PREDICTION OF BUSH BEAN (*PHASEOLUS VULGARIS* L.) YIELDS IN NORTHERN TANZANIA BASED ON SPECTRAL ANALYSIS OF SOILS

Charlotte Mallet

Master Earth Sciences – Environmental Management track

Dr. ir. J. van Heerwaarden (WUR)

Dr. L.H. Cammeraat (UvA)

Dr. A. Tietema (UvA)

University of Amsterdam – Institute for Biodiversity and Ecosystem Dynamics

MSc thesis | 30 ECTS | Student ID: 11574348

May – December 2018

December 25, 2018

PREDICTION OF BUSH BEAN (*PHASEOLUS VULGARIS* L.) YIELDS IN NORTHERN TANZANIA BASED ON SPECTRAL ANALYSIS OF SOILS

Charlotte Mallet

malletcharlotte@hotmail.com

University of Amsterdam

Institute for Biodiversity and Ecosystem Dynamics (IBED)

Master Earth Sciences – Environmental Management track

Master thesis research Earth Sciences | 52641MTR0Y | 30ECTS

Student ID: 11574348

Examiner: Dr. L.H. (Erik) Cammeraat

Co-assessor: Dr. A (Albert) Tietema

Wageningen University & Research

Plant Production Systems (PPS)

Master thesis Plant Production Systems | PPS–80430

Registration number: 940511541060

Supervisor: Dr. ir. J (Joost) van Heerwaarden

Cover photo: Bush bean field in Mawanjeni (Kilimanjaro region, Tanzania).

Table of contents

Acknowledgments

I would like to thank Abba Kamulali and Deogratias Augustine, the field liaison officers of N2Africa in Moshi and Lushoto. They were of a great help for me during my time in Northern Tanzania. Every day of work was an adventure, going around with the “piki-piki” (motorbike), and discovering new villages and people. Thank you for your help, your dedication, your driving talent and above all your friendship. A big thank you also to the extension officers, and to all the farmers that took the time to receive us, answer our questions and show us their field. Thanks to them I got introduced to a whole new world.

A special thanks to my supervisor at Wageningen University, Joost van Heerwaarden, for welcoming me to the Plant Production Systems group and to N2Africa and for his guidance all along my thesis project. I appreciated the support from the other PPS students during the thesis ring sessions. I also want to thank Erik Cammeraat, who supported the project enthusiastically, and Albert Tietema for accepting to be the co-assessor of this thesis. Finally, I am grateful for the support I received from the IITA in Dar Es Salaam, especially Frederick Baijukya and Abubakari Mzanda for the logistical support, from SARI in Arusha and the ICRAF staff in Nairobi, especially Erick Towett and Andrew Sila.

Abbreviation table

DAP	Diammonium Phosphate
EVI	Enhanced Vegetation Index
ICRAF	International Centre for Research in Agroforestry
IITA	International Institute of Tropical Agriculture
Masl	Meters above sea level
MIR	Mid infrared
OC	Organic carbon
OOB	Out-of-Bag
PC	Principal Component
PCA	Principal Component Analysis
PCR	Principal Component Regression
PLSR	Partial Least Squares Regression
RF	Random Forest
RMSE	Root Mean Square Error
SARI	Salien Agriculture Research Institute
SOM	Soil Organic Matter
SSA	Sub-Saharan Africa
TLU	Tropical Livestock Unit

Abstract

Common beans (*Phaseolus vulgaris* L.) are a major subsistence crop in sub-Saharan Africa. Due to their capacity to fix atmospheric nitrogen, they are a very important source of dietary protein for the population and contribute to soil fertility. Nonetheless, common beans are known to have a low productivity and a high variability in yields. A better understanding of what causes this variability is necessary to adopt the right practices that will enhance the productivity of common beans. With this aim in mind, we analysed bush bean (a type of common bean) yields and responses to yield-improving treatments from on-farm try-outs that were led in 2016 and 2017 in two districts of Northern Tanzania as part of the N2Africa project. Detailed information about the trials was contained in a dataset, to which we added soil information during a soil sampling campaign that took place from May to July 2018. Mid-infrared diffuse reflectance spectroscopy was used to analyse the soil samples. It is a relatively recent technique that is faster and cheaper than conventional wet chemistry analyses. It allows large-scale analysis of soils but is still controversial as its accuracy is not always optimal. We aimed to assess the accuracy of soil properties predicted from soil diffuse reflectance spectra, comparing them with soil properties obtained from conventional wet chemistry analyses. The possibility of predicting bush bean yields and responses from soil information was evaluated. Moreover, we assessed the explanatory and predictive capacity of several other variables to find out which were the most important for bush bean yields and responses. This also allowed to get an insight into the importance of soil properties relative to other variables. In addition, a household typology was constructed using multivariate statistical techniques and the available information about household characteristics. The influence of the typology on yields, responses and soil properties was evaluated to get a better understanding of the constraints faced by different farmers.

We found that the accuracy of spectrally-predicted soil properties was lower than what can be found in literature. The accuracy was assessed on an independent sample set, unlike most studies where internal cross validation techniques are used, which is part of the reason for lower accuracy. We also found that the wet chemistry measured soil properties were a better predictor of yields and response than the information provided by spectroscopy. However, when other variables were considered next to soil properties, it appeared that the latter did not have a significant role in explaining and predicting yields and responses. Instead, management variables, such as the fertiliser and bush bean variety used, were found to be the major factors influencing yields and responses, along with environmental variables (temperature, altitude and precipitation) for the control yields. The household typology revealed some patterns in yields, with a low resource endowed category having significantly lower yields than the higher resource endowed categories. Nonetheless, the absolute response and the soil properties did not differ between household types. We conclude from our results that the potential response to yield-improving treatments is the same for all farmers, regardless of their soil and household characteristics. It is therefore important to reduce inequalities between farmers in terms of access to inputs and labour. This will help all farmers to apply good management practices (such as timely planting, weeding and the use of appropriate improved varieties) and fertilisers in their fields, since we have the indication that these are efficient ways to achieve higher yields.

1. INTRODUCTION

1.1 Challenges of food production in sub-Saharan Africa and the role of legumes

In 2010, 30% of the population in sub-Saharan Africa (SSA) was undernourished (FAO, 2010). This figure is not expected to decrease since the annual average population growth rate of 3% occurring in the region is higher than the annual increase in food production (2%, including expansion of harvested area) (Tittonell & Giller, 2013). Smallholders constitute the majority of farmers in SSA, since 80% of all farms are smaller than 2 hectares. Farm sizes are expected to shrink in the near future as the available land is decreasing and farms gets divided among children in the family. Producing enough to sustain their families and livelihoods is thus becoming increasingly difficult for smallholder farmers (Bremner, 2012). Poor soil fertility and poor nutrient availability are widely recognised as the main constraints limiting productivity in smallholder farming systems in SSA (Giller et al., 2011; Tittonell & Giller, 2013; Vanlauwe & Giller, 2006). There is therefore a need for sustainable intensification of agriculture in the region, i.e. to achieve greater food production per unit area while maintaining, if not even replenishing, the soil nutrient stock. Legumes are a precious tool to reach these two goals. They contribute to soil fertility through their capacity to fix atmospheric N_2 in their tissues thanks to a symbiotic association with the N_2-fixing bacteria Rhizobia. Indeed, grain legumes can contribute up to 300 kg N ha^{-1} per growing season, thereby reducing the need for N fertiliser (Franke et al., 2018; Vanlauwe & Giller, 2006). It is proven that significant yield increases are obtained in cereal crops when they are grown in rotation with legumes, in comparison to continuous cereal cropping systems (Franke et al., 2014; Franke et al., 2018). Legumes are able to provide multiple residual benefits, in addition to the increase (or non-depletion) of soil N and impact on subsequent non-N_2-fixing crops (Franke et al., 2018; Giller, 2001).

This study will attempt to understand better the variability in grain legume yields by determining which are the variables with most impact on yields. This will allow to have a better knowledge of how to improve yields of grain legumes, and increase its use among smallholder farmers.

1.2 Importance of common bean

The grain legume common bean (Phaseolus vulgaris L.) exists in two varieties with different growth habits: climbing beans and bush beans. This study is about bush beans, however a lot of information in literature is written about common beans without specification of the variety. Therefore, in this introduction, the information is given on common beans without a distinction between both varieties.

Common bean is the most consumed grain legume in the world and is one of the most important staple and cash legumes in the tropical world (Broughton et al., 2003; Hillocks et al., 2006). In Africa, common bean is mainly grown for subsistence and is the second source of protein after maize (Jones, 1999; Kaizzi et al., 2012). The protein content of bean seeds ranges from 20 to 25%, mostly in the form of Phaseolin. Phaseolin is deficient in amino acids containing sulphur, such as methionine, that are usually found in cereal seed proteins. The latter are deficient in other essential amino acids such as lysine, which are contained in bean seeds. The nutritional complementarity of both crops is essential, and a cereal-legume ratio of 2:1 is required for a balanced diet (Broughton et al., 2003; Giller, 2001). However, legume yields are often lower than cereals yields and this ratio is rarely achieved (Broughton et al., 2003; Franke et al., 2018). Beans are a major source of micronutrients, providing iron, phosphorus, manganese and magnesium in the diet, as well as zinc, copper and calcium in a smaller extent (Broughton et al., 2003). They are also rich in fibre, and contain vitamins B complex (Reichert et al., 2015). The main consumption product of common beans are the dry grains (Broughton et al., 2003; Hillocks et al., 2006).

In Africa, beans are often grown intercropped (Jones, 1999), mostly with maize in Eastern Africa (Wortmann et al., 1998). They are also often grown without sufficient input, if any, resulting in low yields and a declining soil fertility (Jones, 1999; Lunze et al., 2012). In the district of Lushoto in Northern Tanzania, a survey done in 6 villages found that inorganic fertilizers were used in only 5% of the farms, and only 9% were using organic manure (Mowo et al., 2006).

In Tanzania, common bean is the most important grain legume. The country is the second biggest producer of dry beans in SSA, and among the twenty most important producers worldwide. It is mainly grown by smallholder farmers as a staple food crop. In the Arusha and Kilimanjaro regions in the north of the country, a larger part of

the production is intended for export thanks to a suitable climate and the vicinity of an international airport (Hillocks et al., 2006).

1.3 Environmental constraints to common bean production

Despite its great importance in the Tropics, common bean is known as a low-yielding crop. When grown as a sole crop, grain yields of common beans are around 200 kg ha^{-1} in unfavourable environments and around 700 kg ha^{-1} in more suitable conditions (Lunze et al., 2012). For Tanzania, a national average of 500 kg ha^{-1} was reported in 1998 by Amijee and Giller. Around 90% of common bean production worldwide is occurring at low average yields, due to pests, diseases and non-suitable climatic and edaphic conditions (Jones, 1999).

Environmental stresses affecting the survival of rhizobial population in the free-living state, root nodulation or N_2 fixation, and in turn legume yields, can be divided into two categories: physical and chemical factors (Giller, 2001). The physical factors consist of high temperatures, droughts, salinity and waterlogging. The Atlas of common bean production in Africa (Wortmann et al., 1998) indicates that a favourable climate for common beans consists of 450mm of precipitation per growing season and temperatures between 15 and 23°C. Higher temperatures can cause high soil temperatures (> 30°C), that in turn can kill soil bacteria, thus inhibiting N_2 fixation and decreasing yields (Amijee & Giller, 1998; Giller, 2001). Droughts have the same effect of reducing bacteria populations and inhibiting N_2 fixation. Sandy soils are particularly prone to high temperatures and so are not favourable for bacteria, but also prevent the occurrence of waterlogging, which reduces plant growth and has a negative impact on the N_2 fixation process (Giller, 2001). As a result, common beans grow best in a deep well-drained soil with a sandy loam, sandy clay loam or clay loam texture (Lunze et al., 2012). These physical constraints are not expected to be the main factors limiting yields in the areas this study is focusing on. Indeed, the study area is made of highlands where high soil temperatures are unlikely to occur, precipitation is sufficient and the hilly landscape is not prone to waterlogging. Salinity, which affects the survival and growth of rhizobia in the soil as well as the N_2 fixation process (Giller, 2001), is also not expected to occur in the study areas (F. Baijukya, personal communication, August 2018).

Chemical factors affecting legume growth are soil acidity and nutrient deficiencies (Giller, 2001), and thus exclusively concern edaphic conditions. Low soil fertility, a result of these chemical constraints, is even the main cause of low yields of common beans in Eastern, Central and Southern Africa according to Lunze et al. (2012). The major soil fertility problems in Eastern Africa are N and P deficiency, followed by soil acidity and Al/Mn toxicity (Hillocks et al., 2006; Lunze et al., 2012). Specifically, in bean production areas of Eastern Africa, 65% of soils are P deficient, 50% are N deficient and 45 to 50% have a pH lower than 5.2 (Wortmann et al., 1998).

As for soil acidity and toxicity, the optimum pH for common beans ranges between 5.8 and 6.5 and Aluminium saturation should not be higher than 10% (Lunze et al., 2012). Grain legumes are particularly sensitive to the other chemical constraint, nutrient deficiencies, as several nutrients are essential to the nodulation and N_2 fixation processes, including P, Ca and Mg. However, these nutrients are subject to leaching or deficient in acid soils. Other nutrients such as K do not play a direct role in N_2 fixation but are of great importance to the growth of every plant and are commonly deficient in tropical soils (Giller, 2001). In addition, many micronutrients (such as Co, Mo and Fe) play a direct role in the metabolism of rhizobia and hence can affect legume productivity (O'Hara, 2001). Soil properties are therefore likely to be an important limiting factor to bean yields in tropical areas.

In Tanzania, Amijee and Giller (1998) found that the main reason for low nodulation and low vigour in common beans was a lack of soil P. In the Usambara Mountains (North Eastern Tanzania), K deficiency was found to be another major constraint for common bean yields along with P deficiency (Smithson et al., 1993). Ca and Mg are becoming increasingly depleted in tropical soils (Agegnehu & Amede, 2017) and might now be another important constraint to common bean growth in the Tropics. Table 1 gives an overview of different soil attributes, which conditions cause deficiencies and their role in N_2 fixation as well as their deficiency levels.

Table 1. Several soil attributes measured by conventional soil tests, with deficiency level values, factors limiting their availability and their role in N_2 fixation.

Soil attributes	Deficiency level	Comments	Source
Extractable Phosphorus	< 7 mg kg^{-1}	Low availability for plant uptake in acid soils, subject to leaching. Required for the nodule metabolism in legumes.	Giller, 2001; Makoi & Ndakidemi, 2008; Shepherd & Walsh, 2002
Total Nitrogen	< 2 g kg^{-1}	Deficient in acid soils.	Ndakidemi & Semoka, 2006
Exchangeable Calcium	< 5 $cmol_c$ kg^{-1}	Deficient in acid soils, subject to leaching. Role in the early stages of infection of the legume roots by rhizobia.	Giller, 2001; Ndakidemi & Semoka, 2006
Exchangeable Magnesium	< 2 $cmol_c$ kg^{-1}	Deficient in acid soils, subject to leaching. Role in the early stages of infection of the legume roots by rhizobia.	Giller, 2001; Ndakidemi & Semoka, 2006
Exchangeable Potassium	< 0.2 $cmol_c$ kg^{-1}	Mobile nutrient, readily leached; deficiencies common in humid areas with continuous cropping. No specific role in legumes.	Giller, 2001; Ndakidemi & Semoka, 2006; Shepherd & Walsh, 2002
Soil Organic Carbon (SOC)	< 20 g kg^{-1}	Major impacts on soil quality, through provision of cation exchange capacity and water-holding capacity, release of nutrients and sustaining soil microorganisms biodiversity, between other functions.	Ndakidemi & Semoka, 2006; Brady & Weil, 2010

1.4 Variability of bean yields

Variability in legume yields is large (Cernay et al., 2015). This affects the benefits obtained by individual farmers from growing grain legumes (Franke et al., 2016; Ronner et al., 2016). An average yield increase observed on a number of farms after the application of a technology does not mean that a substantial yield increase occurred in every farm. Understanding this variability could help targeting technologies to farms where their efficiency is predicted to be greater and avoid unnecessary risks for farmers where these technologies are likely to be inefficient.

Several studies have attempted to explain the yield variability of legumes. Ronner et al. (2016) studied the variability in soybean yields in Nigeria by comparing on the same field yields after fertiliser and inoculant applications to yields without these inputs, called control yields. They found that the responses to the inputs were related to the control yields: the largest absolute responses occurred when the control yields were higher than 500 kg ha^{-1}. They however also note that small responses occurred at all levels of control yields. Soil properties did not explain much of this variability in control yields and responses. Franke et al. (2016) also found that control yields were significantly correlated with responses to fertiliser in climbing beans in Northern Rwanda, but without finding a link with soil properties. When studying climbing beans in Uganda, Ronner et al. (2018) found a large variability in yields in farmer-managed trials, which compared yields on a climbing bean plot grown with the farmer's practice and one with improved management practices. They even found that on average the latter plot did not have better yields than the farmer's plot. Moreover, it was observed in Northern Tanzania that crop growth was strongly impacted by the position in the slope, which relates to soil depth and drainage (K. Giller, personal communication, April 2018).

It is clear that soil properties have a role in the determination of bean yields, as stated by Lunze et al. (2012), but their impact on, and predictive value for, yields and response are not yet fully understood. The quantification of their importance in explaining and predicting the variability of yields and response in comparison to other variables, such as environmental variables that can be obtained from remote sensing (e.g. rainfall, temperature, slope,…) and agronomic practices, would also help understand further the causes of this variability.

1.5 Soil properties analysis

Finding the relations that can predict yields based on soil properties requires the possession of an abundance of soil data of good quality. This is not feasible without a technology that allows fast and reliable soil analyses at a reduced cost. Conventional wet chemistry soil analyses are not appropriate for this use as they are time-consuming and costly.

To tackle the increasing demand for good and large-scale soil data, alternative techniques for soil analysis have been developed, such as reflectance spectroscopy. This technique projects light at different wavelengths on a sample and measures the light that is reflected from this same sample. The reflectance is defined by the mineral and organic composition of the sample, and the measurements can be shown as a spectrum (see Fig. 1 for an example of soil spectra). It is a non-destructive technique, which allows cheap, rapid and reproducible analysis of soil properties (Shepherd & Walsh, 2002; Towett et al., 2015; Wetterlind et al., 2013). Moreover, a single measurement can analyse several soil properties simultaneously and does not require much sample preparation, nor the use of chemicals that can be hazardous or environmentally harmful (Du & Zhou, 2009; Shepherd & Walsh, 2002; Wetterlind et al., 2013).

There are two possibilities of use of the information provided by spectroscopy: the first is to translate the spectrum into soil properties and then use these soil properties to obtain information about yields; the second is to use the spectral information to directly predict crop yields and responses.

The first possibility requires the selection of a subset of the sample population as reference set. This reference set is chosen to be representative of the spectral variation of the whole population. The reference samples are analysed with conventional soil analyses, and the measured soil properties are then calibrated to their reflectance spectra (Shepherd and Walsh, 2002). The obtained calibration model allows to predict soil properties for the rest of the population, based on their spectra. This is the case of Mid Infrared (MIR) Diffuse Reflectance spectroscopy, which will be used in this study (Shepherd & Walsh, 2002; Vågen et al., 2010; Wetterlind et al., 2013). A current challenge for the development of the technique is to build spectral libraries that are large and varied enough to capture much of the variation in soils globally (Towett et al., 2015). This would permit to not have to perform reference analysis for each soil population while rendering results that are still accurate.

Table 2. Accuracy of predicted soil properties with MIR spectroscopy as presented in two papers. In the first one (Towett et al., 2015), values were obtained from random forest out-of-bag validation. The values come from around 700 samples collected throughout Sub-Saharan Africa. The second one (Soriano-Disla et al., 2014) shows median r^2 obtained from a review of several studies. The number of studies is shown between brackets and the validation methods are a mix of cross-validation and validation on an independent test set.

Soil attribute	Towett et al. (2015)	Soriano-Disla et al. (2014)
	r^2	r^2
OC	0.90	0.93 (20)
Total N	0.86	0.90 (14)
Exch. Ca	0.84	0.82 (8)
pH	0.82	0.75 (18)
Sand	0.74	0.83 (11)
Exch. Mg	0.73	0.74 (7)
Clay	0.73	0.80 (14)
Silt	0.60	0.63 (8)
Exch. K	0.51	0.37 (12)
Extractable P	0.10	0.35 (11)
Exch. Na	0.07	0.32 (6)

Table 2 shows that for some soil properties such as organic Carbon, total N and Ca the predictions with MIR are highly accurate (note that what is meant here by accuracy is the correlation between values predicted from spectroscopy and values measured with conventional wet chemistry analyses), whereas the predictive value of MIR is low for other soil properties such as K and P, even though these two nutrients are of great importance for soil fertility and crop growth. However, Towett et al. (2015) argue that conventional chemical analyses giving measures of extractable nutrients are not directly relevant to soil fertility, as they do not characterize the ability of the soil to re-supply the soil solution with nutrients. In other words, they do not indicate the nutrient buffering capacity of the soil, whereas MIR spectroscopy gives good predictions of several properties that relate to nutrient buffering capacity (e.g. organic carbon, clay and sand content). The hypothesis of the authors is that MIR soil analyses could predict crop yield responses to the application of nutrients as well, or even better, than conventional soil tests based on soil extracts (Towett et al., 2015).

The second possibility of use of spectroscopy, which consists of predicting crop yields and responses directly from the raw soil spectra (Fig. 1), has been researched by Tittonell et al. (2008) and during the AfSIS Diagnostic Trials Projects (Kihara, 2014). These two studies found some relation between maize yields and soil reflectance

spectra. Their results are promising but show the need for more research in this topic, to assess further the value of soil spectra in yield prediction and find out if this also applies to other crops than maize.

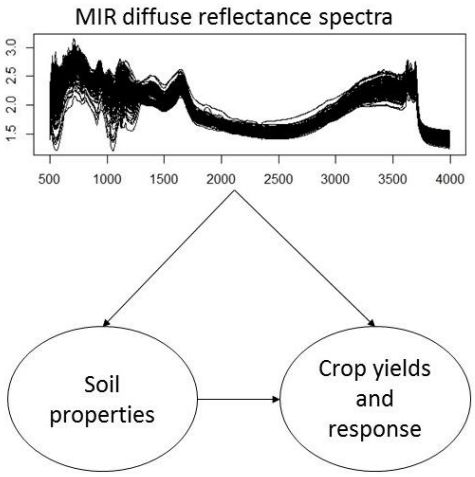


Figure 1. The possibility of a direct route linking soil spectra and crop yields and response. All the arrows present in the figure are tested in this study.

1.6 Household characteristics and soil fertility

Franke et al. (2016) showed that household characteristics of farmers were significantly related with yield variability and responses to fertiliser application. They found that the occurrence of non-responsive soils, i.e. soils with poor control yield and poor response to fertiliser, varied with the household category. Hardly any of the soils in the wealthiest households could be classified as non-responsive, whereas up to 25% of poor households' fields were non-responsive. The study also found that several soil properties were significantly related to household class, showing that wealthier households were located on fields with higher fertility. These differences in soil properties between household categories are likely due to differences in past management, in particular the application of mineral fertilisers and organic manures (Franke et al., 2016). The study also emphasizes that the adoption of technologies in farm households is affected by socio-economic characteristics such as education level, livestock ownership, farm size and diversification of income. Wealthy farmers are more able to invest in fertilisers and improved seed varieties, they tend to own more livestock and can thus apply higher quantity of organic manure to their fields (Mowo et al., 2006). Through the use of hired labour they are also better able to apply best agronomic practices such as weeding and early planting (Chikowo et al., 2014). In a review of several studies on soil fertility and farm typologies, Chikowo et al. (2014) concluded that land and livestock ownership were the two most important household characteristics determining the wealth status of farmers and their farm productivity. They also found that organic carbon and available phosphorus were often good indicators of resource endowment of the household.

A way to approach the diversity in household characteristics among farms and explore how this diversity influences yields and responses to a technology is to build a farm typology. This can be useful to target interventions to farm types with higher potential and in scaling-out a technology. There are several methods to construct a typology, including the use of multivariate statistics to determine farm types based on quantitative information. This approach has the advantage of being reproducible and fast to perform (Alvarez et al., 2014; Kuivanen et al., 2016). The wealth status of households cannot be inferred from one single household variable. It is rather the information brought by several of these variables put together that can give a useful indication of the wealth status of households, and perhaps of the soil fertility and farm productivity. This is why building a household typology is thought to help in revealing how the household characteristics have an impact on yields and soil properties, with poorer households possibly having lower yields and lower soil fertility.

2. RESEARCH AIMS AND RESEARCH QUESTIONS

2.1 Research aims

This study is based on adaptation trials made during the N2Africa project in 2016 and 2017 in two districts of Northern Tanzania. The adaptation trials consisted of comparing the performance of bush beans grown on two types of plots in the farmer's fields: the control plot, where the bush beans were grown with the usual farmer's practice, and the N2Africa plot where N2Africa technologies such as fertiliser and improved varieties were applied. More about these adaptation trials will be said in the method section.

The first aim of this study is to assess the accuracy of spectrally-predicted soil properties, by comparing them to measurements from conventional soil analyses.

Another aim is to determine whether soil information can be used as a predictor for bush bean yields and responses to a treatment. For this, an examination of which soil properties (both from wet chemistry measurements and spectral predictions) are significantly related to yields and responses is made and their predictive value is assessed. The explanatory and predictive value of the raw soil reflectance spectra for bush beans yields and response is also assessed.

The last aim of this study is to explore the relation between the household characteristics, the yields and responses and the fertility status of fields in each farm. Other groups of variables, such as climatic variables and management practices, are also considered in an attempt to explain and predict yields and responses and to study the predictability added by each group of variables.

2.2 Research questions

Question 1: Is Mid-Infrared Diffuse Reflectance spectroscopy an accurate method for estimating soil properties?

Sub-questions: Which soil properties are accurately predicted by spectral methods compared to conventional wet chemistry measurements, and which are not?

Hypothesis: It is expected that the same patterns as shown in Table 2 will be observed, with K and P particularly poorly predicted by MIR spectroscopy.

Question 2: Can soil properties be used as a reliable predictor of yields and responses to N2Africa technologies applied to bush beans?

Sub-questions: Which soil properties are limiting bush bean yields the most in both areas? Which soil properties are significantly related with bush bean yields and responses? Do they have a predictive value for bush bean yields and responses? Can soil reflectance spectra be used as a reliable predictor of bush bean yields and responses to N2Africa technologies?

Hypothesis: It is expected that P and K are the major constraints for bush bean yields and that a clear pattern will be found of low yields and low levels of P and K. Another edaphic constraint that is expected to significantly impact yields is the slope position, as it relates to important soil characteristics for common beans that are soil depth and drainage. As for the direct predictive power of MIR reflectance spectra for yields and response, not enough detailed research has been done in this field before to be able to formulate a precise hypothesis.

Question 3: Which other variables, besides soil properties, have a predictive value for bush bean yields responses to N2Africa treatments?

Sub-question: Does a farm typology based on the exclusive use of multivariate statistics provide sensible information about the productivity and the fertility level of a farm? Which categories of variables give added predictability for yields and efficiency of N2Africa treatments?

Hypothesis: It is expected that some household characteristics such as livestock ownership will relate to soil fertility levels and farm productivity, but that the use of a multivariate statistical analysis exclusively to create farm typologies will not be sufficient to provide sensible information about the farming systems and their relation with soil fertility and crop yields obtained on a farm.

3. MATERIALS AND METHODS

In this chapter, the original dataset resulting from the adaptation trials of 2016 and 2017 is first described. Then the study areas and their own characteristics are presented. Next, it is explained how the original dataset was processed and how the farms that would be sampled during fieldwork were chosen (section 3.3). The next section describes the fieldwork activities and is followed by a description of the laboratory analyses done on the soil samples (section 3.5). The last section of the chapter contains a detailed explanation of the data analysis.

3.1 Description of the dataset

This study is part of N2Africa, a project that aims at improving yields of grain legume crops in smallholder farming systems in Africa. The research was based on an existing dataset that contained information collected during N2Africa trials in 2016 and 2017 in Moshi and Lushoto, two districts of North-eastern Tanzania. This dataset is referred to as the original dataset in the rest of this document.

The original dataset[1] contained data about bush bean yields in the two growing seasons of 2016 and in the first growing season of 2017. The data came from farms that participated during only one of the mentioned seasons in focal adaptation trials of N2Africa technologies. These technologies included the use of improved legume varieties and fertilisers. Adaptation trials were farmer-managed try-outs of different N2Africa technology packages. They were meant to assess the performance of the different technologies compared to the farmer's own practice, and to evaluate the implementation of the technologies by farmers. The farmers were given management recommendations to apply on the adaptation trials, but as these were meant to be farmer-managed trials, the farmers could choose to follow these recommendations or not. The adaptation trials consisted of a 10x10m plot planted with the N2Africa technology package located next to the farmer's legume field, in which a plot of 10x10m was delineated, as shown in Figure 2 ("General Guidelines for 2017 (Focal) adaptations," n.d.). The yields from both plots could be easily weighted and compared after harvest. In the rest of this document, the plot planted with the N2Africa technology package is called the N2Africa plot, and the plot where the farmer's practice was applied is called the control plot or control section.

In addition to yield measurements, information about cropping history, management practices and household characteristics was collected among these focal adaptation farmers. The dataset did not contain information about soil properties but contained information such as the amount and type of livestock owned, the level of education in the household, the household size, the frequency of use of hired labour, the occurrence of months with food shortage and the number of different income sources of the household.

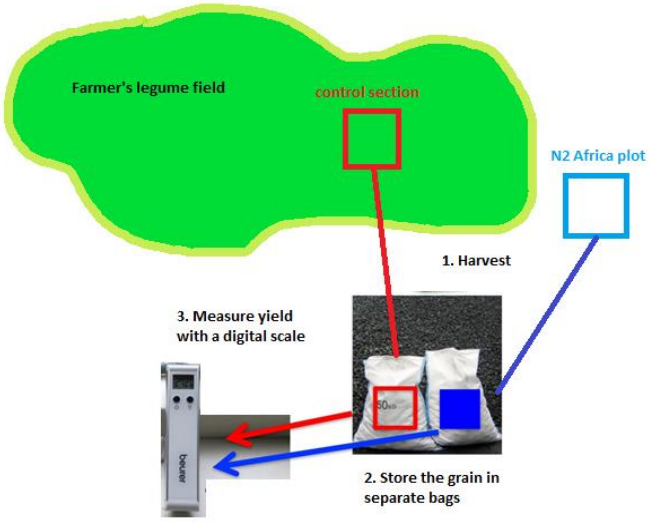


Figure 2. Illustration of an adaptation trial in N2Africa ("General Guidelines for 2017 (Focal) adaptations", n.d.)

[1] The dataset can be downloaded at https://n2africa.shinyapps.io/Agronomy_test/

The N2Africa technology packages (also referred to as treatment in the rest of this document) used in the trials of the study areas in 2016 and 2017 included the use of improved bush bean varieties and fertilisers. In 2016, the fertilisers used were Diammonium Phosphate (DAP) and NPK while in 2017, only NPK was used. Three different improved varieties were used in 2016, and two in 2017. Improved varieties are the result of breeding programs that aim to combine characteristics such as higher yields, resistance to diseases and pests and tolerance to low soil fertility and drought (Hillocks et al., 2006).

It is worth emphasising that the control plot was the plot with the farmer's own practice and thus did not have to be without fertiliser. There could be fertiliser applied if that was already the farmer's practice, and the fertiliser could be different or the same as the fertiliser provided for the N2Africa plot. Moreover, it appeared that in Lushoto, during the trial seasons considered in this study, fertilisers were given to the farmers for both the N2Africa plot and the own plot. This was thus different from the original instructions of the adaptation trials and means that in Lushoto, the yield difference between both plots is solely due to the use of an improved bush bean variety versus a local variety. This information was only revealed during the fieldwork in 2018.

3.2 Description of the study areas

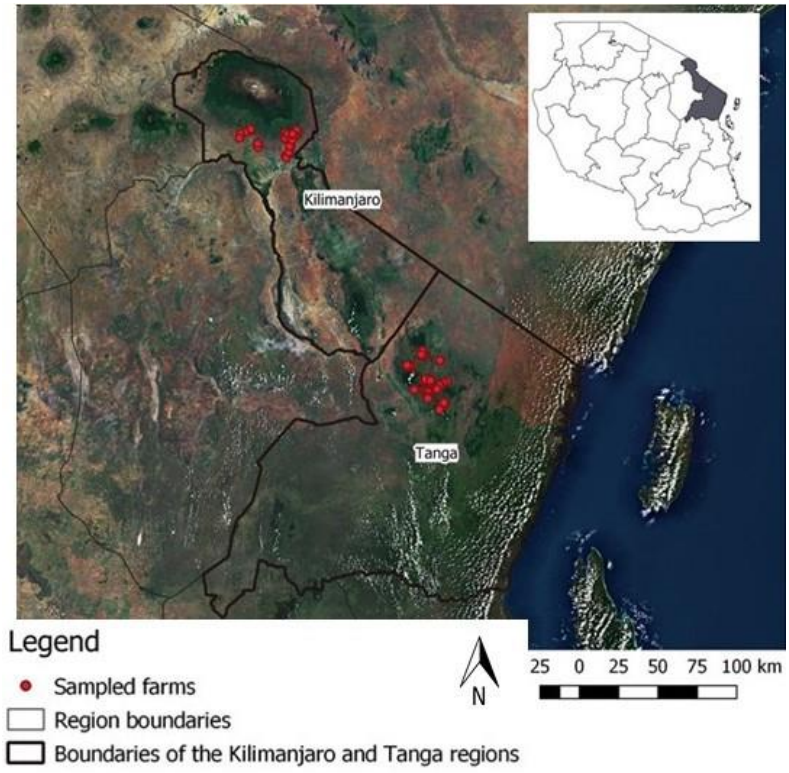


Figure 3. Map of Northern Tanzania, with the location of the sampled farms marked in red.

The study areas are located in the district of Moshi (Kilimanjaro region) and Lushoto (Tanga region). The Lushoto district is located in the Western Usambara Mountains. The Northern part of Tanzania has a bimodal rainfall regime, with a long rainy season from March to May and a short rainy season in November and December (Wickama & Mowo, 2001). The main cropping season depends on the long rainy season and the harvest takes place in July and August. The harvest of the secondary cropping season takes place in January and February (FAO, 2018). Although the country is generally food secure, due to consecutive below average harvests since 2016, poor households are in a stressed phase in some areas of the North-eastern part of the country including the Kilimanjaro and Tanga regions (FAO, 2018).

Mbaga-Semgalawe and Folmer (2000) made a review of the land use history in the North-eastern mountains of Tanzania, which include the Kilimanjaro and Western Usambara mountains. A summary of this evolution is given here. Until the 20th century, natural forests were covering most of the mountains of North-eastern Tanzania. Only a small part of the arable land was cultivated, and practices such as shifting cultivation and fallows were implemented to maintain soil fertility. The population started to increase in the 1920's and the rate of land conversion for cultivation increased as well, until all arable land was cultivated by 1936. The pressure on land was thus very high and soil conservation practices were not anymore in use. In addition to the population increase, the colonial governments (Germany until the end of World War I and Great Britain afterwards) participated to the increased land pressure with the establishment of coffee and tea plantations since the 19th century and with the creation of forest reserves. As a result, farmers started to cultivate on very steep slopes and to encroach on forests, valley bottoms and wetlands. Soil became intensively eroded and land productivity started to decrease. At the same time, farmers adopted other soil conversation practices such as mulching, intercropping, crop rotation and minimum tillage to reverse the decrease in soil productivity. The colonial government also tried to reverse soil degradation and initiated diverse activities to fight erosion and educate the local population about soil degradation. They implemented very strict laws that prohibited cultivation on steep slopes and in forest reserves. After independence in 1961, soil conservation was left behind and formerly prohibited areas were again cultivated. Deforestation intensified as more and more forested land was allocated to cultivation or grazing areas. In the late 1970's-early 1980's, new soil conservation programs were launched by the Tanzanian government to halt soil erosion and even try to restore topsoils. In Lushoto especially, many programs were developed to raise awareness and promote adoption of soil conservation technologies such as agroforestry, afforestation, water harvesting and irrigation.

3.2.1 Lushoto

The Western Usambara Mountains are part of the Eastern Arc Mountains, a chain of mountains that starts in Southern Kenya and goes down to central Tanzania. The Eastern Arc Mountains comprise 13 mountains blocks and are famous for their high levels of endemic animals and plants. These mountains were uplifted at least 30 million years ago and their climate is characterised by the direct influence of the Indian Ocean (Burgess et al., 2007). The eastern and south-eastern slope are facing the Indian Ocean and are thus wetter, and the western and north-western slopes are drier. Before human activity started to have a major influence on the landscape, this difference in slopes could be seen by the difference in forest cover: the wetter slopes had a continuous forest cover at all elevations, and the drier slopes were covered by deciduous woodland at lower elevations and evergreen forest only at higher elevations. A hypothesis to explain the high level of endemicity in the Eastern Arc Mountains is that these high elevation areas have been protected from climate fluctuations and thus from periods of extreme arid climate. This stable climate promoted a persistent forest cover which participated to reduced rates of extinction (Burgess et al., 2007).

The field sites in the Lushoto district are located at an altitude that range from 1150 to 1850 meters above sea level (masl). The annual rainfall in the Usambara Mountains ranges from 1000 to 2000 mm (Vice President's Office: Division of Environment, 2007). Temperatures usually range between 18 and 23°C, with a peak in March, and July is the coldest month (Wickama & Mowo, 2001). The fields are often located on steep slopes that are highly degraded and intensively cultivated. A study found that the main soil fertility constraints were P deficiency, followed by N deficiency (Ndakidemi & Semoka, 2006), whereas other studies found that K is the most yield limiting nutrient in the district of Lushoto (Amijee & Giller, 1998; Mowo et al., 2006; Smithson et al., 1993). Mowo et al. (2006) observed a trend in soil fertility along the slopes in Lushoto: the pH, organic matter, K and P increased down the slope due to erosion causing a downward movement of nutrients.

3.2.2 Moshi

The Kilimanjaro region is the third most densely populated region of the country, and one of the wealthiest. Around 75% of the population lives in rural areas. The annual rainfall is also between 1000 and 2000 mm but is very variable, more than in the Western Usambara Mountains (Vice President's Office: Division of Environment, 2007). The Moshi district is located at the foot of Mount Kilimanjaro. Mount Kilimanjaro is less than 2 million years old and is thus a much younger geological formation than the Eastern Arc Mountains (Burgess et al., 2007;

United Nations Development Program, n.d.). The Kilimanjaro region can be divided in three zones according to land use: the Kilimanjaro Mountain Peak Zone, lying above 1800masl and where is located the Kilimanjaro National Park, the Highlands Zone between 900 and 1800masl and the Lowlands Plains below 900masl. The field sites of this study that are located in Moshi ranged from 700 to 1300masl and are thus located in the Highlands Zone and the Lowlands Zone. The annual average precipitation in the highland zone ranges from 1250 to 2000mm and the temperatures from 15 to 20°C. The lowlands zone receives less rainfall, with an annual average precipitation between 700 and 900mm, and a warmer climate, with temperatures commonly reaching 30°C and above. The two zones are also very different in terms of population density, as the Highlands count 650 inhabitants km^{-2} and the Lowlands count below 50 inhabitants km^{-2}. The lowlands have less favourable climatic conditions compared to the highlands and face frequent droughts and floods. The majority of soils in the Kilimanjaro region are of volcanic origin, which are usually fertile and rich in Ca and Mg. However, these soils have been under continuous cropping for more than two centuries. This, combined with the heavy rains that occur especially in the higher altitudes, caused an intense leaching and a loss of soil fertility. Large areas of the southern slopes have been deforested to give place to coffee plantations since the late 19^{th} century (United Nations Development Program, n.d.).

The southeastern and eastern slopes in the Highlands zone receive more rainfall than the northern and western sides and are characterised by the presence of the Chagga homegardens. The Chagga are one of the two main tribes of the Kilimanjaro region. They have developed in their homegardens a unique farming system with a high species diversity integrating trees, shrubs, food and cash crops and livestock (see Appendix I for a figure showing an example of such garden). The trees provide timber and along with the shrubs provide shade for coffee. This system contributes to a reduced vulnerability to pests or disease and the high diversity of plants ensures a continued source of food and livelihood. With a continuous ground cover and high nutrient cycling, the system allowed the Chagga to keep a sustainable farming system in areas very prone to erosion and is a model for soil conservation practices. Every family also has land in the drier Lowlands zone where this farming system is not applied and is mainly used to grow annual crops (Fernandes et al., 1984).

3.2.3 Soil types of the study areas

According to the most probable soil types predicted in SoilGrids (ISRIC, n.d.), more than 90% of the farms are located on Haplic Acrisols, Haplic Ferralsols and Haplic Nitisols (the latter including a much smaller number of farms than the two first soil types). The rest of the farms are located on four other soil types, non-dominant in the study area. The qualifier “haplic” means that it is not expected to find outstanding features that are worth mentioning, thus that the soil is typical of its Reference Soil Group (i.e. here Acrisols, Ferralsols and Nitisols). Moshi is dominated by Ferralsols and the second most abundant soil type is Nitisols, whereas Lushoto is dominated by Acrisols and Ferralsols are the second most abundant soil type.

Acrisols are strongly acid and weathered soils from acid parent rock. They have a clay-rich horizon in the subsoil dominated by low-activity clays and with a low base saturation (IUSS Working Group WRB, 2015). They have a rather poor physical structure (Giller, 2001). Acrisols are not suitable for sedentary farming. If practiced on such soils, crop cultivation requires liming and fertilisation, and a careful management to preserve the organic matter and avoid erosion (IUSS Working Group WRB, 2015). They are the second most common soil in Tanzania, covering about 9% of the country (Mlingano ARI, 2006).

Ferralsols are the typical soils of the hot and humid Tropics. Their main characteristic is the dominance of kaolinite and oxides. They develop more frequently from basic parent material than acid material. The intense weathering occurring in the humid Tropics causes a deficiency in weatherable minerals and a residual enrichment in resistant primary minerals such as quartz, in low-activity clays and in Fe and Al oxides. The chemical fertility of these soils is virtually null. They contain very little base cations and are usually poor in many micronutrients, in addition to having a strong capacity to fix P. On the other hand, except for low water storage capacity, they have good physical properties. Ferralsols are deep, permeable and have a stable structure that makes them more resistant to erosion than most other tropical soils (IUSS Working Group WRB, 2015).

Nitisols are a very important soil type in the humid tropics as they are one of the most productive soils of the area. More than half of the Nitisols are located in the highlands of tropical Africa. They are characterised by a

high Fe content, the presence of a clay-rich horizon mainly composed of low activity clay and a high P sorption. They are deep and well-drained, with a stable structure that makes them resistant to erosion. They are highly weathered, but also more fertile than other red tropical soils due to a higher nutrient content and a fair water holding capacity (IUSS Working Group WRB, 2015).

3.3 Sampling strategy

3.3.1 Data pre-processing

All the data analyses have been performed in R (version 3.4.3). The original dataset contained 634 plots (i.e. 317 farms since each farm contains two plots, the control and the N2Africa plots) in 2016 and 160 plots (i.e. 80 farms) in 2017. The soil types taken from SoilGrids (ISRIC, n.d.) were added to each farm. Not all plots had the same size, thus yields in each plot were calculated by using the grain weight and the plot area. Plots with missing yield data were removed, as well as farms without coordinates. Only farms that still contained two plots (the control and the N2Africa plot) were kept which resulted in 428 plots in 2016 and 158 plots in 2017. Other farms were removed because of problems in data recording. A linear mixed model was then performed to estimate the grain yield with the treatment as fixed effect and the villages nested within districts as random effects. Plots with studentized residuals higher than 4 in absolute value were removed from the dataset, that is 26 plots in 2016 and none in 2017. In the plots that were left, 4 farms had a recorded plot size of 9m². They were removed because this plot size, if not the result of a recording error, was deemed too small and not comparable to the rest of the dataset. Again, only farms that still contained both plot types were kept. Absolute (1) and relative (2) responses to the treatment were calculated and added as new variables to the dataset. In both years, a farm that had a large relative response (i.e. larger than 1000%) was removed as it seemed to be an unrealistic outcome (Franke et al., 2016). Further, only farms that had both plots either intercropped or monocropped were kept. This resulted in a dataset of 251 farms that were considered reliable and could be used for further analysis (175 in 2016 and 76 in 2017).

$$Response = yield_{N2A} - yield_{control} \quad (1)$$

$$Relative\ response = \frac{(yield_{N2A} - yield_{control})}{yield_{control}}\ x\ 100 \quad (2)$$

3.3.2 Sampling methodology

As it seemed not possible to sample 251 farms during fieldwork, it was decided to aim for a minimum of a hundred farms to sample. Thus, a sampling strategy had to be elaborated to select around 100 farms from the set of 251 farms.

First, the control yields and absolute responses to the treatments were classified in 3 categories: low, medium and high control yields or absolute response. This was done using the 33rd and 66th percentile as threshold values to separate the categories (i.e. values below the 33rd percentile were classified as low, values above the 66th percentile were classified as high, and values in between were classified as medium). Both years were considered separately, as the yields and responses in 2017 were significantly higher ($P < 0.001$) than in 2016. This was to avoid that all observations of 2016 would be classified as low, and those of 2017 as high.

The aim was to ensure that the sampled farms would have contrasting combinations of yields and responses. The observations were thus stratified according to combinations of categories of control yields and absolute responses (i.e. low yield-low response, low yield-high response, high yield-low response, high yield-high response, medium yield-medium response, and the rest of farms with all other combinations of yields and responses, that are called “standard” here), but also according to the year and the presence or absence of intercropping. A stratified sampling was then carried out to reach a sampling subset that would be representative of all strata, i.e. that would contain farms from both years, with intercropping or monocropping, and all types of categories of yields and response. It was decided to keep all the observations from the strata that contained less than 10 observations, and to randomly sample 5 observations from larger strata, which resulted in 97 selected

farms in which to sample soils (Fig. 4). A dataset of 154 spare farms was available to select from in case there were problems in the above described selection or there was time available to sample extra farms.

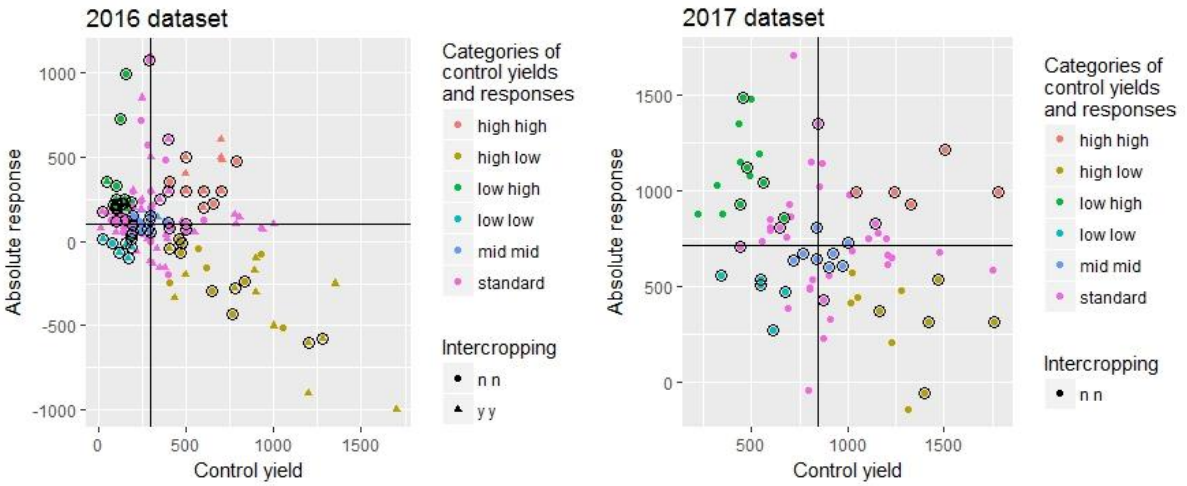


Figure 4. Scatterplot showing the control yield against the absolute response for the 175 and 76 pre-selected farms in 2016 and 2017 respectively. The colours indicate the combination of categories of control yields and absolute response, and the shape indicates the intercropping system. The 97 selected farms for sampling are shown in a black circle. The vertical line shows the median control yield, and the horizontal line shows the median response.

All soil samples had to be analysed with spectroscopy, and a subset of the samples had to be analysed as well through conventional soil analyses. This subset included the reference set and the validation set. The validation set is used to compare spectrally-predicted values of soil properties with values measured with conventional soil analyses. Wetterlind et al. (2013) recommends that this set includes all the soil types present in the population. Therefore, the farms located on non-dominant soil types were all selected for validation, and a random selection was made in the two dominant soil types (Ferralsols and Acrisols) to obtain 25 fields as the validation set. The reference samples should be selected based on their spectral profile, and thus the selection had to be made after the samples were analysed with spectroscopy.

3.4 Fieldwork

All the information collected during the fieldwork was recorded with the ODK Collect application for Android device (https://docs.opendatakit.org/). Topsoils, i.e. from 0 to 20 cm depth, were collected with an auger from the fields where the N2Africa plots were located. The sampling layout is shown in Fig. 5. This layout allowed to get composite soil samples that are representative of topsoils in a 100m² area (M. Walsh, personal communication, April 2018). Topsoils of the four sampling locations were mixed together and put in a bag pre-labelled with a QR code. The coordinates of the central point were recorded in the ODK Collect app and the QR code was scanned. The QR code allows for reliable geo-referencing of all samples. The position in the slope was recorded, as well as if signs of poor drainage such as gleyification were visible. In total, samples have been collected in 156 farms during the soil sampling campaign in 2018.

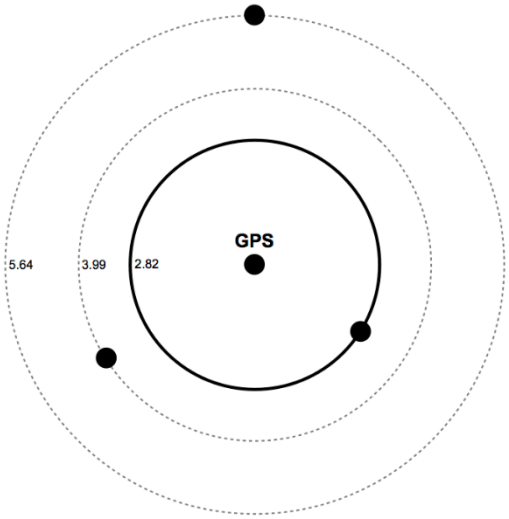


Figure 5. Plot sampling layout. A sample is taken in the center of the plot, and the three other samples are taken at a 2.82, 3.99 and 5.64 meters from the center (from M. Walsh, personal communication, April 2018).

In every farm, some questions were also asked to the farmers to complement or cross-check the information that could be found in the original dataset. It was asked for example which fertilisers were used on the two different plots during the N2Africa trials, and whether the plots were intercropped or not. Some extra questions were asked to the farmers who participated in the trial in 2017 because there was some missing information in the 2017 dataset compared to the 2016 dataset. The form that was filled during fieldwork via the ODK Collect app can be found in Appendix V.

3.5 Laboratory analyses

The technical working principle of infrared spectroscopy is that light is absorbed differently by soil components in each wavelength. The absorption is due to vibrations resulting from stretching and bending of molecules and to electronic transitions of atoms (Vågen et al., 2010). MIR spectral analyses reveal information about the organic and mineral fractions of the soils, as notably soil organic matter, clay minerals, carbonates, iron oxides and silica absorb in the MIR region (i.e. wavelength between 2.5 – 25µm or wavenumber between 400 – 4000cm^{-1}) (Soriano-Disla et al., 2014) and produce very distinctive peaks in the MIR spectra (Reeves, 2010; Viscarra Rossel et al., 2006). In summary, an IR spectrum gives a chemical and physical profile of the soil sample (Nocita et al., 2015).

As mentioned in the introduction, reference samples had to be selected to go through conventional laboratory soil analyses and be used in the calibration model (Fig. 6). Different sizes of reference sets can be found in literature. Gourlay et al. (2017) and the Africa Soil Information Services (AfSIS) (Vågen et al., 2010; Towett et al., 2015) used 10% of the sampled soils as reference samples and Tittonell et al. (2010) used a 20% subset. Of course, the larger the reference set, the more robust the calibration model will be (Wetterlind et al., 2013). On the advice of the Soil-Plant Spectral Diagnostics Laboratory of the International Centre for Research in Agroforestry (ICRAF)

in Nairobi (Kenya), it was decided to use 10% of the samples as reference. The selection of 15 reference samples was made by the ICRAF laboratory based on the MIR reflectance spectra. They used the Kennard & Stone procedure to select reference samples that would be representative of the spectral diversity of the whole population (E. Towett, personal communication, July 2018). This selection of reference samples included some samples that had been already selected for the validation set. Therefore, the validation samples were replaced when possible by other samples from the same soil type, that had not been previously selected either for validation or reference. This resulted in a set of 24 validation samples because one sample could not be replaced.

The MIR diffuse reflectance spectral analyses were carried out on 155 samples (one sample was not analysed) at the Salien Agriculture Research Institute (SARI) in Arusha (Tanzania) with a Bruker Alpha mid-infrared diffuse reflectance spectrometer. The conventional wet chemistry analyses were done in the Analytical Soil and Plant laboratory of the International Institute of Tropical Agriculture (IITA) in Dar es Salaam, Tanzania. The reference and validation samples (15 and 24 samples respectively) were analysed for pH (H_2O, 1:2.5 soil/water suspension), available P (Olsen if pH > 7 and Bray 1 if pH < 7), organic C (Walkley-Black), total N (Kjeldahl), particle size analysis (hydrometer method), and exchangeable K, Ca, Mg and Na (atomic absorption spectrophotometry for Ca and Mg and flame photometry for Na and K).

The 155 spectra and the results of the conventional wet chemistry analyses for the 15 reference samples were sent to the ICRAF laboratory who made the calibrations and the predictions for the 155 samples, following the method described in Sila et al. (2016). The values of soil attributes predicted from the reflectance spectra after calibration with the reference set are called “spectrally-predicted” values in the rest of this document, and the values obtained from the conventional wet chemistry analyses are called “wet-chemistry-measured” values. Out of these 155 samples, seven had very extreme and unlikely values for several soil properties (e.g. values of P ranging from 205 to 7338 mg kg^{-1}). These seven sample were analysed later with another machine by the SARI laboratory, which could be the cause of these unlikely results. It was decided to not use these 7 samples for the rest of the analyses. Out of the 7 removed samples, one was part of the validation set. The validation set was thus reduced again to 23 samples. A value for sand percentage was given by the calibration. However, the wet chemistry values of sand percentage were calculated by the IITA laboratory from the clay and silt percentages (i.e. 100% - (clay + silt)). It was therefore decided to do the same for sand values obtained from spectroscopy and calculate the sand percentage from the spectrally-predicted percentage of clay and silt.

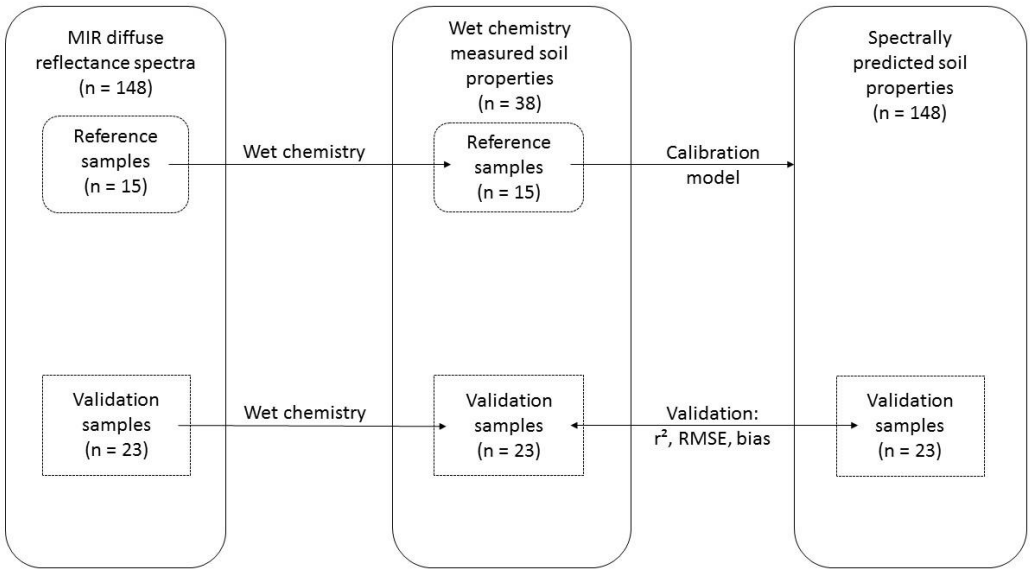


Figure 6. Summary of the different steps of the analyses of the soil properties. The MIR spectra and wet-chemistry-measured soil properties of the reference samples are used to produce a calibration model. This calibration model is then used to predict soil properties for all the samples based on their MIR spectra. The validation samples allow to assess the accuracy of the predictions resulting from the calibration model by comparing the wet-chemistry-measured and the spectrally-predicted soil properties.

3.6 Data analysis

The data analyses necessary to answer the research questions are presented in this section. Figure 7 presents a summary of the data analysis workflow.

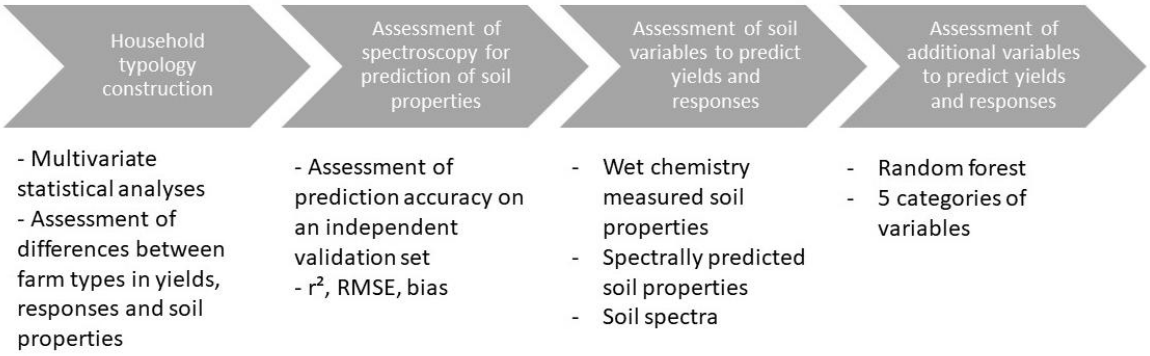


Figure 7. Workflow of the data analysis.

3.6.1 Data preparation

In the Lushoto district, the N2Africa field liaison officer communicated during the fieldwork that for the adaptation trials in the long rainy season in 2016, all fields were intercropped and DAP was applied. In 2017 and in the short rainy season of 2016, all fields were monocropped and NPK was applied. According to him, any different information in the original dataset was a mistake. Some of the farmers interviewed during the first days of fieldwork in Lushoto gave different answers, but the N2Africa field liaison officer instructed to not take their answers into account. Only the two possibilities previously mentioned (intercropped and DAP in 2016 and monocropped and NPK in 2017 and in short rains of 2016) were considered for the rest of the research.

For the Moshi district, the N2Africa field liaison officer advised to not use the results of the farmers' interviews concerning the inputs used during the trial and the intercropping or monocropping of the field, and to rather use the information that appears in the original dataset of 2016 and 2017. Indeed, it was difficult for the farmers to remember which practice they used in a particular field during a particular season and they sometimes didn't even know the names of the fertilisers.

3.6.2 Soil properties in each district

To see which nutrients were most limiting to bush bean yields in the study area, the median values and range for each nutrient in each district were examined. This allowed to see if the edaphic conditions differed between Moshi and Lushoto. This was done twice, once only with the 39 samples with wet chemistry values, and once with the spectrally-predicted soil properties of all the 148 samples. The percentage of farms below deficiency levels were calculated for soil properties with a defined critical level (see Table 1). This was also done once for the 39 wet chemistry samples and once with the spectrally-predicted soil properties for all the samples.

Statistical analysis

The difference in means between both districts was tested for each soil property using an unpaired t-test after validating the assumptions of the test. These are that the two sample groups should be normally distributed and their variances should be equal. When the sample size was lower than 30, the normality of distribution was assessed visually with a quantile-quantile plot and complemented by a Shapiro-Wilk test when necessary; when the sample size was higher than 30, the normality assumption was accepted based on the Central Limit Theorem. The equality of variance was tested with a F-test. If the assumptions could not be accepted, an unpaired Wilcoxon-test was done instead. If only the equality of variance could not be accepted, the Welch's t-test, or unequal variance t-test, was used.

3.6.3 Household characteristics

a. Definition of farm types using multivariate statistics for typology construction

Some household variables contained in the original dataset were selected to be used to build the household typology. Variables used in typology construction can be of two types: structural or functional variables. Structural variables describe wealth and resource endowment of the household, such as land size, household size and livestock ownership. Functional variables describe livelihood strategies and production orientation of the household (Alvarez et al., 2014; Chikowo et al., 2014; Kuivanen et al., 2016). The categorical variables were transformed to turn them into numeric variables, which was necessary for the rest of the analyses. For instance, the use of hired labour was expressed in the original dataset as frequencies (permanently, regularly, sometimes, never) and was turned into a time ratio (1, 0.66, 0.33, 0). The same way, the quantity of the farm production consumed by the household was transformed from categories (all consumed, most consumed, half consumed, most market) to ratios of the production consumed by the household (1, 0.75, 0.5, 0.25). The education level, expressed in the dataset as primary, secondary, post-secondary and university level was turned into total years of education (Education Policy and Data Center, 2018; Unesco Institute of Statistics, 2018). The livestock ownership was turned from numbers of cattle, goats, poultry and pigs to Tropical Livestock Units (TLU) using the conversion factors from Jahnke et al. (1988). Using TLU allows to convert a number of different animals into the equivalents of an animal of 250kg (Jahnke et al. 1988), making easier the comparison between different households with different animals. A description of all the household variables used in the analysis can be found in Table 3. Unfortunately, the land size could not be included in the typology construction as the original dataset contained information about land size but it was not reliable (i.e. two different columns in the same dataset were named farm size and contained different values). The land size was asked again during the 2018 fieldwork (see Appendix V) but it appeared at the end of the fieldwork that the information collected about land size in Moshi was not the correct information and it was thus preferred to not use this variable.

Table 3. Household variables used for the typology construction (n = 156).

Category	Variable	Unit	Mean (min - max)	Code
Structural	Livestock ownership	TLU	1.99 (0 – 13.35)	TLU
	Income index	Number of different income sources	2.28 (1 – 4)	Income_index
	Ratio of hired labour		0.37 (0 – 1)	Hired_ratio
	Household size	Number of members	5.91 (2 – 15)	No_hh_members
Functional	Highest level of education in the household	Years	9.55 (7 – 17)	Years_edu_hh
	Level of education of the household head	Years	7.89 (7 – 17)	Years_edu_hh_head
	Food self-sufficiency of the household	Number of months when the household experiences food shortages	1.85 (0 – 6)	No_months_food_shortage
	Ratio of the production destined to home consumption		0.67 (0.25 – 1)	Perc_home_consump

Statistical analysis

For the typology construction, the method explained in Alvarez et al. (2014) and used in Kuivanen et al. (2016) was followed. This method consisted of using a principal component analysis (PCA) followed by a cluster analysis to create the typology. The PCA results in a limited number of components explaining most of the variation in the household variables. The output of the PCA can then be used in a cluster analysis that separates the farms in groups with similar household characteristics.

Missing values had to be removed before performing the analysis. It is also advised to inspect the data to detect some extreme values and remove them because of their strong influence on the results of the PCA (Alvarez et al., 2014). In this case one farm with TLU much higher than the other farms was removed from the analysis and three farms had a missing value, so this resulted in 152 farms left for the typology construction. The variables were then transformed when possible to obtain a normal or close to normal distribution (Alvarez et al., 2014).

The next step of the typology construction was to perform a PCA on the data. Pearson (1901) was the first to mention principal component analysis, which was further developed in the 1930s by Hotelling (1933). PCA is a dimension reduction method that is useful in cases where the number of predictor variables is large, such as spectral data. It is an unsupervised method as it does not need any response variable, only a set of explanatory variables. The method constructs components that express the directions along which the predictor variables vary the most. The result of a PCA is a set of new variables, the principal components (PC), that explain most of the variability in the original data with a smaller number of variables. The first PC explains most of the variability in the original data, and each subsequent PC explains a smaller part of the variability. Each principal component has scores and loadings. The scores of a PC represent the distance at which every observation stands from the PC, and the loadings express the relative importance that each variable from the original dataset has on the PC (Gareth et al., 2013).

After standardisation of the variables (i.e. centring and scaling), the PCA explained about 70% of the variation in the data with the four first principal components. The inspection of the correlation matrix showed that the food self-sufficiency of the household was not strongly correlated with any of the four first principal components (i.e. no correlation coefficients > 0.5). It was decided to remove this variable from the analysis and perform another PCA. The first four principal components of this second PCA explained 75% of variance in the data and the eigenvalue of the 4^{th} component was 0.95, which is close to the limit value of 1 under which components should not be retained according to Kaiser's criterion (Kuivanen et al., 2016).

The scores of this second PCA were used to perform a clustering analysis of the farms based on their household characteristics. Following the two-step approach of Kuivanen et al. (2016), Ward's method was first used to perform a hierarchical clustering analysis that resulted in a dendrogram and allowed to choose a number of clusters *k*. This result of the hierarchical clustering method was then used in a non-hierarchical clustering analysis (Partitioning Around Medoids), which led to maximise the dissimilarity between clusters. The farm types created by this two-step clustering analysis were then defined by looking at the mean value and the range of each household variable for each group of farms.

b. Farm types and yields

The farm that had been removed before the PCA because of a high TLU was added to the farm category that seemed the most suitable. To get an insight on the influence of the farm types on yields, boxplots of the control yields and responses for each farm type were drawn and Tukey HSD (Honest Significant Difference) was used to check for statistically significant differences between farm types. The Tukey HSD test was used after an ANOVA test, if the latter had shown that there was a significant difference in means between the groups. The assumptions of the ANOVA test (i.e. homogeneity of variance and normality of residuals) were checked visually before accepting the results of the ANOVA test and going on with a Tukey HSD test. Additionally, the control yields and responses of each farm type were compared with the mean control yield and response of the whole population using a t-test when the sample was large enough (i.e. n >= 30). A non-parametric Wilcoxon test was used instead when the sample size was smaller and the combination of a quantile-quantile plot with a normality test (Shapiro-Wilk test) did not show that the sample distribution could be assumed to be normal.

Furthermore, location could be a confounding factor in this analysis, erroneously showing a link between farm types and yields. As a result, an additional analysis was performed to correct for location: a linear mixed model with villages nested within district (Moshi and Lushoto) as random effect and the farm types as fixed effect to predict control yields and responses. Linear mixed models were done with the R package *lmerTest* (Kuznetsova et al., 2017). An ANOVA of the linear mixed model indicated the significance of farm types when the effect of location was accounted for. The R package *predictmeans* (Luo et al., 2018) was used to draw a graph of the predicted means and, if the ANOVA showed a significant effect of farm types, to analyse the differences in predicted means between farm types.

c. Farm types and soil fertility

The same way the effect of the household typology was analysed on yields and response, its effect on soil properties was analysed. First, boxplots and Tukey HSD were used to assess the differences in each soil property between farm types. Next, the confounding effect of location was assessed by doing a linear mixed model for each soil property with location (i.e. villages nested within district) as random factor and farm types as fixed effect. An ANOVA was used to test if farm type still had a significant effect on soil properties after controlling for the effect of location. The spectrally-predicted soil properties were used for this analysis.

3.6.4 Validation of soil properties predicted by MIR diffuse reflectance spectroscopy

The validation set was used to assess the accuracy of MIR diffuse reflectance spectroscopy to predict soil properties. The 23 validation samples were analysed with both spectroscopy and conventional wet chemistry, without being used for the construction of the calibration model. Each validation sample thus had two values per soil property that could be compared using the root mean squared error (RMSE) and the bias (Wetterlind et al., 2013). A linear regression model was built as well for each property with the spectrally-predicted values as explanatory variables and the measured values as response variables. The adjusted r^2 of that model was considered to assess the accuracy of the spectral predictions.

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

$$Bias = \frac{\sum_{i=1}^{N}(\hat{y}_i - y_i)}{N}$$

where $\hat{y}_i$ and y_i are respectively the predicted value (i.e. with spectroscopy) and the measured value (i.e. with wet chemistry) of the sample i, and N the number of samples.

The SARI laboratory, which performed the spectroscopy, also provided an estimation of soil properties based on the MIR reflectance spectra. These values were therefore predictions of soil properties before calibration. To have an insight into the level of improvement in prediction brought by the calibration, these predicted values were also compared with the values measured with wet chemistry.

3.6.5 Yield and response analysis

a. Soil properties, yields and responses

Several analyses were made to explore the performance of soil properties in explaining and predicting yields and responses. Three types of variables represented the soil properties: the wet-chemistry-measured soil properties of 38 samples, the spectrally-predicted soil properties of 148 samples and the 148 spectra themselves. The same analyses were made with each type of variables, so that their performance could be compared.

The only examples in literature of an attempt of linking spectra with yield data came from Tittonell et al. (2008) and (Kihara, 2014), as mentioned in the introduction. Tittonell et al. (2008) used near-infrared diffuse reflectance spectroscopy to analyse soil samples from maize fields in Kenya. They performed a partial least squares regression (PLSR) with the first derivative of the soil spectra as independent variables and the maize yield as

response variable, which gave a model with a cross-validated r^2 of 0.37. Working with the first derivative of the spectra instead of the raw spectra is often recommended to reduce the noise present in the spectra (A. Sila, personal communication, August 2018). Kihara (2014) did similar analyses with near-infrared and MIR spectroscopy to predict maize yields and responses to NPK application in Tanzania and Malawi. They tried different statistical approaches to do so, such as using a combination of a PCA and a multiple linear regression, a combination of a PLSR with a multiple linear regression, and a random forest analysis. These two studies were taken as a starting point to explore further the relation between spectra and yield data.

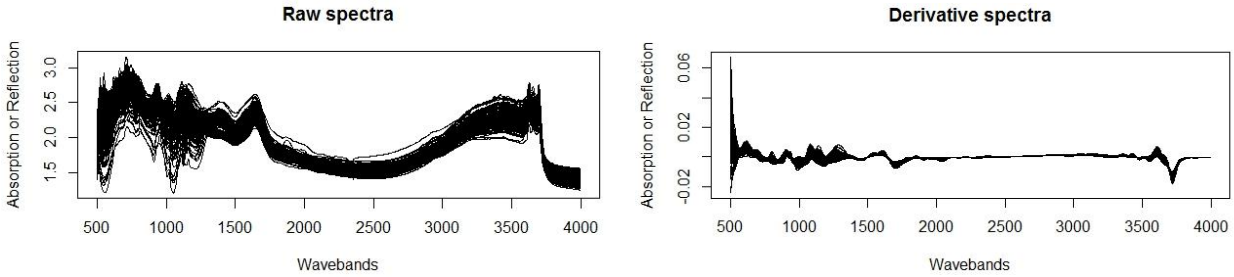


Figure 8. The raw spectra of the 148 soil samples and their first derivative on the right. The y-axis is the absorption.

Statistical analysis

First, the spectral data received from SARI had to be extracted and processed, which was done with the use of the R package *soil.spec* (Sila et al., 2014). The first derivatives of the spectra were extracted (Fig. 8) and a PCA was done on the first derivatives with the function *prcomp*. The first derivatives were centred for the PCA analysis but not scaled. The seven first PC explained 95% of variation in the derivative spectra and the scores of these 7 PC's were kept for later use.

An overview of the different analyses performed is given in Table 4.

Table 4. Statistical methods used to predict control yields and absolute responses to N2Africa treatments from the soil spectral data obtained with MIR diffuse reflectance spectroscopy.

Method	Response variable	Explanatory variables
Partial least squares regression	Transformed control yields / Transformed absolute response	First derivatives of the spectra
Principal component regression	Transformed control yields / Transformed absolute response	First derivatives of the spectra
Linear model	Transformed control yields / Transformed absolute response	Scores of the first 7 PC's
Random forest	Control yields / absolute response	First derivatives of the spectra
Random forest	Control yields / absolute response	Scores of the first 7 PC's

Both the partial least squares (PLSR) and principal component regression (PCR) were done using the *pls* package in R (Mevik et al., 2016). Partial least squares methods were developed by Herman Wold between the 1960's and the 1980's (Sanchez, 2013). PCR and PLSR create principal components the same way as a PCA does, and then use these components as new predictor variables in a linear model that fits the components to the response variable. The difference between a PCR and a PLSR is that the PCR only uses the predictor variables to create the PC's, while a PLSR uses both the predictor and the response variables to create the PC's (Gareth et al., 2013). When doing a PLSR and a PCR, an optimal number of components has to be chosen. With the *pls* package, an internal cross-validation can be performed during the PLSR and PCR analysis. This allows to show the out-of-bag prediction performance of the model with different number of components as root mean square error (RMSE). The model with lowest RMSE can then be chosen. As a reference, the RMSE of a model with no component is also shown (Mevik & Wehrens, 2007).

The control yields were square root transformed, and a Tukey's ladder of powers transformation was used for the absolute response to make their distribution fit the normal distribution as closely as possible, except for the random forest analyses. Random forest is a non-parametric analysis method and there are thus no assumptions made about variable distribution (Jeong et al., 2016). The *randomForestSRC* package (Ishwaran & Kogalur, 2018) was used for the random forest analyses with the default settings (ntree = 1000 trees, nodesize = 5, mtry = p/3 with p = the number of variables).

Further, linear mixed models with soil properties as fixed factors and district as a random factor were performed. This was done to analyse the link between soil properties and control yields and absolute responses. Some of the soil properties were log transformed to reduce the right-skewness of their distribution. Following the method of Ronner et al. (2016), each model was reduced with a backward variable selection method using the function *step*. Linear models in each district were also performed.

b. Random forest for yields and response predictions based on multiple categories of explanatory variables

Statistical analysis

Random forest (RF) is a predictive modelling tool that was developed by Breiman (2001). RF is an evolution of decision trees, that corrects for the high variance of decision trees. RF grows many decision trees, each one using only a random subset of the observations, i.e. a training set that is about 2/3 of all the observations. In the construction of each decision tree, only a subset of the predictor variables is selected randomly each time there is a split in the tree. This allows for the predictions made by the different trees constituting the random forest to be decorrelated. The outcome of the random forest is an average of the predictions of the multiple decision trees. The fact that these predictions are decorrelated give a low variance to the outcome of the RF. Random forest has an internal validation system, called the Out-of-Bag (OOB) error estimation. Every tree grown by the model was built with a random subset of the observations. The observations not used to build a tree are called the out-of-bag observations. For each observation, a single OOB prediction can be obtained by averaging the predictions of all the trees where that observation was not present. Hence, all the observations have a single OOB prediction resulting from this process, from which the OOB error of the whole model can be calculated. An issue with random forest is that even though it has a high prediction power, it is not easily interpretable. A way to interpret the model is to rank the variables by their importance in the construction of the model (Breiman, 2001; Gareth et al., 2013). The importance of a variable is determined by permuting all the values of the variable. The OOB prediction is then calculated again with this permuted variable. The occurrence of a large drop compared to the OOB prediction obtained with the non-permuted variable shows that the variable has an important role in the performance of the random forest model (Ishwaran & Kogalur, 2018).

Random forest has some advantages over more conventional analysis such as linear (mixed) regression models. As mentioned above, no assumptions are made in random forest about the distribution of the dependent and independent variables and hence RF is useful when working with data with highly skewed distributions. Random forest can be used for both regression and classification and accepts a mix of continuous and categorical variables as explanatory variables. It can also be used in large *p* small *n* situations, i.e. when a large number of explanatory variables are used with a small number of observations, and is robust to outliers (Grömping, 2009; Jeong et al., 2016; Tittonell et al., 2008; Towett et al., 2015). Jeong et al. (2016) compared the performance of random forest and multiple linear regression to predict crop yields from a variety of climate and biophysical variables and found that random forest prediction accuracy outperformed the linear regression models in all their analyses.

Random forest was used here (*randomForestSRC* package, same settings as previously mentioned) to disentangle the role of several categories of variables in the variability in control yields and absolute response. The different categories used are presented below (Table 5). The importance of each category for prediction was assessed by creating a full model with all the categories, then comparing the performance of the full model with reduced models containing all categories minus one. A drastic decrease in the model performance after removing a category would show that this category is an important contributor to the model performance. Models with individual categories of variables were also done and compared to the full model. The percent of variance explained and the r^2 and RMSE of the out-of-bag prediction were recorded to compare models and assess their performance. The ranking of variables for their importance in prediction was also recorded. It is common to use partial dependence plots to visualise how each variable impacts the prediction. Partial dependence plots show the marginal effects of a variable on the response variable, i.e. the effect of that variable after the average effects

of all the other variables of the model were accounted for (Elith et al., 2008). Partial dependence plots result in straightforward visualisations but are not always reliable, especially when variables are correlated, which is often the case in agronomic research. To have a visualisation of the effect of the most important variables of the best models on the control yields and absolute responses, linear mixed models with location (village nested within district) as random factor and the explanatory variable to visualise as fixed factor were used and a scatter plot was made. When the predictor variable was a categorical variable, an ANOVA was performed to assess the differences in predicted control yields or absolute responses between the levels of the categories. The model with the best performance was then further assessed by evaluating the capacity to predict across districts and years, i.e. by training the model in one district and testing it in the other, and the same with years. The RF analyses were done twice, with two different sets of observations: once with the subset of samples analysed with wet chemistry, and once with all the samples and their spectrally-predicted soil properties.

Categories of explanatory variables

The explanatory variables could be divided in 5 categories: soil, management, field information, remote sensing and household data (Table 5). The original dataset contained information about management practices during the trials, such as whether the field was sole cropped or intercropped, the planting date and the number of weeding. During the 2018 fieldwork, a variety of information about the cultivated plot was collected: a characterisation of the slope (i.e. flat, moderate, steep or very steep), the position of the field in the landscape (i.e. flat, valley bottom, lower slope, mid slope, upper slope or hilltop), the soil depth, the presence of coarse fragments in the field, the presence of signs of poor drainage and the occurrence of waterlogging and drought problems in the field (see Appendix V). These variables are called “field information”. The household data as shown in Table 4 and the typology were added as a category. Finally, remote sensing information relevant to crop growth was collected. This included Shuttle Radar Topography Mission (STRM) digital elevation data (Farr et al., 2007) from which were derived the slope and aspect. The soil types predicted in SoilGrids (ISRIC, n.d.) were taken, as well as the Enhanced Vegetation Index (EVI) (Didan, 2015). The EVI data was taken from the corresponding growing season in the previous year, i.e. from March to August of the previous year for the long growing season, and from October to January of the previous year for the short growing season. This was done because yields in a particular growing season were being predicted, thus variables that were available before that growing season were needed (Shmueli, 2010). Finally, climate data from WorldClim version 2 was used. WorldClim consists of high resolution average monthly climatic data between 1970 and 2000 (Fick & Hijmans, 2017). Mean temperature and precipitation at a resolution of 30 seconds (1km²) were used to derive the average temperature during the growing season, the temperature seasonality (i.e. coefficient of variation of the average temperature during the months of the growing season), the average and total precipitation during the growing season and precipitation seasonality. WorldClim was used instead of temperature and precipitation data from the studied growing season because the aim of the analysis was to make a prediction, and thus there was a need for climate information that is available before the season in which yields were being predicted.

This amounted to 40 explanatory variables in total for the control yield and 46 variables for the absolute response. The management variables for both the control and the N2Africa plots (e.g. inputs, planting date,…) were used in the analysis of the absolute responses since the intensity of the response depended on what happened in both plots. Additional variables used in the absolute response analysis were the improved bush bean variety and the yield of the control plot. The latter is not a variable that can be known in advance, thus should not be used in predictive modelling. It was still used because in another context, the yields obtained in a previous season could be used as “control yield” and provide the same type of information.

c. Model evaluation of predictive performance

The predictive power of all linear model analyses were checked with cross-validation, using 80% of the data as training set and 20% as validation set and repeating the analysis a hundred times, every time with different randomly picked validation and training sets. The cross-validation r^2 and root-mean square error (RMSE) were calculated as the average of the squared Pearson correlation and as the average RMSE between the predicted and measured values on the validation sets over the 100 random subsets. The training r^2 and RMSE values were calculated the same way on a random subset of the training set the same size as the validation set, i.e. 20% of the whole dataset, to allow for direct comparison with the cross-validated model.

The random forest models were evaluated by calculating the squared Pearson correlation and the RMSE between the out-of-bag predictions and the measured values of all observations. When the predictive value across years

and districts was tested, the number of observations in each year and each district was different, which did not allow for fair comparison between the model performance on the training set and on the test set. Hence, when the training set was larger than the test set (i.e. there were more observations in Lushoto than in Moshi, and in 2016 than in 2017), the cross-validation r^2 and RMSE were obtained from averaging the r^2 and RMSE between the predicted and measured values over a hundred random subsets. The predicted values were obtained from performing the model a hundred times on a different random subset each time, of the size of the smallest group of observations. The predictions from Moshi and 2017 were repeated a hundred times on a random subset of the observations in Lushoto and 2016 respectively. The random subset had the size of the set of observations that was used to make predictions in the new district or trial year. The training performance was calculated as in-bag prediction on the training set, and when necessary (i.e. in Lushoto and in 2016) it was the result of averaging a hundred repetitions of the model on a random subset with the size of the smallest group of observations.

Table 5. Description of the variables used in the random forest analysis.

Category	Variable	Description	Code	Source
Soil	Wet-chemistry-measured soil properties (n = 39)	pH, soil organic C, total N, Clay, Silt, Sand, Textural class, Ca, Mg, K, Na		Laboratory of IITA Dar Es Salaam
	Spectrally-predicted soil properties (n = 148)	pH, soil organic C, total N, Clay, Silt, Sand, Ca, Mg, K, Na		Predictions made by ICRAF Nairobi
Management	Fertiliser	None, DAP, NPK	input_n2a input_own	Data collected during trial (2016 or 2017)
	Intercropping	Yes/No (all farms have or two monocropped fields, or two intercropped fields)	intercrop_own	Id.
	Improved variety used on the N2Africa plot	Jesca, Uyole Njano, Lyamungo 90 *Only used for absolute response*	Pack_variety	Id.
	Relative planting date	Number of days the beans were planted after the 1st of March for the long rainy season, and the 1st of November for the short	rel_planting_date_n2a rel_planting_date_own	Id.
	Number of weeding	Number of weeding on the field during trial	nbr_weeding_n2a nbr_weeding_own	Id.
	Yield of the control plot	*Only used for absolute response*	yield_own	Id.
Field information	Slope	Flat, moderate, steep, very steep	slope	Data collected during 2018 fieldwork
	Slope position	Flat, valley bottom, lower slope, mid slope, upper slope, hilltop	slope_position	Id.
	Coarse fragments in the field	Yes/no	coarse_fragments	Id.
	Signs poor drainage in the field	Yes/no	signs_poor_drainage	Id.

	Soil depth class	<30 cm, 30-60 cm, 60-85 cm, >= 85 cm	soil_depth_class	Id.
	Waterlogging problems in the field	Yes/no	waterlogging	Id.
	Drought problems in the field	Yes/no	drought	Id.
Remote sensing information	Altitude	Meters 30 m resolution	altitude_STRM	(Farr et al., 2007)
	Slope	Degrees 30 m resolution	slope_STRM	Derived from Farr et al. (2007)
	Aspect	Classified in North, East, South, West aspects.	aspect_class	Id.
	Soil type	From SoilGrids.	soil_type	ISRIC World Soil Information (ISRIC, n.d.)
	Enhanced Vegetation Index (EVI)	250m resolution. Average EVI during the previous corresponding growing season.	avg_EVI	(Didan, 2015)
	Average temperature during the growing season	Derived from WorldClim data, i.e. monthly average temperature between 1970 and 2000. 1km² resolution	avgTemp	Derived from Fick & Hijmans (2017)
	Temperature seasonality	Coefficient of variation of temperature during the growing season.	CVTemp	Id.
	Average monthly precipitation during the growing season	Derived from WorldClim data, i.e. monthly average precipitation between 1970 and 2000. 1km² resolution	avgPrec	Id.
	Precipitation seasonality	Coefficient of variation of precipitation during the growing season.	CVPrec	Id.
	Total precipitation during the growing season	Sum of the monthly average precipitation between 1970 and 2000 in the months of the growing season	totPrec	Id.
Household variables				
	Variables in Table 3 + the household typology built as explained in section 3.6.3.			

4. RESULTS

4.1 Data exploration

4.1.1 Exploration of yields and responses

The N2Africa treatment had a strong and significant effect on bush bean yields ($p < 0.001$), with 86% of the studied farms having a yield after N2Africa treatment (called N2Africa yield) greater than the control yield and a third of the farms having more than doubled their yield. A relative increase of minimum 10% is necessary for a farmer to notice the effect of a treatment (Ronner et al., 2016) and this was reached by 81% of the farms. However, all farmers did not benefit in the same way from the N2Africa treatment. Indeed, 22 farmers or 14% of them obtained a negative response to the treatment, i.e. a lower yield in the N2Africa plot than in the control plot (Fig. 9). The correlation between the control yields and the absolute responses was low, with a r^2 of only 0.06.

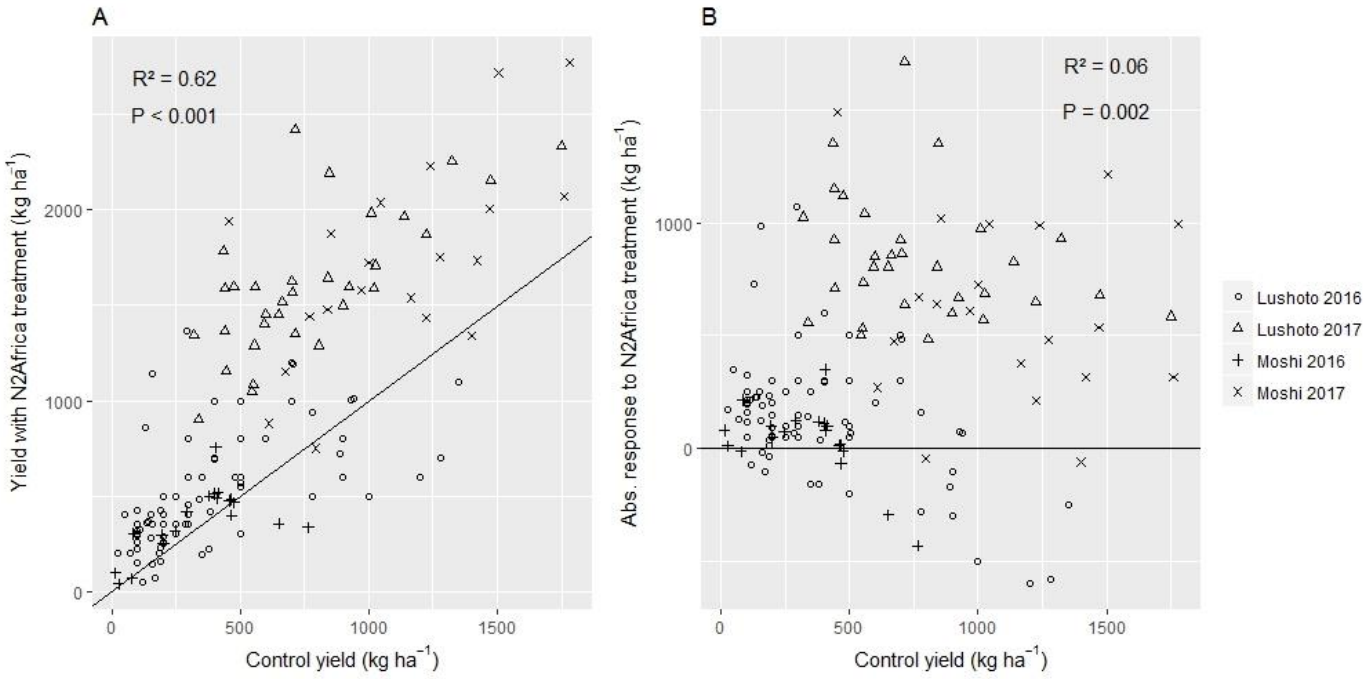


Figure 9. Control yields of bush beans in 156 farms in Northern Tanzania against A) the yield after the application of N2Africa technologies (i.e. improved varieties and fertiliser), shown with the 1:1 line and B) the absolute yield increase after the application of N2Africa technologies. The adjusted r^2 and p-value obtained from a linear model is indicated in each figure.

Control yields in Moshi were greater than in Lushoto ($P = 0.004$) with a mean control yield of 735.4 kg ha^{-1} and 481.4 kg ha^{-1} in Moshi and Lushoto respectively. The absolute response did not differ between both districts (Fig. 10). The control yields and absolute response in the two trial years were very different ($P < 0.001$), with higher control yields and responses in 2017 than in 2016 (Fig. 10). Daily precipitation data from the CHIRPS dataset (Funk et al., 2015) was used to check if this was due to a rainfall difference between both years, but no significant difference in total rainfall was found. Another possible reason is that all farms were monocropped in 2017, and most were intercropped in 2016 (see Table 6). This could also be a reason for lower yields in Lushoto, where intercropping was more widespread (see Table 6). To verify for the difference in control yields and responses between districts when the same intercropping practice was used, the yields and responses in monocropped fields of both districts were compared. It appeared that monocropped fields in Moshi still had a higher control yields than Lushoto ($P = 0.03$), but the absolute responses were higher in monocropped fields of Lushoto ($P = 0.002$).

Table 6. Proportion of use of different management practices in the different districts and trial years.

	Lushoto (n = 117)	**Moshi (n = 39)**	**2016 (n = 104)**	**2017 (n = 52)**
Trial year/Distict	73% 2016	49% 2016	82% Lushoto	62% Lushoto
	27% 2017	51% 2017	18% Moshi	38% Moshi
Intercropping	65% intercropped	18% intercropped	81% intercropped	0% intercropped
	35% monocropped	82% monocropped	19% monocropped	100% monocropped
Fertiliser in control plot	0% none	33% none	10.5% none	4% none
	65% DAP	16% DAP	79% DAP	0% DAP
	35% NPK	51% NPK	10.5% NPK	96% NPK
Fertiliser in N2Africa plot	65% DAP	49% DAP	91% DAP	0% DAP
	35% NPK	51% NPK	9% NPK	100% NPK
Improved variety	23% Jesca	33% Jesca	9% Jesca	60% Jesca
	17% Lyamungo 90	18% Lyamungo 90	6% Lyamungo 90	40% Lyamungo 90
	60% Uyole Njano	49% Uyole Njano	85% Uyole Njano	0% Uyole Njano

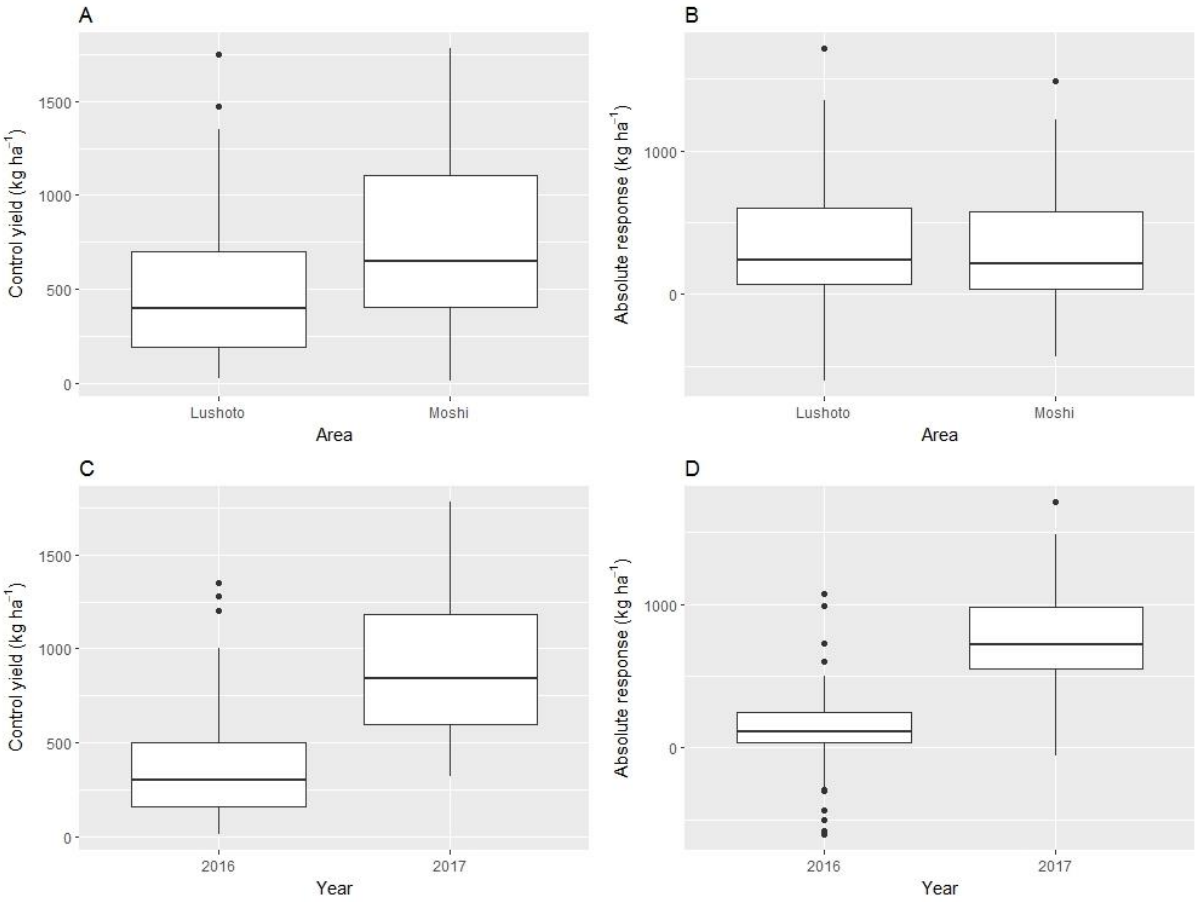


Figure 10. Boxplots of the control yields (A and C) and absolute responses (B and D) against the two study areas (Moshi, n = 39, and Lushoto, n = 117, graphs A and B) and the years of the trial (2016, n = 104, and 2017, n = 52, graphs C and D).

4.1.2 Exploration of soil properties per district

The analysis of soil properties per district revealed that Moshi and Lushoto have different soil characteristics (Table 7 and 8). Overall, Moshi seemed to have more fertile soils than Lushoto, with higher levels of pH, P, Clay, Silt, Mg and K. Soils in Lushoto were more sandy and had higher levels of organic carbon and N. The pH was slightly acidic to neutral in Lushoto, and neutral to slightly alkaline in Moshi. Lunze et al. (2012) stated that the optimum pH for common beans is between 5.8 and 6.5, and the ranges of pH encountered in the sampled fields did not seem to differ much from this range, thus pH should not be a major constraint to bush bean yields.

The percentage of farms that were below the deficiency levels indicated in Table 1 are shown in Table 7 and 8. It is noticeable that the percentage obtained with wet chemistry and spectrally-predicted soil properties were similar for all properties, except Mg and K. When looking at Moshi and Lushoto separately, it appeared that none of the fields in Moshi were deficient in P and that the main deficiency was N and OC (N deficient in 94 or 100% of the fields with wet-chemistry-measured and spectrally-predicted values respectively; OC deficient in 81 or 92% of the fields in the same order). K was not a main constraint in Moshi with only 6% of the fields in deficiency with wet-chemistry-measured values and none with spectrally-predicted values. Spectral predictions showed 20% of the fields of Moshi with Mg deficiency whereas wet chemistry showed none. In Lushoto, for most soil properties and whatever the measurement method, at least a third of the fields were below the deficiency value (P: 43%/31%, N: 48%/61%, Mg: 35%/59%, K: 35%/2%, OC: 56%/71% with wet-chemistry-measured followed by spectrally-predicted values). Only Ca did not seem to be a constraint anywhere.

Table 7. Median values for each soil attribute measured with conventional wet chemistry soil analyses, with the min and max values in brackets when separated by district (Moshi and Lushoto). The p-value of the difference between districts is indicated, as well as deficiency levels found in literature and the percentage of farms below this deficiency level (see Table 1).

Soil attribute	All samples (n = 39)	Moshi (n = 16)	Lushoto (n = 23)	Significance of the difference	Deficiency level	% farms below deficiency level
pH	6.69	7.24 (6.15 - 8.34)	6.5 (4.6 - 7.96)	0.001		
%OC	1.66	1.36 (0.37 - 2.83)	1.97 (0.78 - 5.94)	0.01	< 2 %	67%
Available P (mg kg^{-1})	18.36	39.6 (14.21 - 129.6)	10.8 (0.79 - 112.86)	< 0.001	< 7 mg kg^{-1}	26%
%Total N	0.17	0.13 (0.1 - 0.26)	0.2 (0.1 - 0.54)	< 0.001	< 0.2 %	67%
%Clay	34	36 (16 - 54)	32 (4 - 54)	0.03		
% Silt	17	23 (5 - 39)	13 (4 - 31)	0.002		
% Sand	45	33 (27 - 65)	52 (37 - 87)	< 0.001		
Exch. Ca ($cmol_c$ kg^{-1})	14.92	16.84 (9.87 - 48.59)	13.57 (4.04- 58.08)	0.03	< 5 $cmol_c$ kg^{-1}	2%
Exch. Mg ($cmol_c$ kg^{-1})	2.35	2.51 (2.27 - 6.41)	2.17 (0.85 - 2.58)	< 0.001	< 2 $cmol_c$ kg^{-1}	20%
Exch. K ($cmol_c$ kg^{-1})	0.47	1.62 (0.17 - 4.56)	0.25 (0.07 - 2.07)	< 0.001	< 0.2 $cmol_c$ kg^{-1}	23%
Exch. Na ($cmol_c$ kg^{-1})	0.15	0.15 (0.05 - 0.72)	0.12 (0.03 - 0.27)	n.s.		

n.s.: not significant.

Table 8. Median values for each soil attribute predicted with MIR diffuse reflectance spectroscopy, with the min and max values in brackets when separated by district (Moshi and Lushoto). The p-value of the difference between districts is indicated, as well as deficiency levels found in literature and the percentage of farms below this deficiency level (see Table 1).

Soil attribute	All samples (n = 148)	Moshi (n = 39)	Lushoto (n = 109)	Significance of the difference	Deficiency level	% farms below deficiency level
pH	6.44	6.77 (6.55 - 8.1)	6.23 (5.44 - 7.44)	<0.001		
%OC	1.5	1.39 (0.63 - 2.17)	1.55 (0.89 - 4.51)	<0.001	< 2 %	77%
Available P (mg kg^{-1})	12.01	29.86 (15.23 - 164.97)	9.25 (2.38 - 53.55)	<0.001	< 7 mg kg^{-1}	23%
%Total N	0.17	0.13 (0.09 - 0.19)	0.18 (0.1 - 0.47)	<0.001	< 0.2 %	72%
%Clay	32.99	36.78 (13.36 - 53.89)	30.42 (4.23 - 61.88)	0.003		
% Silt	15.53	19.35 (16.53 - 23.96)	14.74 (9.41 - 19.76)	<0.001		
% Sand	51.66	43.72 (28.8 – 65.66)	54.25 (23.53 – 86.36)	<0.001		
Exch. Ca ($cmol_c$ kg^{-1})	12.27	12.21 (8.88 - 21.04)	12.32 (6.44 - 43.24)	n.s.	< 5 $cmol_c$ kg^{-1}	0%
Exch. Mg ($cmol_c$ kg^{-1})	2.01	2.33 (1.72 - 3.19)	1.93 (1.35 - 2.89)	<0.001	< 2 $cmol_c$ kg^{-1}	49%
Exch. K ($cmol_c$ kg^{-1})	0.45	0.8 (0.54 - 1.77)	0.36 (0.16 - 1)	<0.001	< 0.2 $cmol_c$ kg^{-1}	1%
Exch. Na ($cmol_c$ kg^{-1})	0.14	0.13 (0.06 - 0.18)	0.14 (0.03 - 0.2)	n.s.		

n.s.: not significant.

4.2 Household characteristics

4.2.1 Definition of farm types using multivariate statistics for typology construction

The first principal component (PC) explained 29% of variation in the household data and was strongly correlated with the maximum years of education of both the household head and the whole household, as well as with the TLU (Fig. 11). The second PC explained 18.5% and was dominated by the ratio of the production destined for home consumption and more weakly correlated with the income index. It thus seemed to explain the household strategy in terms of production objectives, i.e. if the household was subsistence or market oriented. The third component was strongly related with both the household size and the hired labour ratio, these two variables having an opposite direction on the component. This indicated that small households tend to use more hired labour. The fourth component described the income index of the household.

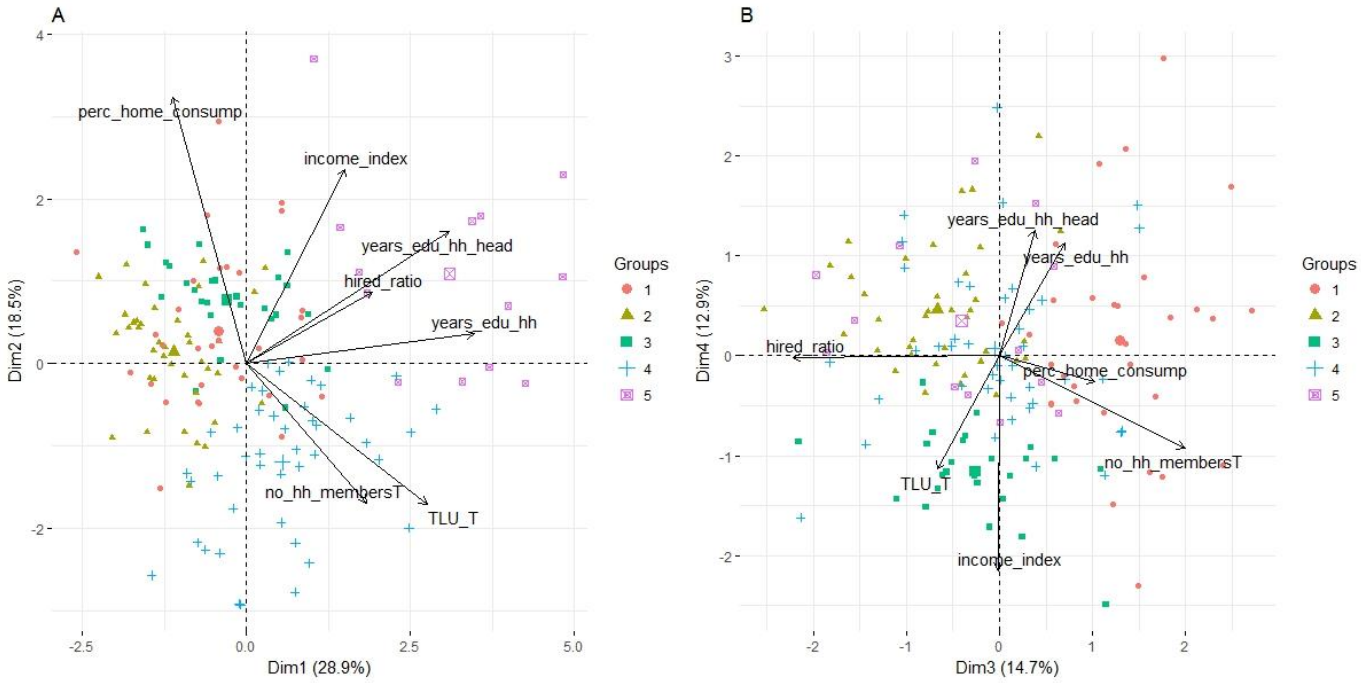


Figure 11. Biplot of the first two principal components (A) and of the third and fourth principal components (B) with the individual farms plotted by farm type. Refer to Table 3 (section 3.6.3) for the meanings of the names in the figure. The "T" at the end of a name means that the variable has been transformed to achieve a (close to) normal distribution.

The hierarchical clustering suggested to choose a cut-off height of 16.5 that partitioned the data in five different clusters. The farms have then been assigned to a final cluster using a non-hierarchical clustering method. Table 9 describes the mean and minimum and maximum values of each variable in each farm type.

Table 9. Mean values of each household variable per cluster, with min – max values in brackets (n = 152). The significance of the difference between farm types for each variable is given as p-values.

Variable	Farm type 1 (n = 31)	Farm type 2 (n = 39)	Farm type 3 (n = 26)	Farm type 4 (n = 43)	Farm type 5 (n = 13)	Significance (p-values)
Income_index	2.23 (1 – 3)	1.9 (1 – 2)	3 (3 – 3)	2 (1 – 3)	2.92 (2 – 4)	< 0.001
no_hh_members	7.06 (4 – 13)	4.26 (2 -7)	5.27 (2 – 12)	6.98 (3 – 15)	6.15 (3 – 10)	< 0.001
TLU	1.15 (0 – 2.79)	1.13 (0 – 2.9)	1.81 (0.15 – 6.1)	2.75 (0.42 – 5.73)	3.57 (0.15 – 7.7)	< 0.001
no_months_food_shortage	2.03 (0 – 6)	2.33 (0 – 5)	2.46 (0 – 6)	1.3 (0 – 5)	0.92 (0 – 4)	< 0.001
years_edu_hh	9.84 (7 – 11)	7.67 (7 – 13)	8.31 (7 – 13)	10.58 (7 – 17)	13.77 (11 – 17)	< 0.001
years_edu_hh_head	8.16 (7 – 11)	7.2 (7 – 11)	7 (7 – 7)	7.46 (7 – 11)	12.38 (11 – 17)	< 0.001
perc_home_consump	0.82 (0.5 – 1)	0.7 (0.25 – 1)	0.75 (0.5 – 1)	0.51 (0.25 – 0.75)	0.63 (0.25 – 1)	< 0.001
hired_ratio	0.12 (0 – 0.33)	0.44 (0 – 1)	0.39 (0 – 0.66)	0.36 (0 – 1)	0.71 (0.33 – 1)	< 0.001

Following is a definition of each farm type based on the information in Table 9:

Farm type 1: low resource-endowed category with large number of household members, little livestock ownership and the most subsistence oriented households.
The first farm type was defined by the largest household size, the smallest use of hired labour and the highest rate of the farm production destined to home consumption. It also had the second lowest TLU and a medium education level.

Farm type 2: subsistence oriented, low resource-endowed category with small household size, little livestock ownership and little income diversification.
The second farm type had the lowest income index, the smallest household size, the smallest TLU, the second lowest food self-sufficiency with on average 2.33 months of food shortage and a low education level. An average of 70% of the farm production was consumed by the household. Farm type 2 had the highest ratio of hired labour among the low resource endowed farm types (i.e. farm types 1, 2 and 3).

Farm type 3: subsistence oriented, low resource-endowed household category with high income diversification but low food self-sufficiency.
The third farm type exclusively contained households with three different income sources and thus had the highest average income index. It had the lowest food self-sufficiency with an average of 2.46 months experiencing food shortage, a low education level and the second highest rate of home consumption of the farm production.

Farm type 4: high resource-endowed category with high livestock ownership, large household size, little problems of food shortage and the most market-oriented households.
The fourth farm type was characterised by the lowest rate of food production destined to home consumption, a medium to high education level and the second highest food self-sufficiency with on average 1.3 months of food shortage. It also had the second largest household size and the second highest TLU.

Farm type 5: market oriented, high resource-endowed category with the largest livestock ownership, the highest education levels and the highest food self-sufficiency.

The fifth farm type was defined by the highest TLU, the highest education level, the largest use of hired labour and the highest food self-sufficiency with an average of 0.92 months with food shortage. It also had the second lowest rate of production going to home consumption and the second highest income diversification.

The fourth farm type was the most present in Moshi (41% of households) followed by the fifth farm type (20% of households). The first farm type was the less represented in Moshi with 10% of the farms belonging to farm type 1. The situation was different in Lushoto, with 30% of farms belonging to the second farm type, followed by farm types 1 and 4 that comprised an equal number of farms, each making up 24% of the farms in Lushoto. The farm type the less present in Lushoto was the fifth farm type, with only 5% of the farms belonging to that category.

4.2.2 Farm types and yields

The farm removed before the PCA was added to the fifth category of farms because it had a high TLU, a hired ratio of 1 and a total food self-sufficiency.

When using the observed values of yields and responses, the farm types differed in terms of control yields, as the first farm type had significantly lower yields than farm types 2, 4 and 5 (P = 0.05, 0.007 and 0.03 respectively). There was no significant difference in response between the different farm types (Fig. 12). Farm type 1 had a significantly lower mean control yield (P < 0.001) than the population mean (i.e. 550.4 kg ha^{-1}), as well as the third farm type (P = 0.01). The other farm types did not significantly differ from the population mean (Fig. 12). The second farm type had a mean absolute response significantly lower (P = 0.003) than the mean absolute response of the population (i.e. 332 kg ha^{-1}).

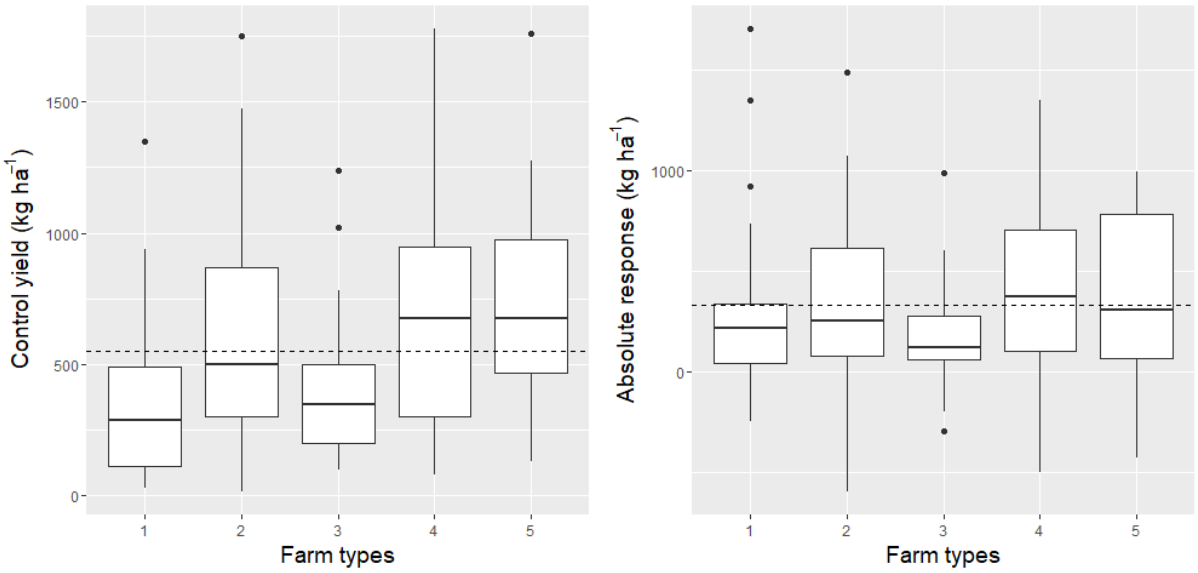


Figure 12. Boxplots of the five farm types against the observed control yields and the observed absolute responses to the N2Africa treatments. The dashed line represents the mean of the population (n = 153; mean control yield = 550.4 kg ha^{-1}, mean response = 332 kg ha^{-1}).

When the effect of location was accounted for (in a linear mixed model with villages nested in district as random factor), it could be seen (Fig. 13) that the farm types still had a significant effect on the control yields (P = 0.006). Indeed, farm type 1 had a significantly lower mean control yield than farm types 2 (P = 0.002), 4 (P = 0.01) and 5 (P = 0.01). Farm types 2 and 5 had a significantly higher mean control yield than farm type 3 (P = 0.02 and 0.03 respectively). There was no effect of farm types on the predicted mean absolute response.

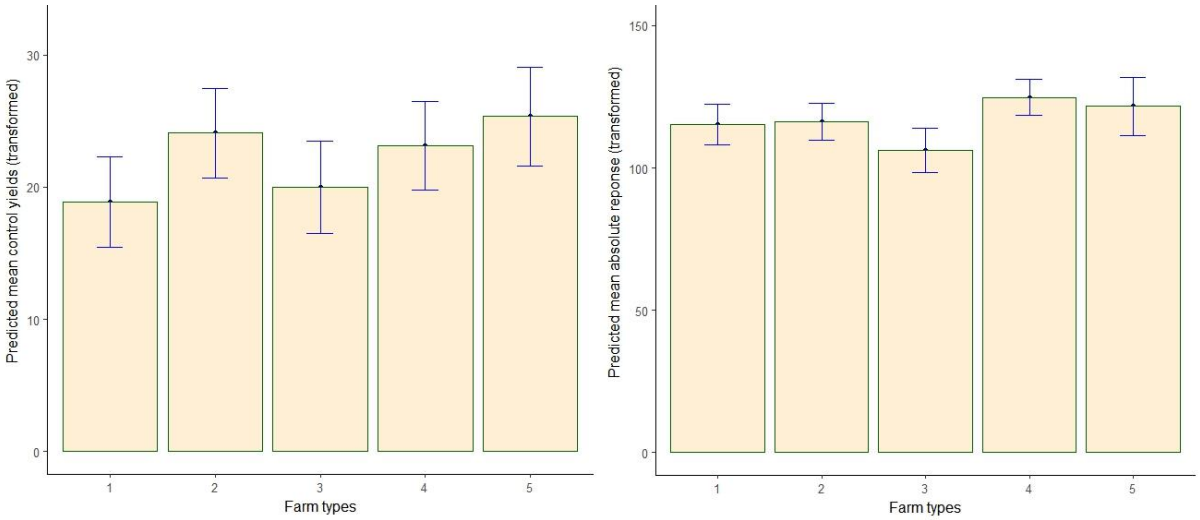


Figure 13. Predicted means of the control yields (left graph) and absolute response (right graph) for each farm type. The prediction is from a linear mixed model with the farm types as fixed factor and villages nested within district (Moshi and Lushoto) as random factor. The bar shows the standard error of the predicted means.

The analysis showed that the household typology built with the information available in this study could give some useful indication about farm productivity, with farm type 1 consistently showing the lowest yields and farm types 4 and 5 having higher yields. However, the fact that the second farm type had one of the highest predicted control yields was not expected based on the household information, as it was part of the low resource-endowed household categories. This showed that the typology was probably too rough to reveal more nuanced implications of the household characteristics on yields. The absolute response did not vary with farm types, indicating that all types of farms, irrespective of their intrinsic household characteristics, could benefit the same way from the N2Africa treatments.

4.2.3 Farm types and soil fertility

Table 10. Overview of the differences in control yields, response and soil fertility between farm types.

Variables	Sign. * (p-value)	Sign ** (p-value)
Control yield	0.002	0.006
Response	n.s.	n.s.
pH	< 0.001	n.s.
OC	n.s.	n.s.
P	< 0.001	n.s.
N	n.s.	n.s.
Clay	n.s.	n.s.
Silt	< 0.001	n.s.
Sand	n.s.	n.s.
Ca	0.03	n.s.
Mg	0.006	n.s.
K	< 0.001	n.s.
Na	0.01	n.s.

*Significance of difference between farm types when not compensated for location; **Significance of difference between farm types when it is compensated for location
n.s.: not significant

When location was not accounted for, the farm types differed in terms of soil fertility. Indeed, the statistical tests indicated significant differences in pH, P, Silt, Mg, K and Na between farm types (see Appendix II for boxplots of all soil properties per farm type). The high resource-endowed categories 4 and 5 had higher pH than the other lower resource-endowed farm types, with farm type 5 having significantly higher pH than farm types 1, 2 and 3 ($P < 0.001$, $P = 0.02$ and $P = 0.02$ respectively) and farm type 1 being significantly different with farm type 4 ($P = 0.01$). This could also be seen with P and silt levels, where farm type 5 had significantly higher levels of P and silt than farm type 1 ($P < 0.001$ for both), farm type 2 ($P = 0.002$ and $P < 0.001$ for P and silt respectively) and farm type 3 ($P = 0.006$ and $P = 0.01$ for P and silt respectively). The fourth farm type also had significantly higher levels of P and silt than farm type 1 ($P = 0.01$ and $P = 0.03$ for P and silt respectively). Ca was higher in farm type 5 compared to farm type 1 ($P = 0.03$). Also for Mg and K, farm type 5 had higher levels of Mg and K than the first farm type ($P = 0.003$ and $P < 0.001$ for Mg and K respectively) and higher levels of K than the second farm type ($P = 0.03$). Farm type 4 had larger values of K than farm type 1 ($P = 0.004$). Finally, the farms of the fifth category had the lowest levels of Na of all the farms ($P = 0.02$ with farm type 1, $P = 0.009$ with farm type 2, $P = 0.06$ with farm type 3 and $P = 0.005$ with farm type 4). None of these differences in soil properties were still significant when the effect of location was accounted for. A summary of the results of the ANOVA analyses is shown in Table 10.

4.3 Validation of soil properties predicted by MIR diffuse reflectance spectroscopy

The comparison of the predictions before and after calibration showed that the calibration improved the prediction accuracy. The bias and RMSE were all smaller after than before calibration, and the r^2 were all greater after calibration. Nevertheless, the performance of MIR diffuse reflectance spectroscopy to predict soil properties was still low, i.e. lower than found in literature (see Table 2 for examples). In literature, mainly silt, K, P and Na have been found to be poorly estimated with MIR spectroscopy. The other soil attributes usually all have r^2 above 0.7 (Table 2). Here, the highest r^2 obtained was 0.6 (Table 11), showing that the overall accuracy of the predictions was much lower than what can be found in literature. Total N was the attribute with the highest r^2, which was in accordance with literature as it usually ranks among the best predicted soil attributes. Similarly, P was very poorly predicted (r^2 = 0.06), which was also in accordance with literature. Nonetheless, some results were highly contrasting with expectations, such as the results for K and OC. Indeed, K is usually poorly predicted but had here the second best prediction with a r^2 of 0.55, while OC usually has a r^2 higher than 0.9 but obtained here a r^2 of only 0.35. The bias were all negative (except for sand), showing that spectral predictions usually tend to underestimate the wet-chemistry-measured values of the soil properties. The r^2 of the reference samples was also obtained from a linear model regressing their wet-chemistry-measured soil properties on their spectrally-predicted soil properties, to compare with the validation r^2. The reference r^2 were all higher, indicating a large overfitting of the model.

Scatter plots of the predicted against the measured value of the attributes with a post-calibration $r^2 > 0.3$ were made (Fig. 14). The plots showed that for pH and K, the general regression was dominated by the samples from Moshi while the samples from Lushoto were following a distinct trend. Predictions of K and pH were of lower accuracy in Lushoto (Fig. 14). For OC and N, samples from both Moshi and Lushoto followed the same general regression line. However, the separate models also had a lower performance than the general model, with no significant r^2 for organic carbon and for total N in Moshi. The model for total N in Lushoto was seemingly dominating the performance of the general model.

Table 11. Evaluation of the accuracy of MIR diffuse reflectance spectroscopy to predict soil properties values compared with conventional wet chemistry soil analyses on 23 samples. The r^2 and p-values come from the linear regression of the wet-chemistry-measured soil properties on the spectrally-predicted soil properties. Between brackets is shown the average r^2 obtained from performing the linear regression a hundred times with different randomly picked subsets of the same size as the reference set (n = 15), to allow for direct comparison with the r^2 values of the reference samples.

Soil attribute	**Spectral prediction on the validation set before calibration**			**Spectral prediction on the validation set after calibration**				**r^2 on the reference samples**
	RMSE	**Bias**	**r^2**	**RMSE**	**Bias**	**r^2**		**r^2**
pH	1.25	-0.718	0.16**	0.7	-0.13	0.4****	(0.39)	0.73****
OC	1.27	-0.43	0.12 *	0.87	-0.07	0.35***	(0.35)	0.97****
Extr. P	54.63	-32.17	-0.02	47	-7.28	0.06	(0.09)	0.2*
Total N	0.16	-0.13	0.23**	0.05	-0.01	0.6****	(0.5)	0.99****
Exch. Ca	20.01	-17.2	0.11*	9.84	-3.753	0.14**	(0.16)	0.91***
Exch. Mg	2.91	-2.65	0.19**	1.09	-0.41	0.26***	(0.24)	0.57****
Exch. K	1.63	-1.13	0.26**	1.01	-0.41	0.55****	(0.54)	0.34**
Exch. Na	0.25	0.05	-0.015	0.14	-0.05	-0.05	(0.02)	0.38***
Clay	n.a.	n.a.	n.a.	12.34	-2.22	0.2*	(0.11)	0.96****

Silt	n.a.	n.a.	n.a.	10.84	-3.51	0.3***	(0.29)	0.48***
Sand	n.a.	n.a.	n.a.	14.82	5.72	0.04	(0.07)	0.89****

Significant level: p-value: * < 0.1; **<0.05; *** < 0.01; **** < 0.001; n.a. = not available.

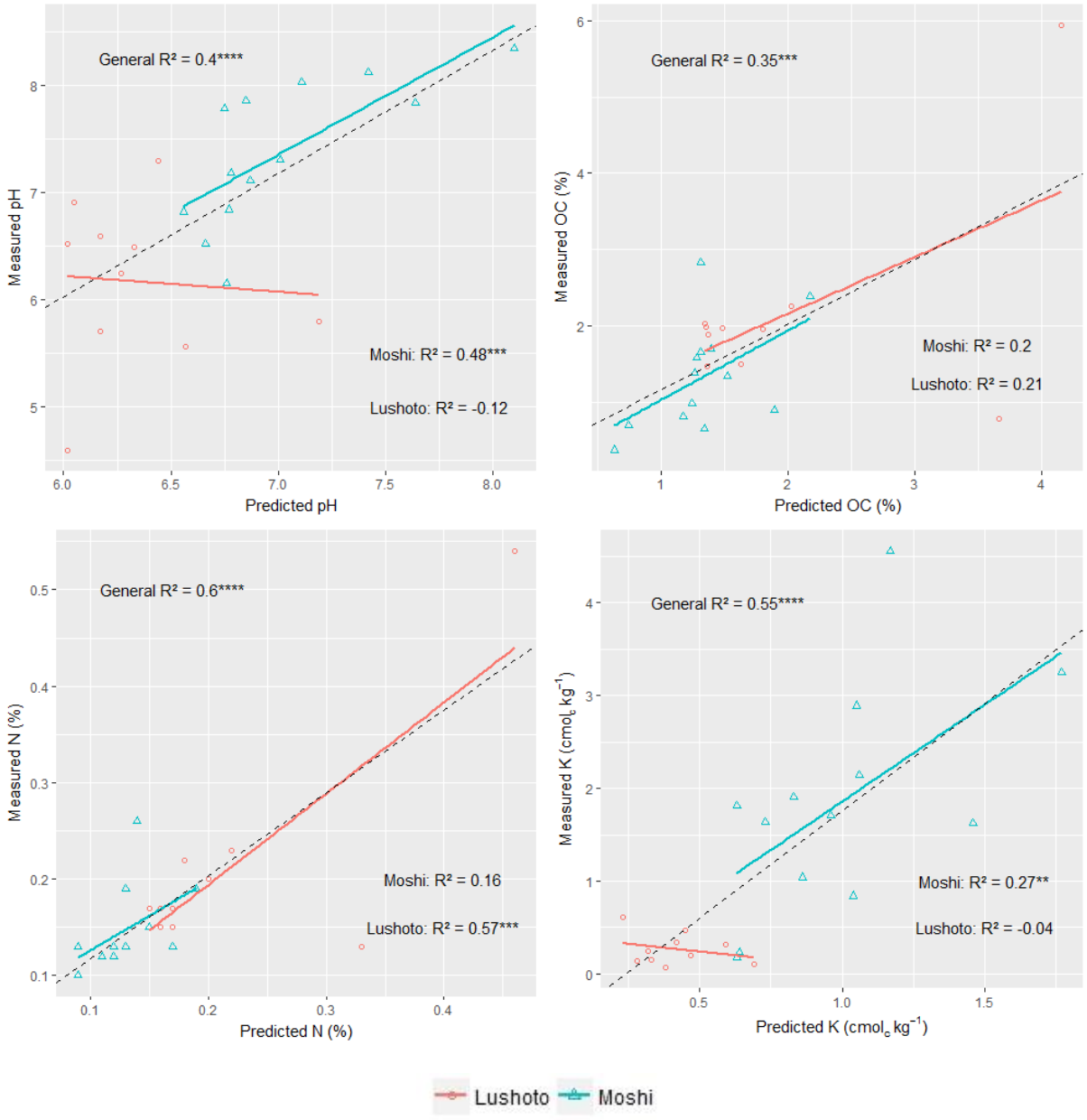


Figure 14. Scatterplot of the spectrally-predicted values of pH, OC, N and K against the values measured with conventional laboratory analyses. The dashed line shows the linear regression fit of the 23 validation samples. The two coloured solid lines show the linear regression fit for samples from Lushoto (n = 10) in red and for samples from Moshi (n = 13) in blue. The r^{2}'s come from the linear regression of the wet-chemistry-measured soil properties on the spectrally-predicted soil properties. Significant levels: p-value: * < 0.1; **<0.05; *** < 0.01; **** < 0.001.

4.4 Yield and response analysis

4.4.1 Soil properties, yields and responses

To explore the predictive value of the soil MIR diffuse reflectance spectra, several statistical methods have been used, as shown in Table 4 (section 3.6.5). In the partial least square regression (PLSR) and principal component regression (PCR) analyses, the reference model with no component always had lower root mean square error (RMSE) than models with principal components, thus indicating that a model without component was performing better than models with. This showed that PLSR and PCR were not appropriate analyses. The results for PLSR and PCR are therefore not shown.

More variation in control yields was explained by the soil properties than in absolute responses. The absolute response was explained poorly by all the types of soil information and all models. Nonetheless, the wet chemistry resulted in higher cross-validated r^2 for both control yields and absolute responses (Table 12 and Table 13). The spectrally-predicted soil properties seemed to be a less powerful predictor, with cross-validated r^2's not higher than 0.03. Similarly, none of the methods used to predict the yield and response from the soil spectra resulted in r^2 superior to 0.03, far below the cross-validated r^2 obtained by Tittonell et al. (2008) and Kihara (2014) from soil spectral data. However, there was a large difference in size between the samples with wet-chemistry-measured soil properties and the samples with spectral data (i.e. 38 and 148 respectively). The observed difference in performance between the different types of soil information to predict yield data could be due to the effect of sample size. Indeed, a smaller sample size might lead to an overoptimistic model (Kelley & Maxwell, 2003). To verify the effect of sample size, the analyses with the spectrally-predicted soil properties and with the spectra were done a second time, with only the 38 samples that also had wet chemistry values (Table 12 and Table 13). It appeared that with the same sample size, the wet-chemistry-measured soil properties still had a higher r^2 than the spectra and spectrally-predicted soil properties, irrespective of the statistical method used. The prediction r^2's obtained for the absolute response were all low.

It should be noted that the performance of the linear models and the random forests should not be compared, as the cross-validation in both models was done a different way and does not reflect a difference in prediction performance between the methods. The fact that variables were transformed for the linear model and not for the random forest could also have been a reason for the difference in r^2 values between both models. Hence, a random forest with the transformed values was also performed to compare the r^2 obtained in both situations. This resulted in similar r^2 to the ones indicated for random forest models in Table 12 and Table 13 (data not presented).

The linear mixed model using district as a random factor and wet-chemistry-measured soil properties of 38 samples and reduced with the help of a backward variable selection method explained 30% of variability in control yields (see Appendix III Table 8.2). The random factor was not retained by the variable selection method. Only N and clay were retained by the model. N had a significant and negative relationship with yields whereas clay had a positive relationship with yields. The negative relationship between N and control yields could be caused by lower nodulation and atmospheric N_2 fixation rates when soil N is higher, as this could result in lower yields (Giller, 2001). As for the absolute response, 10% of the variability was explained by wet-chemistry-measured soil properties, and again the random factor (district) was not retained (see Appendix III Table 8.3). Only pH and K were retained by the variable selection model. K had a negative relationship while pH had a positive relationship with the absolute response. This seemed to show that better absolute responses were obtained on fields with lower acidity. The linear mixed models with the spectrally-predicted soil properties either explained a very low part of the variability in the response, or no variable at all was retained by the variable selection method, even when only 38 samples were used (see Appendix III Table 8.5 and 8.6).

Two conclusions could be drawn from these analyses. First, soil properties predicted a larger part of the variability in control yields than in absolute responses, which was only predicted very poorly by soil properties. Second, wet-chemistry-measured soil properties seemed more relevant to control yields and absolute responses than soil information obtained from spectroscopy.

Table 12. Statistical methods used to predict control yields in N2Africa adaptation trials from MIR diffuse reflectance spectra (n = 148), soil properties predicted from the spectra (n = 148), and soil properties measured with wet chemistry (n = 38). The performance of the cross-validated model is shown. When a linear model is used, the r^2 and RMSE values for the training model (on a random subset of the same size as the test dataset, i.e. 20% of the whole dataset) are shown between brackets for direct comparison with the cross-validated model. When random forest is used, the in-bag predictions are shown between brackets for direct comparison with the out-of-bag predictions. The linear model and random forest cross validation results should not be compared with each other. The models done with the wet-chemistry-measured soil properties are highlighted in bold.

Model	n	Response variable	Explanatory variable	Cross-validation	
Linear model	148	Transformed control yield	7 PC's of the first derivatives of the spectra	r^2 = 0.03 RMSE = 8.95	(0.08) (8.4)
	38	Transformed control yield	7 PC's of the first derivatives of the spectra	r^2 = 0.16 RMSE = 9.5	(0.4) (6.9)
	38	**Transformed control yield**	**(Transformed) wet-chemistry-measured soil properties**	**r^2 = 0.28 RMSE = 9.36**	**(0.61) (5.61)**
	148	Transformed control yield	(Transformed) spectrally-predicted soil properties	r^2 = 0.03 RMSE = 9.1	(0.11) (8.28)
	38	Transformed control yield	(Transformed) spectrally-predicted soil properties	r^2 = 0.14 RMSE = 10.34	(0.46) (6.61)
Random forest	148	Control yield	7 PC's of the first derivatives of the spectra	r^2 = 0.0007 RMSE = 421.55	(0.9) (227.84)
	38	Control yield	7 PC's of the first derivatives of the spectra	r^2 = 0.008 RMSE = 461.72	(0.91) (244.71)
	148	Control yield	First derivatives of the spectra	r^2 = 0.02 RMSE = 412.34	(0.93) (183.27)
	38	Control yield	First derivatives of the spectra	r^2 = 0.01 RMSE = 463.01	(0.83) (21.6)
	38	**Control yield**	**Wet-chemistry-measured soil properties**	**r^2 = 0.12 RMSE = 412.28**	**(0.83) (228.13)**
	148	Control yield	Spectrally-predicted soil properties	r^2 = 0.02 RMSE = 416.57	(0.87) (221.49)
	38	Control yield	Spectrally-predicted soil properties	r^2 = 0.05 RMSE = 434.72	(0.84) (234.46)

PC: Principal component. RMSE: root mean square error.

Table 13. Statistical methods used to predict absolute response to N2Africa treatments from MIR diffuse reflectance spectra (n = 148), soil properties predicted from the spectra (n = 148), and soil properties measured with wet chemistry (n = 38). The performance of the cross-validated model is shown. When a linear model is used, the r^2 and RMSE values for the training model (on a random subset of the same size as the test dataset, i.e. 20% of the whole dataset) are shown between brackets for direct comparison with the cross-validated model. When random forest is used, the in-bag predictions are shown between brackets for direct comparison with the out-of-bag predictions. The linear model and random forest cross validation results should not be compared with each other. The models done with the wet-chemistry-measured soil properties are highlighted in bold.

Model	n	Response variable	Explanatory variable	Cross-validation	
Linear model	148	Transformed absolute response	7 PC's of the first derivatives of the spectra	r^2 = 0.03 RMSE = 39.15	(0.04) (36.71)
	38	Transformed absolute response	7 PC's of the first derivatives of the spectra	r^2 = 0.12 RMSE = 48.13	(0.2) (31.54)
	38	**Transformed absolute response**	**(Transformed) wet-chemistry-measured soil properties**	**r^2 = 0.13 RMSE = 44.78**	**(0.5) (25.15)**
	148	Transformed absolute response	(Transformed) spectrally-predicted soil properties	r^2 = 0.02 RMSE = 39.93	(0.1) (35.71)
	38	Transformed absolute response	(Transformed) spectrally-predicted soil properties	r^2 = 0.08 RMSE = 49.57	(0.32) (28.96)
Random forest	148	Absolute response	7 PC's of the first derivatives of the spectra	r^2 = 0.0001 RMSE = 427.19	(0.91) (230.84)
	38	Absolute response	7 PC's of the first derivatives of the spectra	r^2 = 0.05 RMSE = 445.33	(0.9) (241.09)
	148	Absolute response	First derivatives of the spectra	r^2 = 0.01 RMSE = 419.44	(0.95) (183.16)
	38	Absolute response	First derivatives of the spectra	r^2 = 0.02 RMSE = 445.41	(0.92) (196.98)
	38	**Absolute response**	**Wet-chemistry-measured soil properties**	**r^2 = 0.01 RMSE = 416.18**	**(0.86) (228.81)**
	148	Absolute response	Spectrally-predicted soil properties	r^2 = 0.003 RMSE = 440.89	(0.88) (237.55)
	38	Absolute response	Spectrally-predicted soil properties	r^2 = 0.004 RMSE = 441.96	(0.81) (247.4)

PC: Principal component. RMSE: root mean square error.

4.4.2 Random forests for yields and response predictions based on multiple categories of explanatory variables

In this section, five different categories of variables were used to test their predictive capacity for bush bean yields and responses to N2Africa treatments. This was done once using 37 samples with wet-chemistry-measured soil properties, and once using 145 samples with spectrally-predicted soil properties. Among other categories of variables, household variables and the farm types presented in section 4.2 were used. For the creation of the typology, 3 samples were removed because of missing values in their household data. These three samples were not included as well here. For this reason the sample sizes are reduced compared to the previous section.

The random forest identified remote sensing and particularly management variables as important variables for the prediction of yields and responses (Table 14 and 15). Since soil properties were not retained, neither with wet-chemistry-measured or spectrally-predicted soil properties, the two models differed only by the number of observations used to build them. As it seemed that the model was performing better with a higher number of observations, only the results of the model with 145 samples are further explained.

The variation in control yields was dominated by management variables, with the fertiliser and the intercropping of the plot as most important variables. These were followed by environmental variables: the average temperature during the growing season, the average monthly precipitation during the growing season and the altitude of the field. The sixth most important variable was again a management variable, i.e. the number of times the field was weeded. The variation in absolute responses was completely dominated by management variables. The major variable determining the absolute response was the improved variety used on the N2Africa plot. Following was the fertiliser used on the N2Africa plot, with NPK associated with higher predicted responses than DAP. The following three variables were all associated with management variables on the control plot, and the sixth most important variable was the planting date on the N2Africa plot. Plots showing the marginal effects of each previously cited variable on the control yields and responses can be found in Appendix IV (section b).

The environmental variables, retrieved through remote sensing, were thus retained as important for the control yields only. The regression plot of the average temperature showed that the predicted control yields were raising with temperature, and the regression plot of the altitude showed that the predicted control yields were decreasing when altitude was increasing. However, the temperature and altitude values were not overlapping between both districts, as can be seen in Fig. 8.3 (Appendix IV, section b). The effect of these two variables detected by the model was thus likely to be rather an effect of area than a real effect of temperature and altitude alone. Indeed, the optimal temperature range for bush bean growth is 15-23°C (Wortmann et al., 1998), which was exactly the range in the study areas, although with a difference between both districts (temperatures ranging from 19.5 to 23°C in Moshi and from 15 to 19°C in Lushoto). Lushoto had lower yields and lower temperatures, which resulted in the association between yields and temperatures detected by the model. Since temperatures were in the optimal range for common beans (Wortmann et al., 1998), it is not likely that low temperatures were a major cause of lower yields in Lushoto. The same can be said for altitude: fields in Moshi were located up to 1300masl whereas fields in Lushoto started at an altitude of 1200masl. Therefore, lower yields at higher altitudes were likely due to the majority of fields being located in Lushoto. On the other hand, average precipitations were overlapping between districts and higher average precipitations were associated with higher predicted control yields.

The effect of some management variables on control yields and responses can be observed in Fig. 15. This figure depicts cumulative frequency plots against observed control yields and absolute responses, which show an important difference in the proportion of farmers reaching a certain level of yields and responses. However these figures contained observed values which are partly confounded with location. Regression plots showing the marginal effects of the different management variables retained as important by the model are shown in Fig. 8.3 and Fig. 8.4 (Appendix IV). In these figures the effect of location was controlled for, which allowed to disentangle real effects of management variables from the confounding effect of location. Notably, Fig. 15 shows that DAP seemed to result in the lowest predicted control yields. In Fig. 8.3, when controlled for the effect of location, DAP was still associated with lower predicted yields than when NPK was used ($P < 0.001$). However, intercropping impacted yields as well, with monocropped fields associated with higher predicted yields than intercropped fields ($P < 0.001$), and it was difficult to separate the effect of fertiliser and of intercropping on the yields. Indeed, often

the same combinations of fertiliser and intercrop were used together (see Table 6). Hence, additional analyses were done to solve this issue: the intercropping was added to the model predicting the marginal effect of fertiliser on the control yields, and the fertiliser was added to the model predicting the marginal effect of intercropping. It then appeared that DAP and NPK were associated with significantly higher control yields than when no fertiliser was applied (P = 0.006). In these analyses, the monocropped fields were still associated with higher predicted control yields than intercropped ones (P = 0.01). The difference in response between improved varieties appeared clearly in both Fig. 15 and Fig. 8.4 (Appendix IV), with one variety (Uyole Njano) being consistently less performant than the two other varieties. The difference between DAP and NPK in the absolute response was also consistent between Fig. 15 and Fig. 8.4, with NPK being associated with higher responses than DAP. Good management practices such as timely planting and proper weeding were also found to be important variables, and consistently increased yields and responses.

The absolute response was also associated with management variables that concerned the control plot (Fig. 8.4, Appendix IV). The regression plot of the control yields showed that absolute responses were predicted to be higher on plots with higher control yields. Thus, what was determining the control yields could also be determining the absolute response. Control yields were predicted to be higher on monocropped fields, and this suggested that higher absolute responses could also be found on monocropped fields. The regression plot of the absolute response on the intercropping confirmed this (data not presented). The regression plot of the relative planting date on the control plot showed that the absolute responses were predicted to be higher when the control plot was planted early. While there were no direct links between the absolute response and the planting date of the control plot, early planting of the control plot led to higher predicted control yields (data not presented). The relation between the control plot planting date and absolute response could thus be due to the link between higher control yields and higher absolute response. The planting date of the control plot could also be an indicator of the management practices of the farmer. The early planting of the control plot would thus indicate overall good management practices, leading to higher absolute responses.

Table 14. Random forest predictions of bush bean control yields from different categories of variables with the samples that have wet-chemistry-measured soil properties (n = 37) and spectrally-predicted soil properties (n = 145). Refer to Table 3 and 5 (section 3.6.3 and 3.6.5) for a description of the variables. The r^2 and RMSE of the out-of-bag (OOB) predictions are shown. The soil variables are wet-chemistry-measured soil properties for the set of 37 samples, and spectrally-predicted soil properties for the set of 145 samples. The farm types previously created are included in the household variables.

Predictor variables for *control yields*	**Number of variables**	**% of variance explained**			**OOB predictions**	
		n = 37	**n = 145**		**n = 37**	**n = 145**
All variables	40	28.96%	50.96%	r^2 RMSE	0.29 376.08	0.54 288.67
Without household variables	31	29.9%	50.9%	r^2 RMSE	0.29 373.60	0.53 284.38
Without field variables	33	29.35%	50.67%	r^2 RMSE	0.30 375.06	0.53 288.86
Without soil variables	30	28.07%	53.87%	r^2 RMSE	0.27 378.44	0.57 279.98
Without remote sensing variables	30	31.07%	41.46%	r^2 RMSE	0.32 370.45	0.43 315.39
Without management variables	36	11.33%	23.38%	r^2 RMSE	0.09 420.18	0.24 360.83
Household variables	9	-17.05%	16.14%	r^2 RMSE	0.005 482.74	0.16 377.5
Field variables	7	-9.4%	-1.49%	r^2 RMSE	0.005 466.71	0.06 415.29
Soil variables	10	10.08%	-3.77%	r^2 RMSE	0.09 423.12	0.01 419.93
Remote sensing variables	10	4.7%	28.28%	r^2 RMSE	0.06 435.6	0.28 349.1
Management variables	4	43.35%	38.1%	r^2 RMSE	0.42 335.83	0.38 324.32
Management + remote sensing variables	14	34.49%	57.82%	r^2 RMSE	0.33 361.15	0.59 267.72

OOB: out-of-bag. RMSE: root mean square error.

Table 15. Random forest predictions of bush bean absolute response to N2Africa treatments from different categories of variables with the samples that have wet-chemistry-measured soil properties (n = 37) and spectrally-predicted soil properties (n = 145). Refer to Table 3 and 5 (section 3.6.3 and 3.6.5) for a description of the variables. The r^2 and RMSE of the out-of-bag (OOB) predictions are shown. The soil variables are wet-chemistry-measured soil properties for the set of 37 samples, and spectrally-predicted soil properties for the set of 145 samples. The farm types previously created are included in the household variables.

Predictor variables for *absolute response*	**Number of variables**	**% of variance explained**			**OOB predictions**	
		n = 37	**n = 145**		**n = 37**	**n = 145**
All variables	45	38.15%	53.22%	r^2 RMSE	0.36 328.37	0.53 284.55
Without household variables	36	37.13%	53.76%	r^2 RMSE	0.35 331.07	0.53 282.92
Without field variables	38	37.72%	53.00%	r^2 RMSE	0.36 329.51	0.53 282.25
Without soil variables	35	35.59%	55.02%	r^2 RMSE	0.34 335.11	0.55 279.01
Without remote sensing variables	35	38.05%	52.92%	r^2 RMSE	0.36 328.64	0.53 285.45
Without management variables	35	-2.12%	20.14%	r^2 RMSE	0.003 421.95	0.20 371.80
Household variables	9	-10.11%	6.23%	r^2 RMSE	0.001 439.13	0.07 402.86
Field variables	7	-10.01%	-5.89%	r^2 RMSE	0.002 437.95	0.03 428.11
Soil variables	10	-4.5%	-11.38%	r^2 RMSE	0.006 426.82	0.004 439.08
Remote sensing variables	10	2.13%	26.14%	r^2 RMSE	0.04 413.06	0.26 357.56
Management variables	9	39.83%	57.06%	r^2 RMSE	0.38 323.88	0.57 272.01

OOB: out-of-bag. RMSE: root mean square error.

The higher importance of remote sensing variables that the model detected with 145 samples, compared to the results with 37 samples, for both control yields and absolute response suggested that the role of these remote sensing variables only started to be detected when more observations were used.

The inspection of the scatter plot of the observed values against the predicted control yields and response by the best RF models showed that the prediction is clearly improved when the number of observations used to build the model is increased (Appendix IV, section c). No district or trial year seemed to have more influence on the model and be better predicted than others when 145 samples were used, whereas when only 37 samples were used it was visible that some samples were not well predicted by the model (e.g. samples from Lushoto in 2017 or from Moshi in 2017 for the control yields).

For each of the best model obtained in the random forest analysis, the performance of the model was checked by testing its predictive value across years and districts (Appendix IV, Table 8.7 and 8.8). It appeared that splitting the data in distinct groups (here, according to district or trial year) usually resulted in much lower performance of the model when trying to predict in new groups of farms, as Ronner et al. (2016) already noted. For the absolute response, when the explanatory variables were only management variables, about a third of the

variability in the other district could be explained. This was probably because management variables had a similar effect in both districts. When remote sensing variables were involved for the control yields, predictions in other districts were more limited since variables such as temperature, precipitation and altitude were only slightly overlapping between areas. Using areas with more similar environmental conditions could result in better prediction capacity across areas. The prediction across years was very poor, both for control yields and absolute responses. This was probably due to the management practices being so different between both years, in terms of the intercropping and the varieties and fertilisers used.

A reason for this loss of predictive power also lies in the statistical procedure of OOB validation. When the random forest performs its internal out-of-bag validation, the model is trained on about two thirds of the observations in each run, i.e. each tree (here, the default value of 1000 trees in the random forest was used). The model will therefore likely use observations from all groups (i.e. in this case both districts and both years) as training set, as well as observations from all groups as test set. The OOB error estimate is thus efficient when all observations are similar but does not reflect the loss of predictive power observed when the observations are grouped according to a common characteristic not included in the model, and predictions are made on new groups of observations.

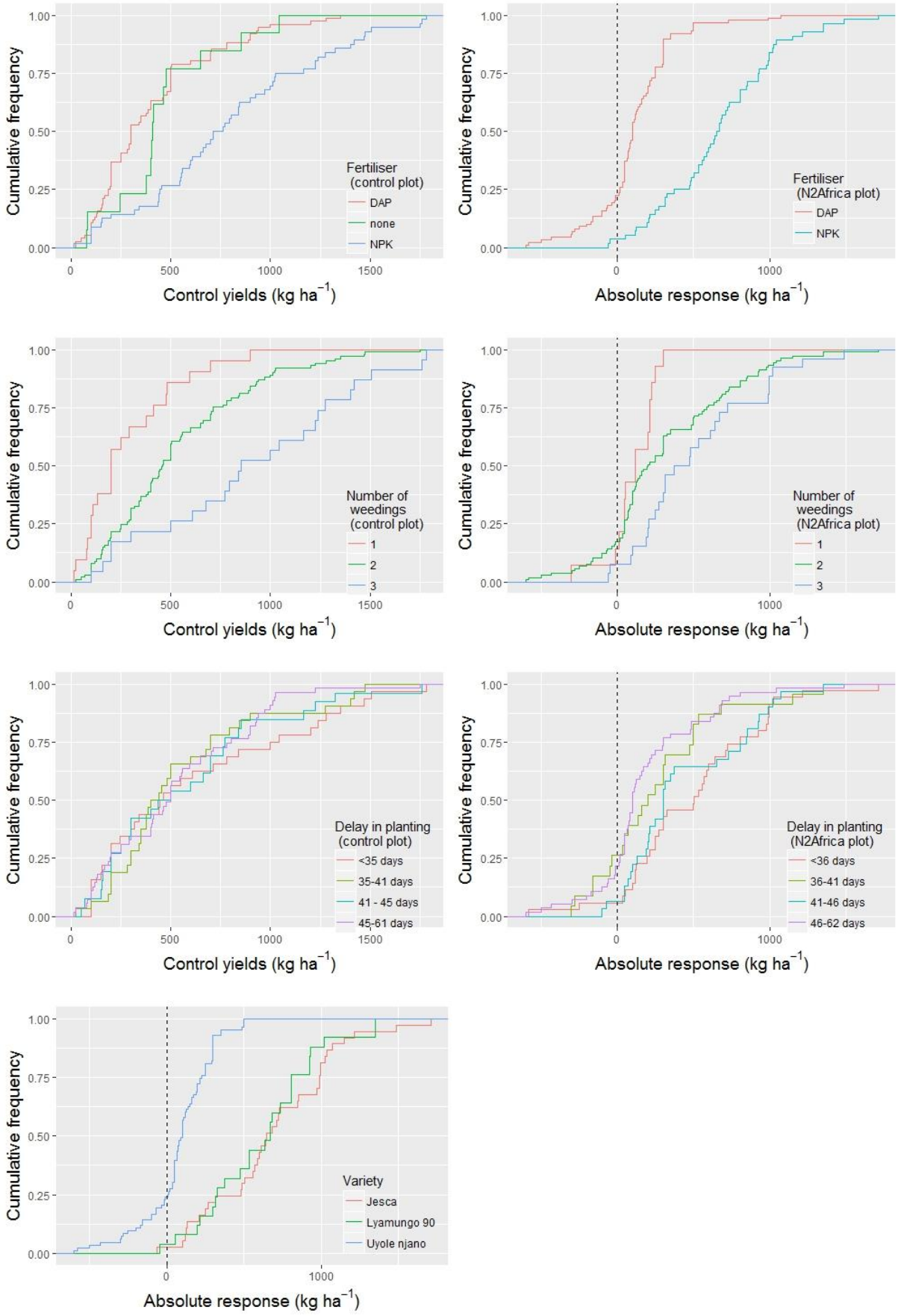


Figure 15. Cumulative frequency plots of different management variables against the observed bush beans control yields and absolute responses.

5. DISCUSSION

In this research, we assessed the accuracy of MIR diffuse reflectance spectroscopy in the prediction of various soil properties, in comparison to conventional wet chemistry measurements. We found that the prediction accuracy of spectroscopy, when assessed on an independent validation set, was lower than what we expected based on literature. Next to this, the role of the different types of soil variables (i.e. wet-chemistry-measured soil properties, spectrally-predicted soil properties and soil MIR reflectance spectra) in the explanation and prediction of the yields and responses of bush beans was evaluated. The wet-chemistry-measured soil properties had a considerably better predictive performance than the other types of soil variables, although it was still limited. Other categories of variables, including household characteristics and farm types built from these characteristics, were then considered next to soil information to find out which type of variables had most explanatory and predictive power. This revealed that soil properties have little role in the determination of yields when other factors are considered, and that the highest explanatory and predictive power came from management variables. These findings are further detailed in the following sections.

5.1 Spectroscopy

The increase in r^2 for all soil properties after calibration shows that the use of reference samples to build a calibration model improves the prediction accuracy of MIR spectroscopy, with r^2 sometimes tripling (e.g. from a r^2 of 0.12 to 0.35 for OC). However, the overall level of accuracy is still very limited even after calibration. It can thus be doubted whether the level of improvement obtained is worth the investment made to perform the reference analyses. Indeed, the r^2 obtained from the independent validation samples (Table 11) indicated that the prediction accuracy was lower than what can be found in literature (see Table 2) (Soriano-Disla et al., 2014; Towett et al., 2015). In addition, when taking a closer look at the four soil properties that were best predicted in this study, it appeared that the performance of the calibration differed between districts (Fig. 14). In most cases, one district was dominating the general calibration resulting in the relatively high r^2, which gave a false impression of the model accuracy. The bias for all soil properties but sand were negative. This showed that spectral predictions of soil nutrients offer a conservative picture of the nutrient levels in a field. Indeed, the negative bias show that the spectral predictions of nutrients seem to usually be underestimations of the wet-chemistry-measured values.

Several reasons can be put forward as to why the prediction accuracy was lower than what was initially expected. One of them is the use of different methods to assess the model accuracy. Indeed, validation of spectrally-predicted soil properties in literature is often done with internal statistical validation methods such as cross-validation (Viscarra Rossel et al., 2006; Tittonell et al., 2008), random forest out-of-bag validation (Towett et al., 2015) or bootstrapping. In these methods the prediction error is repeatedly calculated on subsets of the samples that were used for reference analysis and thus for building the calibration model. The method in this study was different, as we used independent validation samples that were not involved in the calibration. The prediction error was calculated for the independent validation samples, comparing the spectrally-predicted and wet-chemistry-measured values for each soil attribute. This method gives a different picture from what can be found in the previously cited literature (see Table 2 for examples) since the prediction error was greater. The comparison with the r^2 of the reference samples showed that the model was largely overfitting (Table 11). Soriano-Disla et al. (2014) and O'Rourke & Holden (2011) already warned that cross-validation will usually give more optimistic results than assessing the model performance on an independent test set, and hence they recommended the use of independent validation samples. Some studies used independent validation samples as well, but often with a calibration set much larger than 10% of the total population. Shepherd & Walsh (2002) used independent validation sets to validate predictions from NIR spectroscopy, and obtained r^2 higher than 0.7 for most soil attributes. However, they used two third of the samples for calibration and one third for validation. McCarty et al. (2002) used independent validation sets as well to validate the predictions of various forms of carbon, and obtained r^2 all higher than 0.8. In their study the calibration set was again much larger than the validation set. Shepherd & Walsh (2002) also showed that when reducing the calibration set from 67% to 10% of the total population, the predictive performance was decreasing drastically, although not for all soil attributes. These studies show that in order to obtain high r^2 on an independent test set, a larger proportion of the total population should be used to create the calibration model. Indeed, more robust calibration models can be

obtained with larger reference sets that capture more of the variability in the soil population (Soriano-Disla et al., 2014; Wetterlind, 2013). Thus, a drawback of using an independent test set to evaluate the performance of the model is that the costs are increased (because conventional wet chemistry analyses are done on the validation samples) without improving the calibration model. However, using 10% of the total population as reference set is a common method. Our results show that when calibration is made on a small set of 10% of the population, and that internal statistical validation methods are used to assess the prediction accuracy, there is a large risk of overfitting. This is especially true when sample populations are relatively small, as was the case here. In addition, using a reference set larger than 10% of the sample population would increase the analysis costs. This is not desirable since the costs are one of the major benefits of using spectroscopy for soil analysis.

Another possible reason for the low accuracy obtained here is the conventional laboratory analyses. These are as important as the spectral analyses in determining the quality of the spectral predictions. The Salien Agricultural Research Institute (SARI), where the spectral analyses were done, follows standard procedures for spectroscopy developed by the International Centre for Research in Agroforestry (ICRAF). On the other hand, the conventional wet chemistry analyses were done in a laboratory that does not follow standard wet chemistry procedures for reference analyses of MIR spectroscopy (for an overview of the standard reference analyses, see Vågen et al. (2010)). The reference analyses were thus not all the standard ones that the calibration methods were developed for (e.g. Mehlich 3 should be used for Mg, K, P, Ca and Na analysis), and that is likely to alter the final predictions (Soriano-Disla et al., 2014). In addition, the quality of the wet chemistry analyses can be somewhat questioned, especially for phosphorus which had high reported values. Doubts about the accuracy of the results coming from the same laboratory were already reported in the past (Thuijsman, 2017). Nonetheless, these samples were still used in the analyses because they were thought to reflect fields with a higher level of phosphorus than the average, and because removing them would have reduced the reference set size by more than 20%.

A last possible reason is that it is known that calibration of spectral data cannot give very accurate predictions for samples that are not well represented (in their spectra and in their wet-chemistry-measured soil properties) by the ones used for calibration (Nocita et al., 2015; Rourke & Holden, 2011; Soriano-Disla et al., 2014). Hence, samples from new locations are less likely to be correctly predicted by a model built in another location. Some authors, cited by Soriano-Disla et al. (2014), reported that models can successfully predict unknown samples coming from geologically homogenous areas, since they should not greatly differ from the samples used to build the calibration. We could see in this research that the soils of Moshi and Lushoto have rather different properties, due to their respective geological origin and climatic conditions. It is thus likely that taking more samples in both districts (especially Moshi) and doing a separate calibration for both districts would have produced results of higher accuracy. This method is suggested by Soriano-Disla et al. (2014). Figure 14 seems to confirm this, showing that for pH and K samples of the two districts follow a different regression line.

In this research, we continued to work with spectrally-predicted soil properties even though the validation results obtained were indicating a lower accuracy than expected. We assumed that even if the predicted values might not be fully representative of the “true” value of each soil property, they still give a useful indication of the range of values of the soil property, i.e. if the latter is in low or high levels. We also considered that the (even somewhat wrong) picture they gave of the 148 sampled farms was more useful to the research than using only the 39 farms with wet chemistry measurements. Indeed, they might have been a better proxy for crop yields and have more relevance in the explanation and prediction of yields and responses. However, we saw in Table 12 and 13 that the few samples with wet chemistry measurements were more relevant to yields and responses than the spectrally-predicted soil properties. Some studies, on the contrary, argue that the lower accuracy of spectroscopy is compensated by the much larger number of samples that can be collected thanks to this method (Nocita et al., 2015; O’Rourke & Holden, 2011). They claim that it raises the overall accuracy of soil assessments, compared to an assessment done with a few but more accurate measurements. O’Rourke and Holden (2011) did a systematic comparison of several methods for soil organic carbon (SOC) analysis, taking into account not only the accuracy but also the costs and time involved in each analysis. They concluded that in spite of their lower accuracy, the two spectral methods that were assessed outperformed the two conventional laboratory analyses they compared it with. This conclusion was largely due to the fact that the spectral methods appeared to be almost 15 times cheaper than the conventional laboratory analyses, and were much faster. However their observation depends on two important issues: the amount of samples analysed, and the specific soil attribute considered.

Firstly, the reduction in costs is of course scale-dependent: the study was based on a comparison of the analysis of 7000 samples per year with each method. When the sample size increases to such high scales, the costs per sample decrease drastically. In a study likes ours, with almost 150 samples analysed, the costs per sample are less favourable compared to the observed reduction in accuracy. Secondly, OC is always among the attributes best predicted by spectroscopy. If the study of O'Rourke and Holden (2011) was done with phosphorus or potassium, which are known to be poorly predicted by spectroscopy (see Table 2), their conclusion would have likely been less optimistic.

From this discussion we conclude that care should be taken when performance evaluation of spectral predictions is the result of internal statistical validation methods. It should be kept in mind that there is a large risk of overfitting with such methods, especially when the calibration set is made of only a small proportion of the total population. Independent test sets give a more accurate insight into the calibration performance and the quality of the spectral predictions. We found that the accuracy of the prediction was very poor, and it is likely that better results would have been obtained with the collection of more samples and a separate calibration in the two districts. For large-scale sampling we thus recommend to group the samples according to geography, or some other classification criteria, and to perform various separate calibrations rather than only one calibration for the larger area. The benefits of spectroscopy increase with the number of samples, as the costs and processing time per sample are reduced, and the lower accuracy is partly compensated by the large number of samples analysed. However the benefits of spectroscopy also depend on which soil property is being analysed, as some properties do not respond well enough to MIR spectroscopy.

5.2 Explaining variability in yields and response

5.2.1 The role of soil properties

The attempt of linking the different types of soil information that were available with the bush bean yields and responses to N2Africa treatments was summarised in Table 12 and 13. It appeared that wet-chemistry-measured soil properties were more related to yields and responses than spectrally-predicted soil properties and spectra. In the context of smallholder farming systems of sub-Saharan Africa, soil information is mainly sought to know how crop yields will be affected by soils (Sanchez et al., 2003). Our results indicate that wet-chemistry-measured soil properties are more relevant to this issue, and thus that the loss of accuracy of spectrally-predicted soil properties discussed in the previous section is also a loss of meaning and relevance of these values.

The linear mixed models using wet-chemistry-measured soil properties, reduced with a backward variable selection method showed that more variability in control yields than in absolute responses could be explained. Previous studies identified phosphorus and potassium as the main soil constraint to bush beans yields in Tanzania, the second being a particular constraint to Lushoto (Amijee & Giller, 1998; Smithson et al., 1993). Nor phosphorus nor potassium were retained here in the model explaining control yields, but potassium was negatively linked with the response, showing that fields with low levels of potassium have better responses to treatments. This was true in both districts. pH was significantly related to the response, showing that higher responses were obtained on soils with higher pH, this even though very few soils had a pH below 5.5. Nonetheless, with the insights provided by the rest of the results, we found that the role of soil properties in the determination of yields and response is negligible when other variables are considered. The information provided by the linear models should thus be seen as useful but not essential information on what really drives the variability in yields and responses.

5.2.2 Using multiple categories of variables to explain and predict the variability in bush bean yields and responses

We used random forest to disentangle the role of several categories of variables that could have an impact on bush bean yields and responses, in order to quantify the importance of each category. We will first discuss the findings of the household typology construction, and then develop on the meanings of the results from the random forest models.

a. Household typology

Five farms types were defined using the household characteristics that were available. Even after controlling for the possible confounding effect of location, the first farm type, considered as the lowest resource-endowed category (Table 9), had consistently lower yields than the better resource-endowed types. However, the second farm type had unexpectedly high predicted control yields even though it was part of the low resource-endowed category. This could be partly explained by the fact that it had the lowest income diversification, showing that much of the household's efforts could be dedicated to working on their fields. However, the second farm type also had the smallest household size, hence the availability of family labour in this farm type is somewhat limited, which could lead to lower yields. The reason of high yields for the second farm type is therefore not very clear. Perhaps the use of farm size, an important household characteristics identified in several studies (Chikowo et al., 2014; Ronner et al., 2016) would have rendered a different, and more accurate, picture of the typology and would have explained this apparent discrepancy.

The absolute response was not found to be affected by household characteristics. This is a positive result, as it shows that the potential response is the same for all households. However, since all participating farms were given N2Africa technology packages, the differences between farm types in terms of access to fertiliser and improved seeds are not applicable to these results. Franke et al. (2016) observed in Rwanda that wealthier households owned fields with a higher inherent soil fertility, which could also cause a difference in response between farm categories. Nonetheless, we found that when location was accounted for there were no differences in soil characteristics between farm types. Hence, during the trials, the only factors that could have caused lower responses in lower resource-endowed farms would perhaps be poor management practices due to labour constraints, such as delays in planting and in weeding. In practice, not all households have the same possibilities of access to fertilisers and improved varieties (Chikowo et al., 2014; Tittonell et al., 2005). This reduces the possibility of yield increases for certain household categories. Household typologies could possibly be used to give an indication of which farms are able to purchase inputs and which are not. This could help in the development of strategies making it possible for farmers belonging to lower resource-endowed categories to get access to these inputs, and thus to be given the chance to achieve their potential response.

Therefore, we found that the household typology revealed some useful information about the challenges faced by different households. However, the typology did not have any role in the random forest analysis, showing that its implication for yields is only small when other factors are taken into account.

b. Assessing the role of various categories of variables to predict yields and responses with random forest

The random forest was a useful tool to reveal the relative importance of each category of variables introduced in the model. The main conclusion from the models is that soil variables, be it wet-chemistry-measured soil properties or spectrally-predicted soil properties, are not good predictors of the variability in bush beans yields and responses. The management variables were the predictors with most importance in the models, along with environmental variables for the control yields.

This finding suggests that in the study area, soils are not the most important drivers of bush bean yields. Tittonell et al. (2008) studied the relative importance of soil and management variables to explain maize yields in Kenya and found that management variables had most influence on yields, whereas soil variables had some role only when few inputs (organic or mineral) were applied to the field. Franke et al. (2016) found in Rwanda that no soil or management variables were related to the response to fertiliser application, but rather that resource endowment level and the farmer gender were the main drivers of variability in climbing bean yields, and that resource endowment level was also a main determinant of the response to treatments. In the study of Ronner et al. (2016) about soybean yields in Nigeria, it is shown that management variables dominated the variability in yields and response, and some soil and environmental variables had a smaller role. Maman et al. (2018) cites a long list of studies that did not find a relationship between soil test information and responses to fertiliser application in SSA. Jeong et al. (2016) studied the variability of various crops at different geographical scales in the United States, and found that management and environmental variables were the main drivers of yield variability, with clay content being an important variable in just one case. Giller et al. (2011) also found that the role of good management practices and fertiliser application was essential to increase maize productivity. A study in Western Kenya allowed to directly compare three types of plot: a plot with the current farmer's practice (with or without fertiliser), a plot with a better management without fertiliser, and the third plot with a better

management and the use of fertiliser. The better management consistently improved productivity, demonstrating the importance of good management practices. The difference in yields between the farmer's practice and the better managed plot was the first yield gap and was mainly due to labour constraints. The second yield gap was the difference between the second and the third plot, i.e. the additional increase in yields achieved through fertiliser application when good management practices are used. The total increase in yields was greater on fields with poorer soil fertility and poorer yields (Giller et al., 2011). Even though this list is not exhaustive, the previously cited studies show that management variables are usually key in the determination of yields, unlike soils that rarely play an important role in the yield variability when other factors are considered.

However, Ronner et al. (2018) did not find significant relationships in Uganda between climbing bean yields and management practices, except for the number of stakes per hectare. Remarkably, there was no increase in yields after application of fertiliser, and variety only had a weak effect on yields. This shows that the relation that we found in this study between bush bean yields and management practices is not always valid. It is also possible that if more control plots were grown without fertiliser, soil variables would have taken a greater importance in explaining the control yields, as found by Tittonell et al. (2008), especially since two districts with rather different soil properties were considered. However, many other variables have a strong influence on yields. Therefore we doubt that soil variables would have had a high predictive role for control yields even if more control plots were grown without fertiliser.

Predicting variability in control yields

For control yields, the model clearly showed that the input in the control plot was the major determining factor. Both DAP and NPK were efficient in increasing yields, since when the effect of intercropping was taken into account in the model, they were not associated with significantly different predicted control yields. This indicates that the use of fertiliser should be strongly encouraged among smallholder farmers.

Without surprise, the model predicted higher yields when control plots were monocropped. Ideally, monocropped and intercropped yields should be made comparable, using an index such as the Land Equivalent Ratio (LER) or calculating the number of bush beans planted in the plots. However, not enough information was available from the original dataset to do so. Using both monocropped and intercropped fields in the model has the advantage of revealing whether the intercropping may have an influence on response. Even though bush bean yields were higher in pure stands, the intercropped stands allowed to cultivate another crop on the same area of land and possibly achieving higher total yields. Intercropping also provides multiple other benefits: it was shown by Kermah et al. (2017) when studying intercrop of maize and legumes in Ghana that intercropped systems made a more efficient use of environmental resources and were more advantageous agronomically and economically than monocropping. In addition, they found that intercrops gave most benefits when used in fields with low fertility and less favourable environmental conditions. Yields are therefore not the only factor to take into account when considering intercropped and monocropped systems.

The next three variables influencing yields the most were all related to the environmental conditions of the fields, i.e. temperature, precipitation and altitude. It should be noted that these variables had less than half of the influence on control yields compared to the fertiliser and the intercropping (see Fig. 8.3, Appendix IV). As discussed in section 4.4.2, it is difficult to separate the effect of each environmental variable on yields, since the temperature and altitude in both districts hardly overlapped. Therefore, lower temperatures and higher altitudes, that are only found in Lushoto, are associated with lower predicted yields although these might be the result of other factors. Nonetheless, the range of average precipitation was the same for both districts, hence we can be more confident that a higher average precipitation results in higher yields. The sixth variable concerned the number of weedings. Weeding was shown by the model to have some efficiency in increasing yields. It does so by limiting competition for nutrients, water, space and sunlight. However, it can be argued that in such steep terrains, weeding can cause higher erosion risks and thus be negative for soil fertility and yields in the long term (Keesstra et al., 2016). Alternative techniques to weedings that help controlling weeds, such as mulching, could be encouraged in the study areas. Mulching not only reduces weed infestation, but can also increase soil fertility and reduce pests and disease occurrence (Saddiq et al., 2017). However, crop residues are preferably used by farmers for livestock feed and cooking fuel (Agegnehu & Amede, 2017; Andriesse & Giller, 2015) and hence the adoption of mulching practices by farmers is low. Lunze et al. (2012) also suggested the use of cover crops to limit weeds while controlling soil erosion.

Predicting variability in absolute response

The absolute response was best explained with the use of management variables only. As most of the control plots had fertiliser applied, the improved variety was often the major difference between the control and the N2Africa plots and the model found it to be the variable with most impact on the absolute response. In steep terrains such as the ones found in the study areas, especially in Lushoto, some farmers reported that important amounts of fertiliser were flowing down the slope and lost. It was one of the reasons they were not using fertiliser (Wickama & Mowo, 2001). The use of improved varieties instead of fertiliser in these areas could possibly improve yields while avoiding such investment losses, even though the performance of improved varieties without fertiliser is likely to be lower.

After variety and fertiliser, the three following variables that impacted the absolute response were all associated with the control plot. As discussed in section 4.4.2, the model indicated that higher absolute responses were predicted to be on fields with high control yields. Franke et al. (2016) and Ronner et al. (2016) made the same observation in their own study. Nevertheless, the correlation between observed control yields and observed absolute response was very weak (Fig. 9) and it can be observed in Figure 9 that the highest losses also occurred where control yields were higher, i.e. after control yields of 500 kg ha^{-1}. Hence it seems that plots with high control yields are also more at risk of experiencing losses. Since the variability in control yields was dominated by management and environmental variables, some explanation of variability in the response can be found in the same variables. It seems logical that environmental variables that favour higher yields will also favour higher responses, and vice versa. The management variables applied on the control plot that were retained as important variables for the response could be taken as proxies for the level of management in the farm. Farmers that applied good management practices such as timely planting on the control plot probably applied them to the N2Africa plot as well, leading to higher responses. This was also shown by the sixth most important variable for the absolute response prediction: the later the N2Africa plot was planted, the lower the response was predicted to be.

We can conclude from this analysis that the response to a treatment is mainly driven by management variables. Finding the improved variety that will yield the most in the environment studied is a major concern. Where rainfall is irregular or is becoming increasingly irregular, early maturing varieties should be preferred as they are less susceptible to be badly affected by the uncertain rainfall than late maturing varieties (Thuijsman et al., 2017). During trials in Western Tanzania that compared various bean varieties, Jesca and Lyamungo 90 were found to outperform all the other varieties in terms of yields (Bucheyeki & Mmbaga, 2013). Not only that, but they were also ranked higher by the farmers according to their own evaluation criteria. These included yields, but also cooking time, taste, suitability for market and resistance to disease or pests. These criteria are important to take into account, as they influence the adoption potential of varieties by farmers (Bucheyeki & Mmbaga, 2013). This study showed that there is an important adoption potential for Jesca and Lyamungo 90.

Other management variables such as timely planting and weed control are tasks that are labour demanding and require to have enough work force available. It can thus be difficult for small and low resource-endowed households to correctly fulfil these tasks in a timely manner, especially if they are selling their own labour force to other wealthier farmers. Similarly, the use of improved seeds and of fertiliser in sufficient amount depends on the cash availability of the household and on the facility of access to these resources. This shows why the household characteristics may have an impact on such management practices, and thus on yields, since well resource-endowed households are better able to hire labour and to use fertilisers and improved varieties.

Predictions in new areas or growing seasons

The out-of-bag predictions of the RF reached an r^2 of almost 0.6 for both control yields and absolute response, which shows that the relatively few variables used in the model confidently predicted more than half of the variation in yields and responses in new observations. This is a good performance, compared for instance to the study of Ronner et al. (2016) where the highest cross-validated r^2 was 0.47. However, the predictive performance was reduced when using only the data from one district or one trial year to train the RF model and testing it in the other district or trial year. OOB r^2 is an accurate parameter of model performance, that does not require the use of an additional validation method such as cross-validation (Breiman & Cutler, n.d.). Nevertheless, we saw that when there are clusters in the data, like was the case here with district and year, the OOB prediction does not give a good indication of the performance of the model across clusters. In such cases, validating the model on independent data from another cluster likely reduces the prediction performance. Also, some clusters might be less represented among the total number of observations and thus be less well represented by the model.

This was the case for examples with the absolute response in 2017, where only 8% of the variability was explained using the same variables as in the general model that rendered an OOB r^2 of 0.57. Observations from the trial of 2017 formed a third of all observations, and they seemingly did not influence enough the general model to be well represented by it. This stresses the importance of collecting balanced data, with a similar number of observations in each cluster of data. In addition, it shows that models are not able to make predictions outside of the variable range that they have been trained with, as already pointed out by Jeong et al. (2016). Thus, if we want to make predictions in large areas, clusters based for instance on geographical proximity should be defined, and a rather intensive data collection should be made in each cluster. Furthermore, care should be taken that one condition is not present in only one cluster, or over-represented in one cluster, such as was the case here for altitude and temperature and for intercropping. If, for instance, we had at least one area with a range of altitude and temperatures similar to the ones found in Lushoto, we could have concluded with more confidence about the role of these variables on yields and responses.

5.3 Reflections on the research

"The basis of data collection is questionable: farmers, without knowing in advance, are asked to recall numbers and figures about almost everything going on in their lives, while most of them do not keep track of any numbers." Citation from Lotte Klapwijk, in Alvarez et al. (2014).

As summarised by the above quote, there can be discussion about the inherent quality of the data collected from surveys like were done in the adaptation trials in 2016 and 2017, as well as during the fieldwork of 2018. The situation described by Lotte Klapwijk in Alvarez et al. (2014) is a good summary of what was experienced during fieldwork and what possibly occurs often when such interviews are done. The overall impression of these surveys is that it is likely that some questions will get different answers from the same farmers when asked on different days. However, farmers of course have the best knowledge about their farm and their livelihood, and every answer they give supposedly reflects some aspect of their reality. It is thus not suggested to stop involving the farmers in research, but to keep in mind these unavoidable difficulties. An approach was developed by Hammond et al. (2017) to overcome such issues commonly occurring during household surveys.

A good example illustrating the citation above is the farm size: during the fieldwork of 2018, farmers were asked what area of land they owned, but it was easy to see that some did not know what an acre represented and were saying a number quite randomly. Moreover, a language barrier was present, that prevented direct communication and required the field liaison officer to ask the questions and act as a translator. Two different field liaison officers helped during fieldwork, one in Moshi and one in Lushoto. This seemingly introduced a small bias between the two districts as they both had their own way of working and communicating. The fieldwork procedure that had to be done in each farm was discussed with each field liaison officer before starting the fieldwork, but there was still some misunderstanding. For example, an issue arose from the question about farm size. The fieldwork started in Moshi, and when coming to Lushoto it appeared that all farmers there had much bigger farms. The reason was a difference in farm organisation between the two regions: in Moshi all households have some land close to their home as well as other plots of land further away, as described in section 3.2.2. The field liaison officer interpreted farm size as the area of land that was close to their home, and was thus asking only about that and not about the other plots of land. This resulted in much smaller values for farm size in Moshi and was the reason why the farm size was not used in the typology construction. Another example of miscommunication was that simple words such as farm, field and plot were surprisingly problematic as they were used and interpreted differently by different people.

Another issue to keep in mind was that some assumptions were made to carry out this study. Firstly, in an attempt to link soil properties with yields and responses, it was assumed that the soils sampled in 2018 would have the same properties as they had during the trial season one or two years earlier. The soils are not expected to have changed much in such a small time frame, especially for organic matter content and soil texture which are known to change slowly (Vågen et al., 2010). However, fertiliser application in the season of the 2018 fieldwork could have resulted in different nutrient values compared to the situation in the trial season. In addition, control plots and N2Africa plots were assumed to have identical soil properties. This should be quite true in theory, as the N2Africa plots were supposed to be next to the control plots, but in reality it was not always the case, especially in Moshi. It appeared that in Lushoto more care was taken during the adaptation trials to use plots very close to each other, but in Moshi it happened more often that the plots were not in close proximity.

The statistical analyses and the conclusions that can be made from them are another limit of this research. Many results of the random forest models were possibly due to confounding effects of location and/or trial year, but

only guesses can be made about the extent of the real effects and of the confounding effects. A better sampling strategy, with a balanced number of observations in each district and each year studied would have probably reduced this problem. In addition, making sure that one district or year contains fields with a variety of fertiliser or intercropped and monocropped fields, and not one year with only monocropped fields and another year with mostly intercropped fields as was the case here, would allow for a better understanding of the yield variability and an easier application of the statistical analyses. In addition, the study would have probably benefitted from additional variables, such as for instance the exact quantities of fertiliser that were applied during the trial, whether manure was applied, or the distance of the field to the homestead since it was found that fields of a same farm have varying soil fertility and yields (Giller et al., 2011; Tittonell et al., 2007; Vanlauwe et al., 2006).

To finish, we would like to emphasise that soil properties were not found important here does not mean that nothing should be done to improve soil fertility. There is a great problem of poor soil fertility and nutrient depletion in SSA and along with practices increasing yields in short-term, long-term soil management practices should be developed and encouraged. The use of grain legumes is already a step towards soil quality improvements, since they do not deplete soils as much as other crops do. They can even contribute to a positive N balance if residues are left on the field (Vanlauwe & Giller, 2006). Soil organic matter (SOM) is central to soil quality as it has a strong influence on the soil's cation exchange capacity, structure, water-holding capacity, microorganisms and nutrient stocks (Brady & Weil, 2010). SOM management is therefore important to improve soil health. It has been proved that yield increases due to the use of mineral fertilisers participate to increasing the SOM pool through roots and above-ground residues (Vanlauwe & Giller, 2006). Hence, short term yield increases due to fertiliser use and good management practices participate to improved soil quality on the long term.

6. CONCLUSION

Spectroscopy appeared to not only predict soil properties with a low accuracy compared to wet-chemistry-measured soil properties, but also to have low relevance for the explanation and prediction of crop yields and responses to yield-improving treatments. This applies to both the soil properties predicted from the spectra, and to the soil spectra themselves. Conventional wet chemistry analyses, even though they are more expensive and have many disadvantages, have more meaning for crop yields and responses. However, we showed that soil properties do not have any importance when other variables are considered, namely management variables. These results are positive in two ways. Firstly, the main variations in yields and responses were associated with management variables, which are factors that can be controlled and adapted when necessary. The management variables were mainly the improved variety and fertiliser used, but good management practices such as proper weeding and timely planting were also important. Secondly, we found that the potential response to the application of fertiliser and the use of improved varieties is the same for all farmers, irrespective of their household characteristics and soil properties. We thus have the clear indication that research-in-development programs such as N2Africa, which reached thousands of farmers with management practices to increase grain legume production and developed access of households to markets and inputs, are the most efficient ways to make a real impact and increase yields of smallholder farmers in sub-Saharan Africa.

7. REFERENCES

Agegnehu, G., & Amede, T. (2017). Integrated Soil Fertility and Plant Nutrient Management in Tropical Agro-Ecosystems : A Review. *Pedosphere: An International Journal*, *27*(4), 662–680. https://doi.org/10.1016/S1002-0160(17)60382-5

Alvarez, S., Paas, W., Descheemaeker, K., Tittonell, P., & Groot, J. (2014). *Constructing typologies, a way to deal with farm diversity: general guidelines for the Humidtropics. Report for the CGIAR Research Program on Integrated Systems for the Humid Tropics*. Wageningen, the Netherlands. Retrieved from https://cgspace.cgiar.org/bitstream/handle/10568/65374/typology_guidelines.pdf?sequence=1

Amijee, F., & Giller, K. E. (1998). Environmental constraints to nodulation and nitrogen fixation of Phaseolus vulgaris L. in Tanzania. I. A survey of soil fertility, root nodulation and multi-locational responses to Rhizobium inoculation. *African Journal of Crop Science*, *6*(2), 159–169.

Andriesse, W., & Giller, K. E. (2015). The state of soil fertility in sub-Saharan Africa. *Agriculture for Development*, *24*, 32–36.

Brady, N. C., & Weil, R. R. (2010). *Elements of the Nature and Properties of Soils* (3rd ed.). New Jersey: Pearson Education.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Breiman, L., & Cutler, A. (n.d.). Random Forests: Classification/clustering - Description. Retrieved November 3, 2018, from https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm#varimp

Bremner, J. (2012). *Population and Food Security: Africa's Challenge*. Washington. Retrieved from https://assets.prb.org/pdf12/population-food-security-africa.pdf

Broughton, W. J., Hernández, G., Blair, M., Beebe, S., Gepts, P., & Vanderleyden, J. (2003). Beans (Phaseolus spp.) - model food legumes. *Plant and Soil*, *252*(1), 55–128. https://doi.org/10.1023/A:1024146710611

Bucheyeki, T. L., & Mmbaga, T. E. (2013). On-Farm Evaluation of Beans Varieties for Adaptation and Adoption in Kigoma Region in Tanzania. *ISRN Agronomy*, *2013*, 1–5. https://doi.org/10.1155/2013/436064

Burgess, N. D., Butynski, T. M., Cordeiro, N. J., Doggart, N. H., Fjeldså, J., Howell, K. M., … Stuart, S. N. (2007). The biological importance of the Eastern Arc Mountains of Tanzania and Kenya. *Biological Conservation*, *134*(2), 209–231. https://doi.org/10.1016/j.biocon.2006.08.015

Cernay, C., Ben-ari, T., Pelzer, E., Meynard, J., & Makowski, D. (2015). Estimating variability in grain legume yields across Europe and the Americas. *Nature Publishing Group*, 1–11. https://doi.org/10.1038/srep11171

Chikowo, R., Zingore, S., Snapp, S., & Johnston, A. (2014). Farm typologies, soil fertility variability and nutrient management in smallholder farming in Sub-Saharan Africa. *Nutrient Cycling in Agroecosystems*, *100*(1), 1–18. https://doi.org/10.1007/s10705-014-9632-y

Didan, K. (2015). MOD13A1 MODIS/Terra Vegetation Indices 16-Day L3 Global 250m SIN Grid V006 [Data set]. NASA EOSDIS LP DAAC. https://doi.org/10.5067/MODIS/MOD13Q1.006

Du, C., & Zhou, J. (2009). Evaluation of soil fertility using infrared spectroscopy: A review. *Environmental Chemistry Letters*, *7*(2), 97–113. https://doi.org/10.1007/s10311-008-0166-x

Education Policy and Data Center. (2018). Country Profile: Tanzania. Retrieved September 15, 2018, from https://www.epdc.org/country/tanzania

Elith, J., Leathwick, J. R., & Hastie, T. (2008). A working guide to boosted regression trees. *Journal of Animal Ecology*, *77*(4), 802–813. https://doi.org/10.1111/j.1365-2656.2008.01390.x

FAO. (2010). *The State of Food Insecurity in the World - Adressing Food Insecurity in Protracted Crises*. Rome, Italy. Retrieved from http://www.fao.org/docrep/013/i1683e/i1683e.pdf

FAO. (2018). *GIEWS Country Brief: The United Republic of Tanzania*. Rome, Italy. Retrieved from

http://www.fao.org/giews/countrybrief/country.jsp?code=TZA

Farr, T. G., Rosen, P. A., Caro, E., Crippen, R., Duren, R., Hensley, S., … Alsdorf, D. E. (2007). The shuttle radar topography mission. *Reviews of Geophysics*, *45*(2). https://doi.org/10.1029/2005RG000183

Fernandes, E. C. M., Oktingati, A., & Maghembe, J. (1984). The Chagga homegardens: a multistoried agroforestry cropping system on Mt. Kilimanjaro (Northern Tanzania), *1*(1), 73–86.

Fick, S. E., & Hijmans, R. J. (2017). WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. *International Journal of Climatology*, *37*(12), 4302–4315. https://doi.org/10.1002/joc.5086

Franke, A. C., Baijukya, F., Kantengwa, S., Reckling, M., Vanlauwe, B., & Giller, K. E. (2016). Poor farmers - Poor yields: Socio-economic, soil fertility and crop management indicators affecting climbing bean productivity in Northern Rwanda. *Experimental Agriculture*, 1–21. https://doi.org/10.1017/S0014479716000028

Franke, A. C., van den Brand, G. J., & Giller, K. E. (2014). Which farmers benefit most from sustainable intensification? An ex-ante impact assessment of expanding grain legume production in Malawi. *European Journal of Agronomy*, *58*, 28–38. https://doi.org/10.1016/j.eja.2014.04.002

Franke, A. C., van den Brand, G. J., Vanlauwe, B., & Giller, K. E. (2018). Sustainable intensification through rotations with grain legumes in Sub-Saharan Africa: A review. *Agriculture, Ecosystems and Environment*, *261*, 172–185. https://doi.org/10.1016/j.agee.2017.09.029

Funk, C., Peterson, P., Landsfeld, M., Pedreros, D., Verdin, J., Shukla, S., … Michaelsen, J. (2015). The climate hazards infrared precipitation with stations — a new environmental record for monitoring extremes. *Scientific Data 2*, 1–21. https://doi.org/10.1038/sdata.2015.66

Gareth, J., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning with Applications in R* (1st ed.). New York: Springer-Verlag. https://doi.org/10.1007/978-1-4614-7138-7

General Guidelines for 2017 (Focal) adaptations [Guidelines]. (n.d.). Retrieved March 10, 2018, from http://intranet.n2africa.org/intra_guidelines/Agronomy

Giller, K. E. (2001). *Nitrogen fixation in tropical cropping systems* (2nd ed.). New York, NY: CABI Publishing. https://doi.org/10.1079/9780851994178.0000

Giller, K. E., Tittonell, P., Rufino, M. C., van Wijk, M. T., Zingore, S., Mapfumo, P., … Vanlauwe, B. (2011). Communicating complexity: Integrated assessment of trade-offs concerning soil fertility management within African farming systems to support innovation and development. *Agricultural Systems*, *104*(2), 191–203. https://doi.org/10.1016/j.agsy.2010.07.002

Gourlay, S., Aynekulu, E., Carletto, C., & Shepherd, K. D. (2017). *Spectral Soil Analysis & Household Surveys*. Washington, DC. Retrieved from http://siteresources.worldbank.org/INTLSMS/Resources/3358986-1423600559701/SoilGuidebook_full_web_final.pdf

Grömping, U. (2009). Variable Importance Assessment in Regression: Linear Regression versus Random Forest. *The American Statistician*, *63*(4), 308–319. https://doi.org/10.1198/tast.2009.08199

Hammond, J., Fraval, S., van Etten, J., Suchini, J. G., Mercado, L., Pagella, T., … van Wijk, M. T. (2017). The Rural Household Multi-Indicator Survey (RHoMIS) for rapid characterisation of households to inform climate smart agriculture interventions: Description and applications in East Africa and Central America. *Agricultural Systems*, *151*, 225–233. https://doi.org/10.1016/j.agsy.2016.05.003

Hillocks, R. J., Madata, C. S., Chirwa, R., Minja, E. M., & Msolla, S. (2006). Phaseolus bean improvement in Tanzania, 1959-2005. *Euphytica*, *150*(1–2), 215–231. https://doi.org/10.1007/s10681-006-9112-9

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, *24*(6), 417–441. https://doi.org/http://dx.doi.org/10.1037/h0071325

Ishwaran, H., & Kogalur, U. B. (2018). Random Forests for Survival, Regression, and Classification (RF-SRC), R package version 2.7.0.

ISRIC. (n.d.). SoilGrids. Retrieved April 3, 2018, from https://soilgrids.org

IUSS Working Group WRB. (2015). *World Reference Base for Soil Resources 2014, update 2015. International soil classification system for naming soils and creating legends for soil maps. World Soil Resources Reports* (Vol. 106). Rome, Italy: FAO.

Jahnke, H. E., Tacher, G., Keil, P., & Rojat, D. (1988). Livestock production in tropical Africa with special reference to the tsetse-afferent zone. In *Livestock production in tsetse affected areas in Africa - Proceedings of a meeting held 23-27 November 1987*. Nairobi, Kenya: ILCA/ILRAD.

Jeong, J. H., Resop, J. P., Mueller, N. D., Fleisher, D. H., Yun, K., Butler, E. E., … Kim, S. H. (2016). Random forests for global and regional crop yield predictions. *PLoS ONE*, *11*(6), 1–15. https://doi.org/10.1371/journal.pone.0156571

Jones, A. L. (1999). *Phaseolus bean: Post-harvest Operations*. Retrieved from http://www.fao.org/3/a-av015e.pdf

Kaizzi, K. C., Byalebeka, J., Semalulu, O., Alou, I. N., Zimwanguyizza, W., Nansamba, A., … Wortmann, C. S. (2012). Optimizing smallholder returns to fertilizer use: Bean, soybean and groundnut. *Field Crops Research*, *127*, 109–119. https://doi.org/10.1016/j.fcr.2011.11.010

Keesstra, S. D., Bouma, J., Wallinga, J., Tittonell, P., Smith, P., Cerdà, A., … Fresco, L. O. (2016). The significance of soils and soil science towards realization of the United Nations sustainable development goals. *SOIL*, *2*, 111–128. https://doi.org/10.5194/soil-2-111-2016

Kelley, K., & Maxwell, S. E. (2003). Sample Size for Multiple Regression: Obtaining Regression Coefficients That Are Accurate, Not Simply Significant. *Psychological Methods*, *8*(3), 305–321. https://doi.org/10.1037/1082-989X.8.3.305

Kermah, M., Franke, A. C., Adjei-Nsiah, S., Ahiabor, B. D. K., Abaidoo, R. C., & Giller, K. E. (2017). Maize-grain legume intercropping for enhanced resource use efficiency and crop productivity in the Guinea savanna of northern Ghana. *Field Crops Research*, *213*, 38–50. https://doi.org/10.1016/j.fcr.2017.07.008

Kihara, J. (2014). Predicting crop yield and response to nutrients from soil spectra: example from sub-Saharan Africa. Retrieved April 18, 2018, from https://www.slideshare.net/CIAT/kihara-et-al-predicting-crop-yield-and-response-from-soil-spectra-wcss-2014

Kuivanen, K. S., Alvarez, S., Michalscheck, M., Adjei-Nsiah, S., & Descheemaeker, K. (2016). Characterising the diversity of smallholder farming systems and their constraints and opportunities for innovation : A case study from the Northern Region , Ghana. *NJAS - Wageningen Journal of Life Sciences*, *78*, 153–166. https://doi.org/10.1016/j.njas.2016.04.003

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). **lmerTest** Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, *82*(13). https://doi.org/10.18637/jss.v082.i13

Lunze, L., Abang, M., Buruchara, R., Ugen, M., Nabahungu, N. L., Rachier, G. O., … Rao, I. (2012). Integrated Soil Fertility Management in Bean-Based Cropping Systems of Eastern, Central and Southern Africa. In J. K. Whalen (Ed.), *Soil Fertility Improvement and Integrated Nutrient Management - A Global Perspective* (pp. 239–272). Rijeka, Croatia: InTech. https://doi.org/10.5772/29151

Luo, D., Ganesh, S., & Koolaard, J. (2018). predictmeans: Calculate Predicted Means for Linear Models. R package version 1.0.1. Retrieved from https://cran.r-project.org/package=predictmeans

Makoi, J. H. J. R., & Ndakidemi, P. A. (2008). Selected chemical properties of soil in the traditional irrigation schemes of the Mbulu district, Tanzania. *African Journal of Agricultural Research*, *3*(5), 348–356. Retrieved from http://www.academicjournals.org/ajar/PDF/pdf 2008/May/Joachim et al.pdf

Maman, G., Idriss, S., & Wortmann, C. (2018). Crop Yield Response to Fertilizer Relative to Soil Properties in Sub-Saharan Africa. *Soil Science Society of America Journal*, *82*(4), 862. https://doi.org/10.2136/sssaj2018.02.0066

Mbaga-Semgalawe, Z., & Folmer, H. (2000). Household adoption behaviour of improved soil conservation: The case of the North Pare and West Usambara Mountains of Tanzania. *Land Use Policy*, *17*(4), 321–336. https://doi.org/10.1016/S0264-8377(00)00033-8

McCarty, G. W., Reeves, J. B., Reeves, V. B., Follett, R. F., & Kimble, J. M. (2002). Mid-Infrared and Near-Infrared Diffuse Reflectance Spectroscopy for Soil Carbon Measurement. *Soil Science Society of America Journal*, *66*(2), 640–646. https://doi.org/10.2136/sssaj2002.6400}

Mevik, B.-H., & Wehrens, R. (2007). The pls Package: Principal Component and Partial Least Squares Regression in R. *Journal of Statistical Software*, *18*(2). https://doi.org/10.18637/jss.v018.i02

Mevik, B.-H., Wehrens, R., & Liland, K. H. (2016). pls: Partial Least Squares and Principal Component Regression. R package version 2.6-0. Retrieved from https://cran.r-project.org/package=pls

Mlingano ARI. (2006). *Soils of Tanzania and their Potential for Agriculture Development*. Tanga, Tanzania. Retrieved from http://www.ovice.or.kr/filebank/construction/OVUS16039407/OVUS16039407.pdf

Mowo, J. G., Janssen, B. H., Oenema, O., German, L. A., Mrema, J. P., & Shemdoe, R. S. (2006). Soil fertility evaluation and management by smallholder farmer communities in northern Tanzania. *Agriculture, Ecosystems and Environment*, *116*(1–2), 47–59. https://doi.org/10.1016/j.agee.2006.03.021

Ndakidemi, P. A., & Semoka, J. M. R. (2006). Soil fertility survey in Western Usambara Mountains, northern Tanzania. *Pedosphere*, *16*(2), 237–244. https://doi.org/10.1016/S1002-0160(06)60049-0

Nocita, M., Stevens, A., van Wesemael, B., Aitkenhead, M., Bachmann, M., Barthès, B., … Wetterlind, J. (2015). Soil Spectroscopy: An Alternative to Wet Chemistry for Soil Monitoring. *Advances in Agronomy*, *132*, 139–159. https://doi.org/10.1016/bs.agron.2015.02.002

O'Hara, G. W. (2001). Nutritional constraints on root nodule bacteria affecting symbiotic nitrogen fixation: a review. *Australian Journal of Experimental Agriculture*, *41*, 417–433. https://doi.org/https://doi.org/10.1071/EA00087

O'Rourke, S. M., & Holden, N. M. (2011). Optical sensing and chemometric analysis of soil organic carbon - a cost effective alternative to conventional laboratory methods? *Soil Use and Management*, *27*(2), 143–155. https://doi.org/10.1111/j.1475-2743.2011.00337.x

Pearson, K. F. R. . (1901). On lines and planes of closest fit to systems of points in space. *The London, Edinburgh and Dublin Philosophical Magazine and Journal of Science*, *2*(11), 559–572. https://doi.org/10.1080/14786440109462720

Reeves, 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? *Geoderma*, *158*(1–2), 3–14. https://doi.org/10.1016/j.geoderma.2009.04.005

Reichert, J. M., Rodrigues, M. F., Awe, G. O., Riquelme, U. F. B., Kaiser, D. R., & Reinert, D. J. (2015). Common bean in highly variable weather conditions, on sandy soils, and food security in a subtropical environment. *Food and Energy Security*, *4*(3), 219–237. https://doi.org/10.1002/FES3.65

Ronner, E., Descheemaeker, K., Almekinders, C. J. M., Ebanyat, P., & Giller, K. E. (2018). Farmers' use and adaptation of improved climbing bean production practices in the highlands of Uganda. *Agriculture, Ecosystems and Environment*, *261*, 186–200. https://doi.org/10.1016/j.agee.2017.09.004

Ronner, E., Franke, A. C., Vanlauwe, B., Dianda, M., Edeh, E., Ukem, B., … Giller, K. E. (2016). Understanding variability in soybean yield and response to P-fertilizer and rhizobium inoculants on farmers' fields in northern Nigeria. *Field Crops Research*, *186*, 133–145. https://doi.org/10.1016/j.fcr.2015.10.023

Saddiq, A., Ibrahim, A. M., Jada, M. Y., Tahir, A. M., & Umar, I. (2017). Soil Fertility Management, a Tool for Sustainable Disease and Weed Control in Sub-Saharan Africa: A Review. *Recent Research in Science and Technology*, *9*, 18–24. https://doi.org/10.25081/rrst.2017.9.3358

Sanchez, G. (2013). *PLS path modeling with R*. Berkeley, CA: Trowchez Editions. Retrieved from http://gastonsanchez.com/PLS_Path_Modeling_with_R.pdf

Shepherd, K. D., & Walsh, M. G. (2002). Development of Reflectance Spectral Libraries for Characterization of Soil Properties. *Soil Science Society of America Journal*, *66*(3), 988–998. https://doi.org/10.2136/sssaj2002.9880

Shmueli, G. (2010). To Explain or To Predict? *Statistical Science*, *25*(3), 289–310. https://doi.org/10.1214/10-STS330

Sila, A., Hengl, T., & Terhoeven-Urselmans, T. (2014). soil.spec: Soil Spectroscopy Tools and Reference Models. R package version 2.1.4. Retrieved from https://cran.r-project.org/package=soil.spec

Sila, A. M., Shepherd, K. D., & Pokhariyal, G. P. (2016). Evaluating the utility of mid-infrared spectral subspaces for predicting soil properties. *Chemometrics and Intelligent Laboratory Systems*, *153*, 92–105. https://doi.org/10.1016/j.chemolab.2016.02.013

Smithson, J. B., Edje, O. T., & Giller, K. E. (1993). Diagnosis and correction of soil nutrient problems of common bean (Phaseolus vulgaris) in the Usambara Mountains of Tanzania. *Journal of Agricultural Science*, *120*(2), 233–240. https://doi.org/10.1017/S0021859600074281

Soriano-Disla, J. M., Janik, L. J., Viscarra Rossel, R. A., MacDonald, L. M., & McLaughlin, M. J. (2014). The performance of visible, near-, and mid-infrared reflectance spectroscopy for prediction of soil physical, chemical, and biological properties. *Applied Spectroscopy Reviews*, *49*(2), 139–186. https://doi.org/10.1080/05704928.2013.811081

Thuijsman, E., Ronner, E., & van Heerwaarden, J. (2017). *Tailoring and adaptation in N2Africa demonstration trials, www.N2Africa.org*.

Thuijsman, E. S. (2017). *Light and nutrient capture by common bean (Phaseolus vulgaris L.) and maize (Zea mays L.) in the Northern Highlands of Tanzania*. Wageningen University and Research.

Tittonell, P., & Giller, K. E. (2013). When yield gaps are poverty traps: The paradigm of ecological intensification in African smallholder agriculture. *Field Crops Research*, *143*, 76–90. https://doi.org/10.1016/j.fcr.2012.10.007

Tittonell, P., Muriuki, A., Shepherd, K. D., Mugendi, D., Kaizzi, K. C., Okeyo, J., … Vanlauwe, B. (2010). The diversity of rural livelihoods and their influence on soil fertility in agricultural systems of East Africa – A typology of smallholder farms. *Agricultural Systems*, *103*(2), 83–97. https://doi.org/10.1016/j.agsy.2009.10.001

Tittonell, P., Shepherd, K. D., Vanlauwe, B., & Giller, K. E. (2008). Unravelling the effects of soil and crop management on maize productivity in smallholder agricultural systems of western Kenya - An application of classification and regression tree analysis. *Agriculture, Ecosystems and Environment*, *123*(1–3), 137–150. https://doi.org/10.1016/j.agee.2007.05.005

Tittonell, P., Vanlauwe, B., de Ridder, N., & Giller, K. E. (2007). Heterogeneity of crop productivity and resource use efficiency within smallholder Kenyan farms: Soil fertility gradients or management intensity gradients? *Agricultural Systems*, *94*(2), 376–390. https://doi.org/10.1016/j.agsy.2006.10.012

Tittonell, P., Vanlauwe, B., Leffelaar, P. A., Rowe, E. C., & Giller, K. E. (2005). Exploring diversity in soil fertility management of smallholder farms in western Kenya: I. Heterogeneity at region and farm scale. *Agriculture, Ecosystems and Environment*, *110*(3–4), 149–165. https://doi.org/10.1016/j.agee.2005.04.001

Towett, E. K., Shepherd, K. D., Sila, A., Aynekulu, E., & Cadisch, G. (2015). Mid-Infrared and Total X-Ray Fluorescence Spectroscopy Complementarity for Assessment of Soil Properties. *Soil Science Society of America Journal*, *79*(5), 1375–1385. https://doi.org/10.2136/sssaj2014.11.0458

Unesco Institute of Statistics. (2018). United Republic of Tanzania: Education and literacy. Retrieved September 15, 2018, from http://uis.unesco.org/en/country/tz?theme=education-and-literacy

United Nations Development Program. (n.d.). *Reducing Land Degradation on the Highlands of Kilimanjaro Region*. Retrieved from http://www.tz.undp.org/content/dam/tanzania/Sustainable Land Management in Kilimanjaro.pdf

Vågen, T., Shepherd, K. ., Walsh, M. ., Winowiecki, L., Tamene Desta, L., & Tondoh, J. E. (2010). *AfSIS Technical Specifications - Soil Health Surveillance*. Retrieved from http://www.worldagroforestry.org/sites/default/files/afsisSoilHealthTechSpecs_v1_smaller.pdf

Vanlauwe, B., & Giller, K. E. (2006). Popular myths around soil fertility management in sub-Saharan Africa.

Agriculture, Ecosystems and Environment, *116*(1–2), 34–46. https://doi.org/10.1016/j.agee.2006.03.016

Vanlauwe, B., Tittonell, P., & Mukalama, J. (2006). Within-farm soil fertility gradients affect response of maize to fertiliser application in western Kenya. *Nutrient Cycling in Agroecosystems*, *76*(2–3), 171–182. https://doi.org/10.1007/s10705-005-8314-1

Vice President's Office: Division of Environment. (2007). *The United Republic of Tanzania: National Adaptation Programme of Action (NAPA)*. https://doi.org/10.1007/BF02233368

Viscarra Rossel, R. A., Walvoort, D. J. J., McBratney, A. B., Janik, L. J., & Skjemstad, J. O. (2006). Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties. *Geoderma*, *131*(1–2), 59–75. https://doi.org/10.1016/j.geoderma.2005.03.007

Wetterlind, J., Stenberg, B., & Viscarra Rossel, R. A. (2013). Soil analysis using visible and near infrared spectroscopy. In F. J. M. Maathuis (Ed.), *Plant Mineral Nutrients: Methods and Protocols* (pp. 95–107). New York: Humana Press, Springer.

Wickama, J. M., & Mowo, J. G. (2001). *Using local resources to improve soil fertility in Tanzania*. *Managing Africa's Soils*. Retrieved from http://pubs.iied.org/pdfs/9041IIED.pdf

Wortmann, C. S., Kirkby, R. A., Eledu, C. A., & Allen, D. J. (1998). *Atlas of Common Bean Production in Africa*. Cali, Colombia: CIAT.

8. APPENDICES

Appendix I – The Chagga homegarden

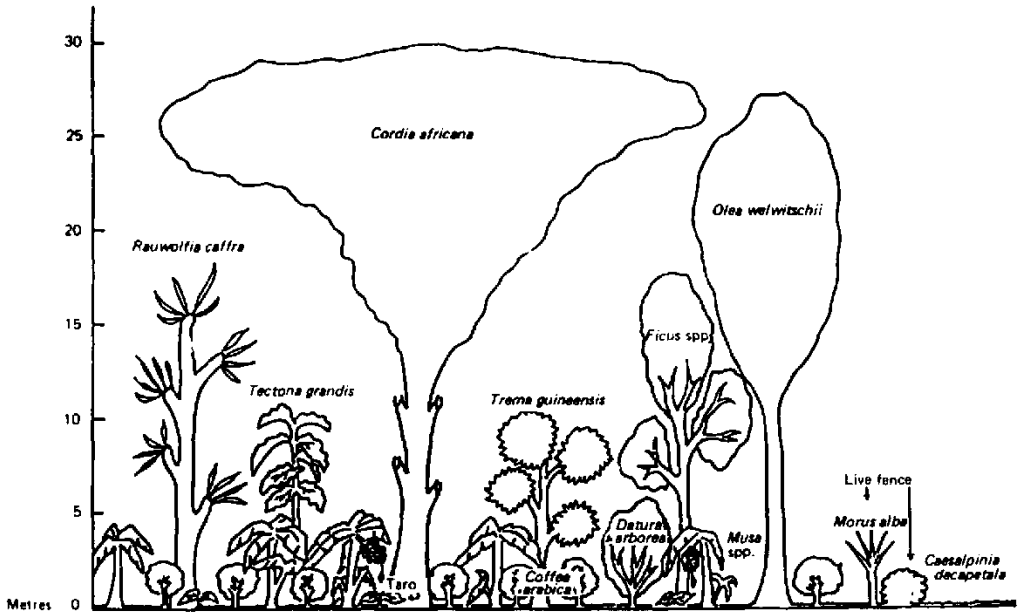


Figure 8.1. An example of the structure of a Chagga homegarden (Fernandes et al., 1984).

Appendix II – Soil properties and farm types

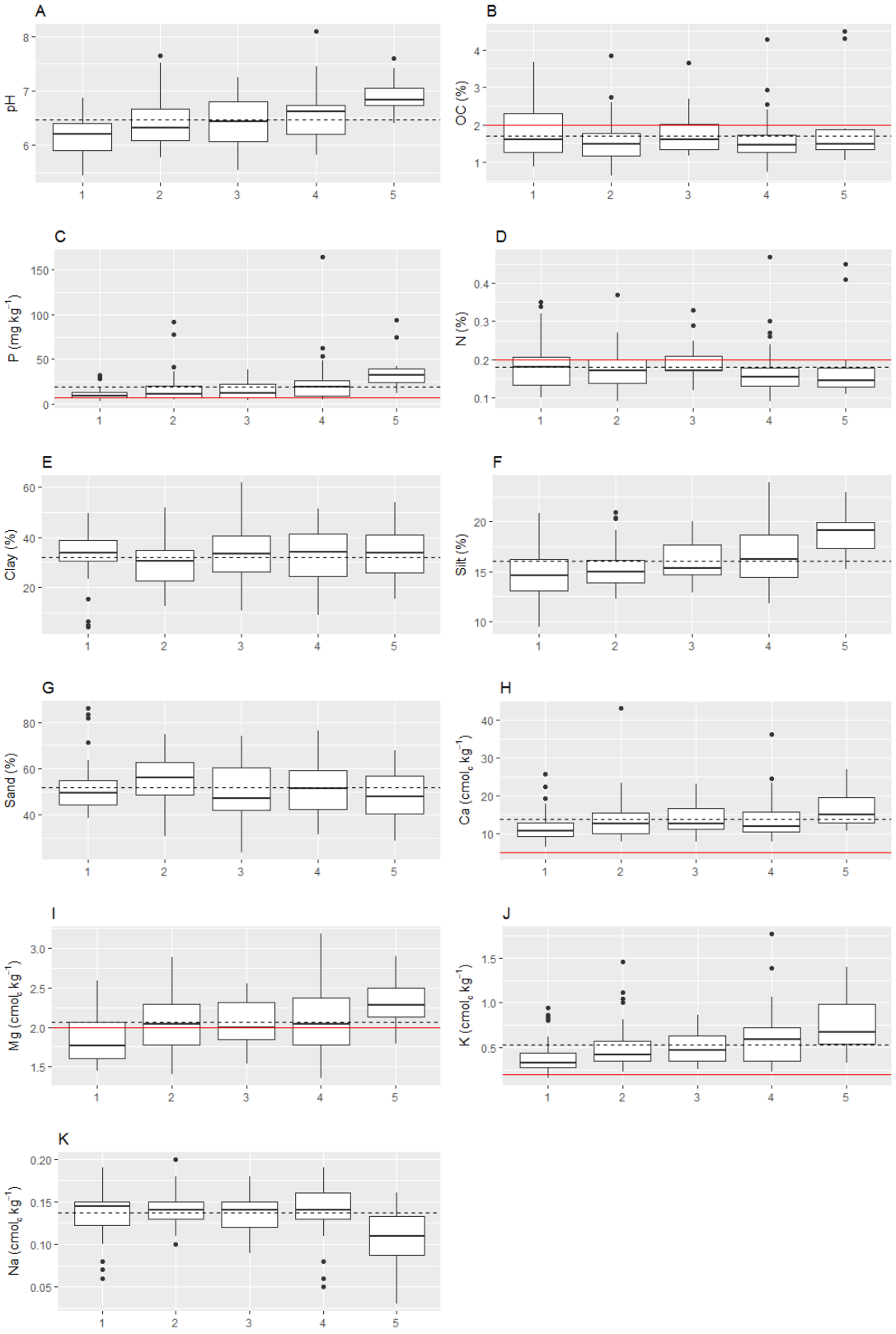


Figure 8.2. Boxplots of the spectrally-predicted soil properties for each farm type, when location is not accounted for. The dashed line shows the mean of the population and the red solid line shows the deficiency level of some soil attributes. There are no significant differences between farm types for OC, N, clay and sand. N = 145.

Appendix III – Linear models linking wet-chemistry-measured soil properties and spectrally-predicted soil properties with control yields and absolute responses

a. Wet-chemistry-measured soil properties

Table 8.1. Correlation matrix of wet-chemistry-measured soil properties and control yields and absolute responses, n = 38.

	pH	%OC	P	%N	%Clay	%Silt	%Sand	Ca	Mg	K	Na	C. yield
%OC	-0.24											
P	0.67	-0.01										
%N	-0.23	**0.94**	-0.03									
%Clay	-0.2	-0.07	-0.12	-0.19								
%Silt	0.54	0.08	0.59	0.1	-0.17							
%Sand	-0.18	0.01	-0.28	0.1	-0.76	-0.51						
Ca	0.61	0.24	0.6	0.2	-0.01	0.27	-0.17					
Mg	0.53	-0.09	0.5	-0.15	0.25	0.22	-0.36	0.61				
K	0.7	-0.12	0.74	-0.16	0.01	0.52	-0.35	0.57	0.77			
Na	0.37	-0.18	0.1	-0.22	0.2	0.19	-0.31	0.33	0.5	0.5		
C. yield	-0.06	-0.25	-0.03	-0.38	0.53	-0.29	-0.27	0.08	0.35	0.1	0.04	
Response	0.1	-0.18	-0.2	0.22	0.1	-0.12	-0.01	-0.05	0.03	-0.18	0.03	0.24

C. yield = control yield

Table 8.2. Linear (mixed) models with soil properties measured with conventional wet chemistry soil analyses as explanatory variables for the square root transformed control yield. Only the variables that were retained by the backward selection variable method are indicated. The random factor used in the mixed model is in italic, and it is indicated whether the random factor was retained by the backward variable selection method. The r^2 and RMSE of the cross-validated model are indicated, with the r^2 and RMSE of the training dataset shown between brackets.

Model/variables retained by step function ***Response variable: transformed control yield***	**Pos (+) or neg (-) effect**	**P-value**	**Model parameters**	**Cross-validated model**	
Linear mixed model (n = 38)					
District	n.r.		Adj. r^2 = 0.3	r^2 = 0.35	(0.36)
log_N	-	0.02	p-value < 0.001	RMSE = 7.64	(7.26)
Clay	+	0.004			
Linear model in Moshi (n = 16)					
pH	+	0.18	Adj. r^2 = 0.67	r^2 = 0.67	(0.88)
log_OC	+	0.21	p-value = 0.03	RMSE = 10.83	(3.35)
log_P	-	0.17			
log_N	-	0.1			
Clay	+	0.05			
log_Ca	-	0.28			
log_Mg	+	0.03			

log_Na	-	0.07			
Linear model in Lushoto (n = 22)					
log_OC	+	0.06	Adj. r^2 = 0.28	r^2 = 0.42	(0.47)
log_N	-	0.01	p-value = 0.02	RMSE = 6.09	(5.2)

n.r.: not retained

Table 8.3. Linear (mixed) models with soil properties measured with conventional wet chemistry soil analyses as explanatory variables for the Tukey-transformed absolute response to the N2Africa treatment. Only the variables that were retained by the backward variable selection method are indicated. The random factor used in the mixed model is in italic, and it is indicated whether the random factor was retained by the backward variable selection method. The r^2 and RMSE of the cross-validated model are indicated, with the r^2 and RMSE of the training dataset shown between brackets.

Model/variables retained by *step* function ***Response variable: transformed absolute response***	**Pos (+) or neg (-) effect**	**P-value**	**Model parameters**	**Cross-validated model**	
Linear mixed model (n = 38)					
District	n.r.				
pH	+	0.03	Adj. r^2 = 0.1	r^2 = 0.16	(0.22)
log_K	-	0.03	p-value = 0.06	RMSE = 38.62	(31.32)
Linear model in Moshi (n = 16)					
log_K	-	0.02	Adj. r^2 = 0.26 p-value = 0.02	r^2 = 0.58 RMSE = 28.90	(0.51) (33.45)
Linear model in Lushoto (n = 22)					
pH	+	0.004	Adj. r^2 = 0.4	r^2 = 0.44	(0.64)
log_OC	+	0.02	p-value = 0.01	RMSE = 33.15	(22.13)
log_N	-	0.009			
log_K	-	0.07			

n.r.: not retained

b. Spectrally-predicted soil properties

Table 8.4. Correlation matrix of spectrally-predicted soil properties and control yields and absolute responses, n = 148.

	pH	%OC	P	%N	%Clay	%Silt	%Sand	Ca	Mg	K	Na	C.yield
%OC	-0.19											
P	**0.83**	-0.22										
%N	-0.16	**0.94**	-0.23									
%Clay	-0.4	0.26	-0.35	0.1								
%Silt	0.6	0.08	0.68	-0.07	0.09							
%Sand	0.24	-0.26	0.17	-0.08	-0.97	-0.32						
Ca	0.62	0.42	0.4	0.52	-0.26	0.29	0.18					
Mg	0.65	-0.08	0.63	-0.1	-0.18	0.65	0.02	0.41				
K	0.78	-0.33	0.87	-0.42	-0.19	0.84	-0.02	0.27	0.74			
Na	-0.07	-0.49	-0.09	-0.48	-0.14	-0.17	0.17	-0.16	-0.04	0.09		
C. yield	0.16	-0.16	0.1	-0.19	0.05	0.13	-0.08	-0.04	0.13	0.16	0.01	
Response	0.01	0.05	-0.01	0.08	0.04	-0.03	-0.03	0.01	-0.04	-0.07	-0.18	0.23

C. yield = control yield

Table 8.5. Linear (mixed) models with spectrally-predicted soil properties as explanatory variables for the square root transformed control yield. Only the variables that were retained by the backward variable selection method are indicated. The random factor used in the mixed model is in italic, and it is indicated whether the random factor was retained by the backward variable selection method. A linear mixed model was also done with only the 38 samples that also have wet chemistry measured soil properties. The r^2 and RMSE of the cross-validated model are indicated, with the r^2 and RMSE of the training dataset shown between brackets.

Model/variables retained by *step* function ***Response variable: transformed control yield***	**Pos (+) or neg (-) effect**	**P-value**	**Model parameters**	**Cross-validated model**	
Linear mixed model (n = 148)					
District	n.r.	/	/	/	/
Linear mixed model with 38 samples					
District	n.r.	/	/	/	/
Linear model in Moshi (n = 39)					
log_N	-	0.04	Adj. r^2 = 0.14	r^2 = 0.21	(0.26)
log_Ca	+	0.02	p-value = 0.04	RMSE = 9.25	(8.62)
log_K	-	0.006			
Linear model in Lushoto (n = 109)					
log_OC	-	0.02	Adj. r^2 = 0.05	r^2 = 0.07	(0.13)
Silt	+	0.17	p-value = 0.05	RMSE = 8.01	(7.29)

log_Mg	+	0.04
log_K	-	0.08

n.r.: not retained

Table 8.6. Linear (mixed) models with spectrally-predicted soil properties as explanatory variables for the Tukey-transformed absolute response to the N2Africa treatment. Only the variables that were retained by the backward variable selection method are indicated. The random factor used in the mixed model is in italic, and it is indicated whether the random factor was retained by the backward variable selection method. A linear mixed model was also done with only the 38 samples that also have wet chemistry measured soil properties. The r^2 and RMSE of the cross-validated model are indicated, with the r^2 and RMSE of the training dataset shown between brackets.

Model/variables retained by *step* function ***Response variable: transformed absolute response***	**Pos (+) or neg (-) effect**	**P-value**	**Model parameters**	**Cross-validated model**	
Linear mixed model (n = 148)					
District	n.r.		Adj. r^2 = 0.02	r^2 = 0.05	(0.06)
Na	-	0.04	p-value = 0.04	RMSE = 37.78	(36.36)
Linear mixed model with 38 samples					
District	n.r.	/	/	/	/
Linear model in Moshi (n = 39)					
pH	+	0.006	Adj. r^2 = 0.24	r^2 = 0.25	(0.33)
log_K	-	<0.001	p-value = 0.005	RMSE = 34.89	(29.49)
Na	+	0.05			
Linear model in Lushoto (n = 109)			Adj. r^2 = 0.04	r^2 = 0.08	(0.08)
Na	-	0.02	p-value = 0.02	RMSE = 37.35	(36.84)

n.r.: not retained

Appendix IV – Random forest models

a. Assessment of the random forest predictions across districts and trial years

Table 8.7. Assessment of the performance of the best model from Table 14 when predictions for control yields across districts and years are done with 145 samples with spectrally-predicted soil properties. The cross-validation r^2 and RMSE are the average of the squared Pearson correlation and RMSE between the predicted and measured values. The predicted values are obtained from performing the model a hundred times on a different random subset each time. The random subset has the size of the smallest group of observations (i.e. between areas: 39 observations; between years: 49 observations). The values between brackets are the in-bag predictions, averaged over a hundred repetitions of the model on a random subset with the size of the smallest group of observations when necessary (i.e. in Lushoto and in 2016).

Predictor variables for *control yields*	**% variance explained**		**OOB predictions in the district/year**	**Cross-validation in other district/year**		**Two first most important variables**
Management and remote sensing variables in Moshi (n = 39, p = 14)	65.05%	r^2 RMSE	0.68 291.68	0.08 366.51	(0.93) (157.51)	Nbr_weeding_own, input_own
Management and remote sensing variables in Lushoto (n = 106, p = 14)	44.18%	r^2 RMSE	0.44 268.53	0.2 478.22	(0.83) (187.75)	Input_own, intercrop_own
Management and remote sensing variables in 2016 (n = 96, p = 14)	32.93%	r^2 RMSE	0.32 236.75	0.06 672.21	(0.83) (152.06)	avgTemp, totPrec
Management and remote sensing variables in 2017 (n = 49, p = 14)	36.72%	r^2 RMSE	0.36 311.07	0.03 640.65	(0.84) (171.65)	avgTemp, avg_EVI

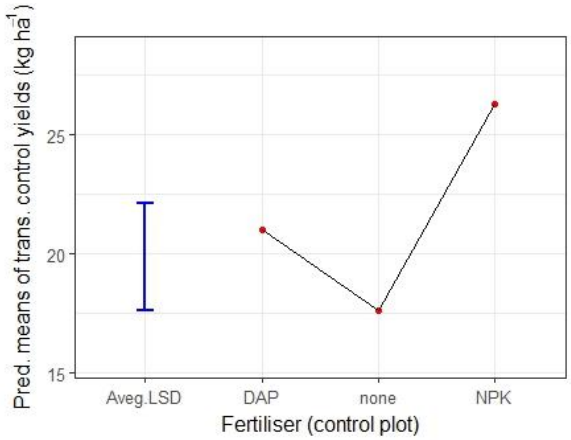
Table 8.8. Assessment of the performance of the best model from Table 15 when predictions for absolute response across districts and years are done with 145 samples with spectrally-predicted soil properties. The cross-validation r^2 and RMSE are the average of the squared Pearson correlation and RMSE between the predicted and measured values. The predicted values are obtained from performing the model a hundred times on a different random subset each time. The random subset has the size of the smallest group of observations (i.e. between areas: 39 observations; between years: 49 observations). The values between brackets are the in-bag predictions, averaged over a hundred repetitions of the model on a random subset with the size of the smallest group of observations when necessary (i.e. in Lushoto and in 2016).

Predictor variables for *absolute response*	**% variance explained**		**OOB predictions in the district/year**	**Cross-validation in other district/year**		**Two first most important variables**
Management variables in Moshi (n = 39, p = 9)	43.33%	r^2 RMSE	0.42 321.9	0.29 361.09	(0.81) (195.73)	Input_n2a, nbr_weeding_n2a
Management variables in Lushoto (n = 106, p = 9)	62.25%	r^2 RMSE	0.62 254.21	0.32 407.02	(0.82) (179.36)	Pack_variety, yield_own
Management variables in 2016 (n = 96, p = 9)	38.1%	r^2 RMSE	0.37 202.59	0.05 736.46	(0.74) (147.73)	Yield_own, pack_variety
Management variables in 2017 (n = 49, p = 9)	6.5%	r^2 RMSE	0.08 343.09	0.12 822	(0.76) (206.35)	Nbr_weeding_own, rel_planting_date_own

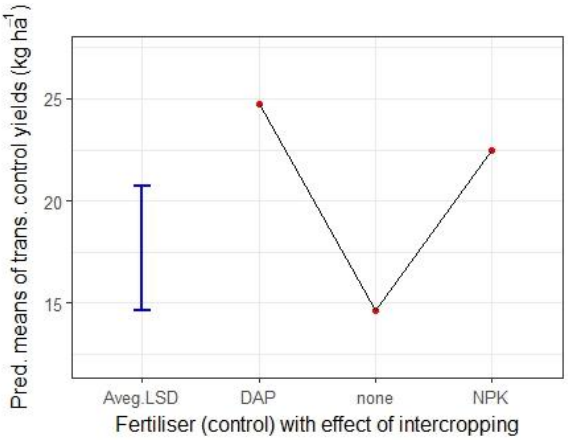
b. Regression plots of the most important variables from the random forest models

Control yields

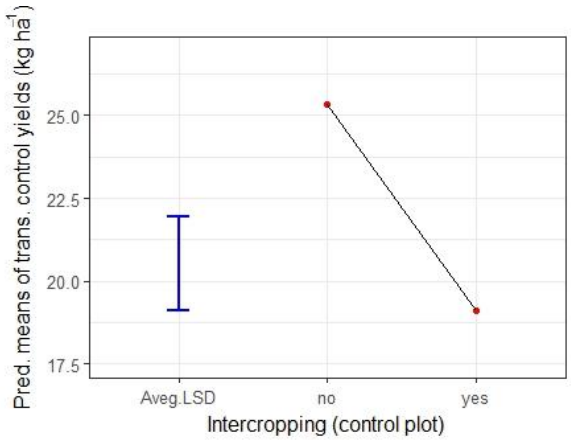
A.1. Fertiliser in control plot against means of control yields predicted with a linear mixed model.
Relative importance: 1



A.2. Fertiliser in control plot against means of control yields predicted with a linear mixed model that includes the effect of intercropping.



B.1. Intercropping of control plot against means of control yields predicted with a linear mixed model.
Relative importance: 0.71



B.2. Intercropping of control plot against means of control yields predicted with a linear mixed model that includes the effect of fertiliser.

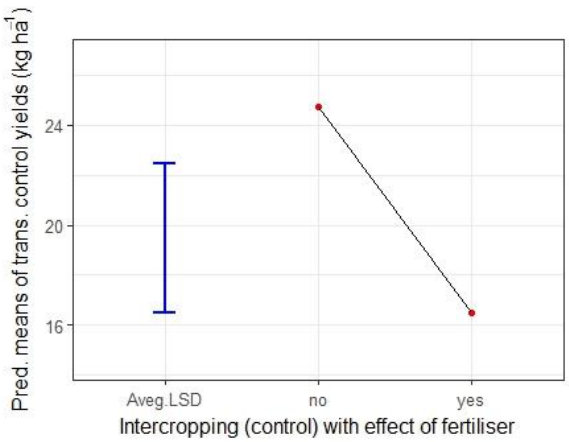
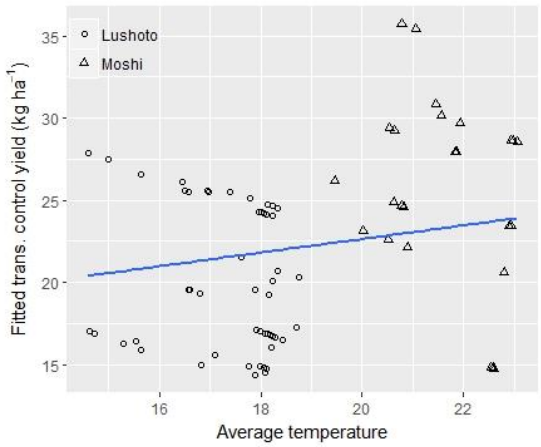
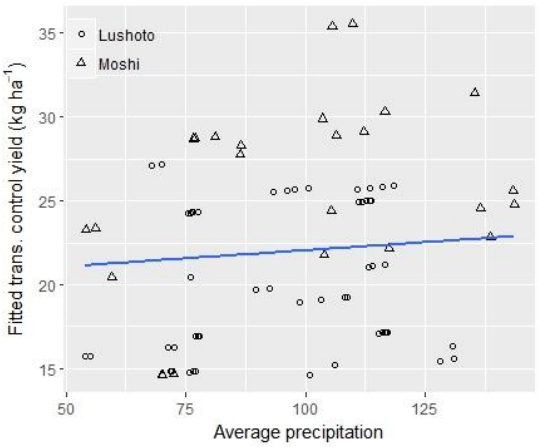


Figure 8.3. Partial dependence plots of the six variables with most importance in bush beans control yields (kg ha^{-1}) predicted by the best random forest model with 145 samples. The relative importance of the variable, given by the random forest model, is indicated. The linear mixed models were performed with location (i.e. village nested within district) as a random effect. The relative importance obtained from the RF model are indicative, but it should be noted that they could not be made reproducible. Hence, re-running the code would give slightly different values, although it would of course stay in the same range of relative importance.

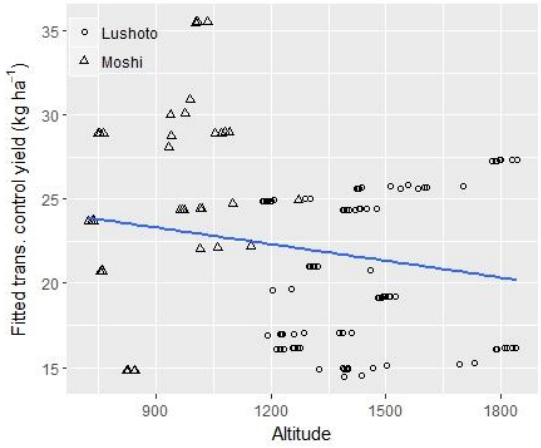
C. Regression of control yields on average temperature with a linear mixed model.
Relative importance: 0.37



D. Regression of control yields on average precipitation with a linear mixed model.
Relative importance: 0.35



E. Regression of control yields on altitude with a linear mixed model.
Relative importance: 0.26



F. Weeding of control plot against means of control yields predicted with a linear mixed model.
Relative importance: 0.23

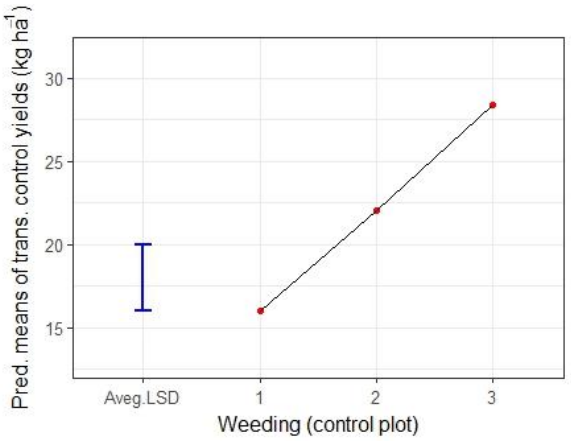
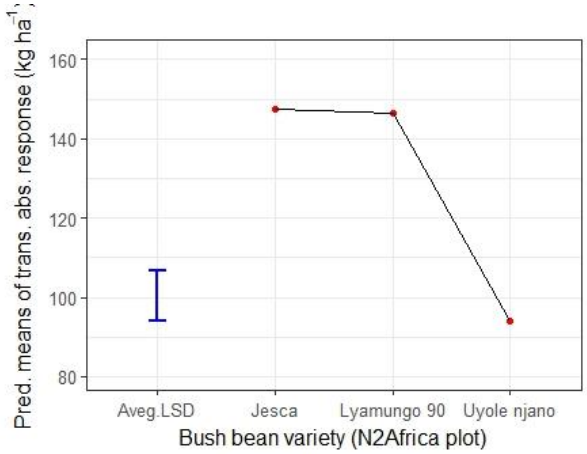


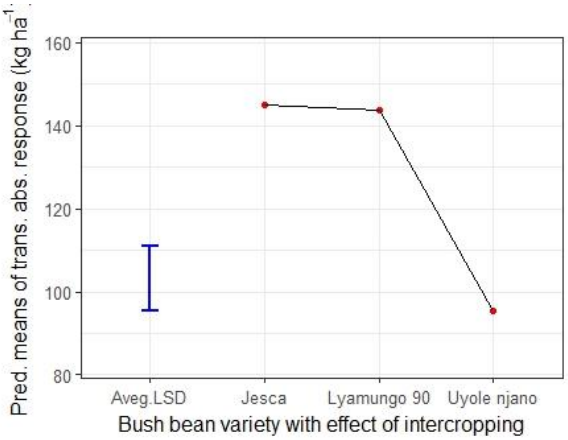
Figure 8.3. (Continued)

Absolute responses

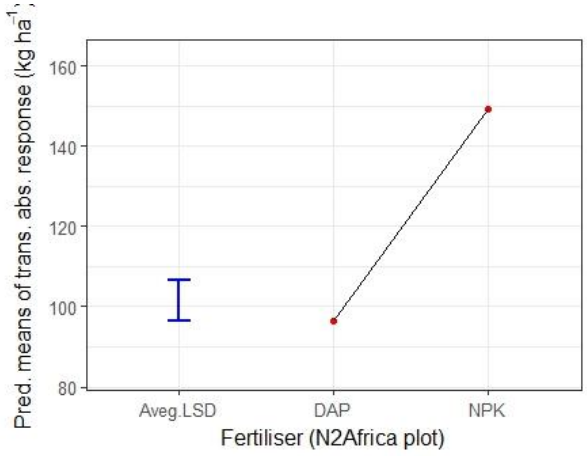
A.1. Variety used in the control plot against means of control yields predicted with a linear mixed model.
Relative importance: 1



A.2. Variety used in the control plot against means of control yields predicted with a linear mixed model that includes the effect of intercropping.



B.2. Fertiliser used in the N2Africa plot against means of absolute response predicted with a linear mixed model.
Relative importance: 0.75



B.2. Fertiliser used in the N2Africa plot against means of absolute response predicted with a linear mixed model that takes into account the effect of intercropping.

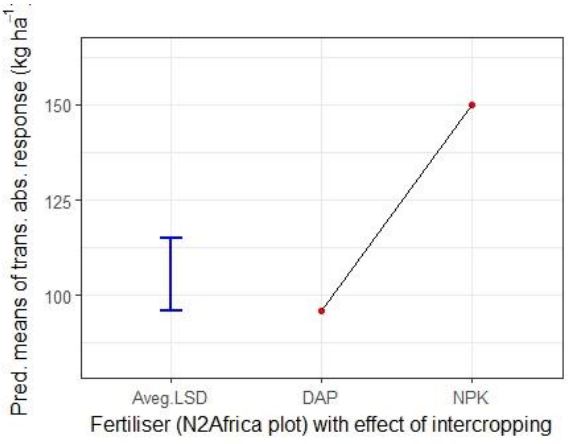
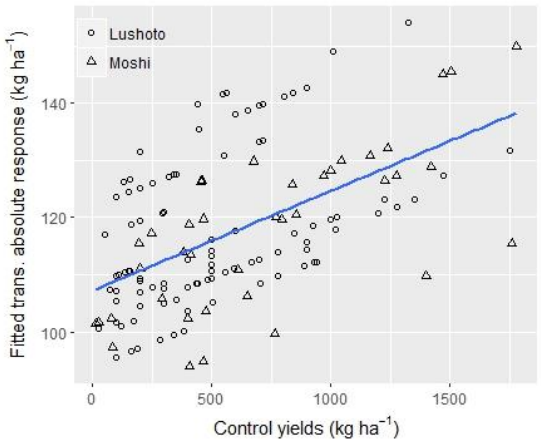
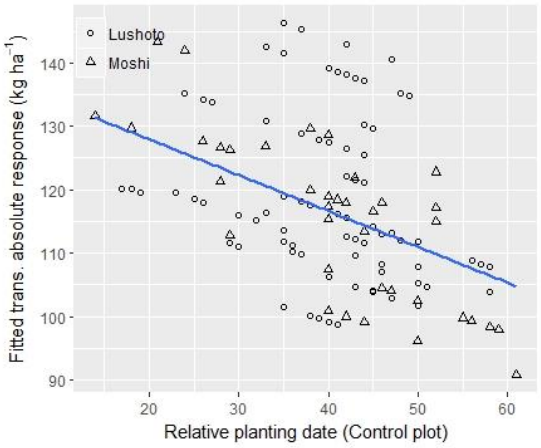


Figure 8.4. Partial dependence plots of the six variables with most importance in bush beans absolute responses (kg ha^{-1}) to N2Africa treatments predicted by the best random forest model with 145 samples. The relative importance of the variable, given by the random forest model, is indicated. The linear mixed models were performed with location (i.e. village nested within district) as a random effect. The relative importance obtained from the RF model are indicative, but it should be noted that they could not be made reproducible. Hence, re-running the code would give slightly different values, although it would of course stay in the same range of relative importance.

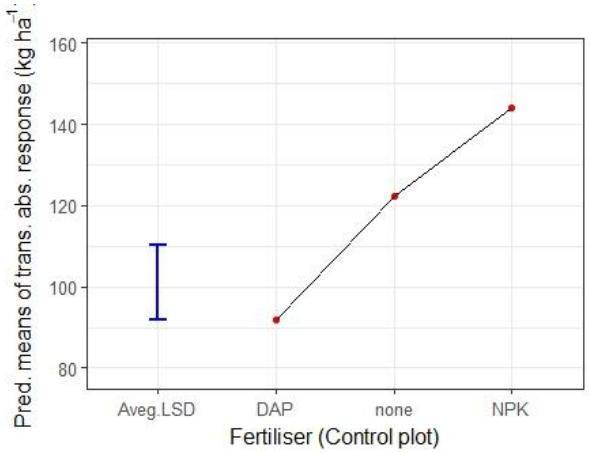
C. Regression of absolute response on the control yields with a linear mixed model.
Relative importance: 0.43



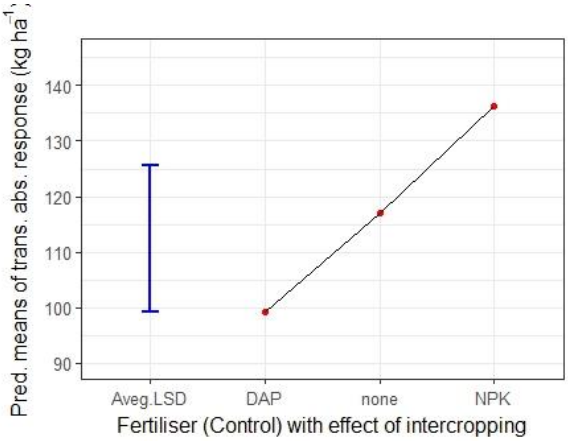
D. Regression of absolute response on the relative planting date of the control plot with a linear mixed model.
Relative importance: 0.24



E.1. Fertiliser used in the control plot against means of absolute response predicted with a linear mixed model.
Relative importance: 0.18



E.2. Fertiliser used in the control plot against means of absolute response predicted with a linear mixed model that includes the effect of intercropping.



F. Regression of absolute response on the relative planting date of the N2Africa plot with a linear mixed model.
Relative importance: 0.12

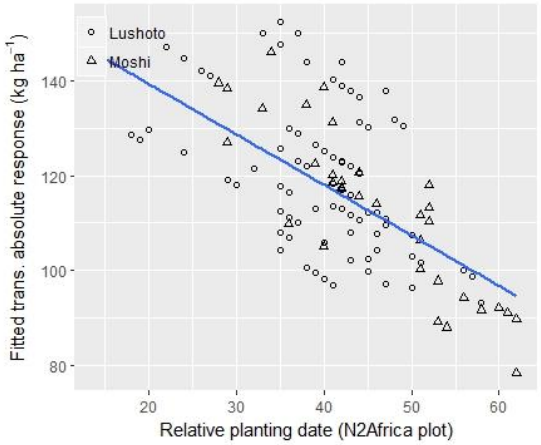


Figure 8.4. (Continued)

c. Scatter plot of the predictions by the best random forest model against observed values

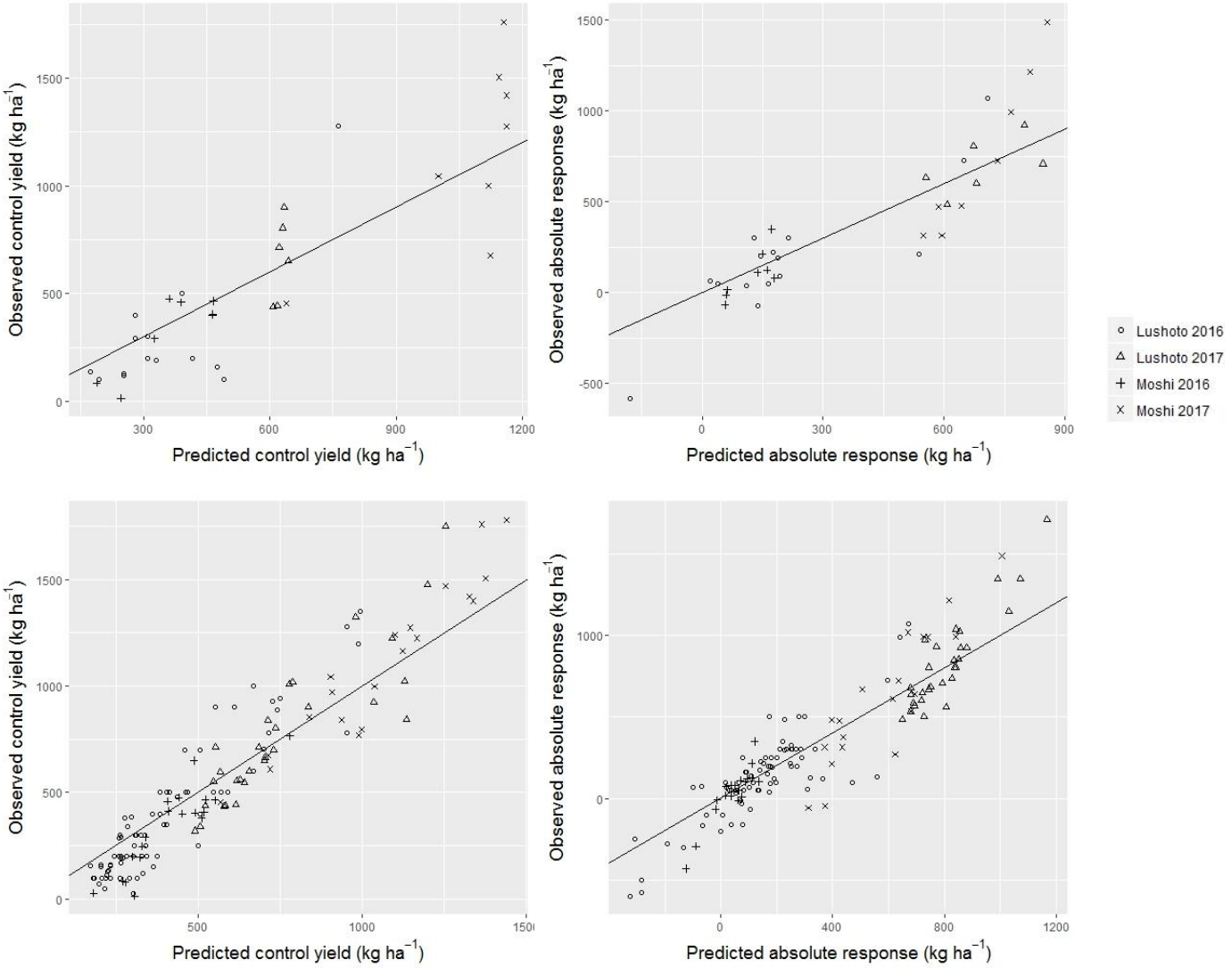


Figure 8.5. Scatter plots of the predicted against observed control yields (left column) and absolute response (right column), separated by district and trial year. The values are predicted with the best random forest model from Table 13 (left column) and Table 14 (right column). The top row contains the 37 samples with wet-chemistry-measured soil properties, and the bottom row contains the 145 samples with spectrally-predicted soil properties. The solid line represents the 1:1 line.

Appendix V – Form filled during the fieldwork of 2018

What is ID of the farm (as marked on the original dataset)?			
Which district?			
Which village?			
What is the size of the farm (all land owned by the farmer)?			Unit of the farm size:
What was the input (e.g. mineral or organic fertiliser) used on the own farmer's plot during the trial season?			
What was the input (e.g. mineral or organic fertiliser) used on the N2Africa plot during the trial season?			
Was the control plot intercropped during the trial season?			0 Yes 0 No 0 Farmer doesn't remember exactly
Was the N2Africa plot intercropped during the trial season?			0 Yes 0 No 0 Farmer doesn't remember exactly
Are there flooding/waterlogging problems in the field?			0 Yes 0 No If yes: How often are there waterlogging problems?
Are there drought problems in the field?			0 Yes 0 No If yes: How often are there drought problems in the field?
If trial was in 2017:	What proportion of the income comes from farming?		0 All the income comes from the farm 0 Most of the income comes from the farm 0 About half of the income comes from the farm 0 Most of the income comes from outside the farm 0 All of the income comes from outside the farm
	Rank of income sources:	Rank the most important sources of income in the household (i.e. 1 is the most important source, 2 is the second most, …) Crops: __ Livestock: __ Trade / business: __ Remittances: __ Salaried job: __ Pension: __ Casual labour off-farm: __ Casual labour in agriculture: __ Other: __ If other: specify other source of income:	
Were the control plot and the N2Africa plot located in the same field?			0 Yes 0 No 0 Not sure
Is the sampling done in the N2Africa plot or in the control plot?			0 N2Africa plot 0 Control plot 0 Other
GPS coordinates of the central sampling point:			
What is the position of the sample in the slope?			
Is the sampled field in a flat, moderate or steep slope?			
What is the soil depth (in cm)?			
Are there visible signs of poor drainage? (e.g. gleyification)			0 Yes 0 No 0 Not clear whether there are signs of poor drainage
Are there coarse fragments in the field (e.g. rocks)?			0 Yes 0 No

What was on the field at the moment of sampling? (e.g.: bare field, fallow, intercrop bean-maize, monocrop bean or maize,…)	
What kind of fertiliser was used on the sampled field this season, if any?	
Is there an auger restriction to the specified depth?	0 Yes 0 No
Scan the QR code of the sample	
Add any general comment on the farm:	