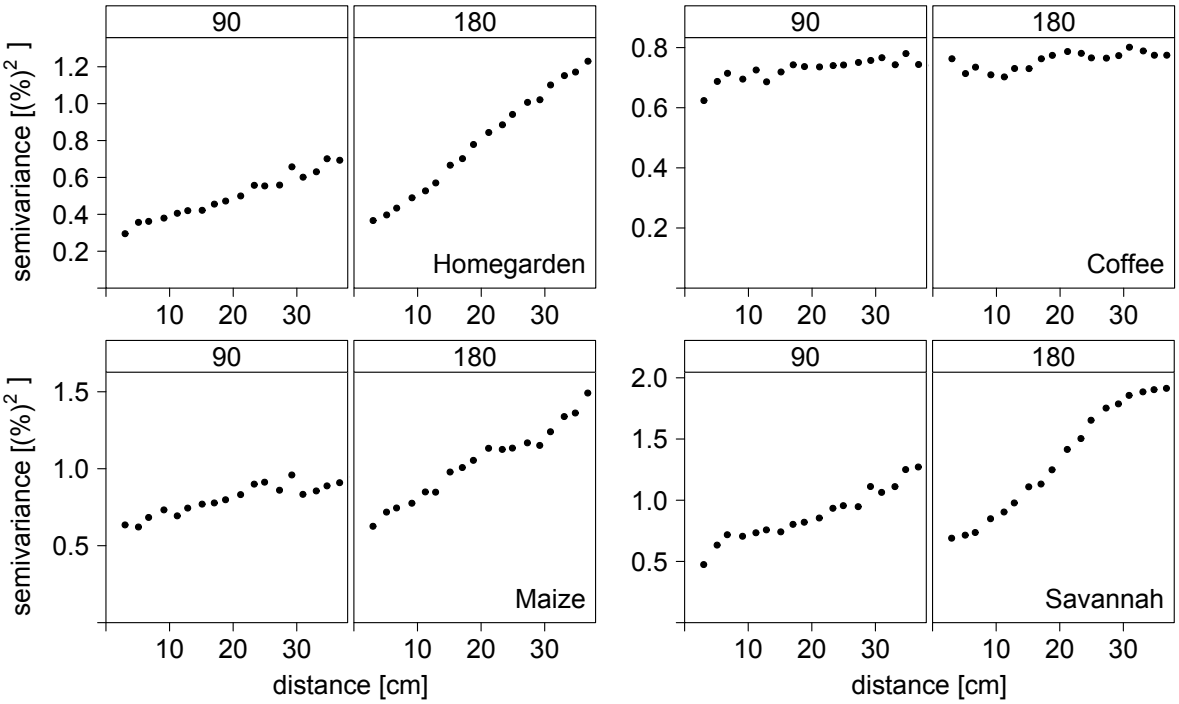



Figure C.11: Anisotropic variograms of $\theta_{1.8}$ as predicted with the random forest model: 90 represents the horizontal direction, 180 the vertical direction within the soil profiles.

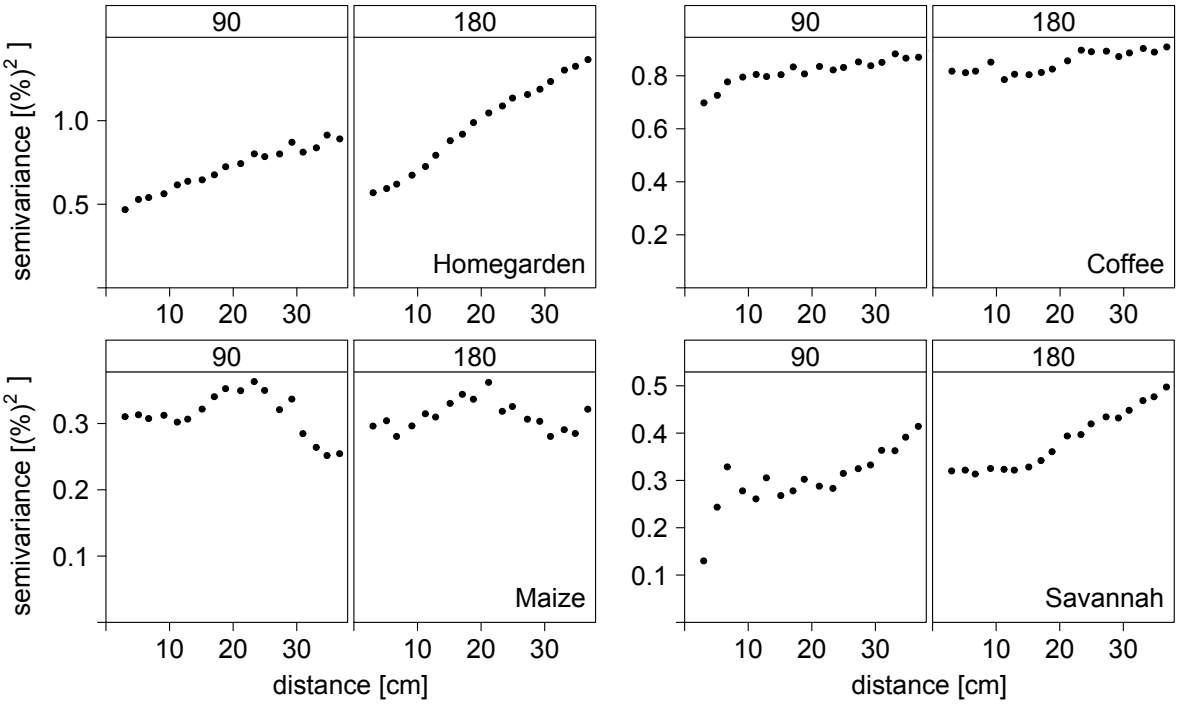



Figure C.12: Anisotropic variograms of $\theta_{4.2}$ as predicted with the random forest model: 90 represents the horizontal direction, 180 the vertical direction within the soil profiles.

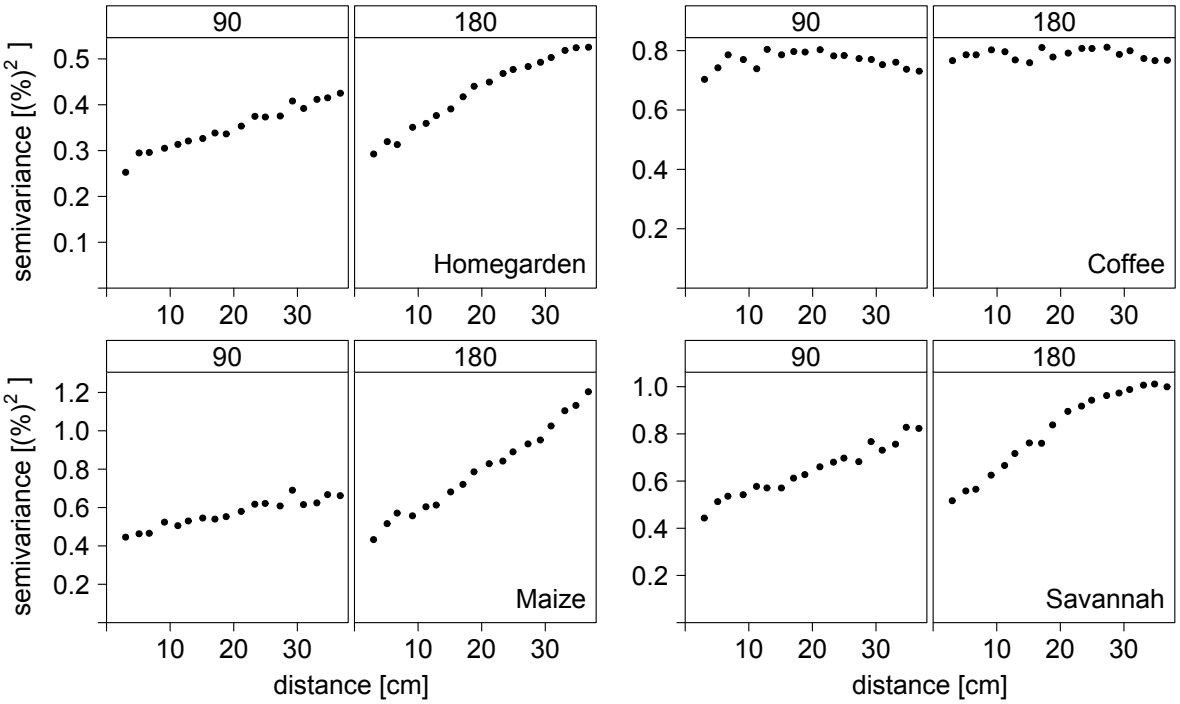



Figure C.13: Anisotropic variograms of available water capacity ($AWC = \theta_{1.8}$ - $\theta_{4.2}$): 90 represents the horizontal direction, 180 the vertical direction within the soil profiles.

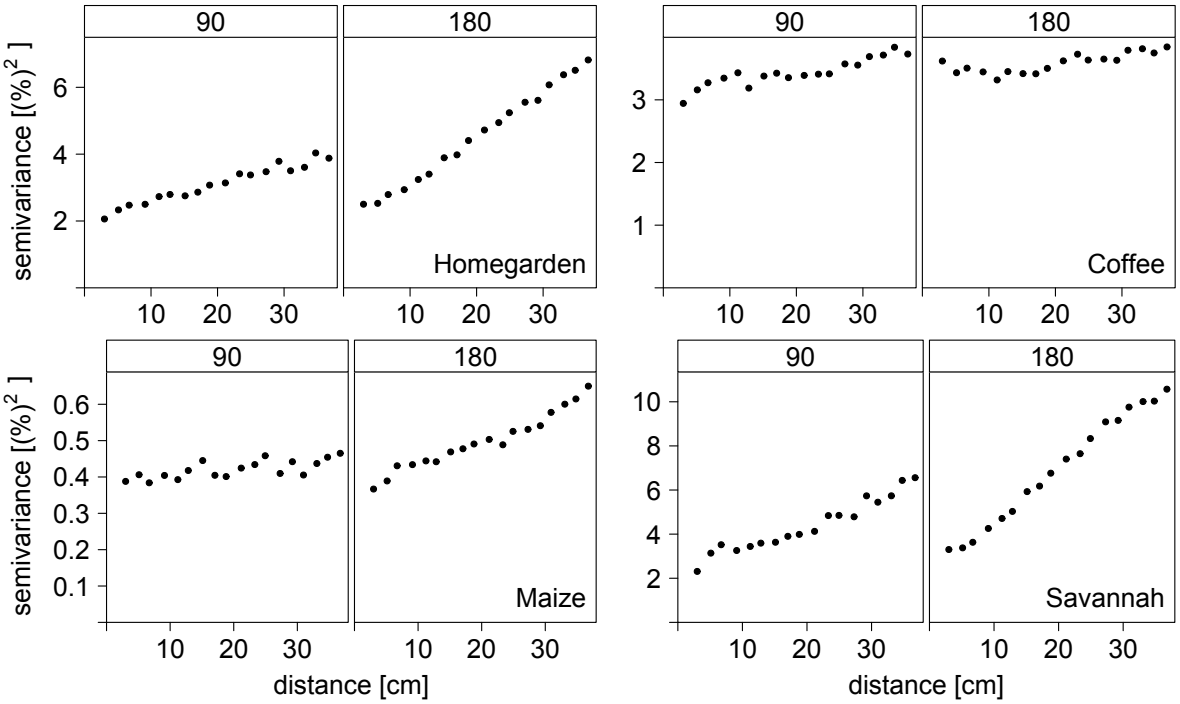



Figure C.14: Anisotropic variograms of field air capacity ($FAC = \phi - \theta_{1.8}$: 90 represents the horizontal direction, 180 the vertical direction within the soil profiles.

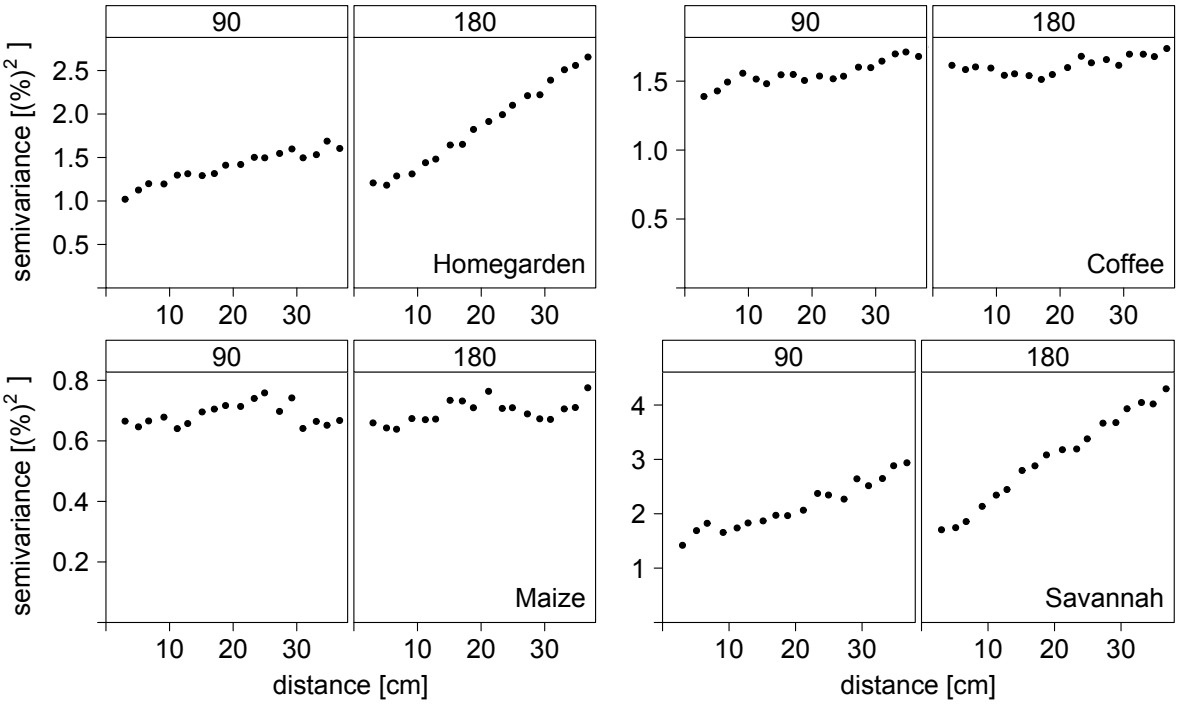



Figure C.15: Anisotropic variograms of $\phi$ as predicted with the random forest model: 90 represents the horizontal direction, 180 the vertical direction within the soil profiles.

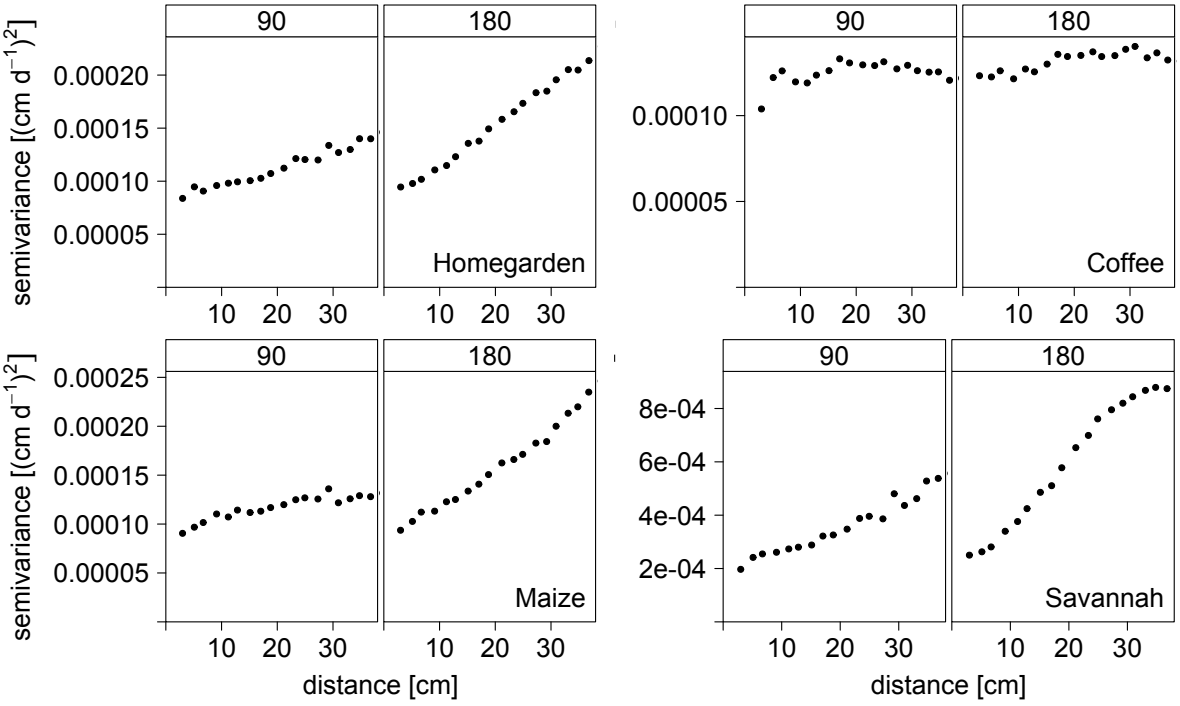



Figure C.16: Anisotropic variograms of unsaturated hydraulic conductivity ($K_r(1.8)$) as predicted with the random forest model: 90 represents the horizontal direction, 180 the vertical direction within the soil profiles.

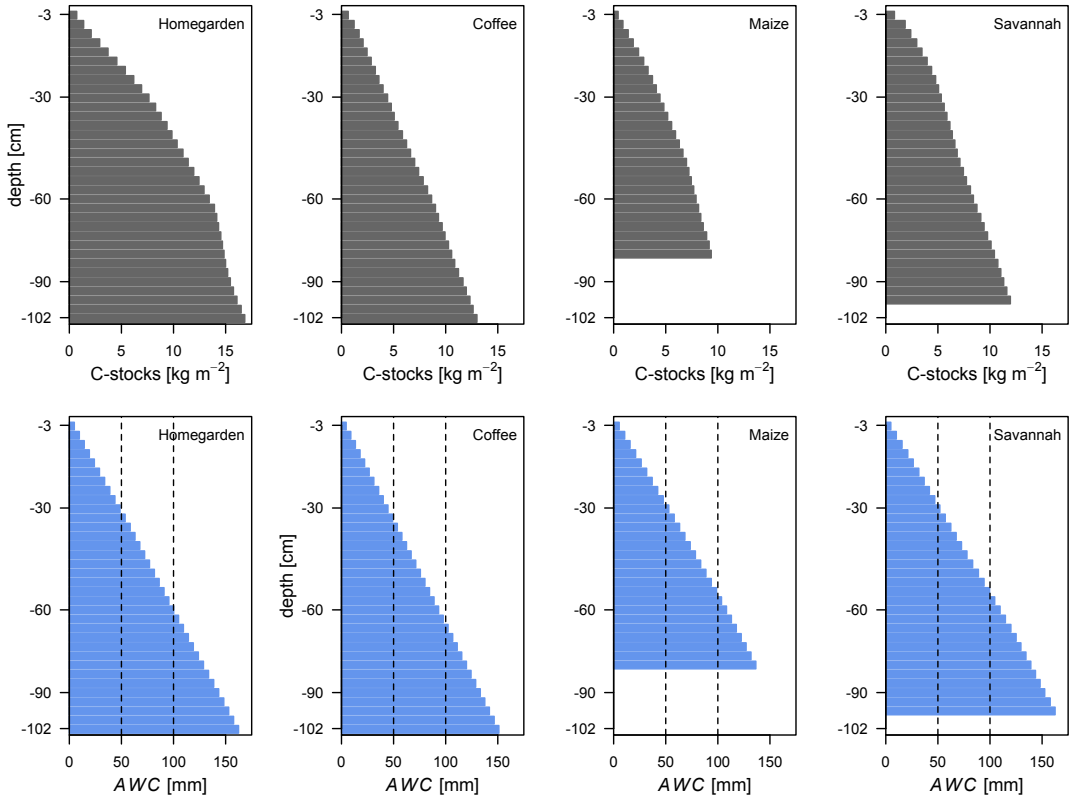



Figure C.17: Cumulative C stocks (kg m$^{-2}$) and available water (mm) of the different profiles.

The cumulative C stocks visualize the importance of the contribution of the lower soil horizons to the overall C stocks (Figure C.17). As the available water was increasing with depth for most profiles, the lower soil horizons play an important role, as a high amount of water was still available in the lower parts of the profile.

# Appendix D

## Supplementary Material to Manuscript 5 "Spatial patterns of microbial biomass and fauna activity in savannah soils at Mt. Kilimanjaro"

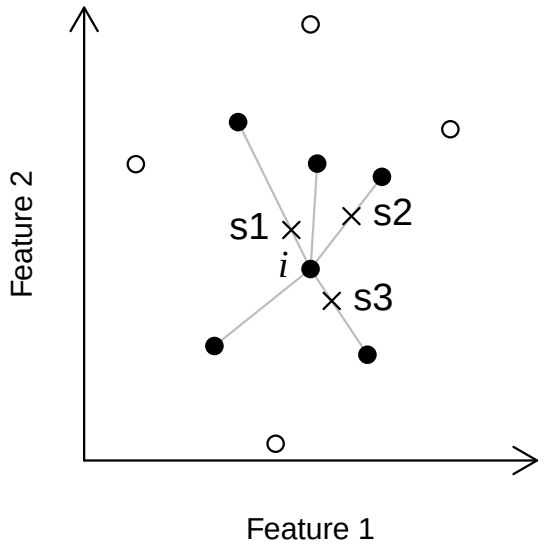



Figure D.1: Illustration of the synthetic minority oversampling technique (SMOTE) in two dimensions. The $k$ nearest neighbours (black dots) are chosen for an existing point $i$ to generate synthetic points (crosses denoted $s1$ through $s3$) along the connection lines between $i$ and its nearest neighbours. Circles show samples that are not the $k$ nearest neighbours of $i$.

**Algorithm:** SMOTE

**Input**: $T$ original samples to be smoted
Amount of SMOTE $N$%
Number of nearest neighbours $k$

**Output**: $(N/100) \times T$ synthetic samples with their target values (i.e. concentrations)

**if** $N < 100$ **then**
  Randomize the $T$ original samples:
  $T = (N/100) \times T$
  $N = 100$
**end**

$orig.s[i]$: original sample $i, i = 1, \ldots, T$

$orig.t[i]$: target value of original sample $i$

$new.s[j]$: synthetic sample $j, j = 1, \ldots, (N/100) \times T$

$new.t[j]$: target values of synthetic sample $j$

$ng \leftarrow N/100$: number of synthetic samples to compute for each original sample

Generate synthetic samples:

**for** $i$ *in* $1$ *to* $T$ **do**
  $nns \leftarrow$ compute $k$ nearest neighbours for $orig.s[i]$
  **for** $\ell$ *in* $1$ *to* $ng$ **do**
    randomly choose $x \in nns$
    $diff = x - orig.s[i]$
    $new.s[(i-1) \times ng + \ell] = orig.s[i] + \text{RANDOM}(0,1) \times diff$
    $d_1 = \text{DIST}(new.s, orig.s[i])$
    $d_2 = \text{DIST}(new.s, x)$
    $target = \frac{d_2 \times orig.t(orig.s) + d_1 \times orig.t(x)}{d_1 + d_2}$
  **end**
**end**

**return** $new.t \cup new.s$

Table D.1: Error parameters of the different partial least squares regression models for VIS-NIR-DRS

| | | | Calibration | | | Validation | | |
|---|---|---|---|---|---|---|---|---|
| Plot | Parameter | m | n | $RMSE$ | $R^2$ | n | $RMSE$ | $R^2$ |
| $P_{plain}$ | $C_{org}$ | 7 | 91 | 4.48 | 0.72 | 16 | 2.61 | 0.70 |
| | N | 7 | 91 | 0.31 | 0.76 | 16 | 0.23 | 0.63 |
| | clay | 10 | 146 | 8.0 | 0.84 | - | - | - |
| $P_{slope}$ | $C_{org}$ | 7 | 91 | 6.17 | 0.70 | 16 | 4.79 | 0.83 |
| | N | 8 | 91 | 0.49 | 0.75 | 16 | 0.40 | 0.85 |
| | clay | 10 | 146 | 8.0 | 0.84 | - | - | - |

$m$ = number of model parameters; $n$ = number of samples; $RMSE$ = root mean squared error

# Appendix E

## List of publications

Holger Pabst, Anna Kühnel, Yakov Kuzyakov, *Effect of land-use and elevation on microbial biomass and water extractable carbon in soils of Mt. Kilimanjaro ecosystems*, 2013, Applied Soil Ecology 67, pp 10–19 (available at: 10.1016/j.apsoil.2013.02.006)

Michael Zech, Claudia Hörold, Katharina Leiber-Sauheitl, Anna Kühnel, Andreas Hemp, Wolfgang Zech, *Buried black soils on the slopes of Mt. Kilimanjaro as a regional carbon storage hotspot*, 2014, Catena 112, pp. 125–130 (available at: 10.1016/j.catena.2013.05.015)

Anna Kühnel, Christina Bogner, Bernd Huwe, *In situ prediction of soil chemical properties with visible and near infrared spectroscopy in an African savannah*, 2014, GlobalSoilMap: Basis of the global spatial soil information system, CRC Press, pp 409–413 (see Appendix A)

# List of conference contributions

Anna Kühnel and Bernd Huwe, *Charakterisierung des Wassertransports im Boden in Savannenökosystemen*, „Neue messtechnische Entwicklungen zur Erfassung des Wasser- und Stoffhaushalts in Pflanzen und Böden “ UGT, 25 May 2012, Freising (talk)

Anna Kühnel and Bernd Huwe, *Spatial patterns of clay content in soils around Mount Kilimanjaro*, Eurosoil 2012, 2-6 July 2012, Bari (poster)

Anna Kühnel and Bernd Huwe, *Spatial patterns regarding water storage in soils around Mount Kilimanjaro, 2$^{nd}$ International Conference on Hydropedology*, 22-27 July 2012, Leipzig (talk)

Anna Kühnel and Bernd Huwe, *Spatial patterns of soil physical parameters in Savannah soils around Mount Kilimanjaro*, TR32-HOBE International Symposium on “Patterns in Soil-Vegetation-Atmosphere-Systems: Monitoring, Modelling & Data Assimilation”, 11-14 March 2013, Bonn (poster)

Anna Kühnel, Christina Bogner, Holger Pabst and Bernd Huwe, *Spatial structure of soil properties at different scales of Mt. Kilimanjaro, Tanzania*, EGU General Assembly 2013, 07-12 April 2013, Vienna (talk)

Anna Kühnel, Christina Bogner, Holger Pabst and Bernd Huwe, *Visualizing small scale variability of clay content on soils at Mt. Kilimanjaro by Vis-NIR spectroscopy*, 3$^{rd}$ Global Workshop on Proximal Soil Sensing, 26-29 May 2013, Potsdam (talk)

Anna Kühnel, Holger Pabst, Juliane Röder and Bernd Huwe, *Savannah soils at Mount Kilimanjaro: Small scale spatial patterns of microbial biomass and soil fauna activity*, Jahrestagung der Deutschen Bodenkundlichen Gesellschaft, 7-12 September 2013, Rostock (poster)

Anna Kühnel, Christina Bogner and Bernd Huwe, *In situ prediction of soil chemical properties with visible and near infrared spectroscopy in an African savannah*, Global Soil Map Conference, 7-9 October 2013, Orléans (talk)

Anna Kühnel, Christina Bogner and Bernd Huwe, *Spatial structure of soil properties at the profile scale, Tanzania*, DBG Workshop on “Soil processes – is the whole system regulated at ‘hot spots’?”, 4-6 May 2014, Freising (talk)

## (Eidesstattliche) Versicherungen und Erklärungen

(§5 Nr. 4 PromO)

Hiermit erkläre ich, dass keine Tatsachen vorliegen, die mich nach den gesetzlichen Bestimmungen über die Führung akademischer Grade zur Führung eines Doktorgrades unwürdig erscheinen lassen.

(§8 S. 2 Nr. 5 PromO)

Hiermit erkläre ich mich damit einverstanden, dass die elektronische Fassung meiner Dissertation unter Wahrung meiner Urheberrechte und des Datenschutzes einer gesonderten Überprüfung hinsichtlich der eigenständigen Anfertigung der Dissertation unterzogen werden kann.

(§8 S. 2 Nr. 7 PromO)

Hiermit erkläre ich eidesstattlich, dass ich die Dissertation selbständig verfasst und keine anderen als die von mir angegebenen Quellen und Hilfsmittel benutzt habe.

(§8 S. 2 Nr. 8 PromO)

Ich habe die Dissertation nicht bereits zur Erlangung eines akademischen Grades anderweitig eingereicht und habe auch nicht bereits diese oder eine gleichartige Doktorprüfung endgültig nicht bestanden.

(§8 S. 2 Nr. 9 PromO)

Hiermit erkläre ich, dass ich keine Hilfe von gewerblichen Promotionsberatern bzw. -vermittlern in Anspruch genommen habe und auch künftig nicht nehmen werde.

. . . . . .

Ort, Datum, Unterschrift