Digital soil mapping using uncertain soil observations to support agricultural intensification in West and Central Africa

Bertin Takoutsing

Propositions

1. Sustainable agricultural intensification requires increased investment in accurate soil information.
   (this thesis)

2. Quantification of prediction uncertainty is as important to end-users as making the prediction itself.
   (this thesis)

3. The practical value of research findings is contingent on their effective implementation to improve decisions.

4. Transforming the current land usage and management practices is imperative for any viable solution to agricultural intensification.

5. A map can always tell additional stories for a given dataset.

6. A PhD program is a set of requirements rather than a learning resource.

7. Equal rights and opportunities for men and women are not sufficient to ensure gender equality.

Propositions belonging to the thesis entitled:

Digital soil mapping using uncertain soil observations to support agricultural intensification in West and Central Africa

Bertin Takoutsing
Wageningen, 30 April 2024

Digital soil mapping using uncertain soil observations to support agricultural intensification in West and Central Africa

Bertin Takoutsing

Thesis committee

Promotors
Prof. Dr Gerard B.M. Heuvelink
Special Professor, Pedometrics and Digital Soil Mapping
Wageningen University & Research

Co-promotors
Dr Ermias Betemariam
Land Health Scientist
World Agroforestry (CIFOR-ICRAF), Nairobi, Kenya

Dr Jetse J. Stoorvogel
Associate Professor, Soil Geography and Landscape Group
Wageningen University & Research

Other members
Prof. Dr Wim de Vries, Wageningen University & Research
Dr Madlene Nussbaum, University of Utrecht
Prof. Dr David Rossiter, Cornell University, New York, United States of America
Dr Francis Silatsa, Wageningen University & Research

This research was conducted under the auspices of the C.T. de Wit Graduate School for Production Ecology and Resource Conservation (PE&RC).

Digital soil mapping using uncertain soil observations to support agricultural intensification in West and Central Africa

Bertin Takoutsing

Thesis
submitted in fulfilment of the requirements for the degree of doctor
at Wageningen University
by the authority of the Rector Magnificus,
Prof. Dr C. Kroeze,
in the presence of the
Thesis Committee appointed by the Academic Board
to be defended in public
on Tuesday 30 April 2024
at 11 a.m. in the Omnia Auditorium

Bertin Takoutsing
Digital soil mapping using uncertain soil observations to support agricultural intensification in West and Central Africa,
178 pages.

PhD thesis, Wageningen University, Wageningen, the Netherlands (2024)
With references, with summary in English

ISBN 978-94-6469-890-9
DOI https://doi.org/10.18174/653541

Table of contents

Chapter 1

General introduction

1.1. The role of soil information for sustainable agricultural intensification

The rapid population growth has become a consistent global issue for the agricultural sector to address the food security challenge. To meet the global food demand and achieve agricultural intensification, some major constraints limiting agricultural yields have to be overcome, including soil nutrient deficiencies, suboptimal management of resources, and limited availability of information on soil nutrients (Giller *et al.*, 2009). The importance of soil information in agriculture can hardly be overlooked as several global challenges such as land degradation affecting food production systems cannot be understood and subsequently addressed without up-to-date and accurate soil information (Vanino *et al.*, 2023). Agricultural intensification necessitates the sustainable management of soil, including strategies to prevent negative nutrient balances, minimize leaching through soil erosion, foster the accumulation of soil carbon, and the retention of thresholds of soil biological diversity. These practices are essential for maintaining crucial soil functions that contribute to the overall and productivity of agricultural ecosystems (Kopittke *et al.*, 2019).

The demand for soil information is in the increase to support relevant scientific research and evidence-based decision-making related to land management interventions and sustainable agricultural intensification (Sanchez *et al.*, 2009; Arrouays *et al.*, 2014). However, in several countries in Sub-Sahara Africa (SSA), access to soil information remained limited, with only few regions having undergone detailed soil mapping surveys (Brevik *et al.*, 2019). The limited access to soil information also limits the opportunities for smallholder farmers to improve soil health through appropriate and sustainable land management practices (e.g., targeted fertilizer application). The significance of soil in food production stands in contrasts to the inadequacy of available soil information (Pozza and Field, 2020), a challenge earlier referred to as the "soil data crisis" (McBratney *et al.*, 2006). The path to soil data collection and mapping is indeed plagued by many challenges including field accessibility and limited financial resources. Moreover, farming systems are characterised by high spatial variability in soil nutrients both within farming communities and at the level of individual farms (Tittonell *et al.*, 2015). The availability of soil nutrient supplies, fertilizer efficiency, and productivity vary significantly even within short distances. These variations arise from differences in in environmental factors such as topography, parent material, and climate, which have exerted influence on the formation of soil properties over time (Takoutsing *et al.*, 2018). Spatial heterogeneity in soil across different scales represent a challenge in providing comprehensive soil information. Moreover, this amplifies the cost associated with representative sampling, rendering large-scale sampling impractical due to the substantial time and cost expenditures required for laboratory analysis (Padarian *et al.*, 2019).

To overcome this issue, cost-effective methods to measure soil properties for large areas have been developed. Soil spectroscopy has increasingly been utilised as a cost-effective, environmental-friendly, non-destructive, reproducible, and repeatable soil analytical technique (Poppiel *et al.*, 2022; Viscarra Rossel *et al.*, 2022). Proponents have demonstrated that conventional laboratory methods can ultimately be supplemented by the less laborious and more rapid spectroscopic methods (Nocita *et al.*, 2015). In addition, global datasets such as World Soil Information Service (WoSIS) are also playing a major role by providing quality-assessed and standardised soil profile data useful for the application of digital soil mapping and other environmental assessments at broadscale levels (Batjes *et al.*, 2017). Pedotransfer functions (PTFs) have been developed to derive some of the properties from more easy-to-measure soil properties (Van Looy *et al.*, 2017; Zhao *et al.*, 2020; Perreault *et al.*, 2022). Nonetheless, the application of PTFs on larger scales still requires large high-quality and spatially distributed soil data. Digital soil mapping (DSM) has emerged over the last two decades to change

this trend; however, most initiatives have focused on national and continental scales not readily usable by small-scale farmers (Chen *et al.*, 2022a). Soil maps play a vital role in characterizing the spatial variability of soil conditions. By providing a comprehensive representation of the distribution and characteristics of soils across a given area, these maps enable a nuanced understanding of the distribution of soil properties. This detailed insight helps land and agricultural managers make informed decisions regarding the implementation of targeted management practices, as they can identify areas where specific interventions or adjustments are most suitable based on the site-specific soil attributes of the locations. Soil maps depict spatial variability using conventional or digital mapping techniques. Conventional methods involve field surveys, while digital mapping uses advanced technologies like remote sensing and GIS. The choice depends on scale, accuracy, and resources, highlighting the versatility of soil mapping for informed land management.

1.2. Soil mapping

Soil maps have been widely viewed as a potential solution to fill the soil information gaps in agriculture, resulting in continued efforts to improve the spatial resolution and accuracy of soil information across targeted landscapes (Brevik *et al.*, 2016). These maps are often required for specific purposes by farmers, policy makers, researchers, and extension agents to implement land management interventions. Soil spatial information can be produced either using conventional soil mapping, or digital soil mapping methods.

1.2.1. Conventional soil mapping

Conventional soil mapping (CSM) systems were dominant from the 1950s to the early 2000s (Heuvelink and Webster, 2001; Minasny and McBratney, 2016). Conventional soil maps typically take the form of paper maps and are often difficult to understand by non-experts and experts from other disciplines and were mostly created in the pre-computer era through intensive fieldwork. In the CSM approach, the landscape is partitioned into finite circumscribed regions (i.e., soil map units), where sharp boundaries delineate clear differences in soil types (Heuvelink and Webster, 2001; Kempen *et al.*, 2012). CSM uses empirical, expert-based models to delineate the location and extent of soil types in the areas of interest. These empirical models are often based on local geomorphology and vegetation patterns and calibrated by direct observations in the field. The typical ranges of soil properties encountered for each soil type are established based on representative soil profiles and expert knowledge. Main criticisms of CSM include irreproducibility, because the production processes used to derive soil mapping units is neither recorded nor documented; soil bodies that are strictly represented as discrete, homogeneous entities, which in many cases is unrealistic; lack of quantified measures of accuracy; and 'static' representation of soil variation (Goovaerts and Journel, 1995; Brevik and Hartemink, 2010).

However, despite the criticisms, there are also advantages of CSM over digital soil mapping (DSM) methods. As compared to DSM, which typically provides specific-purpose soil maps, CSM maps are general-purpose maps because they provide information on the three-dimensional spatial distribution of a wide range of soil properties that are inferred from representative soil profile descriptions associated with the map units (Kempen *et al.*, 2012). Similarly, complex soil forming processes that are difficult to quantify and represent by environmental explanatory variables in DSM, fins more feasible representation in CSM. Notably, DSM methods for soil type mapping are

limited by the number of soil types that can be handled (Brus *et al.*, 2008), while CSM is able to deal with a larger number of soil types.

1.2.2. Digital soil mapping

As the way to produce conventional soil maps is labour- and cost-intensive, and time-consuming, they are difficult to be updated. DSM (McBratney *et al.*, 2003) has been developed based on Jenny's theory of soil formation (Jenny, 1941) to provide a cost-effective way of producing soil information. DSM is defined as the 'computer-assisted production of digital maps of soil types and soil properties, by use of mathematical and statistical models that combine information from soil observations with information contained in explanatory environmental variables' (Minasny and McBratney, 2016; Arrouays *et al.*, 2017). Under the conceptual soil formation framework, soil properties can be predicted at unvisited positions by their relationships to environmental covariates, namely other soil information, climate, organisms, relief, parent materials, age, and spatial position (Minasny and McBratney, 2016). Despite the significant advances in GIS, remote sensing, geostatistics, machine learning, and high-performance computing capacity, soil information is still lacking in SSA because DSM techniques are mostly applied at national and continental levels, and not at local scale where they are often most needed (Chen *et al.*, 2022a).

As DSM needs soil observations to train statistical models, it is as much limited as CSM by the availability of existing soil information. However, DSM offers several notable advantages over CSM: i) it can provide higher resolution soil data over larger mapping extents; ii) it exploits a large range of spatial data on landscape variation provided by spatial data infrastructures; iii) it provides a better evaluation of the quality of the outputs and understanding of mapping uncertainties (e.g. prediction error variance for kriging, coefficient of determination for regression, prediction interval), which allows making a realistic use of the outputs; and iv) its outputs can be easily updated if new data are collected. Providing a better evaluation of the accuracy of the outputs and understanding of the associated uncertainties can help users to take rational decisions on whether a DSM map is good enough for its intended use. This can only be done if the prediction accuracy is quantified and communicated. Following the concept of DSM, large efforts, and developments in the generation of spatially distributed soil maps have been observed in recent years. Globally there has been a drive to fulfil the need for soil information, with projects such as the GlobalSoilMap (Arrouays *et al.*, 2020), the SoilGrids initiative (Poggio *et al.*, 2021) and continental and national soil maps which have been created for various soil parameters in SSA (Grundy *et al.*, 2015; Hengl *et al.*, 2015; Hengl *et al.*, 2021).

1.3. Uncertainty quantification in soil mapping

Soil maps generated using either CSM or DSM are not perfect, and the associated uncertainty is a result of the contribution of uncertainty from many sources. Uncertainty can arise from errors in soil measurement in the laboratory (van Leeuwen *et al.*, 2022), geolocation, digitization, data generalization, and interpolation (Arrouays *et al.*, 2014). Other sources may include modelling bias, parameterization errors, a too-small training sample size, and predictive modelling errors associated with modelling the relationship between the predictors and response variables (Minasny and McBratney, 2016; Schmidinger and Heuvelink, 2023).

All these inherent processes to generate soil maps lead to prediction errors, since soil spatial variability is a very complex phenomenon, and the mapping exercise can never be able to capture

this variability perfectly. But only the recognition of uncertainty is not enough. It is good practice in DSM to provide spatially explicit estimates of the uncertainty associated with the predictions and properly communicate it to the end-users. In many circumstances, especially within the decision-making process, it is just as important to quantify prediction uncertainty as it is to make the prediction itself, thus uncertainty maps are necessary (Nikou and Tziachris, 2022).

In recent studies, the term 'uncertainty quantification' is used to describe the process of quantifying and analysing the uncertainty in models, simulations, and data. The primary objective of uncertainty quantification is to assess the accuracy of predictions by providing a thorough examination of the potential uncertainties inherent in the underlying processes and information (Heuvelink, 2018; Wadoux, 2019; Nikou and Tziachris, 2022). Several methods have been proposed in the literature to estimate DSM prediction uncertainty. Cross-validation is one of the common approaches used to estimate the accuracy of a given model performance. The goal of cross-validation is to test the model's ability to predict new data that were not used in model calibration and prediction. The integration of new or 'unseen' data is crucial, in order to flag problems like overfitting or selection bias and to give an insight on how the model will generalize to an independent dataset. Validation measures are then computed by comparing the predictions with the put-aside observations for all observation locations (Heuvelink, 2018). While incorporating the relationships between soil properties and environmental covariates by means of various linear and non-linear regression models, regression kriging includes a kriging of the regression residuals. Kriging provides an estimate of the probability distribution of the true value of a particular soil property at any location. Quantile regression forests (Meinshausen, 2006) allows spatially explicit non-parametric (i.e., an empirical distribution function) estimates of model uncertainty. Similar to kriging, quantile regression forests also provide information of the full conditional distribution of the target soil property, not only about the conditional mean, but also the conditional probability distribution at each prediction location from which prediction intervals can be constructed. Therefore, the lower and upper limits of the 90% prediction interval can be identified by the 5th and 95th percentiles of the conditional distribution (Vaysse and Lagacherie, 2017). This method has been adopted as the standard for the GlobalSoilMap project for reporting uncertainties (Arrouays *et al.*, 2014).

Uncertainty quantification in soil mapping is important for many reasons: i) it help modellers to compare the performances of mapping approaches and determine the one that performs best, and hence to decide which approach to use in any particular case; ii) reporting and communicating uncertainties is crucial in maintaining credibility in the model and modellers community; iii) it enables end-users to know how accurate the results are, and in turn enables them to take into account the uncertainty when making decisions; iv) it is needed for uncertainty propagation analysis, a technique that helps modellers and users to understand how uncertainty propagates through environmental models and identify where the main sources of uncertainty reside (Brown and Heuvelink, 2005). This, in turn helps users to direct efforts toward the main sources of uncertainty, as well as taking rational decisions on how to reduce uncertainty about the model outputs; v) quantified uncertainty can be included in decision-making processes, as well as in risk analysis (Breure *et al.*, 2022b). Many users would like to avoid risks, but since it cannot be completely avoided, a best alternative is to be aware of the uncertainty of soil maps, so that decision-making can take this into account (Lark *et al.*, 2022); and 6) uncertainty analysis also involves acknowledging the model's limitations, which is an important step towards model interpretability (Wadoux *et al.*, 2020a).

1.4. Quantification of soil measurement errors

Soil properties are commonly determined by the field sampling, samples processing and analysis in laboratories following operational procedures. During the measurement process, several factors frequently contribute to measurement error, including the analyst, wet chemistry methodologies, varying measurement conditions (e.g., temperature and humidity), a variety of different sample preparation and subsampling methods, and the measurement instrument itself (Libohova *et al.*, 2019; van Leeuwen *et al.*, 2022). The term 'error' in soil measurements can be defined as the difference between the 'true' of a soil property and its measured value (Hibbert, 2007). Because of error, a measurement on a soil sample for a given property is considered to be only an approximation of the 'true' value. The 'true' value of a soil property must be known to quantify the error. However, we rarely have this knowledge and therefore we are uncertain about the 'true' value. The term 'uncertainty' on the other hand is defined as the range of estimated or predicted values within which the true value is expected to lie. One way to express uncertainty is therefore to represent the 'true' value by a probability distribution. As described in Section 1.3, this for example entails the computation of the 0.05 and 0.95 quantiles of the distribution to derive a 90% prediction interval as a measure of the uncertainty. The terms 'error' and 'uncertainty' are often used in the same context, with 'uncertainty' being the quantification of confidence in our knowledge of the true value of a specific soil property (Heuvelink *et al.*, 2007). Since in practice the error is not known, it is common to work with uncertainty.

One way to quantify measurement errors in soil observations is by repeated measurements. In this case the measurement error can be divided into a systematic and a random component. The systematic error will be the same for all repeated measurements, whereas the random error may vary between replicates. Therefore, repeated measurements can only detect the random error component, and not the systematic error component. The random error characterizes the precision of repeated measurements. Precision is defined as the closeness of agreement between independent test results that were obtained under stipulated conditions and includes terms such as repeatability and reproducibility (Ebentier *et al.*, 2013). Measurement error variances in soil observations analysed using conventional laboratory methods can be derived using repeated measurements on soil samples (Libohova *et al.*, 2019). Repeated measurements are required to quantify the laboratory error, as for instance done in costly laboratory proficiency testing programs (van Leeuwen *et al.*, 2022). In this context, repeatability describes the variation of a mean result obtained in successive measurements of the same soil sample analysed in the same laboratory under comparable conditions. The repeatability process requires that sufficient duplicates be taken, and the sample order randomized before analysis in the laboratory.

There is need for innovative methods to help laboratories to rapidly characterize the soil conditions and adopt internal quality controls with lower costs (Poppiel *et al.*, 2022). As previously noted, over the last decades, soil spectroscopy has evolved into a rapid, affordable, and environmentally-friendly technique, emerging as a viable alternative for measuring soil properties (Nocita *et al.*, 2015). Determining soil properties from soil reflectance data requires the development of statistical models, i.e., regression models that relate the measured spectral signal to the soil property of interest measured using conventional laboratory methods (analytical data). These relationships can then be used to predict the properties of new samples from the spectral signal (Sila *et al.*, 2016; Vågen *et al.*, 2016). Accurate and reliable analytical data are key for calibration and validation of such models (Dangal *et al.*, 2019).

An important consideration is the uncertainty associated with the final spectral predictions. Firstly, the spectral measurements are affected by the sample preparation, e.g. drying, sieving, grinding (Nduwamungu *et al.*, 2009). Secondly, sensor noise and other spectrometer internal sources (electronic and mechanical) can affect the measurements. Thirdly, the spectra can only explain

part of the variation of the soil property since the predictive power of the spectra may be limited. Fourthly, the required calibration data are determined with analytical data from laboratory analysis, with associated uncertainties. Spectrally estimated soil data are therefore not normally as accurate as analytical data because the statistical model error that predicts the soil property values from spectra add so the laboratory soil measurement error. Measurement error in spectral data can be estimated from the residual variance of the regression models used to predict the soil properties of interest. Addressing the various sources of uncertainty and their implications in spectral modelling is thus becoming increasingly important, such as to contribute to the development of standardized protocols for soil spectroscopic measurements and calibration development (Semella *et al.*, 2022).

Generally, DSM models tend to ignore errors in soil measurements used for model calibration and validation. This is a cause for concern because of the increasing use of low-cost techniques such as infrared spectroscopy instead of more expensive techniques (Nocita *et al.*, 2015). There is also a growing trend in using soil data acquired through citizen science initiatives (Rossiter *et al.*, 2015). However, there are concerns regarding the accuracy of such data generated by non-soil experts as it may contain substantial errors (Pino *et al.*, 2022). All these errors in soil measurements propagate into the outputs of DSM and should be accounted for. With a growing demand for more detailed and accurate soil maps, which require higher sampling densities, the need to account for measurement errors is further amplified (van der Westhuizen *et al.*, 2022).

1.5. Crop modelling and sustainable agricultural intensification

1.5.1. Soil information for crop modelling

Anticipated population growth and rising food consumption in SSA are expected to drive an ongoing increase in future food demand (Mesfin *et al.*, 2021). The main challenge lies on optimizing crop productivity to ensure food security whilst concurrently preserving the degradation of land resources and striving for sustainable agricultural intensification (Ramankutty *et al.*, 2018). Sustainable agricultural intensification englobes the impact on overall farm productivity, profitability, stability, market risks, resilience, as well as the participation of the land-users (Musumba, 2017). By this definition, it is not limited to environmental concerns, but also includes social and economic criteria such as improving livelihoods, equity and social capital (Wezel *et al.*, 2015).

The fundamental role of soil in sustaining productivity and environmental sound agricultural systems is widely recognised, acknowledging its capacity to provide a range of terrestrial ecosystem services and functions including biodiversity conservation, climate change mitigation and adaptation (Keesstra *et al.*, 2016; Trap and Blanchart, 2023). In addition, the recent recognition of the soil security concept is motivated by sustainable development, emphasising on the maintenance and improvement of the global soil resource to support its production capacity, and to maintain the biodiversity and the overall protection of the ecosystem (McBratney *et al.*, 2014). In this context, the term sustainable intensification characterizes the importance of linking agricultural production with sustainability and recognizes the need to maintain or increase food production per unit area without increasing the use of land or external inputs (Rees *et al.*, 2018). Soil is not only the major source of nutrients that are essential for plant growth, soil water infiltration and storage, and resistance to erosion also play an important role. Soil fertility is defined as the capacity of a soil to provide plants with nutrients, and it is greatly dependable on

physical, chemical, and biological properties of the soil. Some soil characteristics that are determinants of soil fertility, and commonly measured in the laboratory are soil texture, pH, soil organic carbon (SOC), nitrogen (N), phosphorus (P), potassium (K) and micronutrients (Zuber *et al.*, 2020). How much and in what proportion macro- and micro-nutrients are needed by plants, and how nutrient supply influences yield is of great importance to achieve sustainable agricultural production.

One way to manage nutrients efficiently to improve agricultural production is either through nutrient omission trials or the use of decision support tools such as crop models. Crop models have the ability to estimate crop yields as a function of available nutrients in the soil, as well as recommending fertilizer application rates to achieve a target yield. Examples of such models include WOFOST (van Diepen *et al.*, 1989), DSSAT (Jones *et al.*, 2003) , APSIM (Keating *et al.*, 2003) and QUEFTS (Janssen *et al.*, 1990). These crop models are vital tools in decision-making in SSA, in assessing the impacts of input and management practices on crop productivity, and in assessing the economic/environmental performance of alternative cropping systems, allowing to promote better and sustainable agricultural production (Kadiyala *et al.*, 2015).

1.5.2. Spatialisation of crop models: challenges and opportunities

Crop modelling can be applied to support agricultural production as it can quantify the complex, non-linear and interdependent interactions between the various components of a cropping system. This capacity enables the evaluation of its economic and environmental performance (Li *et al.*, 2015). One of the main aims of a crop model is the ability to simulate the functioning of a cropping system, providing a simplified representation of its behaviour and responses to variation of management practices and pedo-climatic conditions (Nassiri Mahallati, 2020). Considering the complexity of cropping systems in heterogenous landscapes, it seems very difficult, perhaps impossible, and unnecessary to capture all cropping system processes and sub-processes in mathematical terms, therefore models are only simplified versions of the reality (Heuvelink, 1998). Crop models are available in various types with applications for both science and policy (van Ittersum *et al.*, 2013). They can be classified based on objective (scientific understanding/ policy-making), purpose (simple/complex), and outcome (deterministic/stochastic) (Jones *et al.*, 2017a). Deterministic models are able to provide a clear cause-and-effect relationship between inputs and outputs, thus simplifying interpretation. However, they have limitations in capturing uncertainty and randomness. On the other hand, Stochastic models, conversely, consider uncertainty and provide a spectrum of possible outcomes.

A crop model can be applied over an area much larger than that over which it was developed. This has been referred to as "spatialising" a crop model (Faivre *et al.*, 2004; Shiva Shankar *et al.*, 2023), This requires that the model has been calibrated using data from the same area or an area with similar conditions, and that all the input data are available at the locations where the model is run. There are many incentives for implementing crop models in large areas and previous studies have reported examples of crop model spatialisation for a large range of purposes and at various scales (Ginaldi *et al.*, 2019; Kumar *et al.*, 2023). However, a constant challenge to crop model implementation and spatialisation, especially for future crop performance projections, is the unavailability of reliable data related to the initial conditions of the crop system and the period over which the model is run, which can be used as inputs for model calibrations (Jones *et al.*, 2017a; Kephe *et al.*, 2021). In some cases, available input data may not be in the quantity and quality needed to drive a crop model (Jones *et al.*, 2017b). Even when a suitable choice of a crop simulation model is selected, data limitations might hamper the models' effective role for simulating cropping systems. This challenge only increases when the study is upscaled to larger areas beyond the calibration area. Addressing the challenge of data availability is crucial

for successfully spatializing crop models, and it stands as a key solution for developing strategies that can significantly improve the use of crop models for sustainable agricultural production (Kephe *et al.*, 2021).

Typical input data required for most crop models are climate (e.g. temperature), soil (e.g. soil properties), crop varieties, initial conditions, crop management practices and their interactions (Ojeda *et al.*, 2021; Chapagain *et al.*, 2022). Available data often fall short of the expectation of crop models, which require information that is up-to-date, sufficiently accurate for the intended modelling purpose, accompanied with uncertainty information, easily integrated with other digital spatial data, and readily available for interested stakeholders (Buenemann *et al.*, 2023). The problem is that some of the model inputs are difficult to measure, and consequently, their estimation may require a statistical/geostatistical derivation, which introduces uncertainty. Crop model predictions are often derived under uncertainty, and this should be communicated effectively so as to enable the end-users to draw proper conclusions and so make sound decisions (Breure *et al.*, 2022b; Lark *et al.*, 2022). Among the different data used as inputs in crop model spatialization, the estimation of soil data is one of the most critical. As indicated in Section 1.2, the rapid advance of DSM offers the opportunity to fill the soil information gap by providing crop models with high quality and fine-resolution soil spatial datasets that can be more cost-effective and faster to produce than conventional soil maps. Unavoidably, the DSM generated soil maps are not error-free and are associated with uncertainty originating from various sources. These soil data therefore suffer from prediction and interpolation errors which can be quantified using approaches presented in Section 1.3.

Following error propagation techniques, such uncertainty in DSM maps could be propagated through model simulations to quantify their effect on the accuracy of the crop model outputs (Heuvelink, 2018; Ran *et al.*, 2022). Several methods for uncertainty propagation analysis have been developed and applied at various scales (Heuvelink, 2014). Monte Carlo analysis has often been used to compute output statistics by repeated model simulations with input variables randomly sampled from their probability distribution. The method has been successfully applied to propagate input uncertainty to model outputs in crop modelling, agronomy, and precision agriculture (Liu *et al.*, 2017; Ran *et al.*, 2022).

Using crop models to assess the impacts of agronomic practices on model predictions is also of specific importance to farmers in SSA. These farmers have been 'tagged' as vulnerable to food insecurity since they usually operate under suboptimal agricultural production conditions, often attributed to limited resources. More reliable estimations of model outputs such as predicted yield and fertilizer recommendations are therefore essential to farmers, to take decisions on input investments. The general principles of crop model application and spatialisation have been well developed as described above, but incorporating uncertainty information in decision making is less developed. The practical implementation of that requires working out the details of the various processes for quantification of uncertainty in soil measurements and in DSM maps, application of uncertainty propagation analyses in spatialized crop models, and communication of uncertainty to end-users to support their decision-making under uncertainty.

1.6. Problem statement, research objectives and questions

1.6.1. Problem statement

The growing demand for accurate soil information has led to the increased use of proximal soil sensing (PSS) methods such as spectral data for the generation of soil data (Stoorvogel *et al.*, 2015). Despite the success of PSS in providing soil data, one aspect that has received little attention so far is the quantification of the errors in soil measurements used for DSM and crop model calibration and prediction (Section 1.2). Since the calibration of PSS models depends on wet chemistry measurements which are not error free, the measurement error in analytical data add up to that of the PSS model and eventually propagate to the DSM and crop model final predictions (Heuvelink, 2018; Somarathna *et al.*, 2018). Unfortunately, measures of uncertainty associated with compilations of such analytical and spectral datasets are seldom provided, leaving the modellers with limited knowledge about the quality, or uncertainty associated with these data. Along with this recognition, the lack of consideration of measurement errors in soil data used for model training may lead to suboptimal models and systematic overestimation of prediction accuracy (Malone *et al.*, 2015; Poggio *et al.*, 2016; Heuvelink, 2018). Decisions that are based on poor quality maps whose accuracy is overestimated may have extensive and profound impacts on end-user decisions related to land management interventions. End-users may increase their investments in obtaining accurate soil data, for instance either by taking more samples to increase spatial sampling density or taking more accurate measurements. However, this is only possible if they are reliably informed about the accuracy of the DSM outputs and the potential sources of errors associated with the information. Currently, the solution for incorporating measurement errors into DSM model training and prediction is not obvious and has been the subject of limited research. This is particularly important when the magnitude of error varies between measurements, as more accurate measurements should carry more weight than less accurate measurements in DSM and other applications. Possible solutions for quantifying measurement errors in analytical and spectral data have been discussed in Section 1.4, and if applied, laboratories can report the measurements with associated uncertainties.

Despite the many uses of DSM products, the availability of high-resolution digital soil data for SSA, especially at local scales – remains limited. Currently, the digital soil dataset with the most comprehensive coverage of Cameroon is iSDAsoil (Miller *et al.*, 2021) SoilGrids (Poggio *et al.*, 2021) and the Africa Soil Profiles Database (Leenaars *et al.*, 2014). Nowadays, various DSM approaches including geostatistics and machine learning, are in use to model and quantify the uncertainty of DSM products (see Section 1.3). However, few studies in SSA have evaluated and compared the performance of regression kriging and random forest with respect to their accuracy in predicting soil properties and their success in modelling prediction uncertainty.

Equally, there are numerous challenges associated with undertaking large-scale soil sampling in SSA. These challenges stem from the intricate and remote nature of the terrain, creating difficulties in accessing specific areas within the targeted landscape. Additionally, the constrained allocation of resources for sampling exacerbates the complexities involved in this process. Cluster sampling design has been favoured in this context to provide soil data for DSM approaches (Winowiecki *et al.*, 2016; Vågen *et al.*, 2018). While there are numerous benefits to cluster sampling, such as cost and time efficiency, as well as explanation of variability at short distances, disadvantages such as the under-representation of some parts of the targeted landscape could affect the DSM modelling. Extrapolation of DSM models and algorithms from a relatively easily accessible area could overcome these challenges, as an alternative to conventional DSM, especially in remote areas. However, mapping in areas where there are no training data is challenging because extrapolation in geographic space often induces extrapolation in feature

space and can seriously deteriorate prediction accuracy (Hateffard *et al.*, 2024). The problem of spatial extrapolation for different DSM algorithms has not been fully investigated (Neyestani *et al.*, 2021), particularly in SSA, despite the predominant use of sparsely distributed and highly clustered soil data. The significance of using a well-thought-out sampling design, such as the stratification of soils based on soil legacy and taxonomy data, holds considerable importance (Brus, 2019).

Opinions converge on the fact that increase in food production to feed the anticipated increase in population cannot be attained without an increased use of fertilizers (ten Berge *et al.*, 2019), particularly in SSA where application rates are low (Sheahan and Barrett, 2017). This, therefore, presents a scenario where the information needed for agricultural decision-making at all levels from farm management to policy are also increasing and methods of supplying such information in relatively short timeframes are crucially needed. Traditional agronomic research based on field experiments has been and is being used as a reliable information source for establishing causal relationships between agricultural system patterns and the real-world (Durr *et al.*, 2016; Liang *et al.*, 2016). However, such research methods are becoming insufficient to meet the rate of increasing needs and demands for data to guide agricultural policies and decision-making (Kephe *et al.*, 2021). Field experiments often generate results obtained only from trials conducted at points in time and place, thereby creating plot-specific results that cannot easily be extrapolated to other plots or areas (Oteng-Darko *et al.*, 2013). Also, these trials may not provide sufficient data in space and time to identify appropriate and effective management practices (Jones *et al.*, 2017b, a).

Furthermore, agricultural landscapes in SSA are usually managed under "blanket" agronomic recommendations (Rurinda *et al.*, 2020), even though there is often considerable intrinsic spatial variation in soil nutrients (Kihara *et al.*, 2016; Shehu *et al.*, 2018). This increases the risk of over- and under-application of fertilizers, leading both to undesirable environmental effects and increase in production costs. Also, the composition and quality of the recommended fertilizer may present challenges that can impact yield responses post-applications. Site-specific fertilizer recommendations have been proposed as a solution but current strategies to formulate site-specific fertilizer are still based primarily on the results of experimental plots. Due to these shortcomings, there is a need for tools that can combine new data and research findings and make the results available to stakeholders for decision-making. Hence, there is a need for the development and efficient use of tools, such as crop models, to project cropping systems under various scenarios, conditions, and scales. Since these crop models, just as other models, are only an imperfect approximation to the real-farming systems, recent developments in crop modelling have acknowledged the need to quantify the uncertainty associated with the outputs (Ewert *et al.*, 2015; Wallach and Thorburn, 2017).

1.6.2. Research objectives and questions

The objective of this thesis is to apply and extend state-of-the-art DSM approaches to analyse and quantify the spatial patterns of soil properties while accounting for uncertainty in soil measurements. Moreover, it involves evaluating the repercussions of using uncertain DSM maps as inputs in crop modelling to predict yield and formulate fertilizer recommendations. The impact of realistic quantification of accuracies of soil measurements and soil maps could improve the performance of DSM and crop models and support the development of strategies and policies that enhance sustainable agricultural intensification. Accurate measurement of soil observations, soil spatial information, and fertilizer recommendations, along with associated uncertainties, involves a comprehensive integrated process. This encompasses soil sampling, laboratory analysis, evaluating uncertainty in soil measurements, collecting environmental co-variables,

employing geostatistical and machine learning algorithms, quantifying uncertainty in soil maps, and conducting uncertainty propagation analysis in crop modelling. This multifaceted approach aims to improve information delivery to end-users, aiding in informed decision-making processes.

These requirements to achieve the overall objective result in four sub-objectives with associated research questions as follows:

1) Description and quantification of the spatial variation of soil properties using simple geostatistical methods (ordinary kriging) (**Chapter 2**).
   - What is the importance of describing and quantifying soil properties?
   - Can simple geostatistical techniques be used to display spatial variability of soil properties?
2) Quantification of errors in soil measurements and incorporation of the associated uncertainty into a state-of-the art geostatistical method (regression kriging) for spatial interpolation and comparison with a case in which measurement errors are ignored (**Chapter 3**).
   - How can measurement errors in analytical and proximal soil sensing soil data be quantified?
   - Can the quantified measurement errors be incorporated into a state-of-the-art geostatistical method for spatial interpolation?
   - How do the results of spatial interpolation accounting for measurement errors compare with a case in which measurement errors are ignored?
3) Extension of the calibration and prediction of DSM models using uncertain soil measurements from linear kriging with external drift, i.e. regression kriging (RK) to non-linear machine learning algorithms-based DSM models, i.e. random forest (RF) (**Chapter 4**).
   - Do RK and RF have similar prediction accuracy?
   - What are the performances of RK and RF in modelling prediction uncertainty?
   - How sensitive are RK and RF predictions to soil measurement errors?
   - Does spatial extrapolation of a DSM model lead to a bigger or smaller deterioration when using RF instead of RK?
4) Quantification of uncertainty of QUEFTS model predictions considering uncertainty in soil input variables using Monte Carlo simulation (**Chapter 5**).
   - How does uncertainty in soil inputs propagate through the QUEFTS crop model?
   - Which of the uncertain soil inputs are the main contributors to the overall uncertainty of the QUEFTS model outputs?
   - How large is the uncertainty in QUEFTS model outputs given uncertainty of soil information?
   - How can uncertainty in model outputs be incorporated in decision-making?

1.7. Study area and sampling framework

The research was conducted in the west region of Cameroon, which is largely an agrarian area and spanning over 13,892 km² in the western highlands of the country (Fig. 1.1). The climate is

tropical humid with two seasons: a long, wet season of eight months from March to October, and a short, dry season of four months from November to February. The average annual temperature ranges between 20 °C and 28 °C, while annual average rainfall ranges from 1,200 mm to 2,300 mm (Neba, 1999). The area is characterized by accidented relief of massifs and mountains that consist of plains, undulating hills, and gentle sloping areas. The elevation in the study area ranges between 450 and 2400 meters above sea level (masl). The dominant soil types are classified as Ferralsols and Nitisols of the World Reference Base system (IUSS, 2015), and high variability of soil can be observed at short distances. The area is dominated by subsistence agricultural systems, with highly skilled farmers utilizing nearly every available strip of land to cultivate a variety of annual and perennial crops. Maize stands out as the primary staple crop, cultivated either in association with or in rotation with other crops.

The multiple spatial scales sampling frame used in this research was based on the land degradation surveillance framework (LDSF) (Vågen and Winowiecki, 2020). The LDSF framework is built around a spatially stratified field survey and sampling protocol using sentinel sites that are 100 km2 (10 x 10 km). Within each LDSF site, 16 tiles (2.5 x 2.5 km in size) are created and random centroid locations for clusters within each tile are generated. Each cluster consists of 10 plots, with randomized centre-point locations falling within a 1 km2 area. Each plot is 0.1 ha and consists of 4 subplots, 0.01 ha in size. This is particularly important in predictive modelling where we need to understand uncertainty (or accuracy) of predictive models at different scales. Four LDSF sites were established across the study area.

A total of 640 topsoil samples (0 – 20 cm) were collected between 2015 – 2017 within the framework of the implementation of multiples land restoration projects and used for the study. Sub-areas were delineated within the study area to accommodate the various objectives (Fig. 1.1). Area 1 consists of two LDSF sites of 100 km^2 each with 160 sampling locations and contrasting land use types. Each LDSF site was suitable for implementation of a basic geostatistical method in **Chapter 2** that requires sampling that covers the geographical space well. Area 2, which spans over 1,053 km^2, encloses three LDSF sites with 480 sampling plots. This sub-area was chosen for the implementation of regression kriging and random forest based on environmental variables and uncertain soil measurements in **Chapter 3** and **Chapter 4**. Area 3 covers the administrative boundaries of the west region of Cameroon with four LDSF sites. Area 3 was used to assess the transferability and the spatialisation of the QUEFTS model over large areas and quantification of the uncertainty in predicted crop yield and fertilizer recommendations in **Chapter 5**.

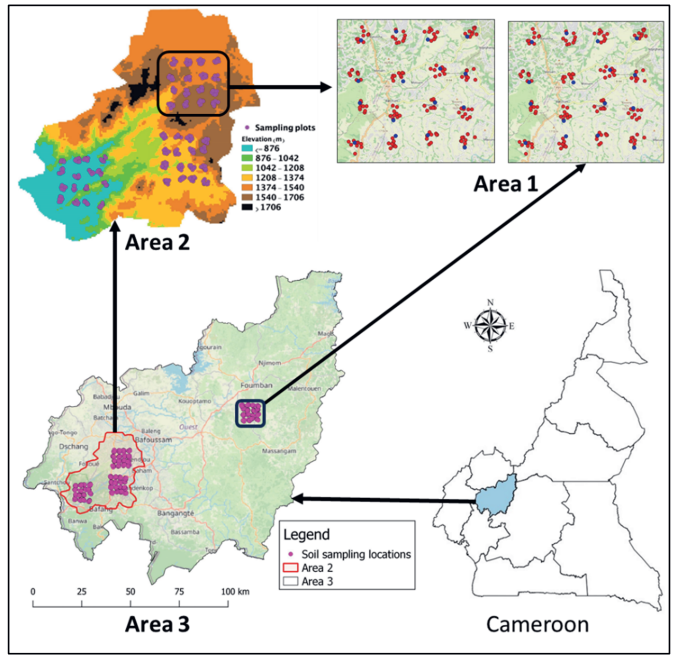


Fig. 1.1: Map of the study areas of this thesis and sampling locations: Area 1 (upper-right panel), Area 2 (upper-left panel) and Area 3 (lower-left panel).

1.8. Overview and structure of the thesis

The thesis is written in an article format with each chapter presented with a separate introduction, methodology and results adjusted to address aspects of the objectives of this thesis. Moreover, sections on discussion and conclusion are also presented for each chapter. In order to enhance the readability of the thesis, a *General introduction,* **Chapter 1**, is presented to introduce the topic, define the research objectives and questions, and provide a description of the study area.

Chapter 2, *Landscape approach to assess key soil functional properties in the highlands of Cameroon: Repercussions of spatial relationships for land management interventions*, directly addresses the first objective of this thesis, and presents a brief introduction of the geostatistical techniques used in this research, an analysis of the spatial variation of selected properties using basic geostatistical methods and assesses the implications of spatial relationships for land management interventions.

Chapter 3, *Accounting for analytical and proximal soil sensing errors in digital soil mapping*, was designed to demonstrate how to quantity the measurement errors in analytical and proximal soil sensing soil data, incorporate them in a geostatistical method for spatial interpolation, and compare the results with a case in which measurement errors are ignored.

Chapter 4, *Comparing the prediction performance, uncertainty quantification and extrapolation potential of regression kriging and random forest while accounting for soil measurement errors*, addresses a recommendation that was made in **Chapter 3** to extend DSM accounting for uncertainty in soil observations to non-linear machine learning regression methods. To meet objective 3 of the thesis, this chapter compares the performance of random forest and regression

kriging with respect to the accuracy of their predictions, their successes in quantifying prediction uncertainty, the sensitivity of their predictions to measurement errors, and their extrapolation potential.

Chapter 5, *Modelling and mapping maize yields and fertilizer recommendations with uncertain soil information*, acknowledges the impacts of uncertainty of soil information on model outputs, carries out an uncertainty propagation analysis and quantifies uncertainty sources of QUEFTS model outputs. Other objectives include the formulation of strategies to communicate uncertainty to end-users and advising on how uncertainty can be incorporated in decision-making.

The final **Chapter 6,** *Synthesis*, provides a summary of the key findings, the implications and practical applications, and recommendations for improving DSM and crop modelling approaches, as well as an outlook on future studies.

Chapter 2

Landscape approach to assess key soil functional properties in the highlands of Cameroon: Repercussions of spatial relationships for land management interventions

Understanding spatial variability of soil properties is essential to support land management decisions. However, despite the growing worldwide emphasis on integrated landscape management, soil variations resulting from land use changes have rarely been documented. The study used the land health surveillance concept in combination with simple geostatistical approaches to describe selected soil properties among land use types and characterize their spatial variability in the highlands of Cameroon. A total of 320 soil samples were collected in two sites with contrasting landscape attributes and land uses (agricultural and pasture) and were analysed for granulometric fraction, soil organic carbon (SOC), nitrogen (N), pH, phosphorous (P), calcium (Ca), potassium (K), magnesium (Mg), aluminium (Al) and zinc (Zn) using diffuse reflectance mid-infrared spectroscopy (MIRS). Variogram models were used to quantify the spatial variation of soil properties, while ordinary kriging was applied to generate the respective maps of SOC, N, and clay content. The accuracy of the prediction performance of models was assessed using cross-validation, and the implications of spatial relationships for land management interventions and restoration was evaluated. The results showed that soil properties differed considerably across the area, with significant positive and negative correlation coefficients among many pairs of soil properties. The coefficients of variation (CV) helped in comparing the degree of variation of soil properties relative to the mean within the two sites. For Bamendjou, the most variable properties (CV >38%) were P, Mn, and Ca. Moderate variability (2.8% < CV < 38%) was observed for sand, SOC, N, K and Mg, while properties with very low variability (CV < 2.8%) were clay, soil pH, and Al. For Koutaba, the most variable properties were sand, Ca, K, Mg, Mn, P and Zn. Moderate variables were N, SOC, and silt, while clay, Al and pH were the least variables. Spherical variogram models were chosen as the best-fitted models for the investigated soil properties as attested by cross-validation. The spatial correlation ranges were significantly larger for SOC and N in Bamendjou than in Koutaba. Land use with more vegetation cover (forest, grassland, and fallow) exhibited the highest concentration of soil properties, attesting that land use types had significant impacts on spatial patterns and distribution of the soil properties. Well-defined patterns of higher concentrations of SOC and N were observed in the lowlands, valleys, and areas dominated by annual vegetation. Kriged maps provided a detailed visualization of soil properties at the landscape scale and helped to identify 'hotspots' of land degradation and critical areas in need of specific land management practices to improve land productivity. These findings can be a helpful tool in achieving efficient site-specific land management interventions, that lead to better decisions aimed at enhancing the efficient use of agricultural inputs, such as fertilizers, in the context of limited resources. Although the spatial models could explain a large part of the spatial variation of soil properties, they may be improved by quantifying the measurement errors in soil observations, and expanding the analysis with relevant covariates that represent the soil forming factors.

A version of this chapter has been published in Takoutsing, B., Martín, J.A.R., Weber, J.C., Shepherd, K., Sila, A. & Tondoh, J. 2017. Landscape approach to assess key soil functional properties in the highlands of Cameroon: Repercussions of spatial relationships for land management interventions. *Journal of Geochemical Exploration*, 178, 35-44.

2.1. Introduction

Land management has received considerable attention during the past years with focus on increasing nutrient input efficiency, improving crop productivity, and reducing the environmental risks (Reza *et al.*, 2010). Soils are characterized by a high degree of spatial variability due to the combined effects of physical, chemical, and biological processes that operate with different intensities and at different scales (Awal *et al.*, 2019). Consequently, soils exhibit marked spatial variability both at the macro- and micro-scales (Amirinejad *et al.*, 2011; Shukla *et al.*, 2016). Soil spatial variability is a function of land-use type, topographic features, soil formation factors, soil depth, anthropogenic activities, and time (McBratney *et al.*, 2003; Guan *et al.*, 2015). Changes in soil properties and processes, in turn, impact plant growth and the environment and require accurate knowledge of these properties for efficient utilization of agricultural inputs. Understanding spatial variability of soils and nutrient is essential for devising site-specific nutrient management strategies with the aim of improving sustainability in crop production.

Agricultural landscapes in Cameroon have been managed as being homogeneous, although presenting considerable spatial variations in soil properties concentrations. The current fertilizer recommendation to small-scale farmers is uniform for the entire targeted areas irrespective of intrinsic variations in soil, cropping history and management practices. There is therefore a risk of over- or under-application of the required nutrients which can lead to undesirable environmental effects and increased in crop production cost. Practically, understanding soil spatial variations is important for farmers attempting to enhance land productivity and to support soil management decisions such as fertilizer recommendations. The spatial information on key soil functional properties can be a powerful decision-making tool in achieving efficient site-specific land management practices (Guedes Filho *et al.*, 2010), that lead to better land management decisions aimed at remedying nutrient deficiencies and maintains sustainable productivity of the soils (Özgöz, 2009). Despite the growing worldwide emphasis on integrated landscape management with emphasis on spatial information, large investments are still being made in the highlands of Cameroon to improve land productivity without considering soil variations.

Furthermore, with the increasing concern about food security and sustainable development, agricultural land has to face great pressure to be able to provide not only food for the growing population, but also a range of multiple ecosystem services (Schulte *et al.*, 2014). For a given agricultural landscape, soil properties and their spatial patterns are primarily controlled by land use type and landscape physical attributes (Bünemann *et al.*, 2018). Moreover, the variation in landscape attributes is likely to play an important role in the spatial heterogeneity of soil properties (Takoutsing *et al.*, 2018; Chen *et al.*, 2022b). These changes in soil properties caused by land use and other landscape attributes inevitably lead to the changes in soil function (Mandal *et al.*, 2021; Vrebos *et al.*, 2021). The knowledge gap of spatial variation in soil properties hinders us to specifically understand the spatial pattern of soil degradation so as to best formulate restoration strategies.

The western highlands of Cameroon have experienced dramatic land use changes in the past decades due to intense anthropogenic activities, and soil variations resulting from such changes have rarely been documented. Soil characteristics generally show spatial dependence, which means that samples close to each other are more similar than samples farther apart (Webster and Oliver, 2001). Geostatistical methods have been used to predict soil variables at unknown locations using a property measured at a given place and time (Reza *et al.*, 2010). These mapping methods take into account the spatial autocorrelation of a given variable in predicting its value in another unsampled location. There is therefore a need to use some of these techniques to quantify the variations of soil properties in order to aid accurate estimation of nutrient budgets, cycling rates and actual demand for inputs. In recent years, Ordinary Kriging (OK) has been

widely used by many researchers for preparation of the maps of spatial distribution of soil properties (Reza *et al.*, 2016). The aims of this Chapter were to use simple geostatistical analysis methods to a) assess the spatial variation characteristics of selected soil properties, b) display spatial variability of selected soil properties using simple geostatistical techniques and c) determine factors that influence the spatial variability of soil properties across the targeted landscape.

2.2. Materials and Methods

2.2.1. Study area and soil sampling

The study area is located in the Western highlands of Cameroon, where two sentinel sites of 100 km2 each with contrasting land uses were established; one dominated by agricultural activities and the other intensively used as pasture for livestock rearing (Fig. 2.1). The area is characterized by a tropical climate: mean daily minimum and maximum temperatures are 20 and 28° C, respectively; mean annual rainfall ranges from 1,200 to 2,300 mm per year, and the rainy season extends from March to October. The topography is undulating with altitude ranging from 1,000 to 1,800 masl, and the vegetation is predominantly savannah with patches of gallery and montane forests. The soil is mainly composed of Ferralsols according to the World Reference Base (WRB) for soil resources classification system (IUSS, 2006). Due to anthropogenic activities, the landscape is a mosaic of land use types made up of cropland, fallow, forest, grassland, pasture and shrubland.

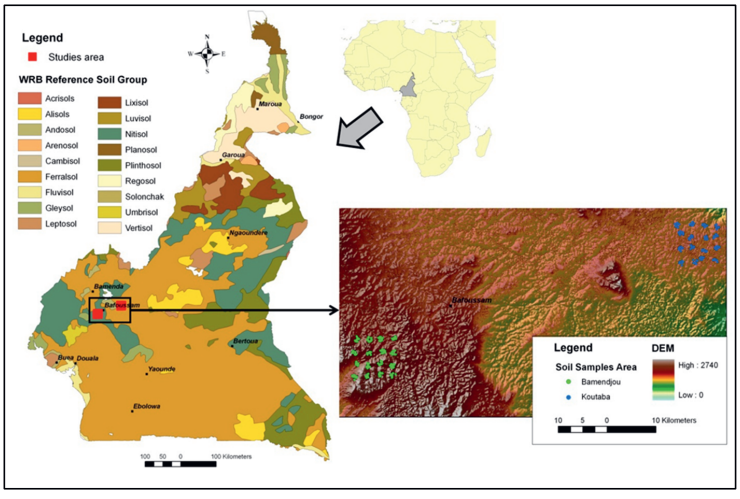


Fig. 2.1: Map of Cameroon showing the WRB Reference soil group and location of study sites.

This Chapter used the land degradation surveillance framework (LDSF), which is a spatially stratified, random sampling design framework based on the concept of sentinel site (Vågen *et al.*, 2006; Vågen *et al.*, 2012). A sentinel site is a demarcated landscape of 100 km2 (10 X 10 km) that is representative of a larger area, and from which in-depth data and information are gathered and the resulting analysis can be used to inform land management programs and policies (Shepherd *et al.*, 2015). A sentinel site has 160 randomized sampling design that allow

for statistical modelling. Within each site, 16 tiles (2.5 x 2.5 km in size) are created and random centroid locations for clusters within each tile are generated. Each cluster consists of 10 plots, with randomized centre-point locations falling within a 5.64 m radius from each cluster centroid (Fig. 2.2a). Thus, the LDSF has two levels of randomization, which minimize local biases that may arise from convenience sampling. Each plot is 0.1 ha (1000 m2) and consists of 4 subplots, 0.01 ha in size (Fig. 2.2b). The two sites namely Bamendjou and Koutaba were established and surveyed following procedures described in (Vågen *et al.*, 2006; Vågen *et al.*, 2013). A total of 320 topsoil (0 – 20cm) composite samples were collected across the two sentinel sites (160 samples per site).

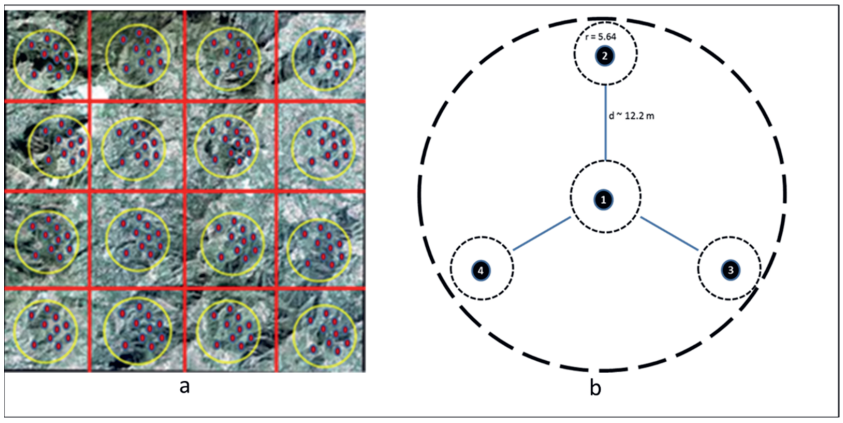


Fig. 2.2: a) Schematic illustration of a sentinel site of 10 x 10 km divided into 16 grids. b) Sampling plot and subplots. Source (Vågen and Winowiecki, 2020).

2.2.2. Soil laboratory and spectral analyses

The soil samples (n = 320) were air-dried and then passed through a 2 mm sieve to remove coarse and stones prior to laboratory analysis. A subset (n = 32) representing ~ 10% of the total samples were analysed for a range of properties at the Crop Nutrition (Cropnut) Laboratory in Nairobi using conventional wet chemistry methods. Then the 320 soil samples were ground to < 100 µm using agate mortar and pestle and were analysed by diffuse reflectance mid-infrared spectroscopy (MIRS) using the procedure described elsewhere (Terhoeven-Urselmans *et al.*, 2010; Vågen *et al.*, 2016). Soil MIRS is a non-destructive, rapid, and cost-effective methodology for characterising a large number of soil samples, and for a range of soil properties, therefore enabling landscape-level assessments of spatial variability of soil health indicators (Shepherd and Walsh, 2002; Viscarra Rossel *et al.*, 2006; Vasques *et al.*, 2008). Regression used MIRS data as independent variables and the laboratory data as dependent variables. Then the fitted regression models were used to predict properties of all the samples based on the calibration samples (Vågen *et al.*, 2016).

2.2.3. Statistical and geostatistical analysis

A standard statistical analysis (mean, median, standard deviation, coefficient of variation, etc.) was carried out to describe the soil properties involved in two sentinel sites in the highlands of Cameroon. Significant differences between land use types were assessed by non-parametric Kruskal-Wallis tests. In order to study the relationship between soils properties, canonical

correlation analyses (CCorA) used considerably in ecology were applied between explanatory soil granulometric fraction, SOC and pH and response N, P, Ca, K, Mg, Al, Mn and Zn. The relationships between the two groups were estimated using quadrants of the CCorA (Odumo *et al.*, 2014). However, these classical statistical approaches ignore spatial dependence between observations (Nourzadeh *et al.*, 2012; Nanos *et al.*, 2015). Simple geostatistical techniques based on semivariograms have increasingly been used to analyse the spatial pattern of environmental variables (Goovaerts, 1997; Rodríguez Martín *et al.*, 2016). An experimental semivariogram, which is the plot of semi variance as a function of the distance, was developed to evaluate the degree of spatial continuity and to determine the range of spatial dependence for each soil property. The variogram γ was calculated using the equation defined in Oliver and Webster (2014) as follows:

$$\gamma(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} [Z(u_i) - Z(u_i + h)]^2 \quad (2.1)$$

where $\gamma(h)$ is the empirical semivariogram value at the lag interval distance h, $\mathrm{Z}(u_i)$ the Z value at location u_i, $Z(u_i + h)$ the Z value at a location separated from u_i by distance h, and $N(h)$ the number of point pairs for lag h. The spherical models were used to fit the experimental semivariograms as the best model. One of the main applications of geostatistics has been the estimation and mapping of soil properties. Kriging estimates are calculated as the weighted sums of the adjacent sampled values. There are many different kriging algorithms including ordinary kriging (OK), and most have been reviewed in Goovaerts (1997) with references to soil applications. In this Chapter, three soil properties namely SOC, N, and clay content were mapped by OK. OK assumes that we do not know the mean. The most likely value $\hat{\mathrm{Z}}(\boldsymbol{x}_0)$ that one could expect to encounter in a particular grid cell when using m nearby observations was defined as:

$$\hat{z}(\boldsymbol{x}_0) = \sum_{j=1}^{m} \lambda_j \cdot z(\boldsymbol{x}_i) \quad (2.2)$$

Where $\hat{\mathrm{z}}(\boldsymbol{x}_0)$ is the estimated value and λ_{j} is the kriging weight of the determined value at location $\boldsymbol{x}_0$. These weights are obtained by solving a system of linear equations known as an 'ordinary kriging system' (Goovaerts 1997). The prediction accuracy of the soil property maps was evaluated by the cross-validation technique, which removes each data location, one at a time, and predicts the associated data value. In the determination of errors, one point is omitted, and its value is estimated based on the remaining values. Afterwards the kriged value is compared with the real value in the situation of the omitted point and indicates whether the accuracy of the Kriged results can be accepted (Marchant *et al.*, 2010). This procedure is repeated for all the values. All the statistical analyses were carried out using the SPSS statistical package version 21.9 (SPSS Inc., Chicago, USA), XLSTAT (Addinsoft Version 2012.2.02), ISATIS V10.0 and the Geostatistical Analyst extension for ArcGis 9.3.

2.3. Results and discussion

2.3.1. Descriptive statistics of soil properties

Means and coefficients of variation (CV) of soil physico-chemical properties in different land use types in Bamendjou and Koutaba sites are presented in Tables 2.1 and 2.2 respectively. Bamendjou had higher values for SOC, N, pH, Ca, and sand, while Koutaba had higher value for clay and Al. Differences in essential elements (N, P and K) were also observed between the two sites. Mean P content in Bamendjou (5.48 mg kg^{-1}) was two times higher than that found in Koutaba (2.42 mg kg^{-1}), and the mean K content in Bamendjou (0.423 mg kg^{-1}) was three times higher than that found in Koutaba (0.131 mg kg^{-1}). The same trend was observed for micronutrients (Mg, Mn, and Zn), with the highest levels found in Bamendjou. Textural analysis revealed that the entire study area is dominated by clay-rich soils. The soil reaction was also found to be slightly acid in both sites (pH < 6.2). The general acidity of the soils is mainly due to the chemical composition of parent materials; soils in the study area are Ferralsols according to the World Reference Base (WRB) for Soil Resources classification system (IUSS, 2006). These are strongly-weathered soils in which acidity, high exchangeable aluminium, and a low ratio of basic to total cations could constitute some of the main limiting factors to permanent cropping systems.

Soil organic carbon (SOC) is considered among the most important bio-chemical parameters to assess functional capacities of soils (Rodríguez Martín *et al.*, 2016). The SOC concentration fell within the range of 0.80 - 7.72 % (mean = 3.75 %) in Bamendjou and 1.15 - 5.98 % (mean = 2.60 %) in Koutaba. A review of critical SOC levels in soils (Loveland and Webb, 2003) has suggested that 2% SOC [ca. 3.4% soil organic matter (SOM)] is a threshold value, below which potentially serious decline in soil quality will occur. The differences observed between the two sites are ascribed to the fact that Bamendjou is dominated by agricultural activities while Koutaba is predominantly used for pasture. Bamendjou site has a long history of intense agricultural production, and most cultivated plots are managed to sustain productivity. Farmers improve soil fertility by ploughing and adding organic and inorganic fertilizers (Takoutsing *et al.*, 2013). Koutaba has been under intensive grazing for many years, and that has considerably reduced vegetation cover and has probably contributed to nutrient depletion.

Table 2.1: Descriptive statistics of soil properties in Bamendjou site and mean values for different land use types (n = 160).

	Mean	Min	Max	CV (%)	Cropland	Fallow	Pasture	Forest	Shrubland	Grassland
Clay (%)	67.67	52.36	84.73	8.06	67.93 (a)	66.83 (a)	76.50 (a)	57.17 (a)	68.85 (a)	61.34 (a)
Silt (%)	21.60	9.77	51.76	17.45	20.88 (ab)	22.55 (b)	14.77 (a)	36.01 (b)	19.05 (ab)	28.22 (b)
Sand (%)	13.62	2.47	55.66	21.17	12.70 (a)	14.83 (a)	7.54 (a)	25.05 (a)	11.19 (a)	20.77 (a)
SOC (%)	3.757	0.806	7.722	27.84	3.790 (b)	3.90 (b)	1.63 (a)	5.79 (b)	3.58 (ab)	4.48 (b)
N (%)	0.262	0.051	0.589	29.30	0.260 (b)	0.274 (b)	0.107 (a)	0.439 (b)	0.262 (ab)	0.336 (b)
pH	5.893	5.107	6.563	4.00	5.905 (b)	5.903 (b)	5.559 (a)	6.181 (b)	5.818 (ab)	6.191 (b)
P (mg kg^{-1})	5.480	1.480	13.212	38.40	5.385 (b)	5.707 (b)	2.753 (a)	9.721 (b)	4.451 (ab)	8.036 (b)
Ca (mg kg^{-1})	8.313	0.704	47.694	38.47	7.753 (b)	8.819 (b)	2.353 (a)	24.349 (b)	5.231 (ab)	17.494 (b)
K (mg kg^{-1})	0.423	0.140	1.078	33.63	0.399 (ab)	0.446 (bc)	0.242 (a)	0.934 (c)	0.397 (bc)	0.685 (bc)
Mg (mg kg^{-1})	3.773	0.284	16.743	31.68	3.652 (b)	3.929 (b)	1.467 (a)	8.386 (b)	6.205 (b)	5.500 (b)
Mn (mg kg^{-1})	12.70	4.700	29.60	44.03	11.9 (a)	13.30 (a)	8.50 (a)	26.50 (a)	11.20 (a)	20.50 (a)
Zn (mg kg^{-1})	1.341	0.768	2.397	21.28	1.297 (a)	1.346 (a)	1.440 (a)	1.869 (a)	1.275 (a)	1.695 (a)
Al (mg kg^{-1})	1406	1079	1608	5.3	1420 (b)	1421 (b)	1192 (a)	1393 (ab)	1447 (b)	1319 (ab)

Value in parentheses after the land use indicates the number of soil samples. Significant differences amongst the different land use types ($P < 0.05$) are indicated by different letters, based on the non-parametric Kruskal–Wallis test SD.

The coefficients of variation (CV) helped in comparing the degree of variation within the two sites. The CV values for all soil properties ranged from 4 to 45 % in Bamendjou and from 2.9 to 95 % in Koutaba, which indicated low to high variation in both sites (Nielsen and Bouma, 1985). For Bamendjou, the most variable properties (CV > 38%) were P, Mn, and Ca. Moderate variability (2.8% < CV < 38%) was observed for sand, SOC, N, K and Mg, while properties with very low variability (CV < 2.8%) were clay, soil pH, and Al. For Koutaba, the most variable properties were sand, Ca, K, Mg, Mn, P and Zn. N, SOC, and silt were moderately variable while clay, Al and pH were the least variable. The wide range observed for most of the properties may be associated with historical land use, land cover and varying management practices. Farmers used various methods to improve the productivity of the soil, and this probably has an influence on the concentration of soil nutrients. Similar findings have been reported at various scales (Di Virgilio *et al.*, 2007; Fu *et al.*, 2010). Most skewness coefficients for soil properties (not tabled) in Koutaba were positive and extremely high, while the majority of skewness coefficients in Bamendjou were moderate and negative. Highly skewed parameters indicate that these elements have a local distribution, that is, high values were found for these elements at some points, and low values at some other points (Grego *et al.*, 2006).

Table 2.2: Descriptive statistics of soil properties in Koutaba site and mean values for different land use types (n = 160).

	Mean	Min	Max	CV (%)	Cropland	Fallow	Pasture	Forest
Clay (%)	78.13	58.96	99.33	7.79	71.41 (a)	73.47 (a)	80.05 (b)	76.77 (ab)
Silt (%)	14.33	5.49	33.33	24.23	19.54 (b)	17.21 (b)	12.98 (a)	15.44 (ab)
Sand (%)	4.95	1.07	22.25	42.55	7.21 (b)	7.39 (b)	4.12 (a)	5.47 (ab)
SOC (%)	2.600	1.151	5.989	20.36	3.198 (b)	3.059 (b)	2.417 (a)	2.780 (ab)
N (%)	0.143	0.065	0.410	27.22	0.185 (b)	0.179 (b)	0.129 (a)	0.153 (ab)
pH	5.298	4.870	5.921	2.90	5.327 (a)	5.435 (a)	5.273 (a)	5.222 (a)
P (mg kg^{-1})	2.423	0.883	7.787	40.28	3.108 (b)	3.242 (b)	2.170 (a)	2.397 (ab)
Ca (mg kg^{-1})	1.030	0.171	10.285	95.60	1.612 (b)	1.974 (b)	0.774 (a)	0.888 (ab)
K (mg kg^{-1})	0.131	0.060	0.578	66.92	0.185 (b)	0.193 (b)	0.111 (a)	0.130 (ab)
Mg (mg kg^{-1})	0.350	0.108	2.290	77.71	0.561 (b)	0.617 (b)	0.269 (a)	0.336 (ab)
Mn (mg kg^{-1})	4.99	2.02	18.12	43.55	7.24 (b)	6.87 (b)	4.30 (a)	5.15 (ab)
Zn (mg kg^{-1})	0.931	0.572	1.600	41.43	1.120 (b)	1.029 (b)	0.885 (a)	0.955 (ab)
Al (mg kg^{-1})	1467	1243	1647	2.98	1454 (a)	1480 (a)	1464 (a)	1482 (a)

Value in parentheses after the land use indicates the number of soil samples. Significant differences amongst the different land use types (P < 0.05) are indicated by different letters, based on the non-parametric Kruskal–Wallis test SD.

2.3.2. Relationships between soil properties

There were significant positive correlations among most pairs of soil properties and some correlations (P/K, P/N, P/Ca, K/Ca, Ca/Mg, K/Mg, P/Mn, or K/Mn) were very strong (Table 2.3). The strongest correlation was between SOC and N in Bamendjou (r = 0.989) and Koutaba (r = 0.971). These correlations were expected because the two properties are related to the amount of organic matter in the soil (Kahle *et al.*, 2002; Takoutsing *et al.*, 2013).

Previous studies have demonstrated the influence of N on carbon stock through plant growth and litter quantity, as well as its role in stabilizing soil carbon decomposition particularly in tropical areas (Nave *et al.*, 2009). The dynamics of N in the soil are closely related to that of carbon: thus, any change in the level of one property may cause a change in the level of the other in the soil (Qi *et al.*, 2007; Moges and Holden, 2008). Since the two elements are highly

correlated, their concentrations in the soil will both decrease through the same processes such as erosion, crop harvesting and leaching.

Table 2.3: Spearman's correlation coefficients among soil properties in Koutaba and Bamendjou areas (above and below diagonal, respectively).

	Clay	Silt	Sand	SOC	N	pH	P	Ca	K	Mg	Al	Mn	Zn
Clay		-0.942**	-0.771**	-0.762**	-0.750**	-0.191*	-0.797**	-0.544**	-0.646**	-0.467**	0.252*	-0.807**	-0.811**
Silt	-0.935**		0.739**	0.778**	0.749**	ns	0.729**	0.514**	0.619**	0.437**	-0.373**	0.824**	0.903**
Sand	-0.758**	0.884**		0.520**	0.638**	0.464**	0.882**	0.727**	0.819**	0.683**	ns	0.787**	0.515**
SOC	-0.725**	0.626**	0.309**		0.971**	0.173*	0.739**	0.669**	0.716**	0.647**	-0.290**	0.860**	0.784**
N	-0.774**	0.696**	0.402**	0.989**		0.351**	0.848**	0.811**	0.853**	0.797**	-0.164*	0.926**	0.729**
pH	-0.768**	0.608**	0.441**	0.664**	0.705**		0.652**	0.666**	0.622**	0.606**	0.341**	0.389**	ns
P	-0.937**	0.907**	0.744**	0.775**	0.838**	0.842**		0.906**	0.936**	0.824**	ns	0.907**	0.590**
Ca	-0.765**	0.731**	0.536**	0.723**	0.798**	0.811**	0.908**		0.957**	0.943**	ns	0.862**	0.453**
K	-0.851**	0.861**	0.675**	0.739**	0.822**	0.723**	0.919**	0.930**		0.942**	ns	0.928**	0.562**
Mg	-0.606**	0.587**	0.421**	0.622**	0.704**	0.596**	0.679**	0.802**	0.851**		0.156*	0.837**	0.393**
Al	-0.255**	ns	ns	0.492**	0.449**	0.230*	0.239*	ns	ns	ns		-0.212*	-0.665**
Mn	-0.839**	0.855**	0.653**	0.655**	0.733**	0.667**	0.870**	0.882**	0.967**	0.793**	ns		0.794**
Zn	-0.591**	0.684**	0.494**	0.384**	0.457**	0.380**	0.592**	0.616**	0.751**	0.614**	-0.492**	0.822**	

Sample size = 160 in each site; significant levels *P < 0.05, **P < 0.01, ns P > 0.05

Clay was negatively correlated with all the soil properties in both sites, while sand and silt showed positive and significant correlations. Most likely this is due to the differences in the controlling factors of microbial biomass, such as soil organic matter, management practices, and plant species composition, that influence the effects of clay content on soil properties. Previous studies have attributed the negative correlations of clay to soils that are well-weathered and formed under high precipitation (Côté *et al.*, 2000; Vejre *et al.*, 2003; McLauchlan *et al.*, 2006). In contrast, other studies have reported positive relationships between clay content and soil properties such as N and deduced that N is protected in aggregates rich in clay, which explains the high concentrations of N on clay particles (Waswa *et al.*, 2013).

Canonical correlation analysis (CCorA) was used to evaluate the multivariate relationships between two sets of soil variables: those that have high spatial dependence (SOC, N, and clay) and the other properties (P, K, Ca, Mg, Mn, pH and Zn) (Fig. 2.3). Factors 1 and 2 explained 64% of the total variance in soil properties. Results showed that the concentration of nutrients and micronutrients (P, K, Ca, Mg, Mn, and Zn) were positively related to SOC and pH, while P, K, Ca, Mg, and Zn content were negatively related to clay content. As expected, SOC showed a strong and positive correlation with total N.

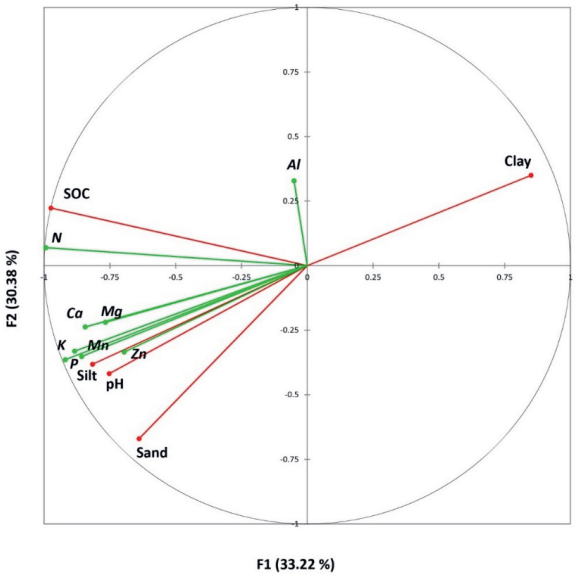


Fig. 2.3: CCA Ordination diagram based on the CCorA of the soil properties (soil granulometric fraction, soil organic carbon and pH) and data for soil elements contents (N, P, Ca, K, Mg, Al, Mn, and Zn).

2.3.3. Spatial distribution of key soil properties

A standard geostatistical analysis was carried out to describe SOC, N, and clay as main soil properties to describe the spatial patterns in the two study sites. The semivariogram parameters are presented in Fig. 2.4 and Table 2.4 and suggested that the soil parameters in both sites were best fitted with the theoretical spherical isotropic model. The range (A0) is considered as the distance beyond which the value of two soil samples can be statistically independent (Goovaerts, 1997; Rodríguez Martín *et al.*, 2009), then the correlation becomes negligible at a separation distance of about 3 km for clay in both sites (Table 2.4). The spatial correlation ranges were significantly wider for SOC (6 km) and N (5 km) in Bamendjou (Fig. 2.4a) than in Koutaba (Fig. 2.4b). Soil properties with wide ranges infer large area of influence and can be attributed to intrinsic factors such as soil formation (McKenzie, 2012; Liu *et al.*, 2013; Rodriguez Martin *et al.*, 2014; Rodríguez Martín *et al.*, 2014). On the other hand, soil properties with short distances are associated with extrinsic factors often attributed to anthropogenic activities such as agriculture and livestock rearing (Rodríguez Martín *et al.*, 2014). The results of this Chapter contrasted with the general geostatistical theory. We expected shorter rather than wider ranges in Bamendjou because of the influence of human activities. This may imply that the selected properties are more influenced by intrinsic factors such as parent materials. In Koutaba however, these soil properties showed shorter ranges due to influence of grazing activities. This led to the partial conclusion that grazing has more influence on the spatial structure of soil properties than agricultural activities. The nugget was higher for clay than for N and SOC in both sites indicating that the selected sampling distance may not have well captured its spatial dependence. On the other hand, the lowest nugget observed for N attests its low spatial variability within small distances.

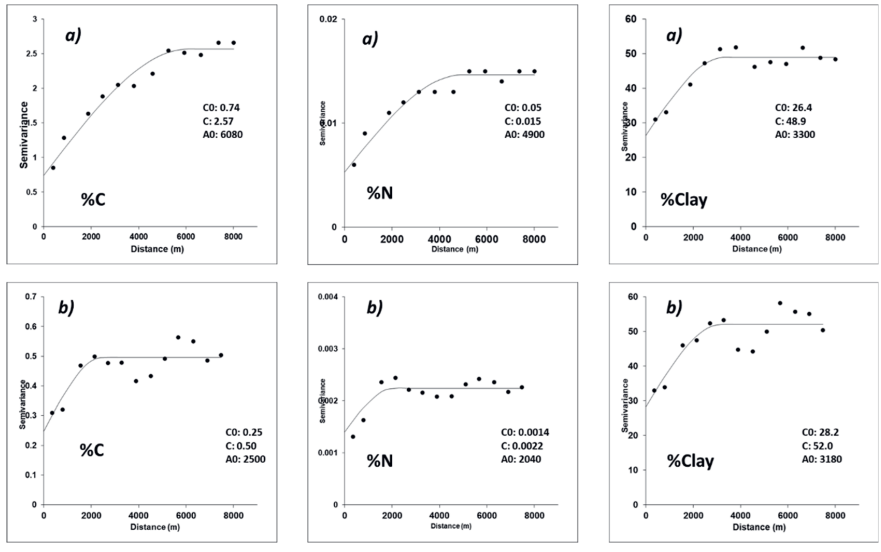


Fig. 2.4: Experimental variogram and spatial models for soil organic carbon (%C), soil nitrogen content (%N) and clay percentage (%Clay) in Bamendjou area (a) and Koutaba area (b).

The moderate spatial dependence of soil properties may be also controlled by extrinsic variations such as organic and inorganic fertilizer application, grazing, tillage and soil

management practices (Cambardella *et al.*, 1994; Mueller *et al.*, 2003). Previous studies have demonstrated that even in the case of weak spatial structure of soil properties, accurate maps are still obtainable but at the expense of intensive sampling exercises (Parfitt *et al.*, 2009). In this context, ordinary kriging (OK) was used for the spatial interpolation and production of maps of SOC, N, and clay.

Table 2.4: Parameters for semivariogram models for selected soil properties in Bamendjou and Koutaba sites.

Soil property	Site	Model structure	nugget (C_0)	Sill (C_0+C_1)	Range (A_0)	DSD (%)
SOC	Bamendjou	Nugget + Spherical	0.74	2.57	6080	28%
N	Bamendjou	Nugget + Spherical	0.005	0.015	4930	36%
Clay	Bamendjou	Nugget + Spherical	26.4	48.9	3300	53%
SOC	Koutaba	Nugget + Spherical	0.25	0.50	2500	49%
N	Koutaba	Nugget + Spherical	0.0014	0.0022	2040	61%
Clay	Koutaba	Nugget + Spherical	28.2	52.0	3180	54%

DSD = degree of spatial dependence; strong DSD (DSD≤ 25%); moderate DSD (25 < DSD ≤75%); weak DSD (DSD > 75%).

Soil parameters (SOC, N and Clay) maps obtained using the ordinary kriging are presented in Fig. 2.5 (Bamendjou) and Fig. 2.6 (Koutaba). The maps indicated areas with low SOC and N values in the central and western parts (A in Fig. 2.5), and high values in the eastern and southern parts (B in Fig. 2.5) of Bamendjou.

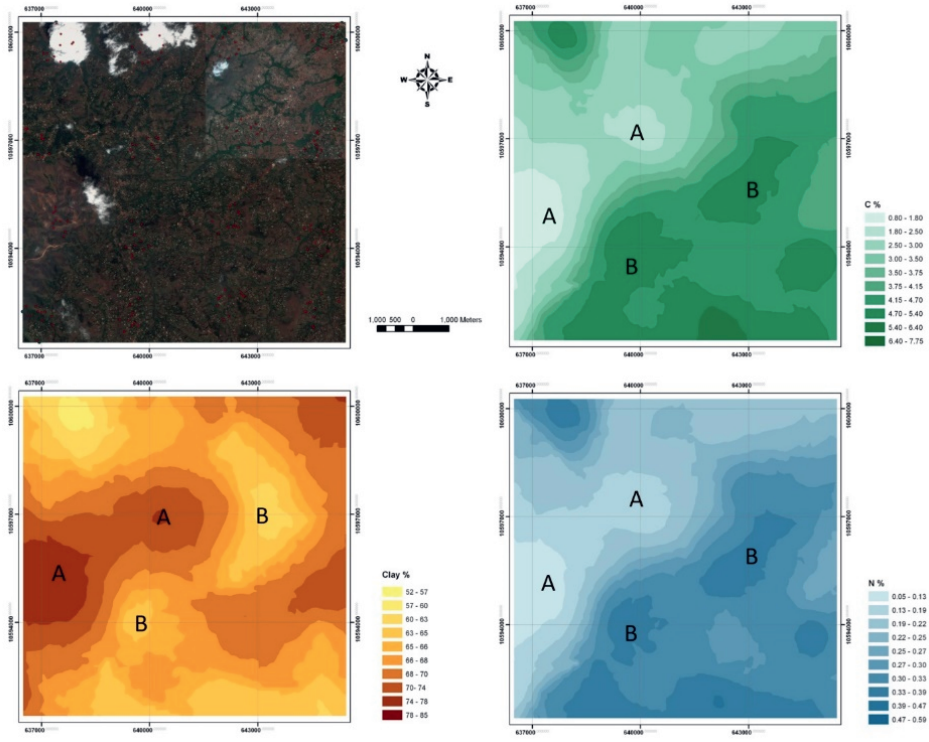


Fig. 2.5: Spatial distribution of C, N and clay interpolated by ordinary kriging and aerial photography showing the sampling points for the Bamendjou site.

In Koutaba, the maps indicated four areas of high concentration levels of SOC and N in the northern and eastern parts (A in Fig. 2.6) and low concentration in the central and southern parts of Koutaba. The low values of SOC and N in each site were found to be associated with high clay content (A in Fig. 2.5; B in Fig. 2.6) and agree with the correlation presented in Table 2.3 that revealed the negative relationship between clay and other soil properties.

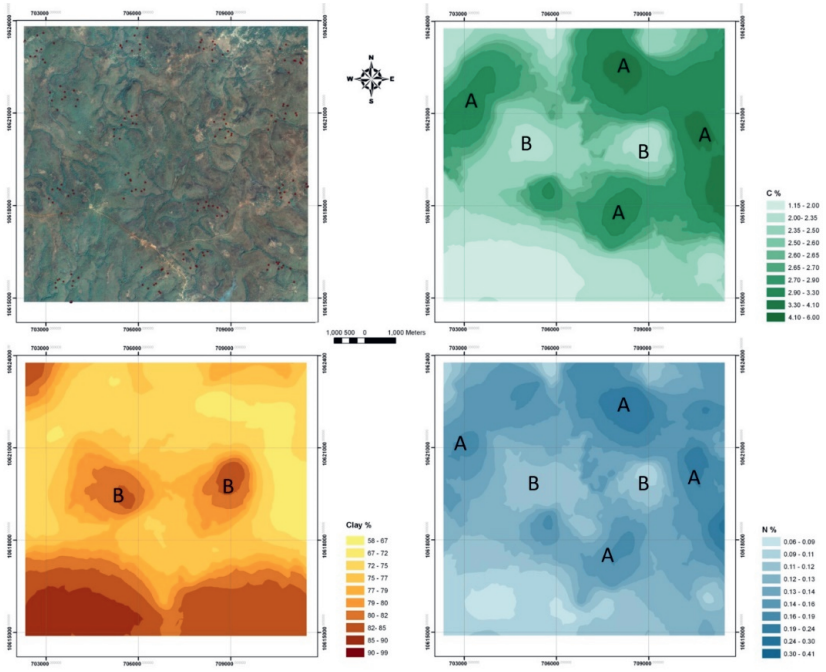


Fig. 2.6: Spatial distribution of C, N and clay interpolated by ordinary kriging and aerial photography showing the sampling points for the Koutaba site.

These variations in soil spatial patterns in both sites appeared to be associated with differences in land use types, landforms, and vegetation cover. Changes in land use types have the potential to generate corresponding change of microenvironment and consequently influence soil nutrients (McGrath *et al.*, 2001; Takoutsing *et al.*, 2015). The concentrations of SOC and N were ranked as forest > grassland > fallow > cropland > pasture in both sites. The land use types with more vegetation cover exhibited the highest SOC and N values attesting that land use types had a significant impact on spatial patterns and distribution of the properties. SOC and N concentrations in the soils are influenced by above-ground biomass, ground litter accumulation and decomposition, below-ground root mass and distribution, physical and biological conditions in the soil utilization (Yong-Zhong *et al.*, 2005; Han *et al.*, 2008; Li *et al.*, 2008). Additionally, the low SOC and N concentration in pastures is attributed to reduced vegetation cover following intensive grazing.

There was a smaller difference in clay content across land use types because the property is relatively little affected by changes in land management, supporting the assumption that soil conditions were more or less similar prior to the shifts in land use. However, clay content was found to be highest in pasture lands in both sites, because of low vegetation cover caused by overgrazing and soil compaction by animals. Our result showed that the effects of land use types, as well as their interaction on soil properties are significant, and corroborates with those previously reported under different conditions and countries (Chen *et al.*, 2007; Meersmans *et al.*, 2011; Meersmans *et al.*, 2012; Berhongaray *et al.*, 2013; Minasny *et al.*, 2013). This implies that land use is one of the main determinant factors for the spatial distribution of soil properties.

The spatial distribution of SOC and N showed similar spatial patterns and their concentrations decreased towards the uplands. In general, and from field observations, the spatial distribution of selected soil properties showed a well-defined pattern of higher concentrations in the lowlands and valleys and areas with permanent vegetation cover. This could be partly attributed to rainfall and erosion responsible for carrying nutrients from uplands to lowlands areas through runoff and leaching (Stone *et al.*, 1985; Haregeweyn *et al.*, 2008). With site-specific management in mind and considering the Bamendjou site, more attention should be paid to the central and western parts to increase the concentration of soil properties and to the eastern parts for protection and conservation of the current status of soil properties. For Koutaba, proper measures should be put in place to conserve the fertility potentials of the four key hotspot areas of high nutrient concentrations, as well as soil improvement measures to increase nutrient levels in low concentration areas. The spatial patterns of soil properties produced could be used as a guide to assess land degradation at landscape level, evaluate land suitability for crops, quantify the amount and types of fertilizers to be used by farmers and design appropriate soil and water conservation measures.

2.3.4. Implication of soil properties variability for smallholder farming systems

The results of the Chapter signposted that spatial variations observed with soil properties are influenced by inherent soil differences, management practices and topographic factors. The fact that Bamendjou was dominated by agricultural activities as opposed to Koutaba dominated by pasture implies that soil properties in both sites could differ spatially. From field observations, high concentrations of soil properties were also found in croplands located close to settlement. These farms tend to be more intensively managed with additional inputs due to their ease of access, and their proximity to households.

Most soil properties that are very sensitive to management practices usually have a shorter spatial range (Özgöz *et al.*, 2007). The differences observed in the range of soil properties in this Chapter, particularly SOC and N could therefore be attributed to both natural and anthropogenic factors (Tesfahunegn *et al.*, 2011). Previous studies have demonstrated the influence of the two factors on spatial patterns of soil properties (Tsegaye and Hill, 1998; Özgöz *et al.*, 2007). This implies that there is a need to develop and disseminate site-specific land management practices, instead of making blanket agronomic recommendations for all fields as is the case now. The results of this Chapter are very useful for smallholder farmers with low financial capacities who are called upon to improve soil productivity and crop yields using additional inputs. The findings could be used to enhance the efficient use of agricultural inputs such fertilizers by giving an indication of the concentration of nutrients (N, P and K) at precise location, or they can be used in selecting areas targeted for specific interventions such as large-scale reforestation.

Across the study site, some plots may not necessarily need additional nutrients as such, but need practices that improve soil structure, water holding capacity and limit nutrient leaching. Some other plots may have deficiency in one element (N) but not in others (P and K). The current fertilizer (NPK) recommendation is uniform for all agroecological zones in the country and has never been revised for decades with updated and reliable soil information. This has resulted in underapplication in areas with low nutrient levels and overapplication in areas with high nutrient levels. Areas receiving under-application of fertilizers will definitely not produce optimal crop yields; and for areas receiving over application, it is an unnecessary expense for farmers and an increase in environmental pollution (Bouma, 1997).

Sustainable intensification called for production systems that are able to increase crop yield, mitigate the risk of environmental pollution, maintain soil health and reduce cost of production for smallholder farmers. For areas in the study site with low N and SOC concentrations, appropriate N fertilizer should be applied to meet the crop nutrient requirements in order to achieve optimum yield. While for areas with high concentrations, application of N fertilizers may not be required for a period of time for both economic and environmental reasons. Management of soil nutrients based on site specific information is one of the important steps in achieving sustainable intensification of agriculture, efficient use of inputs, increased in crop yields and contribute to the fight against food insecurity at household and community levels.

2.4. Conclusions

Geostatistical analyses are very important for assessing the spatial structure of a given soil properties. The maps of properties revealed the variations across the study site by indicating areas of low and high nutrient concentrations. The major conclusions from this Chapter are the following: (1) Strong and positive correlations between soil properties and nutrients were observed except for clay content which negatively affects soil properties, (2) Land use types had significant effects on the concentration of soil properties, and particularly SOC and N decreased in the following order: forest > grassland > fallow > croplands > pastureland, (3) Bamendjou site had higher values for most of the soil properties compared to Koutaba, (4) a simple geostatistical analysis allowed detailed visualization and quantification of soil properties variation at landscape scale and provided a baseline for the comparison and evaluation of the effectiveness of land management interventions. Therefore, management of soil nutrients based on site specific soil information is one of the important steps in achieving sustainable intensification of agriculture.

Chapter 3

Accounting for analytical and proximal soil sensing errors in digital soil mapping

Digital soil mapping (DSM) approaches provide soil information by utilising the relationship between soil properties and environmental variables. Calibration of DSM models requires measurements that may often have substantial measurement errors which propagate to the DSM outputs and need to be accounted for. This Chapter applied a geostatistical-based DSM approach – regression kriging (RK) that incorporates quantified measurement error variances in the covariance structure of the spatial model, weights measurements in accordance with their measurement accuracies and assesses the effects of measurement errors on the accuracies of the resulted DSM outputs. The method was applied in the Western Cameroon, where soil samples from 480 locations were collected and analysed for pH, clay, and soil organic carbon (SOC) using conventional and mid infrared spectroscopy methods. Variogram parameters and regression coefficients were estimated using residual maximum likelihood under two scenarios: with and without taking measurement errors into account. The performance of the spatial models in the two scenarios were compared using validation metrics obtained with three types of cross-validation namely leave-one-out, leave-cluster-out and leave-sentinel-site-out cross-validation. Acknowledging measurement errors significantly impacted the regression coefficients and influenced the variogram parameters by reducing the nugget and sill variance for the three soil properties. Validation metrics including mean error, root mean square error and model efficiency coefficient were quite similar in both scenarios, but the prediction uncertainties were more realistically quantified by the models that account for measurement errors. There were relatively small absolute differences in predicted values of soil properties of up to 0.1 for pH, 1.6% for clay and 2 g kg^{-1} for SOC between the two scenarios. While differences in prediction maps and cross-validation metrics of predictions did not differ significantly between the two scenarios, substantial differences were obtained in prediction error standard deviation maps and in the evaluation of the prediction uncertainty. Relative differences in standard deviations were observed in some areas of the study area between the two scenarios of up to 0.08 for pH, 2.7% for clay and 0.5 g kg^{-1} for SOC. The best modelling approach would therefore be the one that accounts for measurement errors in soil observations. The findings overall indicated that the additional investment in quantification of measurement error and the incorporation in the spatial models is worth the effort, as shown by the improvement in the quantification of the prediction uncertainties. These findings also emphasised the need of incorporating measurement errors in further studies not only in a geostatistical-based DSM approach, but also when using non-linear machine learning regression methods to improve uncertainty quantification, particularly when spectral data are used as the main soil data source.

A version of this chapter has been published in Takoutsing, B., Heuvelink, G.B.M., Stoorvogel, J.J., Shepherd, K.D., Aynekulu, E., 2022. Accounting for analytical and proximal soil sensing errors in digital soil mapping. *European Journal of Soil Science* 73(2), e13226.

3.1. Introduction

Soil spatial information is crucial to address global issues such as food security, climate change and land degradation (McBratney *et al.*, 2003; Shepherd *et al.*, 2015). Soil information is also important at various scales in helping policy makers, extension agents, and land users whose decisions impact land management interventions, particularly those designed to support agricultural production (Stoorvogel *et al.*, 2015). The successes of digital soil mapping (DSM) in providing such information are ascribed to recent technological and computational advances, availability of high-resolution remote sensing data, advancement of proximal soil sensing, and the development of machine-learning algorithms (MLA) (Minasny and McBratney, 2016). The quest for large soil datasets for the development and application of DSM models, as well as the increasing demand for soil spatial information to efficiently manage agronomic inputs such as fertilizers (Stoorvogel *et al.*, 2015), has led to the increased use of proximal soil sensing (PSS) (Viscarra Rossel *et al.*, 2016a). Diffuse reflectance spectroscopy is a rapid and low-cost method to generate soil measurements for use in DSM (Shepherd *et al.*, 2015). Despite the potentials of PSS in generating larger amounts of spatially explicit soil data, soil spectral estimation of soil properties tends to have larger measurement errors than wet chemistry measurements that eventually propagate to DSM outputs (Heuvelink, 2018; Somarathna *et al.*, 2018).

One aspect of DSM that has received little attention so far is the errors in soil measurements used for calibration and prediction. Although modellers may be aware that measurements are not error-free, most DSM studies ignore this fact and consider only the limited predictive power of environmental covariates and spatial interpolation error as sources of uncertainties. Analysis and quantification of uncertainties in soil measurements is a subject of interest and should be incorporated in DSM, to weigh measurements in accordance with their accuracy and provide end-users with reliable information about the accuracy of the prediction maps (Arrouays *et al.*, 2017; Heuvelink, 2018). The lack of consideration of uncertainties may lead to suboptimal models and systematic underestimation or overestimation of the uncertainties of DSM outputs (Poggio *et al.*, 2016; Heuvelink, 2018). Decisions based on suboptimal models and poor-quality maps whose accuracy is overestimated may have extensive and profound impacts on the design of land management interventions (**Chapter 2**), as well as on soil amendment practices, such as fertilizer application. End-users may increase their investments in obtaining accurate soil maps, for instance by increasing soil sampling density or getting better covariates, if they are reliably informed about the accuracy of the available maps. Recent studies have demonstrated that measurement errors may have significant impacts on subsequent spatial analyses (Somarathna *et al.*, 2018). However, to the best of our knowledge, there are no published studies that explicitly considered how uncertainty in PSS data affects DSM outputs.

The recent expansion of DSM approaches has resulted in the shift from geostatistics to machine learning (ML). Although ML has overtaken kriging to become the most popular DSM method due to its flexibility and tendency to improve predictions (Hengl *et al.*, 2015; Veronesi and Schillaci, 2019), kriging has important advantages over ML. First, kriging can better account for spatial autocorrelation than ML, which is not a spatial model (Hengl *et al.*, 2018b). Second, it yields an interpretable parametric model of the soil spatial variation. Third, kriging does not need as large a dataset as ML for calibration and can be used in a case of just 100 measurements or more. Fourth, kriging does not only characterize the prediction uncertainty using a prediction error variance, but it also quantifies the spatial correlation in the kriging prediction errors. At best, ML characterizes the prediction error at prediction locations, for example using Quantile Regression Forests (Vaysse and Lagacherie, 2017), but not the spatial correlation of that error, which is needed to quantify uncertainties of spatial averages. Fifth, from a statistical perspective, it is feasible to incorporate measurement errors in model

calibration and prediction with kriging (Knotters *et al.*, 1995; Chilès and Delfiner, 2012; Viscarra Rossel *et al.*, 2016b; Viscarra Rossel and Brus, 2018).

There is need for DSM approaches that account for uncertainties in soil measurements generated using analytical and PSS methods. While geostatistical methods are available to handle this in a realistic manner, they have not so far been used. Therefore, the objectives of this Chapter were to quantify measurement errors in analytical and PSS soil data, incorporate them into a state-of-the art geostatistical method for spatial interpolation and compare the results with a case in which measurement errors are ignored. We illustrate the methods with a case study in which we map pH, clay, and soil organic carbon for a study area in the Western Highlands of Cameroon. More specifically, we use regression kriging supported by restricted maximum likelihood parameter estimation.

3.2. Materials and methods

3.2.1. Study area

The study area for **Chapter 3** covers parts of the West region of Cameroon that spans 1,053 km² and features the major characteristics of the highlands of Cameroon (Fig. 3.1), dominated by subsistence agricultural systems. The climate is of tropical humid mountain type with average rainfall that varies from 1000 to 2000 mm per year. The mean daily minimum and maximum temperatures are 18 and 30 °C, respectively. The topography is undulating with altitudes ranging between 600 and 1800 meters above sea level, and the vegetation is of savannah type with patches of gallery and montane forests. The soils are predominantly Ferralsols, of volcanic origin and suitable for the production of a range of annual and perennial crops, though soil tends to be generally acidic (Takoutsing *et al.*, 2016).

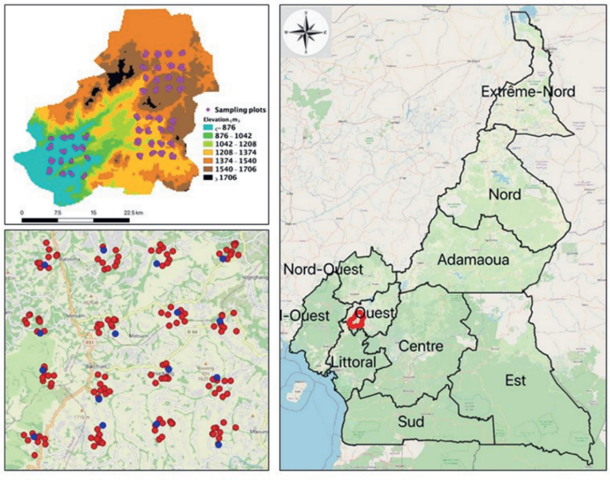


Fig. 3.1: Map of Cameroon showing the study area. Soil sampling was done in three 10 km × 10 km sentinel sites. Each sentinel site has 160 sampling locations (violet dots). Bottom-left panel zooms in on the most northern sentinel site (red dots represent spectral data, blue dots analytical data).

3.2.2. Sampling design

The study area was sampled using a spatially stratified and hierarchical sampling approach based on the concept of 10 km × 10 km sentinel sites (Vågen and Winowiecki, 2020). The soil sampling was limited to the sentinel sites, which is suboptimal for kriging since parts of the study area are poorly covered (Fig. 3.1), but this was done to save travel time and accessibility costs. The site locations were established using convenience sampling, while accounting for land cover/land use and topography of the study area to capture the variation of landscape conditions (i.e., feature space coverage). The sampling design for establishing the sentinel sites was initially conceived for a larger area that covers the entire southern parts of the Republic of Cameroon, but for this **Chapter 3**, we used only three sites namely Bamendjou, Bana and Kekem located in the study area. As explained in detail in Vågen and Winowiecki (2013), each site was subdivided into 16 square 2.5 km × 2.5 km tiles within which random centroid locations for clusters within each tile were generated but buffered to avoid overlapping with neighbouring tiles. Each cluster consists of 10 circular sampling plots (1,000 m^2 each) randomly located within a 1 km^2 circular area randomly placed within the tile, giving 160 sampling plots per sentinel site.

Within each of the sampling plots, four subplots were established, one at the centre of the plot and the three others surrounding the centre plot at 12.2m and disposed at 120°. Topsoil (0 – 20cm) samples (approx. ~ 500g) were collected at the four subplots, pooled together and thoroughly mixed to obtain a composite sample for each plot, yielding 160 soil samples per site and 480 for the entire study area.

3.2.3. Soil data

Four hundred and eighty (480) soil samples were collected (10 samples from each cluster, 160 samples from each site) between 2015 and 2017 within the framework of the Cameroon land health project (**Chapter 2**). One out of ten soil samples in each cluster were randomly selected and subjected to conventional laboratory analyses for pH, clay content and SOC and referred to as "reference samples" (n=48). Soil pH was determined using a pH meter (1:2.5 soil to water ratio), clay by the hydrometer method, and SOC concentration using the potassium dichromate oxidation method.

Next, all samples (n=480) were processed and analysed by mid-infrared spectroscopy (MIRS), following standard procedures described in (Terhoeven-Urselmans *et al.*, 2010). The measured MIR reflectances were first converted to apparent absorbance units [log (1 / Reflectance)] and then pre-processed with the Savitzky–Golay smoothing method (Sila *et al.*, 2016). The reference samples were used for both calibration and validation of the prediction models. All spectral replicates for each sample were averaged and regression models were built to relate the processed spectra to the reference samples using Partial Least Squares Regression (PLSR). The PLSR used spectra data as independent variables and the analytical data as dependent variables. The fitted regression models were used to predict the targeted properties of all the samples. All the 48 paired observations of analytical and spectral data were used for the calibration of the PLSR model and validation metrics were computed using leave-one-outcross-validation. The accuracies of the models were assessed using the mean error (ME), the root mean squared error (RMSE) and the model efficiency coefficient (MEC). See Section 3.2.5.6 for definitions of these accuracy metrics. The fitted PLSR models were applied to obtain soil property predictions based on MIR data at all 480 locations. Soil measurements obtained through conventional laboratory methods are referred to as analytical data while predicted soil values using MIR spectroscopy/PLSR are referred to as spectral data.

3.2.4. Environmental variables

Soil spatial variation is influenced by environmental factors including climate (e.g., precipitation and temperature), organisms (e.g., land cover), relief (e.g., terrain attributes), and parent materials (McBratney *et al.*, 2003). In this Chapter, we derived these factors from several spatial datasets to effectively represent each key soil-forming factor. We initially considered an extensive stack of over 170 environmental layers downloaded from the ISRIC repository. The relief represented by a digital elevation model was obtained from SRTM, from which various topographic parameters were derived (e.g., elevation, slope, and topographic wetness index). Land use/cover classes were obtained from the global land cover map (GlobeLand30) for the year 2015. The MODIS near and mid infrared reflectance (NIR, MIR), and Enhanced Vegetation Index (EVI) products were derived using a stack of MOD13Q1 products. Climatic data made up of annual temperature and precipitation averages were obtained from the CHELSA Bioclimatic images (https://chelsa-climate.org/bioclim/). Landform classes (breaks/foothills, flat plains, high mountains/deep canyons, hills, low hills, low mountains, smooth plains) were based on the USGS's Map of Global Ecological Land Units. Since the environmental layers were from different sources, they were all resampled to 250 m spatial resolution before they were used as independent variables in the spatial models.

3.2.5. Statistical modelling

The DSM model selection, calibration and prediction were fully implemented in the R environment for statistical computing (R Core Team, 2016). The process consisted of the following main steps (Fig. S3.1): 1) model definition, 2) quantification of measurement errors in analytical and spectral data, 3) model selection, 4) model calibration (parameter estimation), 5) spatial prediction, and 6) cross-validation. The sub-sections below explain these six steps for two Scenarios. In Scenario 1, measurement errors are ignored, while Scenario 2 accounts for measurement error variances.

3.2.5.1. Model definition

In Regression kriging (RK), the dependent variable is modelled as the sum of a deterministic trend and a spatially correlated stochastic residual as described in Chapter 9 of (Webster and Oliver, 2007)

$$Z(\boldsymbol{s}) = m(\boldsymbol{s}) + \varepsilon(\boldsymbol{s}) \; = \; \sum_{j=0}^{p} \beta_j \cdot x_j(\mathbf{s}) + \; \varepsilon(\boldsymbol{s}), \qquad \boldsymbol{s} \in A \qquad (3.1)$$

Here, $Z(\boldsymbol{s})$ represents the soil property of interest at any location $\boldsymbol{s}$ in the geographic domain A, m is the trend, taken as a linear combination of covariates x_j $(j = 1, \ldots, p)$, the β_j are regression coefficients (β_0 is the intercept, by setting $x_0(\boldsymbol{s}) = 1$ for all $\boldsymbol{s} \in A$), and ε is a zero-mean, normally distributed, stationary stochastic residual, whose spatial covariance structure is defined by a variogram γ. The normal distribution and residual stationarity are stringent assumptions that need to be justified and possibly adjusted in real-world applications. For instance, in the case study we will apply log-transformation to SOC before invoking the normal distribution assumption.

We also have observations y_i of the dependent variable at a finite number of locations $\boldsymbol{s}_i$ $(i = 1, \ldots, n)$ in A. These are interpreted as realizations of an observation process:

$$Y_i = Z(\boldsymbol{s}_i) + \delta_i, \qquad i = 1, \ldots, n \tag{3.2}$$

where δ_i are measurement errors, assumed jointly normally distributed with zero mean and having an $n \times n$ variance-covariance matrix $\boldsymbol{V}$. Note that the zero-mean assumption signifies that we ignore systematic errors in soil measurements.

3.2.5.2. Quantification of measurement errors in analytical and spectral data

In this **Chapter 3**, we considered errors in the measurement of analytical and spectral data of the soil samples. By assuming that measurement errors of different soil samples are uncorrelated, the variance-covariance matrix V reduces to a diagonal matrix so that errors in soil measurements are completely summarized by their variances. These variances were assumed constant for a given measurement method but were different for analytical and spectral measurements. They were assessed as follows. Let $\boldsymbol{Z_T}$ be the 'true' value of the soil property, $\boldsymbol{Z_A}$ the value of the soil property obtained through laboratory analysis and $\boldsymbol{Z_S}$ the value of the soil property obtained using proximal soil sensing (i.e., the spectroscopy model predictions of ZA). We now have:

$$\sigma_S^2 = var\,(Z_S - Z_T) = var(Z_S - Z_A + Z_A - Z_T)$$

$$= var(Z_S - Z_A) + var(Z_A - Z_T) + 2\,cov(Z_S - Z_A, Z_A - Z_T)$$

$$= var(Z_S - Z_A) + var(Z_A - Z_T) \tag{3.3}$$

where the latter equality holds because the PLSR fitting error is not correlated with the laboratory measurement error.

For laboratory data, the measurement error variances $\sigma_A^2 = var(Z_A - Z_T)$ were derived using laboratory repeatability procedures (Libohova *et al.*, 2019). In this context, repeatability describes the variation of a mean result obtained in successive measurements of the same sample analysed in the same laboratory under the same conditions (Libohova *et al.*, 2019). For this **Chapter 3**, soil samples were analysed following standard processing procedures for pH (10 duplicates), clay (33 duplicates) and SOC (10 duplicates). Next, the standard deviation of the analytical measurement error σ_A was estimated from the differences between the measurements on the same sample. We found unrealistic values (outliers) for clay which might be caused by other sources of errors (i.e., blunders, gross errors). These outliers were removed before computing the estimate of σ_A so that our focus remained on real measurement errors.

For spectral data, the measurement error variances σ_S^2 were obtained using Eq. 3.3 by adding up σ_A^2 and $var(Z_S - Z_A)$. The latter was estimated from the residual variance of the PLSR models. For the case of SOC that showed positive skewness, the analytical data were first log-transformed (logSOC) before running the PLSR. The PLSR residual variance was estimated using:

$$var(Z_S - Z_A) \cong \frac{1}{n}\sum_{i=1}^{n}(z_A(\boldsymbol{s}_i) - \hat{z}_A(\boldsymbol{s}_i))^2 \quad (3.4)$$

where $\hat{z}_A(\boldsymbol{s}_i) = z_S(\boldsymbol{s}_i)$ is the PLRS predicted soil property at location $\boldsymbol{s}_i$ and n is the number of paired observations of analytical and spectral data (i.e. $n = 48$ in the case study). Both the analytical and spectral measurement error variances estimated above were incorporated in the kriging model as shown in the sections below.

3.2.5.3. Model selection

Based on literature, pedological information, and their relevance to specific soil properties, 50 layers were selected out of 170 initial layers, to represent key soil-forming factors. These covariates were processed and overlaid with sample locations to construct a matrix of covariate values for each sample point. Initially, a correlation analysis was performed to reduce redundancy between the selected layers. Some pairs of environmental variables were highly correlated with each other. For statistical models it is preferred that environmental variables retained are weakly correlated with each other, because it increases the potential for fitting a combination of environmental variables to explain the variation in the soil properties. Only layers with a correlation coefficient <= 0.75 with all other layers were retained for subsequent analysis (Hanchuan *et al.*, 2005). For each pair of covariates correlated above the set threshold, we arbitrarily retained the first one in alphabetical order for inclusion in the model. This reduced the number of covariates to 23. Next, the best combination of covariates for each soil property were selected by combined forward and backward stepwise regression using the Bayesian Information Criterion (BIC) (Gao and Song, 2010). Regression models, their coefficients, and p-values were examined to derive quantitative data on the relative roles and behaviour of each covariate in the model. During the model selection procedure, measurement errors in soil data and spatial correlation were ignored.

3.2.5.4. Model calibration by REML

Using matrix notation for compactness, Equations 3.1 and 3.2 combined can be written as $\boldsymbol{Y} = \boldsymbol{X} \cdot \boldsymbol{\beta} + \boldsymbol{\varepsilon} + \boldsymbol{\delta}$, where $\boldsymbol{Y}$ is an n-vector, $\boldsymbol{X}$ an $n \times (p+1)$ matrix, $\boldsymbol{\beta}$ a vector of $p+1$ regression coefficients, and $\boldsymbol{\varepsilon}$ and $\boldsymbol{\delta}$ vectors containing the stochastic residual and measurement errors at the n observation locations, respectively. The parameters of this model are $\boldsymbol{\beta}$, the parameters of the variogram γ (i.e., the nugget, sill, and range, assuming a known shape), and the error variance-covariance matrix $\boldsymbol{V}$ (i.e., σ_A^2 and σ_S^2). Estimation of σ_A^2 and σ_S^2 was explained in Section 3.2.5.2. The variograms were all fitted with exponential models. Note that the measurement error parameters were estimated and fixed prior to estimation of other parameters through REML.

It is a common practice in geostatistics to estimate γ using the method of moments. However, this is suboptimal and has additional bias problems in case of regression kriging, where $\boldsymbol{\beta}$ also needs to be estimated (Lark *et al.*, 2006). For this Chapter, variogram parameters and regression coefficients were estimated using restricted maximum likelihood (REML). We give a brief description and refer to Section 9.2.1 in Webster and Oliver (2007) for details. REML is computationally demanding in a case of large datasets but was quick in our case study, where we have 480 observations.

Since all stochastic components of the model are normally distributed, $\boldsymbol{Y}$ has a multivariate normal distribution, with probability density:

$$f_Y(\boldsymbol{y}) = (2\pi)^{-n/2} \cdot |\boldsymbol{C} + \boldsymbol{V}|)^{-1/2} \cdot exp\left(-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{X\beta})^T(\boldsymbol{C} + \boldsymbol{V})^{-1}(\boldsymbol{y} - \boldsymbol{X\beta})\right) \quad (3.5)$$

where $\boldsymbol{y}$ is the vector of observations y_i $(i = 1, \ldots, n)$ and $\boldsymbol{C}$ is the variance-covariance matrix of $\boldsymbol{\varepsilon}$, derived from the variogram γ and the distances between the observation locations. The idea of maximum likelihood is to choose the model parameters (i.e., $\boldsymbol{\beta}$ and the nugget, sill, and range of γ) such that the probability density $f_Y(\boldsymbol{y})$ is maximized. REML is a particular form of maximum likelihood estimation that estimates the model parameters in two steps. First the variogram parameters are estimated by maximising a conditional likelihood, in which the dependence of the variogram parameters on the regression coefficients is removed (Lark and Cullis, 2004). For this a numerical optimisation technique is used. In a second step the regression coefficients are estimated, conditional on the already estimated variogram parameters. This can be done analytically, because given the variogram parameters, the maximum likelihood estimate of the vector of regression coefficients equals the conventional generalized least squares solution:

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T(\boldsymbol{C} + \boldsymbol{V})^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T(\boldsymbol{C} + \boldsymbol{V})^{-1}\boldsymbol{y} \quad (3.6)$$

Note that the model calibration returns only estimates of the 'true' regression coefficients and variogram parameters. To simplify the subsequent analysis, we will ignore these estimation errors and assume that all model parameters are perfectly known. While it is not difficult to include estimation errors of regression coefficients in kriging (Brus and Heuvelink, 2007), it is much more difficult to account for variogram estimation errors in prediction.

3.2.5.5. Spatial prediction

Our aim is to predict $Z(\boldsymbol{s}_0)$ given the measurements y_i and the covariates $x_j(\boldsymbol{s}_0)$. Note that the prediction location $\boldsymbol{s}_0$ could be any location in the study area for which covariates are available. In practice, the prediction locations are the nodes of a fine grid that are visited one by one. Under the assumptions made, the best prediction (i.e., the one that has the smallest expected squared prediction error) is the conditional mean:

$$\hat{Z}(\boldsymbol{s}_0) = E[Z(\boldsymbol{s}_0)|\boldsymbol{Y} = \boldsymbol{y}] = \boldsymbol{x}_0{}^T\boldsymbol{\beta} + \boldsymbol{c}_0^T(\boldsymbol{C} + \boldsymbol{V})^{-1}(\boldsymbol{y} - \boldsymbol{X\beta}) \quad (3.7)$$

where $\boldsymbol{x_0}$ is a $(p + 1)$-vector of covariates at $\boldsymbol{s}_0$ and $\boldsymbol{c_0}$ is the vector of covariances between $Z(\boldsymbol{s}_0)$ and the $Z(\boldsymbol{s}_i)$. The prediction is unbiased and has prediction error variance:

$$\sigma_K^2(\boldsymbol{s}_0) = Var\left(\hat{Z}(\boldsymbol{s}_0) - Z(\boldsymbol{s}_0)\right) = c_{00} - \boldsymbol{c}_0^T(\boldsymbol{C} + \boldsymbol{V})^{-1}\boldsymbol{c}_0 \quad (3.8)$$

where c_{00} is the variance of $Z(\boldsymbol{s}_0)$ (i.e. the sill of γ). Note that regression coefficient estimation errors are ignored in Equations. 3.7 and 3.8.

Spatial interpolation techniques such as kriging are sensitive to skew distributions due to the high impact of extreme values on variogram parameter estimation that may render outputs unstable. From the summary statistics presented in Section 3.3.1, soil pH and clay satisfactorily met the assumption of a normal distribution. SOC exhibited a positively skewed distribution, and the normal distribution assumption was made after a log-transformation to logSOC. The back-transformed estimate of SOC and local variance for each interpolated location was

obtained as described in Section 8.10 of Webster and Oliver (2007) and in Laurent (1963). RK uses the regression model and the variogram parameters to estimate the values of soil properties at all locations and generate maps of kriging predictions, and those of the kriging standard deviations for both scenarios. The kriging standard deviation was obtained by taking the square root of the kriging variance. In addition, we subtracted the final prediction results of Scenario 1 from Scenario 2 using raster calculation and generated the prediction difference maps between the two scenarios. Recall that Scenario 1 ignores measurement errors. In other words, it uses the model calibration and prediction approach described above but enforces that both σ_A^2 and σ_S^2 are zero.

3.2.5.6. Cross-validation

Because the sampling locations were clustered in three sentinel sites within the study area, a conventional cross-validation might produce overoptimistic results (Roberts *et al.*, 2017). The accuracy of the model predictions was therefore assessed using leave-one-out, leave-cluster-out and leave-sentinel-site-out cross-validation. Density scatter plots were used to compare the predicted values in the two scenarios at validation points. For each soil property, we derived three validation metrics: the Mean Error (ME), the Root Mean Squared Error (RMSE) and the Model Efficiency Coefficient (MEC) (Janssen and Heuberger, 1995). The MEC is equal to one minus the ratio between the residual sum of squares and the total sum of squares, as defined in Equation. 3.11. In hydrology it is known as the Nash-Sutcliffe Model Efficiency (Nash and Sutcliffe, 1970). The MEC equals 1 in case of a perfect model, while it is 0 for a model that is as good as taking the mean of all observations as a prediction. MEC can be negative for models that are severely biased. To evaluate the kriging standard deviation, the prediction interval coverage probability (PICP) was computed and used to derive accuracy plots for the leave-one-out cross-validation case. The section below describes how these cross-validation metrics were computed in case of uncertain validation data.

As before, let $z_T(\boldsymbol{s})$ be the true value of the soil property at location $\boldsymbol{s}$, and let $z_M(\boldsymbol{s})$ be the value of the soil property at $\boldsymbol{s}$ obtained through a measurement (using either laboratory or spectral analysis). Note that we use lower-case notation here because we now treat these as the actual values, not as random variables. Let $\hat{z}(\boldsymbol{s})$ be the prediction of the soil properties obtained using regression kriging in cross-validation mode and $\sigma_K^2(\boldsymbol{s})$ the associated kriging variance. In other words, $\hat{z}(\boldsymbol{s})$ and $\sigma_K^2(\boldsymbol{s})$ are derived as explained in Section 3.2.5.5, using that part of the measurements that were not put aside for validation. This is done for all measurement locations $\boldsymbol{s}_i$, $i = 1,..,n$.

Given n validation locations we derive the Mean Error (ME) as

$$ME = \frac{1}{n}\sum_{i=1}^{n}\hat{z}(\boldsymbol{s}_i) - z_T(\boldsymbol{s}_i) = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{z}(\boldsymbol{s}_i) - z_M(\boldsymbol{s}_i)\right) + \frac{1}{n}\sum_{i=1}^{n}\left(z_M(\boldsymbol{s}_i) - z_T(\boldsymbol{s}_i)\right)$$

$$\approx \frac{1}{n}\sum_{i=1}^{n}\left(\hat{z}(\boldsymbol{s}_i) - z_M(\boldsymbol{s}_i)\right) \quad (3.9)$$

where the latter approximation holds because we assume that the measurement method has no systematic error.

For the Mean Squared Error (MSE) we get

$$\frac{1}{n}\sum_{i=1}^{n}\left(\hat{z}(\boldsymbol{s}_i) - z_M(\boldsymbol{s}_i)\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{z}(\boldsymbol{s}_i) - z_T(\boldsymbol{s}_i) + z_T(\boldsymbol{s}_i) - z_M(\boldsymbol{s}_i)\right)^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left(\hat{z}(\boldsymbol{s}_i) - z_T(\boldsymbol{s}_i)\right)^2 + \frac{1}{n}\sum_{i=1}^{n}\left(z_T(\boldsymbol{s}_i) - z_M(\boldsymbol{s}_i)\right)^2$$

$$+ \frac{2}{n}\sum_{i=1}^{n}\left(\left(\hat{z}(\boldsymbol{s}_i) - z_T(\boldsymbol{s}_i)\right)\cdot\left(z_T(\boldsymbol{s}_i) - z_M(\boldsymbol{s}_i)\right)\right)$$

$$\approx \frac{1}{n}\sum_{i=1}^{n}\left(\hat{z}(\boldsymbol{s}_i) - z_T(\boldsymbol{s}_i)\right)^2 + \frac{1}{n}\sum_{i=1}^{n}\left(z_T(\boldsymbol{s}_i) - z_M(\boldsymbol{s}_i)\right)^2 \quad (3.10)$$

where the latter approximation holds because the kriging prediction error and the measurement error are uncorrelated. Eq. 10 shows that an estimate of the $MSE = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{z}(\boldsymbol{s}_i) - z_T(\boldsymbol{s}_i)\right)^2$ is obtained by subtracting the measurement error variance (i.e., a weighted average of σ_A^2 and σ_S^2, with weights equal to the fraction of analytical and spectral validation measurements, respectively) from the MSE computed on error-contaminated validation data. In practice, we are more interested in the RMSE than the MSE. This is derived by taking the square root after the measurement error variance has been subtracted from the MSE computed on error-contaminated validation data.

Similarly, the MEC under uncertain validation data can be derived as

$$MEC = 1 - \frac{\sum_{i=1}^{n}\left(\hat{z}(\boldsymbol{s}_i) - z_T(\boldsymbol{s}_i)\right)^2}{\sum_{i=1}^{n}(z_T(\boldsymbol{s}_i) - \bar{z}_T)^2} \approx 1 - \frac{\sum_{i=1}^{n}\left(\hat{z}(\boldsymbol{s}_i) - z_M(\boldsymbol{s}_i)\right)^2 - \sum_{i=1}^{n}\left(z_M(\boldsymbol{s}_i) - z_T(\boldsymbol{s}_i)\right)^2}{\sum_{i=1}^{n}(z_M(\boldsymbol{s}_i) - \bar{z}_M)^2 - \sum_{i=1}^{n}\left(z_T(\boldsymbol{s}_i) - z_M(\boldsymbol{s}_i)\right)^2} \quad (3.11)$$

where $\bar{z}_T = \frac{1}{n}\sum_{i=1}^{n} z_T(\boldsymbol{s}_i)$ and $\bar{z}_M = \frac{1}{n}\sum_{i=1}^{n} z_M(\boldsymbol{s}_i)$ and where as before $\sum_{i=1}^{n}\left(z_T(\boldsymbol{s}_i) - z_M(\boldsymbol{s}_i)\right)^2$ is derived from a weighted average of σ_A^2 and σ_S^2.

The PICP evaluates how often the validation data are within a $(1-\alpha)$ prediction interval for various values of α (Shrestha and Solomatine, 2006). Assuming a normal distribution for the kriging prediction error and analytical error, these prediction intervals can be derived from the variances of both errors. Since:

$$Z_M(\boldsymbol{s}) = \hat{Z}(\boldsymbol{s}) + (Z_M(\boldsymbol{s}) - Z_T(\boldsymbol{s})) + \left(Z_T(\boldsymbol{s}) - \hat{Z}(\boldsymbol{s})\right) \quad (3.12)$$

and measurement and kriging errors are uncorrelated we have that $Z_M(\boldsymbol{s})$ should lie between $\hat{Z}(\boldsymbol{s}) - z_{1-\alpha/2}\cdot\sqrt{\sigma_M^2 + \sigma_K^2(\boldsymbol{s})}$ and $\hat{Z}(\boldsymbol{s}) + z_{1-\alpha/2}\cdot\sqrt{\sigma_M^2 + \sigma_K^2(\boldsymbol{s})}$ in $(1-\alpha)\times 100\%$ of all cases. Here, $z_{1-\alpha/2}$ refers to the $1-\alpha/2$ quantile of the standard normal distribution. Note that σ_M^2 equals σ_A^2 in case of analytical measurements and equals σ_S^2 in case of spectral measurements. A plot of the PICP against α boils down to an accuracy plot, which visualizes the assessment of the quality of the estimated prediction uncertainty (Goovaerts, 2001; Wadoux *et al.*, 2018).

3.3. Results

3.3.1. Mid infrared spectroscopy models

Fig. 3.2 shows PLSR predictions against wet chemistry observations for the 48 soil samples where both types of analysis were carried out. Note that SOC data were log-transformed to logSOC prior to running the PLSR (see also Section 3.3.2). Accurate predictive models were obtained for soil pH (ME = 0.004, RMSE = 0.219, MEC = 0.87, clay (ME = 0.216, RMSE = 5.47, MEC = 0.83) and logSOC (ME = 0.006, RMSE = 0.192, MEC = 0.82). Note that these metrics were computed on only 48 observations and are only approximations of the population validation metrics.

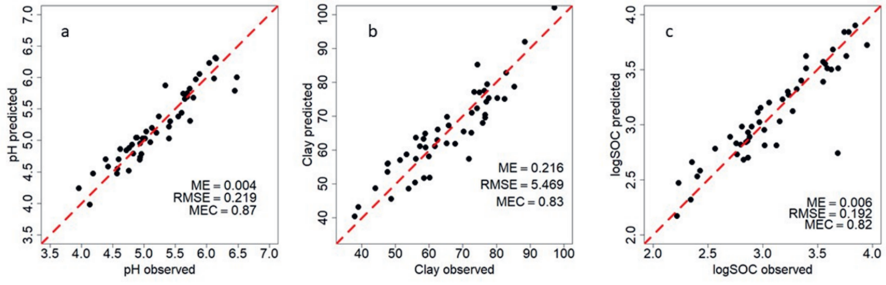


Fig. 3.2: Scatter plots of PLSR predictions against observations: a) pH, b) clay, c) logSOC. Red dashed lines represent the 1:1-line.

The soil analytical and spectral data put together constitute the dataset for the **Chapter 3**. For samples analysed with both conventional and MIRS, only analytical data were retained, resulting in 48 analytical and 432 spectral observations and making a total of 480 observations. Although the accuracies of the predictive models are acceptable, spectrally soil estimated data are not as accurate as the analytical data because the PLSR prediction error adds to the uncertainty.

3.3.2. Descriptive statistics

The basic statistical parameters for both datasets and the merged dataset are summarized in Table 3.1. Considering the total dataset, soil pH was low and varied from strong acidity to near neutral (3.54 to 6.91) with a mean of 5.23, which falls within the optimum range for the production of priority crops in the tropics, such as maize (i.e., pH between 5.5 and 6.5). Textural analysis revealed that the study area is dominated by clay-rich soils with a mean of 65.5% and values ranging from 37.9 to 100%. SOC concentrations ranged from 6.7 to 84.5 g kg^{-1} with a mean of 26.4 g kg^{-1}. The analysis of the distribution of the soil properties indicates that pH and clay approximated normality, while SOC values were positively skewed (skewness coefficient of 1.12), as shown by the histogram (Fig. S3.2). SOC was therefore log-transformed to logSOC, which had a much more symmetric distribution. The statistical modelling hereafter was applied to pH, clay and logSOC.

Table 3.1: Summary statistics of soil properties for the analytical (n = 48), spectral (n = 432) and merged data sets (n = 480).

Variable	Min.	Mean	Max.	St. dev.	CV (%)	Skewness
pH						
Analytical	3.96	5.21	6.48	0.61	11.7	0.13
Spectral	3.54	5.23	6.91	0.56	10.8	0.43
Analytical + Spectral	3.54	5.23	6.91	0.57	10.9	0.37
Clay (%)						
Analytical	37.9	65.4	97.3	13.3	20.3	- 0.01
Spectral	41.7	65.4	100.0	10.9	16.7	0.36
Analytical + Spectral	37.9	65.5	100.0	11.2	17.1	0.33
SOC (g kg^{-1})						
Analytical	9.0	24.7	52.0	11.1	45.1	0.64
Spectral	6.7	26.8	84.4	12.9	48.1	1.15
Analytical + Spectral	6.7	26.4	84.4	12.6	47.9	1.12

CV = coefficient of variation, St.dev = standard deviation.

3.3.3. Quantification of measurement errors in analytical and spectral data

The measurement error standard deviations for the analytical and spectral data were obtained using the methodology described in Section 3.2.5.2 and are given in Table 3.2. The estimated analytical measurement error standard deviation for pH was 0.083, for clay 3.33% and for logSOC 0.038. The PLSR prediction error variance was added to the analytical error variance to get a total measurement error standard deviation for spectral data of 0.234, 6.40%, and 0.196 for pH, clay and logSOC, respectively. As expected, the PLSR prediction errors for the three soil properties had substantially larger standard deviations than the analytical data (Table 3.2).

Table 3.2: Standard deviation of analytical and spectral soil measurement errors.

Soil properties	σ_A	PLSR prediction error standard deviation	σ_S
pH	0.083	0.219	0.234
Clay	3.33	5.47	6.40
LogSOC	0.038	0.192	0.196

3.3.4. Model selection

The stepwise model selection procedure using BIC resulted in the selection of 9 variables for pH, 4 variables for clay and 5 variables for SOC. A brief description of the 12 covariates retained for the three models is summarized in Table 3.3. pH and logSOC were primarily influenced by precipitation, terrain morphology, landform classes, MODIS net productivity and land cover. The selected covariates for clay were climate variables and landform classes.

Table 3.3: Description of environmental variables (covariates) used in the stepwise linear regression models. The plus (+) and minus (–) signs indicate whether a covariate was selected for a soil property.

	Covariate codes	Descriptions	Sources	pH	Clay	logSOC
1	CLM_CHE_PYRSUM	Total annual precipitation	CHELSEA (Karger et al., 2016)	+	+	+
2	CLM_MOD_CCYRAVG	Mean annual cloud cover	EarthEnv (Wilson and Jetz, 2016)	+	–	+
3	CLM_MOD_LSTDYRAVG	Mean annual surface temperature	MODIS (Wan et al., 2006)	–	+	–
4	MOR_MRG_CRU	DEM-parameters: Local upslope Curvature	SRTM (Rabus *et al.*, 2003)	+	–	–
5	MOR_MRG_TPI	DEM-parameters: Topographic Position Index	SRTM (Rabus *et al.*, 2003)	+	–	–
6	MOR_MRG_VDP	DEM-parameters: Valley depth	SRTM (Rabus *et al.*, 2003)	+	–	+
7	MOR_USG_F02	Landform class: Flat plains	USGS (Sayre *et al.*, 2014)	+	+	–
8	MOR_USG_F04	Landform class: Hills	USGS (Sayre *et al.*, 2014)	–	–	+
9	MOR_USG_F06	Landform class: Low mountains	USGS (Sayre *et al.*, 2014)	+	+	–
10	SAT_L07_B4NIR14	Band 4 (NIR) for year 2014	Landsat (Zanter, 2019)	+	–	–
11	LUC_GFC_BARLY10	30 Meter Global Land Cover: Bare soil	ESA (Hansen *et al.*, 2013)	–	–	+
12	VEG_MOD_NPPY00	Net Primary Productivity in 2000	MODIS (Savtchenko et al., 2004).	+	–	–

3.3.5. Model calibration

The linear regression models fitted using REML showed significant correlations between soil properties and the retained covariates. The relationships were of moderate statistical strength for pH (R^2=0.60) and logSOC (R^2=0.49) and of weak statistical strength for clay content (R^2=0.21). The model residuals had a symmetric distribution and were fairly normally distributed (Fig. S3.3).

The regression coefficient estimates without (Scenario 1) and with (Scenario 2) accounting for measurement errors, as well as the accompanying p-values computed using Wald tests and the relative change in coefficient estimates between scenarios are presented in Table 3.4. Note that the coefficients represent the mean change in the dependent variable for one unit of change in the covariate while holding other covariates in the model constant.

Table 3.4: Estimated regression coefficients for the environmental variables under Scenarios 1 and 2.

Covariates	Scenario 1 (without measurement errors)		Scenario 2 (with measurement errors)		Changes per covariate (%)
	Estimate	p-value	Estimate	p-value	
pH					
Intercept	7.27	5.99E-09	7.6	6.74E-09	-4.5
CLM_CHE_PYRSUM	-8.172E-04	5.45E-06	-8.244E-04	5.82E-06	-0.9
CLM_MOD_CCYRAVG	-1.446E-04	1.10E-01	-1.804E-04	1.13E-01	-24.8
MOR_MRG_CRU	8.370E-05	4.04E-01	5.917E-05	4.06E-01	29.3
MOR_MRG_TPI	-3.759E-04	4.78E-03	-3.517E-04	4.81E-03	6.4
MOR_MRG_VDP	-1.132E-04	2.67E-10	-1.102E-04	2.93E-10	2.7
MOR_USG_F02	-1.567E-03	1.41E-02	-1.936E-03	1.43E-02	-23.5
MOR_USG_F06	1.289E-03	3.11E-02	1.413E-03	3.13E-02	-9.6
SAT_L07_B4NIR00	1.845E-02	7.21E-04	1.867E-02	7.13E-04	-1.2
VEG_MOD_NPPY00	-4.133E-05	1.20E-01	-4.664E-05	1.20E-01	-12.9
Clay					
Intercept	653.5	3.64E-12	694.9	3.70E-12	-6.3
CLM_CHE_PYRSUM	-8.691E-03	3.22E-01	-8.585E-03	3.21E-01	1.2
CLM_MOD_LSTDYRAVG	-1.918E-01	6.04E-12	-2.058E-01	6.15E-12	-7.3
MOR_USG_F02	-5.544E-02	1.30E-02	-5.592E-02	1.31E-02	-0.9
MOR_USG_F06	-2.668E-02	1.30E-01	-2.758E-02	1.32E-01	-3.4
logSOC					
Intercept	3.547	6.57E-03	3.568	6.71E-03	-0.6
CLM_CHE_PYRSUM	-5.592E-04	8.12E-04	-5.637E-04	8.38E-04	-0.8
CLM_MOD_CCYRAVG	1.361E-04	2.09E-01	1.344E-04	2.12E-01	1.2
LUC_GFC_BARLY10	-4.213E-02	5.32E-02	-4.316E-02	5.33E-02	-2.5
MOR_MRG_VDP	-8.827E-05	5.84E-09	-8.804E-05	6.71E-09	0.3
MOR_USG_F04	1.340E-04	6.37E-01	1.384E-04	6.39E-01	-3.3

Precipitation, cloud cover, valley depths and net primary productivity had negative regression coefficients for pH, indicating that areas with high rainfall and rich in biomass tend to have lower pH. This is typical for soils of humid climates which are commonly acidic. Increase in precipitation contributes to leaching many of the alkaline basic cations from the topsoil, leading to soil acidification (Chytrý *et al.*, 2007). Clay was also negatively influenced by precipitation and tends to be lower in hills and low mountains. Cloud cover positively influenced logSOC, attesting high values of SOC in areas with high vegetation cover rate. Since SOC is related to organic matter content, areas rich in biomass, humus, and associated organisms responsible for biological activities tend to have higher SOC values (Takoutsing *et al.*, 2018; Lei *et al.*, 2019).

The inclusion of measurement errors modified the regression coefficient estimates. Particularly for pH, coefficients changes were of up to 29% for some of the covariates. This may be because pH has the largest number of covariates and is more sensitive to incorporation of measurement error standard deviations, because of collinearity effects (Fig. S3.4).

The variogram parameters estimated using REML in both scenarios are presented in Table 3.5. The fitted variograms are shown in Fig. 3.3. While the variograms in both scenarios exhibited similar structures and patterns, the nuggets and sills are much smaller in Scenario 2. This is because the nugget represents only spatial variation at short distances and does not include

measurement error variance (Chilès and Delfiner, 2012). Part of the observed variation in soil properties is therefore explained by measurement errors, meaning that the spatial variation of the true (error-free) soil properties is lower than that of the observations.

Table 3.5: Parameters of exponential variogram models of pH, clay and logSOC fitted using REML for Scenarios 1 and 2.

	pH		Clay		logSOC	
Model parameters	1	2	1	2	1	2
Nugget (C_0)	0.077	0.037	51.98	18.34	0.067	0.039
Partial sill (C)	0.08	0.08	61.78	59.93	0.07	0.07
Total sill (C_0 + C)	0.157	0.117	113.76	78.27	0.137	0.109
$C_0/(C_0$ + C) (%)	49.04	31.62	45.69	23.43	48.91	35.78
Range parameter (m)	3,000	3,000	1,577	1,644	3,000	3,000

1 = Scenario1: 2 = Scenario 2

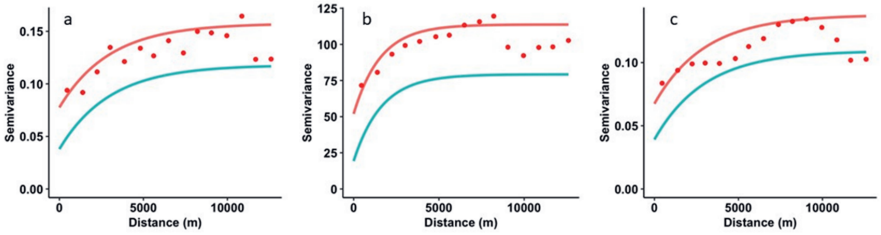


Fig. 3.3: Residual variograms for a) pH, b) clay and c) logSOC. Red lines represent Scenario 1, where measurement errors are not explicitly modelled. Blue lines represent Scenario 2, where measurement uncertainty is accounted for. Red dots are the sample variogram values.

All residual variograms indicated the presence of spatial structure, which means that there is added value in residual kriging. The range parameters for clay slightly increased from 1.57 to 1.64 km from Scenario 1 to Scenario 2, while those of pH (3 km) and logSOC (3 km) remained constant and were not affected by the incorporation of measurement errors. Note that the effective variogram range, which is about three times the range parameter for the exponential model (Webster & Oliver, 2007, Section 5.2.2), is fairly small compared to the extent of the study area for all three soil properties. In other words, the residual spatial structure is somewhat limited and residual kriging will only improve prediction in the neighbourhood of measurement locations.

3.3.6. Model validation

Three validation methods were used in this Chapter: leave-one-out (LOO), leave-cluster-out (LCO) and leave-site-out (LSO) cross-validation. Validation metrics computed for both scenarios show that the models in general provided good (acceptable) predictive ability except for clay (Table 3.6). ME values were close to zero for the three soil properties. Among the soil properties, clay content was poorly modelled with the lowest MEC value, as also revealed by the lowest coefficient of determination of the regression models (see Section 3.3.5).

Table 3.6: Statistical validation metrics obtained by leave-one-out, leave-cluster-out, and leave-site-out cross-validation.

Cross-validation methods	Scenario 1 (without measurement errors)			Scenario 2 (with measurement errors)		
	ME	RMSE	MEC	ME	RMSE	MEC
pH						
LOO	-0.001	0.213	0.834	-0.002	0.214	0.832
LCO	0.012	0.428	0.695	0.010	0.428	0.696
LSO	-0.029	0.435	0.675	-0.031	0.434	0.678
Clay						
LOO	0.020	5.73	0.621	0.021	5.73	0.621
LCO	0.177	11.69	0.297	0.160	11.68	0.300
LSO	-0.495	12.11	0.182	-0.458	12.11	0.184
logSOC						
LOO	0.000	0.217	0.729	0.001	0.218	0.729
LCO	-0.004	0.375	0.594	-0.003	0.374	0.594
LSO	0.015	0.388	0.536	0.018	0.387	0.537

LOO = leave one out cross validation, LCO = leave cluster out cross validation, LSO = leave site out cross validation.

Regression kriging was able to explain the spatial variation between 68 and 83% for pH, between 18 and 62% for clay, and between 53 and 72% for logSOC. RMSE values increased while MEC values decreased from LOO to LCO to LSO cross validation, especially for clay. The decrease in model performances from LOO to LCO to LSO cross validation is as expected and due to the decrease in neighbouring values used when making predictions in each case. For LSOCV, this effectively led to spatial extrapolation rather than spatial interpolation, which is more challenging and susceptible to larger prediction errors. In all cases, validation metrics were practically the same between scenarios, attesting no significant change with the incorporation of measurement uncertainty.

As revealed by the PICP plots shown in Fig. 3.4, which were based on LOO cross-validation, the curves deviate from the 1:1 line and show that both scenarios tend to overestimate the prediction interval widths. However, the deviation from the 1:1 line is much larger for Scenario 1 than for Scenario 2. For example, for pH, we find that for Scenario 1, 64% of the observations is included in the 50% prediction interval, while it is only 57% for Scenario 2. This indicates that Scenario 2 has a more realistic assessment of prediction uncertainty than Scenario 1. For clay, Scenario 2 has a negligible deviation from the 1:1 line as compared to other soil properties.

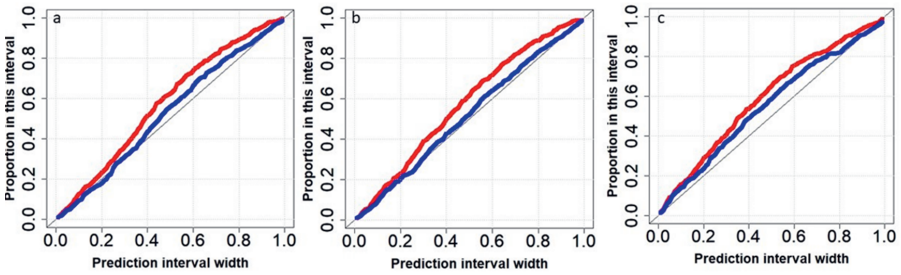


Fig. 3.4: Accuracy plots for a) pH, b) clay and c) logSOC obtained using leave-one out cross validation. Red line represents Scenario 1, blue line Scenario 2.

Overall, both scenarios are similar in their predictive performance (Table 3.6), but the prediction uncertainty is more realistically quantified in Scenario 2 than in Scenario 1. The best modelling approach would therefore be the one that accounts for measurement errors in soil observations.

3.3.7. Spatial prediction

The fitted parameters of the regression models (covariates and regression coefficients) and the variograms (nugget, sill, and range) were used by RK to predict the values of the soil properties at all locations. The logSOC predictions and prediction error standard deviations were back-transformed to SOC values following Laurent (1963). The differences between the predicted values in both scenarios assessed by the scatter density plots (Fig. 3.5) showed no systematic differences between predictions. The absolute differences were never bigger than 0.1, 1.6% and 2 g kg^{-1} for pH, clay, and SOC respectively.

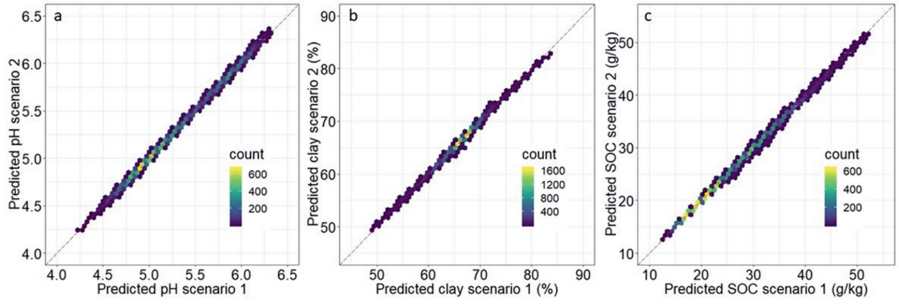


Fig. 3.5: Scatter density plots of predicted values for Scenario 1 and Scenario 2 for a) pH, b) clay, and c) SOC. The dashed red line is the 1:1 line.

The maps of the predicted values for the three soil properties at 250 m spatial resolution, as well as maps of the kriging standard deviations for Scenarios 1 and 2 are presented in Figs. 3.6 and 3.7, respectively. Generally, there are similarities in the spatial distribution of predicted values as the maps showed comparable ranges of predicted values, and similar spatial patterns and features such as areas of low and high concentrations. Therefore, only the maps of Scenario 2 and the maps differences between scenarios are presented.

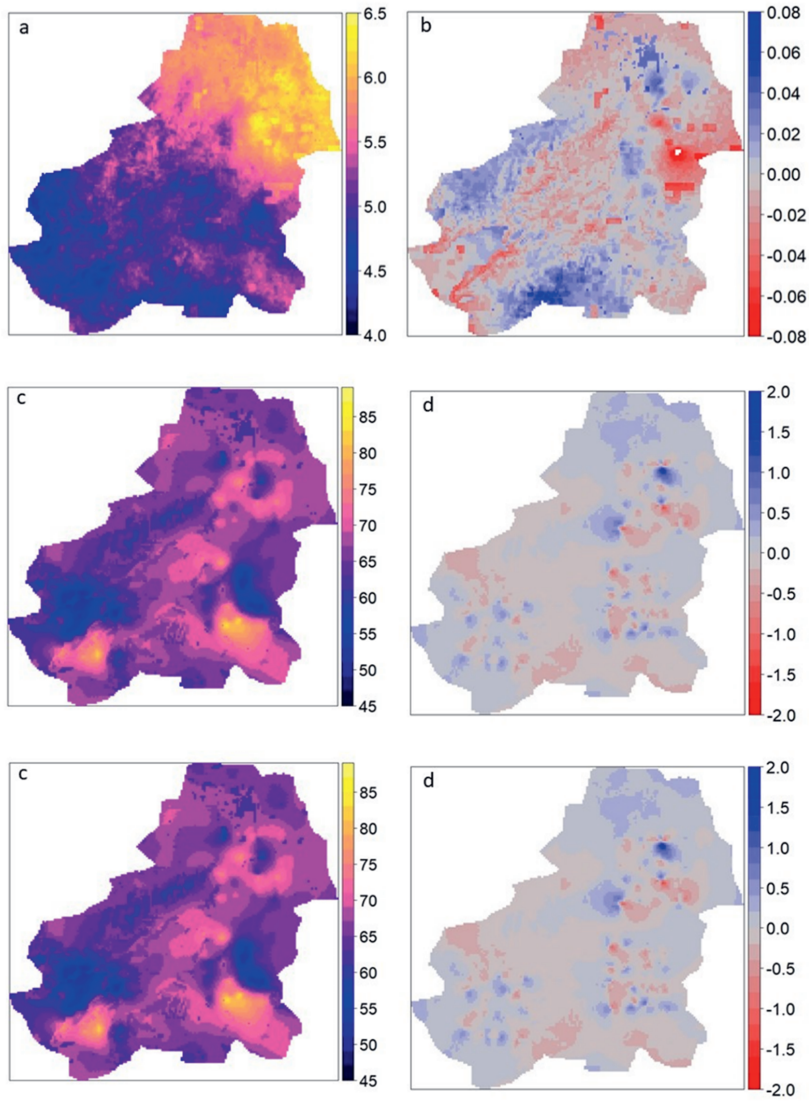


Fig. 3.6: Maps of soil property predictions and prediction differences: a) pH Scenario 2, b) difference between pH Scenario 1 and pH Scenario 2, c) clay Scenario 2 (%), d) difference between clay Scenario 1 and clay Scenario 2 (%), e) SOC Scenario 2 (g kg^{-1}), f) difference between SOC Scenario 1 and SOC Scenario 2 (g kg^{-1}). Note: prediction maps of Scenario 1 not shown because these are very similar to those of Scenario 2.

The kriging standard deviation maps for pH and clay clearly showed the spatial sampling design that was used, which was expected from the fairly small effective variogram ranges, which implies that the kriging standard deviation is small only close to measurement locations (Fig. 3.7). For SOC, the spatial sampling design does not appear in the standard deviation map

because after back-transformation this map is more influenced by the logSOC prediction than the logSOC standard deviation.

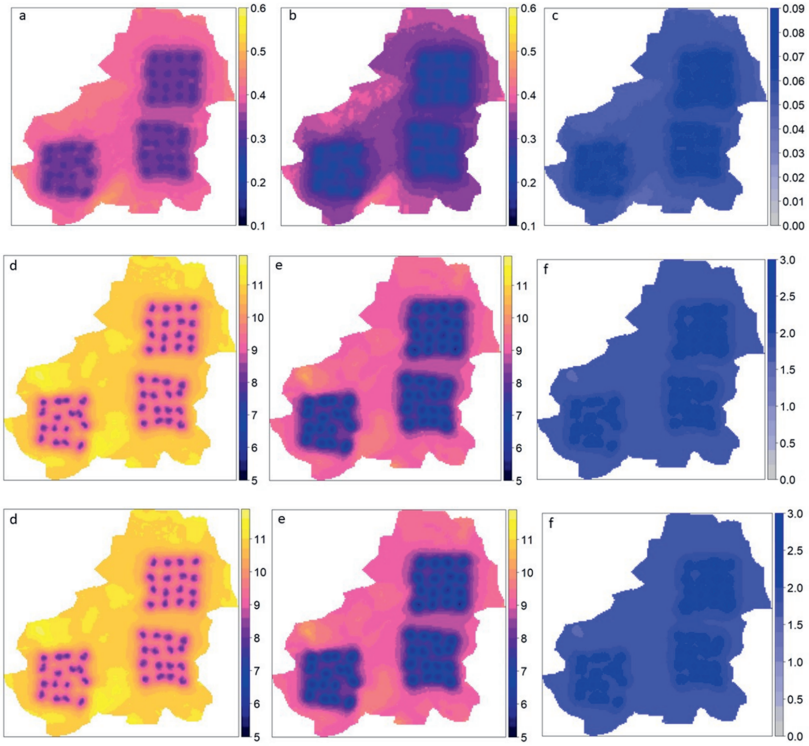


Fig. 3.7: Maps of kriging standard deviations and differences: a) pH Scenario 1, b) pH Scenario 2, c) difference between pH Scenario 1 and pH Scenario 2, d) clay Scenario 1 (%), e) clay Scenario 2 (%), f) difference between clay Scenario 1 and clay Scenario 2 (%), g) SOC Scenario 1 (g kg^{-1}), h) SOC Scenario 2 (g kg^{-1}), i) difference between SOC Scenario 1 and SOC Scenario 2 (g kg^{-1}).

We observed relative differences in standard deviations in some areas of the study area between the two scenarios of up to 0.08 for pH, 2.7% for clay and 0.5 g kg^{-1} for SOC between the two scenarios (Fig. 3.4). The pH and clay standard deviation maps for Scenario 2 had lower values than those for Scenario 1 for the entire study area, and this corroborates well with the results obtained using the accuracy plots (Fig. 3.4). For SOC the standard deviation differences between Scenarios 1 and 2 are both positive and negative.

3.4. Discussion

3.4.1. Measurement errors and implications for spatial modelling

The primary aim of this Chapter was to quantify the uncertainty in soil measurements, incorporate measurement error variances in the covariance structure of spatial models, and evaluate the influence of this on the prediction and prediction uncertainties. We illustrated the

methodology and applied it to a case study, where we mapped three soil properties using soil data obtained using conventional laboratory analytical methods (analytical data) and MIRS (spectral data). Among the three soil properties, the clay content model had poor performance compared with pH and SOC, with the lowest MEC value. This could be attributed to one or a combination of the following: (a) the spatial resolution of some of the environmental variables was not detailed enough to capture the variation (Maleki *et al.*, 2020); (b) the set of covariates retained was not suitable and other environmental variables need to be included; or (c) the sampling protocol was too clustered to capture variability across the study area, since observations were taken from three main clusters and various processes operate at different spatial scales (Hendriks *et al.*, 2021).

Results showed that taking measurement uncertainty into account had a small to moderate effect on the estimated regression coefficients of the regression kriging model (Table 3.4) and a large effect on the residual variograms (Fig. 3.3), which had much smaller nugget variances and sills when measurement errors were explicitly accounted for. This is as expected because measurement error variance was separately modelled in Scenario 2 and hence not included in the residual variogram. Depending on the soil property and the type of cross-validation used (leave-one-out, leave-cluster-out and leave-site-out), the amount of variance explained varied between 18 and 83% and showed that resulting maps are useful in assessing the spatial variation in pH and SOC and provided a first approximation for clay. Large differences were observed between the three types of cross-validation applied, particularly for clay, with LOO > LCO > LSO cross-validation. This is largely due to the decrease in the number of nearby available observations as we move from one type of cross-validation to another (Lagacherie *et al.*, 2020; Loiseau *et al.*, 2021). The reduction in the number of observations may have also weakened the relationships between environmental variables and soil properties in LSO cross-validation. LOO cross-validation likely gives a too optimistic view of model performance, especially when the data are spatially clustered as in our case study, while LSO cross-validation is likely too pessimistic, because it applies spatial extrapolation instead of interpolation. LCO cross-validation may be the best compromise for evaluating prediction performance in this Chapter. But there is a need to investigate how best to carry out cross-validation in the case of clustered data, to get a realistic estimate of model performance that does not lead to biased estimates of the validation metrics (Chartin *et al.*, 2017; Roberts *et al.*, 2017; Poggio *et al.*, 2021; Styc and Lagacherie, 2021).

Spatial models with and without measurement errors were comparable in predictive performance (Table 3.6). The ME, RMSE and MEC values between scenarios were quite similar, and so were the kriging prediction maps (Fig. 3.6). This was contrary to our expectations, also taking into consideration the influence of the measurement error variances on the model parameters and regression coefficients. The insignificant differences between the validation metrics and prediction maps of the two scenarios could perhaps be explained as follows. If all observations had the same measurement error variance, then the same performance would have been achieved because in such case all observations would carry equal weight and Scenarios 1 and 2 are effectively the same. In this Chapter we used data from two different sources (analytical and spectral) with very different measurement error variances. The analytical data had much smaller measurement error variances than the spectral data, and therefore get much larger weights in the estimation of regression coefficients and in kriging. However, analytical data represented only 10% of the data set, and the spatial distribution of the analytical data was similar to that of the spectral data. As shown in Fig. 3.1, the analytical data were in the same clusters (one observation per cluster) as the spectral data. If the analytical and spectral data had been located in different parts of the study area, we likely would have obtained larger differences between the two scenarios (Meyer *et al.*, 2018). The

results that we obtained refer to just one case study, and it is worthwhile to investigate the sensitivity of the DSM models and maps to incorporation measurement errors in other studies and in other contexts.

While differences in prediction maps and cross-validation metrics of predictions did not differ much between the two scenarios, we did get substantial differences in prediction error standard deviation maps and in the evaluation of the prediction uncertainty. As shown by the accuracy plots (Fig. 3.4), ignoring measurement error variances led to a large deviation from the 1:1 line. The line for Scenario 1 was much above it, which indicates that the kriging prediction error standard deviations were unrealistically high. Though we obtained deviations in both Scenarios, the problem is much more pronounced for the variance models in Scenario 1, when measurement errors are ignored. Prediction intervals were larger in Scenario 1 than in Scenario 2, attesting that the quantification of uncertainties had significantly improved and was more realistic when measurement errors were accounted for.

The kriging standard deviations maps for pH and clay in Scenario 2 had lower values than those in Scenario 1, and this corroborates well with the findings derived from comparison of the accuracy plots. Accounting for measurement errors decreased the kriging variance because the 'true' soil properties had less spatial variation than the measured soil properties, which means that they are easier to predict, even in case of presence of measurement errors (Chilès and Delfiner, 2012). For SOC this did not occur, even though the logSOC kriging standard deviations were all smaller for Scenario 2 than for Scenario 1 (results not shown). This was because the back-transformation of logSOC not only depends on the kriging standard deviations of logSOC but also on the logSOC predictions (Laurent, 1963; Webster and Oliver, 2007).

In practice, the usefulness of DSM lies in its ability to quantify and map prediction uncertainties (Malone *et al.*, 2015), and ignoring measurement errors leads to poor assessment of the accuracy of digital soil maps (**Chapter 3** and **Chapter 4**). Soil scientists have made considerable efforts in quantifying prediction uncertainties in their work, much more than in other disciplines, but this has not always been as systematic as it should be (Piikki *et al.*, 2021). Whenever uncertainties are quantified with standard deviation maps, one of course has to make sure that these are realistic assessments of the map error. In this Chapter, this does not happen in Scenario1, while Scenario 2 does it much better. In other words, this stressed the importance of taking measurement error into account to accurately quantify the prediction uncertainties. It is unfortunate that most end-users of DSM products are only interested in the prediction maps from which soil information are extracted, often ignoring the uncertainty maps, meanwhile prediction maps with large errors could have important economic and environmental consequences for the design and implementation of land restoration initiatives (Styc and Lagacherie, 2021).

The results of this Chapter overall indicate that the additional investment in quantification of measurement error variances and the incorporation in the spatial models is worth the effort, as shown by the improvement in the quantification of the prediction uncertainties. There is a need to create awareness among end-users on the importance of realistic and reliable uncertainties of the maps they intend to use, so that in case of large uncertainty, investments can be made to obtain more accurate soil information (Heuvelink, 2014).

3.4.2. Limitations and recommendations for future research

Despite the successful incorporation of measurement error variances in RK and the improvement in the quantification of prediction uncertainties, there are several aspects worthy of attention and further development.

The soil sampling design used was not initially intended for geostatistical mapping, but rather to provide a biophysical baseline, and a monitoring and evaluation framework for assessing processes of land degradation and the effectiveness of rehabilitation measures (Vågen and Winowiecki, 2013). Consequently, the design was not optimised to properly account for spatial dependence over large distances (Brus *et al.*, 2011). In Africa, cluster sampling is often favoured due to accessibility problems and limited resources, however, the method is prone to biases and large prediction errors in unsampled areas. For future sampling, it is highly recommended to combine cluster sampling with other sampling methods to account for variation at both short and larger distances. Moreover, sampling should also cover the feature space well (Brus and Heuvelink, 2007; Wadoux *et al.*, 2019).

One important motive that measurement errors variances have not been systematically incorporated in DSM is the challenge in their quantification in laboratories. These are rarely reported systematically with the results of analyses, probably due to the lack of interests from clients (Li *et al.*, 2019). In this Chapter, we had one sample analysed in duplicate under the same conditions to quantify the measurement error variances of analytical data. We assumed constant measurement error variances for each of the soil properties, but in many practical cases measurement errors are proportional to the measured values (Libohova *et al.*, 2019). Since measurement error variance for SOC was estimated on logarithmic scale, we assumed a proportional error model for SOC, but not for pH and clay. Quantification of measurement errors in analytical data can be improved if laboratories pay more attention and systematically quantify the uncertainties of their measurements and benchmark against standards to minimise systematic bias.

There are many sources of errors that propagate during the DSM process, and each contributes to uncertainty in the final prediction (Robinson *et al.*, 2015). The fact that this Chapter focused on the uncertainty in soil observations does not mean that the influence of other sources of errors can be ignored. From the modellers and end-users' perspectives, possible improvements would be to quantify these other sources of errors and assess their implications so that measures are taken to improve the uncertainties of the DSM. Efforts have already been made in this line and error quantification methods have already been broadly discussed for some of the sources in Nelson *et al.* (2011); Bishop *et al.* (2015); Robinson *et al.* (2015).

This Chapter used resampled environmental variables at 250 m resolutions as covariates, which might be too coarse and not able to capture some of the variation across the study area (Taylor *et al.*, 2013). The use of high spatial resolution covariates that can provide more detailed information and reflect the distribution characteristics of the targeted soil properties, particularly for small study area is recommended (Samuel-Rosa *et al.*, 2015).

This Chapter has shown that it is relatively easy to incorporate measurement error variances in regression kriging once these are quantified. Further development of the approach is the extension to machine learning DSM models (Hengl *et al.*, 2018a; Wadoux, 2019). This is critical because of the rapid uptake for the machine learning algorithms in digital soil mapping that is transforming the process of spatial modelling and generating more accurate predictions (Wadoux *et al.*, 2020a).

3.5. Conclusions

We applied a geostatistical DSM approach to derive prediction and prediction uncertainty maps after quantifying and incorporating measurement error variances in the covariance structure of the spatial model. Accounting for measurement errors resulted in changes in regression coefficients of up to 29% and influenced the variogram parameters by reducing the nuggets and sill variances. Validation metrics were quite similar in the two scenarios, but prediction uncertainties were more realistically quantified by the models that account for measurement errors, as indicated by accuracy plots. Prediction maps were similar between scenarios, but we observed slight differences in predicted values in some parts of the study area of up to 0.1 for pH, 1.6% for clay and 2 g kg^{-1} for SOC. Differences in regression kriging standard deviations were up to 0.08 for pH, 2.7% for clay and 0.5 g kg^{-1} for SOC. For pH and clay the kriging standard deviations were systematically smaller when measurement errors were explicitly accounted for. The Chapter stressed the importance of quantifying prediction uncertainties, particularly when the issue of uncertainty propagation in the modelling processes becomes essential. This will help end-users to be aware of the real prediction uncertainties and their implications for the design and implementation of land restoration interventions. It is advised that the methodology used in this work is also tested in other case studies and further developments of the approach should include its extension to non-linear machine learning regression methods, such as Random Forest.

3.6. Supplementary information

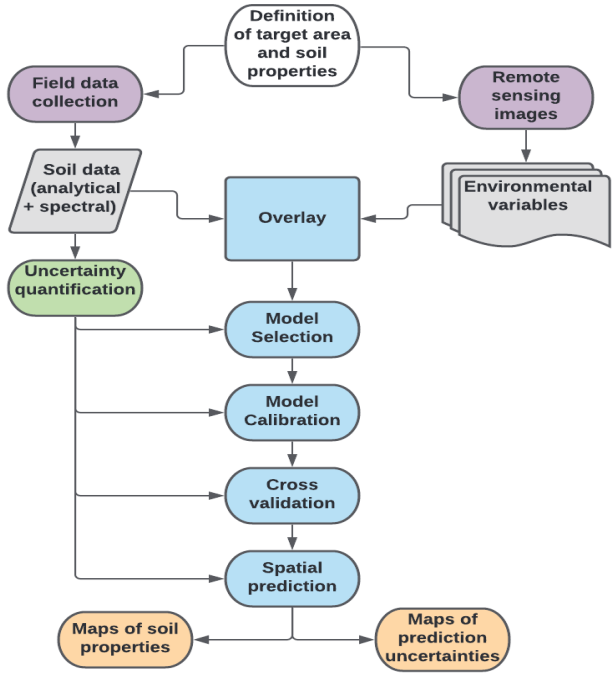


Fig. S3.1: Flowchart accounting for measurement errors uncertainties in digital soil mapping.

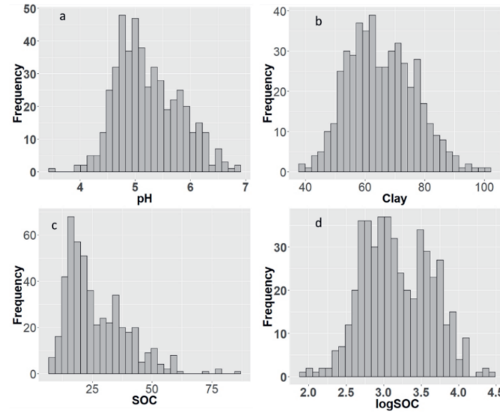


Fig. S3.2: Histograms of a) pH, b) clay (%), c) SOC (g kg^{-1}) and d) log-transformed SOC.

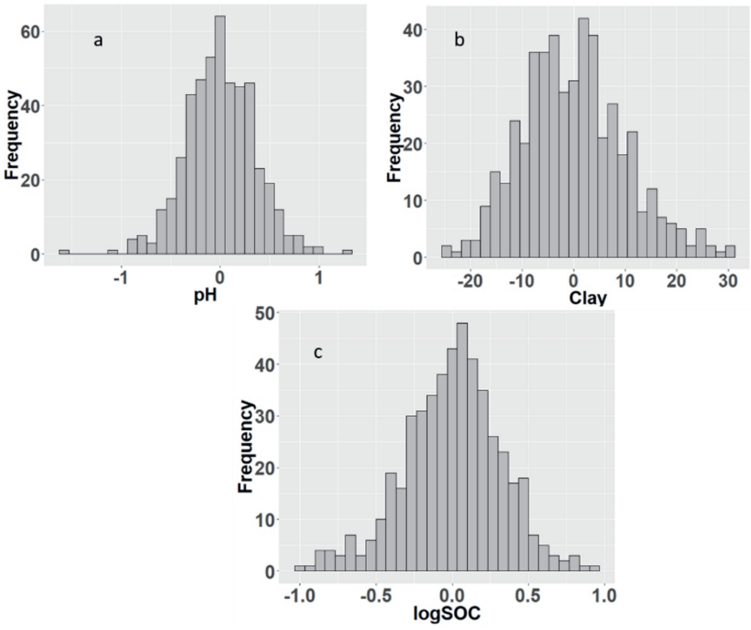


Fig. S3.3: Histogram of the regression model residuals of a) pH, b) clay (%), and c) log-transformed SOC.

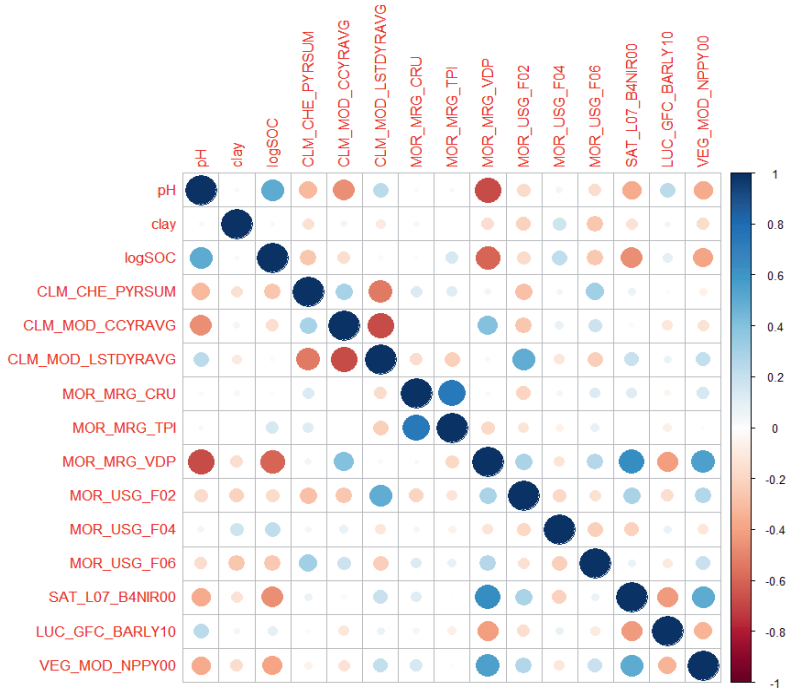


Fig. S3.4: Correlation plot between soil properties and environmental variables.

Chapter 4

Comparing the prediction performance, uncertainty quantification and extrapolation potential of regression kriging and random forest while accounting for soil measurement errors

Geostatistics and machine learning have been extensively applied for modelling and predicting the spatial distribution of continuous soil variables. In addition to providing predictions, both techniques quantify the uncertainty associated with the predictions, although geostatistics is more developed in this respect. Despite the increased use of these techniques, most algorithms ignore that the soil measurements are not error-free. Recently, concern has also arisen about the extrapolation risk of these techniques, be it in geographic space, feature space, or both. In this chapter, regression kriging (RK) and random forest (RF) were compared with respect to their ability to deliver accurate predictions and quantify prediction uncertainties, while accounting for measurement errors in soil measurements. The sensitivity of results of both models to soil measurement errors was also evaluated, as well as their spatial extrapolation potential. This was done for a case study in Cameroon where soil pH, clay, and soil organic carbon (SOC) were mapped from measurements obtained using both conventional and proximal soil sensing methods. The results showed that both models produced comparable ranges and maps of predicted values for the soil properties of interest. Compared to RF, RK outperformed RF by presenting generally a higher Model Efficiency Coefficient (MEC), lower Root Mean Squared Error (RMSE) values and better extrapolation performance. The improvement in RMSE was about 10, 12 and 2 % while the improvement in MEC was on average 5, 22 and 1 % for pH, clay, and SOC, respectively. Overestimation of the local uncertainty observed for RK was larger than that of RF as shown by the accuracy plots, indicating that prediction uncertainties were better quantified by the RF model. Better extrapolation performance was obtained with RK that derived better predictions than RF at unsampled locations as shown by cross-validation metrics and scatter plots, particularly when RK and RF were used for spatial extrapolation. The effects of incorporating measurement errors appeared not significant for predictions and prediction uncertainties. This was partly attributed to the fact that this chapter assumed constant measurement error variances for each of the soil properties, and for analytical and spectral data used for calibration instead of considering measurement errors that are proportional to the measured values. An important finding of this chapter was that model comparison should go beyond using only common validation metrics to evaluate prediction accuracy of DSM approaches but should also account for their ability to quantify prediction uncertainty at unsampled locations. The findings of this chapter show that soil maps are not error free, and if use as soil data inputs, the uncertainty can propagate to affect the results of further modelling. Hence, it is essential to evaluate the repercussions of uncertain soil information in other modelling processes, particularly in crop modelling.

A version of this chapter has been published in Takoutsing, B., Heuvelink, G.B.M., 2022. Comparing the prediction performance, uncertainty quantification and extrapolation potential of regression kriging and random forest while accounting for soil measurement errors. *Geoderma* 428, 116192.

4.1. Introduction

Digital soil mapping (DSM) is the use of mathematical and statistical models to combine soil measurements with spatially explicit environmental variables to generate soil spatial information (Sanchez *et al.*, 2009). One of the advantages of DSM lies in its ability to determine the uncertainty of the final products (Arrouays *et al.*, 2020). Geostatistical-based methods such as regression kriging (RK) have been for decades the major DSM techniques. The main strength of RK lies in its ability to account for spatial correlation, benefit from relationships with covariates (i.e., explanatory environmental variables), minimize the prediction error variance and provide prediction uncertainties through the kriging standard deviation (Odeh *et al.*, 1995; Hengl *et al.*, 2007; Hengl *et al.*, 2015). The assumptions of RK are based on a linear mathematical relationship between soil properties and covariates (Wadoux *et al.*, 2018). However, in reality soil properties are often governed by complex relationships with soil forming factors (Behrens *et al.*, 2014) and landscape attributes (Takoutsing *et al.*, 2018) that make them non-linear and involve complex interactions between covariates (Lamichhane *et al.*, 2019).

Recent years have witnessed an increase in the use of machine learning (ML) techniques for the spatial prediction of soil properties (Hengl *et al.*, 2018a). ML techniques are able to represent complex nonlinear relationships between dependent variables and covariates and can handle a large number of cross-correlated covariates as predictors (Nussbaum *et al.*, 2018). In addition, ML makes no assumptions about the underlying distribution of the data, unlike geostatistical methods where transformation of the original measurements is often required to satisfy the normality assumption (Wadoux *et al.*, 2020a). Since ML algorithms are not required to satisfy strict statistical assumptions, they are more easily used. Furthermore, their flexibility with respect to the structure of the relationship between the soil property of interest and covariates makes them often more accurate than regression kriging (Hengl *et al.*, 2015). While there are numerous ML techniques, one that has attracted a lot of attention particularly for regression purposes is the random forest (RF) algorithm (Breiman, 2001). Several studies (Hengl *et al.*, 2015; Vaysse and Lagacherie, 2017; Hengl *et al.*, 2018a; Nussbaum *et al.*, 2018) have shown that RF is an efficient and valuable DSM technique for soil spatial prediction.

However, ML algorithms are not as good at quantifying the prediction uncertainty as kriging, while the performance of a DSM model must also be judged on its ability to accurately and realistically quantify the prediction uncertainty (Arrouays *et al.*, 2014; Heuvelink, 2018). In this context, Meinshausen (2006) modified the RF algorithm to propose quantile regression forests (QRF) that yields all quantiles of the conditional distribution. Hence it is able to quantify the prediction uncertainty at all prediction locations, for example by the limits of a 90% prediction interval. Note that it does not assess the spatial correlation of the prediction uncertainty because it does not model spatial dependence (Heuvelink and Webster, 2022). This implies that it cannot assess the prediction uncertainty of spatial averages and totals. For this to be accomplished, a geostatistical approach is required (Szatmári *et al.*, 2021). Previous studies have successfully used RF and QRF for prediction of different soil properties (Hengl *et al.*, 2015; Forkuor *et al.*, 2017; Hengl *et al.*, 2021).

Several studies have compared the performance of RK and RF and observed substantial differences in prediction maps. Fouedjio (2020) found that RF can obtain equally accurate and unbiased predictions as RK. Hengl *et al.* (2015) demonstrated that the RF algorithm consistently outperformed RK across a range of soil properties. Veronesi and Schillaci (2019), in a study carried out in the semiarid island of Sicily in Italy, reported kriging as the best performer in comparison with RF. But only few studies (Vaysse and Lagacherie, 2017; Szatmári and Pásztor, 2019) have been carried out to compare RK and RF prediction uncertainty maps and assessed whether these maps are realistic representations of the prediction uncertainty. This calls for extra case studies to get more insight into the ability of RK and RF to quantify the uncertainties well.

The generation of soil information at national and regional levels is often constrained by limited data availability and many algorithms rely on their extrapolation ability to be able to produce the required information. As opposed to spatial interpolation, which uses point values in a study area to predict values at other points enclosed by observation points and located within that same area, spatial extrapolation refers to transferring the model beyond the area from which the training data were taken (i.e., to new geographic space). In other words, while in spatial interpolation predictions are made using data coming from all geographic directions, in spatial extrapolation predictions are based on data from one or a few directions only, possibly from far away. The extrapolation of a model from one area to another faces several challenges due to the differences in type and intensity of soil-forming factors (Angelini *et al.*, 2020). ML models can result in poor and unreliable predictions when they are applied to an area where the feature space is very different from that of the calibration data (Meyer and Pebesma, 2021). Few studies have investigated the problem of spatial extrapolation for digital soil mapping (Neyestani *et al.*, 2021) and comparison of the extrapolation potential of different DSM algorithms also has not been fully investigated. In **Chapter 3**, we noticed a deterioration of performance when RK spatial interpolation is replaced by RK spatial extrapolation. It is not known whether similar results are obtained for RF, or whether RF is better or worse at geographic extrapolation than RK.

Despite the focus of DSM on uncertainty quantification, most algorithms typically do not account for the fact that the soil measurements used to calibrate the models are not error-free. Accounting for measurement errors seems especially important when the magnitude of error varies between measurements, because more accurate measurements should carry more weight than less accurate measurements. Accounting for measurement errors may be important as soil measurements can have large uncertainties (Laslet and McBratney, 1990; Libohova *et al.*, 2019; van Leeuwen *et al.*, 2021), which propagate through the modelling process and affect the DSM results. Tackling this challenge is critical for DSM to gain more credibility and scientific consistency among soil science communities of practice (Wadoux *et al.*, 2020a).

For a case study in Cameroon where we compared RK with and without accounting for measurement errors (**Chapter 3**), relatively small differences were observed between the two scenarios, while the prediction uncertainties were better quantified with the model that accounted for measurement errors as shown by cross-validation metrics and accuracy plots. For ML, the solution how to account for measurement errors in model training and prediction may not be obvious. Some attempts have been made using weights to integrate measurement errors in RF models (Hengl *et al.*, 2018a; Naul *et al.*, 2018; Wadoux, 2019), but these approaches are heuristic and lack a theoretical underpinning. In light to this drawback, van der Westhuizen *et al.* (2022) developed a theoretical framework that extends the ML approach for DSM and accounts for uncertainties in soil measurements. It makes use of a two-stage maximum likelihood framework that incorporates measurement error variances in the penalty function that is optimized in ML calibration. So far, this novel method has not yet been tested in other soil mapping case studies. The aims of this Chapter were to: 1) compare the performance of RF and RK with respect to their accuracy in predicting three soil properties (pH, clay, and SOC) and their success in modelling prediction uncertainty; 2) assess and compare the sensitivity of RF and RK predictions to measurement errors; and 3) evaluate whether spatial extrapolation leads to a bigger or smaller deterioration when using RF instead of RK. These aims were studied using a case study from a 1,053 km^2 area located in the western highlands of Cameroon.

4.2. Materials and methods

The study area and data sets have been described in detail in **Chapter 3**, and we only repeat the essentials here. Handling measurement uncertainty in RK is also well-described in the literature, but for

ML this is relatively unknown, and we therefore summarize it in this section, while referring to van der Westhuizen *et al.* (2022) for details. Application of RK to the case study was also described in detail in **Chapter 3**. Application of RF and QRF to the case study is described in Section 4.2.3.

4.2.1. Study area and sampling design

The study area is located in the western highlands of Cameroon and was sampled using a spatially stratified and hierarchical sampling approach based on the concept of sentinel sites, each covering a square area of 100 km^2 (Vågen and Winowiecki, 2020). Three of such sites were established across the study area (Fig. 4.1). The distribution of the sampling locations shows that the sampling plots are strongly clustered in three sites and that the eastern part of the study area was particularly sparsely sampled. These are unfavourable conditions for spatial interpolation and can lead to poor assessment of spatial structures. One other implication of a clustered sampling design is that model accuracy metrics as obtained using conventional cross-validation might be overoptimistic (Roberts *et al.*, 2017; Wadoux *et al.*, 2021; de Bruin *et al.*, 2022). This Chapter however used a clustered sampling design to reduce accessibility costs (Yang *et al.*, 2018). We applied three different cross-validation strategies to address the problem of overoptimistic results, as explained in detail in **Chapter 3**. These strategies were also used to evaluate the extrapolation potential of RK and RF, as explained in Section 4.2.7.

4.2.2. Soil data and environmental variables

Four hundred and eighty (480) topsoil samples were collected and analysed using mid-infrared spectroscopy (MIRS) to obtain soil spectral data. Of these samples, 10% were analysed using conventional laboratory methods for pH, texture, and soil organic carbon (SOC) and referred to as reference samples. Spectra were scanned from soil samples that were air-dried, passed through a 2-mm sieve, and finely ground to powder (<100 µm) using a sample mill. Samples were loaded into four replicate wells, where each sample was scanned 32 times in MIR reflectance mode using a Fourier-transform MIR spectrometer (FT-IR; Tensor 27, Bruker Optics, Karlsruhe, Germany). All spectral replicates for each sample were averaged to a single spectrum for each soil sample. Then, the measured MIR reflectances were converted to apparent absorbance units [log (1 / Reflectance)] and pre-processed with the Savitzky–Golay smoothing method (Sila *et al.*, 2016). Regression models were built to relate the processed spectra to the reference samples using Partial Least Squares Regression (PLSR). The fitted regression models were used to predict the targeted properties of all samples. SOC analytical data showed positive skewness and were first log-transformed before running the PLSR. Accurate predictive models were obtained for soil pH (r^2 = 0.87), clay (r^2 = 0.83) and logSOC (r^2 = 0.82). For soil samples analysed using both conventional and MIRS, only analytical data were retained, which resulted in a final data set made up of analytical (n = 48) and spectral (n = 432) data. In this Chapter, the measurement error variances for analytical data were derived using laboratory repeatability procedures (Libohova *et al.*, 2019; van Leeuwen *et al.*, 2021), while those of spectral data were estimated by adding the residual variance of the PLSR models. We assumed constant measurement error variances for each soil property and each method of analysis across the study area (**Chapter 3**).

Over 170 environmental layers in the form of raster images were initially considered. These were all obtained from freely available databases and were considered proxies of key soil-forming factors. The environmental variables were extracted for the study area and resampled to a common coordinate reference system with 250 m spatial resolution, either through downscaling of coarser resolution layers, or upscaling for the finer ones. Based on literature, pedological information and expert knowledge, 46 layers were retained to constitute a final environmental data set used for further analyses (**Chapter 3**). A full list of these 46 remaining environmental variables is presented in supplementary information (Table S4.1).

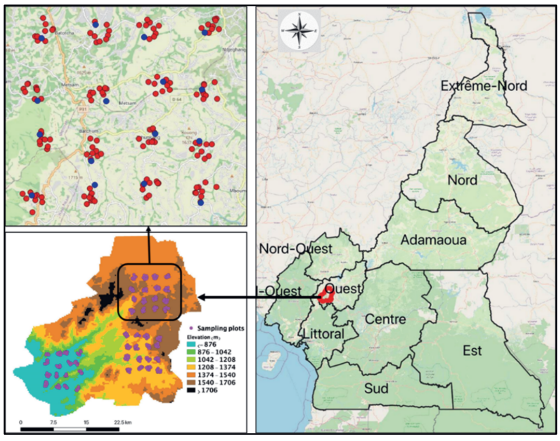


Fig. 4.1: Map of Cameroon showing the study site and the distribution of sampling locations for analytical and spectral data. Soil sampling was done in three 10 km x 10 km sentinel sites. Each sentinel site has 160 sampling locations (violet dots). Top-left panel zooms in on the most northern sentinel site (red dots represent spectral data, blue dots analytical data).

4.2.3. Random Forest modelling and prediction

Random forest is an ensemble tree-based ML technique widely used in DSM (Vaysse and Lagacherie, 2017; Hengl *et al.*, 2018a; Nussbaum *et al.*, 2018). An ensemble of trees is built based on a bootstrap sample of the training data, and all the tree predictions are averaged as a final prediction (Breiman, 2001). We applied the RF models with and without measurement uncertainty to three soil properties (pH, clay and log-transformed SOC, here abbreviated as logSOC) to obtain two models for each soil property, namely Scenario 1 (measurement uncertainty ignored) and Scenario 2 (measurement uncertainty accounted for).

4.2.3.1 *Model selection*

Firstly, we evaluated the importance of environmental variables by fitting a full RF model using the 46 covariates selected in Section 4.2.2 (Table S4.1). Then a selection procedure was performed using the Recursive Feature Elimination (RFE) algorithm as implemented in the *caret* package (Kuhn, 2008). The RFE algorithm is an iterative process that starts by fitting a model using all covariates, then assesses its performance and ranks the covariates according to their importance (Hounkpatin *et al.*, 2018; Gomes *et al.*, 2019). The least important covariates are removed from the pool, and the model is fitted and assessed again. This process is repeated until there are no or only a few covariates left. From all evaluated models the one with the most favourable cross-validation statistics is selected. For each soil property, an optimal set of covariates was identified. The set of covariates used by each model could obviously differ between soil properties.

4.2.3.2 *Model calibration and fitting*

The calibration of the RF model relies on three user-defined parameters (Probst and Boulesteix, 2017): the number of trees (ntree), the number of covariates selected at each split (mtry) and the size of terminal nodes (nodesize). In this Chapter, we fixed ntree = 200, which is the value at which the out-

of-bag overall error stabilized. In many cases, 150 trees have been reported as sufficient to obtain stable outcomes, particularly when the number of covariates is smaller than the calibration sample size (Biau and Scornet, 2016; Lopes, 2019). By default, we used mtry as set to the rounded down square root of the total number of covariates and equally set the nodesize default value to 5. We used all 480 measurements, the retained covariates, and the selected fine-tuned parameter values to fit a final RF model for each soil property (in both scenarios) using the R package *ranger* (Wright and Ziegler, 2017).

4.2.3.3 *Spatial prediction*

The prediction of a single decision tree of RF for a new location $\boldsymbol{s}_0$ is the average of all measurements that are contained in one of the end nodes of the tree. The end node is found by branching through the tree based on the covariate values at $\boldsymbol{s}_0$. The RF prediction is the average of all tree predictions. It is a weighted linear combination of the measurements and can therefore be written as:

$$\hat{y}(\boldsymbol{s}_0) = \sum_{i=1}^{n} w_i \cdot y_i \tag{4.1}$$

where $\hat{y}(\boldsymbol{s}_0)$ is the prediction, n is the number of measurements, the w_i are weights and the y_i are the soil property measurements. Although not made explicit in Eq. 4.1, note that the weights are derived from the covariates at the observation and prediction locations (Meinshausen, 2006).

4.2.3.4 *Quantification of prediction uncertainty*

In this Chapter, we used the quantile regression forest (QRF) approach of Meinshausen (2006) to estimate the uncertainty of the RF predictions. This method derives quantiles of the conditional probability distribution for every prediction location, which together completely define the distribution at the location. In DSM practice, often only the 0.05 and 0.95 quantiles are derived. Subtracting the 0.05 quantile map from the 0.95 quantile map provides a map of the 90% prediction interval width, which is a common summary measure for the prediction uncertainty. The quantiles are derived from the conditional cumulative probability distribution $F(y|\boldsymbol{x})$, with $\boldsymbol{x}$ a vector of covariates at the prediction location, which in turn is estimated from the data using a modification of Eq. 4.1:

$$\hat{F}(y|\boldsymbol{x}) = \sum_{i=1}^{n} w_i \cdot 1_{\{y_i \leq y\}} \tag{4.2}$$

where $1_{\{y_i \leq y\}}$ is a binary variable that equals 1 if the condition is satisfied and 0 otherwise. Note that QRF uses the same weights as RF, meaning that the RF model needs to be trained only once (see Section 4.2.3.2). QRF is also implemented in the *ranger* package and is invoked instead of RF by setting 'quantreg = TRUE'. The 0.05 and 0.95 quantiles are derived by identifying those values of y for which $\hat{F}(y|\boldsymbol{x})$ equals 0.05 and 0.95, respectively.

In the case study, we characterized uncertainty by means of the 90% prediction interval width. We used RF to estimate the conditional mean (i.e., the prediction) and QRF to compute the lower and upper limits of a 90% prediction interval for each of the three soil properties. Note that we only did the latter for Scenario 1, because QRF has not yet been extended to a case of uncertain training data. For SOC, the logSOC quantiles were back-transformed to SOC quantiles using the exponential function.

In the case of regression kriging, the lower and upper limits of the 90% prediction interval for pH and clay were derived by subtracting and adding 1.64 times the kriging standard deviation to the kriging

prediction. It was thus assumed that the kriging prediction error was normally distributed. The same was done for logSOC, and the resulting prediction interval limits were back-transformed using the exponential function. The SOC prediction was obtained by back-transforming the logSOC prediction and making a variance correction as described in Section 8.10 of Webster and Oliver (2007) and in Laurent (1963). Without variance correction the back-transformation would return the median instead of the mean.

4.2.3.5 *Random Forest with uncertain data*

In case of RF modelling with uncertain training data, we used the approach proposed by van der Westhuizen *et al.* (2022) to account for soil measurement errors. In regular RF, every observation is equally important because measurement errors are ignored, and the residual variance of the RF model is assumed constant. Let the true value of the target variable at measurement location $\boldsymbol{s}_i$, $Y(\boldsymbol{s}_i)$, be defined by:

$$Y(\boldsymbol{s}_i) = f(\boldsymbol{x}(\boldsymbol{s}_i)|\theta) + \varepsilon(\boldsymbol{s}_i) \tag{4.3}$$

where f is a statistical RF model, $\boldsymbol{x}(\boldsymbol{s}_i)$ the covariates at location $\boldsymbol{s}_i$, θ represents the parameters of the model and $\varepsilon(\boldsymbol{s}_i)$ is the residual, or model error. Because the residual is assumed to have constant variance the model is trained by minimizing the sum of squared prediction errors of the training data, where every observation contributes equally to this sum. In the case of measurement error filtered RF, the measurement error variances are incorporated in the loss function of the model and each measurement is assigned a weight based on the inverse of the sum of the residual variance and the measurement error variance. The implementation of measurement error filtered RF therefore requires an estimate of the residual variance, which is obtained using an iterative procedure (van der Westhuizen *et al.*, 2022). In R this can be done in the *ranger* package because it allows assigning different weights.

For each soil property, we first applied ordinary RF without accounting for measurement errors (Scenario 1). Then, we applied the measurement error filtered RF that incorporates the measurement error variances (Scenario 2). As indicated above, in Scenario 2 we used the 'case.weights' argument of the *ranger* package to assign weights and to define the probability for each observation to be used in the bootstrapped samples used to build the trees. Less accurate measurements are given a lower probability to be sampled in each tree, effectively lowering their weight in the fitting of the model. To evaluate the effect of accounting for measurement errors on predictions we subtracted the final prediction results of Scenario 1 from those of Scenario 2 for each soil property.

4.2.4. Regression kriging model selection and calibration

As was the case with the RF model, 46 layers were retained to represent key soil-forming factors. Firstly, a correlation analysis was performed to reduce redundancy between the layers and detect highly correlated pairs of layers. Only layers with a correlation coefficient ≤ 0.75 with all other layers were considered for later analyses. Then, the set of covariates for each of the three soil properties were selected by combined forward and backward stepwise regression using the Bayesian Information Criterion (*BIC*) (Gao and Song, 2010). During the model selection procedure, we ignored spatial correlation and measurement errors in soil data.

We used the restricted maximum likelihood (REML) to estimate the variogram parameters and regression coefficients, as explained in more detail in **Chapter 3**. REML estimates the variogram parameters by maximising a conditional likelihood, after removing the dependence of the variogram parameters on the regression coefficients (Lark and Cullis, 2004), and using a numerical optimisation

technique. Then, the regression coefficients are estimated, based on the already estimated variogram parameters.

4.2.5. Assessment of RF and RK model performance

RF and RK were compared not only for their ability to predict the three soil properties of interest in the study area, but also for their ability to accurately quantify the prediction uncertainty. The performance of both RF and RK models was assessed by three types of cross-validation techniques, namely leave-one-out (LOOCV), leave-cluster-out (LCOCV), and leave-site-out (LSOCV) cross-validation, using common validation metrics mean error (ME), root mean squared error (RMSE) and model efficiency coefficient (MEC) (Wadoux *et al.*, 2018). The performance of the models was also visualized using scatter density plots of observed versus predicted values.

Since the three classical validation metrics above do not provide information about the accuracy of the prediction uncertainty (Vaysse and Lagacherie, 2017), the performance of RF and RK in terms of prediction uncertainty assessment was evaluated and compared using the prediction interval coverage probability (PICP) (Shrestha and Solomatine, 2006; Malone *et al.*, 2011). We used the PICP90, which evaluates the proportion of validation data that fall within the 0.05 and 0.95 quantiles of the conditional distribution. A PICP90 value close to 0.90 indicates a good assessment of the uncertainty, while a PICP90 significantly greater or smaller than 0.90 suggests that the uncertainty is significantly overestimated or underestimated, respectively. To provide a visual assessment of the quality of the estimated uncertainty for each model, we also plotted PICP values for all prediction interval widths in accuracy plots against the actual proportion of measurements in each prediction interval. Ideally all points in the plots are close to the 1:1 line (Malone *et al.*, 2011).

4.2.6. Effects of measurement errors on predictions and prediction uncertainties of RK and RF

The ability of RF to account for measurement errors was discussed in Section 4.2.3.4. In RK, measurement error variances were quantified and incorporated in the covariance structure of the RK model as described in Section 8.2.7 of Webster and Oliver (2007). For details of how this method was applied to our data we refer to **Chapter 3**. The estimated parameters were used to predict at new locations and generate maps of soil properties under the two scenarios, together with those of the prediction uncertainty. The validation metrics used to evaluate the performance of the RF and RK models without (Scenario 1) and with (Scenario 2) accounting for measurement errors were the same as used before and based on a LOOCV. However, note that under Scenario 2 the validation data are uncertain too. This was accounted for as explained in **Chapter 3**.

4.2.7. Evaluating the extrapolation potentials of RF and RK

RK and RF models were also assessed for their ability to handle spatial extrapolation while predicting the three soil properties of interest. The assessments were made using LOOCV, LCOCV and LSOCV. As we move from LOOCV to LCOCV to LSOCV, the number of neighbouring values used when making the prediction decreases (Fig. 4.1), in such a way that we move from interpolation to extrapolation in geographical space. As a result, prediction accuracy assessed by cross-validation statistics may drop substantially. In **Chapter 3**, we found that it matters a lot for RK and found LCOCV as the best compromise among the three cross-validation techniques for evaluating prediction performance. We used the same approach to evaluate the performance drop for RF when moving from LOOCV to LCOCV to LSOCV and compared it with that of RK.

4.3. Results and Discussion

4.3.1. Descriptive statistics of soil properties and estimation of measurement errors

Descriptive statistics for the analytical and spectral data set are presented in Table 4.1. Soil pH varied from strong acidity to near neutral (3.54 to 6.91). The study area is dominated by clay-rich soils with clay values ranging from 37.9 to 100% with a mean of 65.5%. SOC ranged from 6.7 to 84.5 g kg^{-1} with a mean of 26.4 g kg^{-1}. Standard deviations of pH and clay were lower for spectral than for analytical data, which is likely caused by the smoothing effect of PLSR. Interestingly, for SOC the standard deviation of the spectral data was larger than for the analytical data. Similar to the PLSR analysis reported in Section 4.2.2 and in **Chapter 3**, given the positive skewness of SOC we did the RF and RK modelling on log-transformed SOC (i.e., logSOC).

Table 4.1: Summary statistics of soil properties for the analytical (n = 48), spectral (n = 432) and merged data sets (n = 480).

Variable	Min.	Mean	Max.	St.dev.	CV (%)	Skewness
pH						
Analytical	3.96	5.21	6.48	0.61	11.7	0.13
Spectral	3.54	5.23	6.91	0.56	10.8	0.43
Analytical + Spectral	3.54	5.23	6.91	0.57	10.9	0.37
Clay (%)						
Analytical	37.9	65.4	97.3	13.3	20.3	- 0.01
Spectral	41.7	65.4	100.0	10.9	16.7	0.36
Analytical + Spectral	37.9	65.5	100.0	11.2	17.1	0.33
SOC (g kg^{-1})						
Analytical	9.0	24.7	52.0	11.1	45.1	0.64
Spectral	6.7	26.8	84.4	12.9	48.1	1.15
Analytical + Spectral	6.7	26.4	84.4	12.6	47.9	1.12
LogSOC						
Analytical	2.21	3.11	3.95	0.46	14.79	0.07
Spectral	0.19	3.18	4.44	0.46	14.47	0.15
Analytical + Spectral	0.19	3.17	4.44	0.46	14.51	0.13

St.dev. = standard deviation, CV = coefficient of variation.

As reported in **Chapter 3**, the estimated analytical measurement error standard deviations were 0.083 for pH, 3.33 % for clay and 0.038 for logSOC, while those of the PLSR prediction error estimated from the residual variance were 0.219 for pH, 5.47 % for clay and 0.192 for logSOC. The PLSR prediction error variance added to the analytical error variance gave a total measurement error standard deviation for spectral data of 0.234, 6.40 %, and 0.196 for pH, clay and logSOC, respectively.

4.3.2. Random Forest modelling and comparison with regression kriging

4.3.2.1 Model selection and calibration

RF and RK selected a different number of environmental variables used as covariates. For RF, the RFE algorithm selected 13, 15 and 23 covariates for pH, clay and logSOC respectively, while the RK model selection procedure resulted in 9, 4 and 5 covariates for pH, clay and logSOC respectively (**Chapter 3**). The selected covariates are listed in Tables S4.2 (RF) and S4.3 (RK). For both RF and RK, DEM derivatives, climate parameters, land cover and MODIS net primary productivity emerged as the most relevant variables. RF models selected more covariates than RK. This was expected, because their

prediction performance is less influenced by collinearity, which is partly reduced by the random selection features (Dormann *et al.*, 2013). Furthermore, RF is recognized for its good predictive ability when using many covariates, whereas stepwise regression reduces the number of covariates by mitigating multicollinearity. RK also benefits from residual spatial autocorrelation (Jin *et al.*, 2018).

4.3.2.2 Cross-validation

The LOOCV cross-validation statistics of the RF and RK models for the three soil properties are presented in Table 4.2. Note that Table 4.2 also presents LCOCV and LSOCV cross-validation statistics, which will be discussed and interpreted in Section 4.3.4. For this case study, the LOOCV ME, RMSE and MEC statistics show that both models had good predictive ability for pH and logSOC, but less so for clay. This indicates that for pH and logSOC the selected covariates successfully captured an important part of the spatial variation of the soil property. RK also benefited from spatial interpolation of the stochastic residual, because the residual semivariograms showed that there was significant residual spatial structure (**Chapter 3**).

RK outperformed RF by presenting generally higher MEC and lower RMSE values (Table 4.2). The change in RMSE values (ΔRMSE) and MEC values (ΔMEC) show the relative improvement in prediction accuracy of the RK models over RF for all three properties. Table 4.2 shows this only for LOOCV, since LCOCV and LSOCV had similar trends. The improvement in RMSE is about 10, 12 and 2 % for pH, clay and logSOC respectively, while the improvement accuracy in MEC is on average of 5 %, 22 % and 1 % for pH, clay and logSOC, respectively.

Table 4.2: Random Forest and regression kriging statistical validation metrics obtained by three cross-validation methods.

	Random Forest						Regression Kriging							
Cross validation	ME		RMSE		MEC		ME		RMSE		MEC		ΔRMSE %	ΔMEC %
	1	2	1	2	1	2	1	2	1	2	1	2		
							pH							
LOOCV	-0.003	-0.003	0.237	0.240	0.794	0.790	-0.001	-0.002	0.213	0.214	0.834	0.832	10.5	-5.2
LCOCV	0.001	0.000	0.307	0.306	0.656	0.657	0.012	0.010	0.289	0.288	0.695	0.696		
LSOCV	0.047	0.039	0.605	0.696	-0.338	-0.299	-0.029	-0.031	0.298	0.297	0.675	0.678		
							Clay							
LOOCV	0.005	0.005	6.260	6.253	0.547	0.548	0.020	0.021	5.730	5.730	0.621	0.621	12.4	-22.7
LCOCV	-0.048	-0.013	9.426	9.380	-0.026	-0.016	0.177	0.160	7.805	7.787	0.297	0.300		
LSOCV	-0.930	-0.993	9.582	9.528	-0.060	-0.048	-0.495	-0.458	8.420	8.410	0.182	0.184		
							logSOC							
LOOCV	0.002	0.005	0.223	0.224	0.717	0.715	0.000	0.001	0.217	0.218	0.729	0.729	1.8	-1.1
LCOCV	-0.001	0.004	0.257	0.258	0.624	0.621	-0.004	-0.003	0.267	0.267	0.594	0.594		
LSOCV	0.010	0.021	0.380	0.391	0.180	0.132	0.015	0.018	0.285	0.285	0.536	0.537		

1 = ignore measurement errors (Scenario 1); 2 = account for measurement errors (Scenario 2), LOOCV = leave one out, LCOCV = leave cluster out, LSOCV = leave site out cross-validation, ΔRMSE = change in RMSE values (%), ΔMEC = change in MEC values (%).

The plots of predicted against measured values from LOOCV provide a visual guide of the deviation of the predicted values of soil properties from the collocated measured values (Fig. 4.2). Both models generally present a similar pattern, but the superiority of RK over RF is demonstrated by less scatter around the 1:1 line. This corroborates well with the lower RMSE and the higher MEC values of RK for the three soil properties. As with any interpolation method, RK and RF overpredicted lower and underpredicted higher values. Interestingly, values around the mean were underpredicted for pH and logSOC and overpredicted for clay.

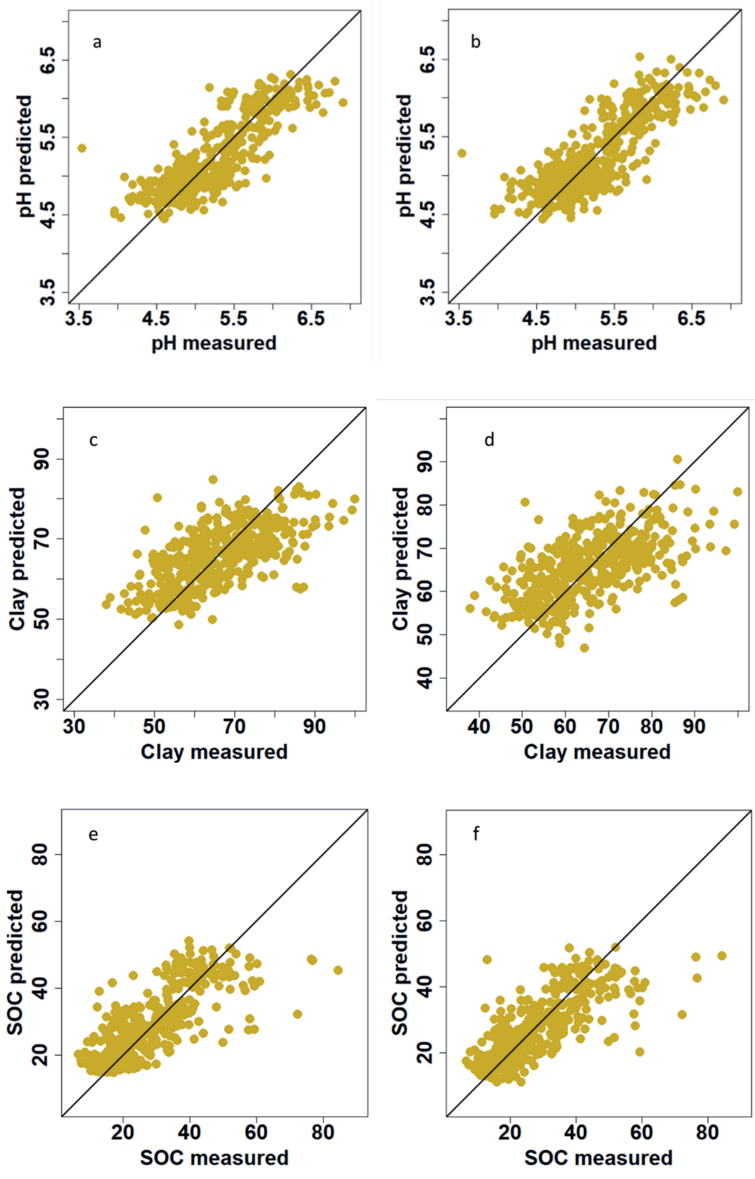


Fig. 4.2: Scatter plots of predicted vs measured from LOOCV; a) pH with RK; b) pH with RF; c) clay with RK (%); d) clay with RF (%); e) SOC with RK (g kg^{-1}); f) SOC with RF (g kg^{-1}).

The relative improvement in prediction accuracy of RK over RF is remarkable because most studies report that RF has higher performance than RK (Hengl *et al.*, 2015; Mariano and Mónica, 2021) with some exceptions (Pouladi *et al.*, 2019; Makungwe *et al.*, 2021). The superiority of RK over RF in this Chapter could be due to the fact that in RF, the spatial dependency of data is ignored. In essence, RF is a non-spatial method that ignores the geographical locations of the measurements (Heuvelink and Webster, 2022). This can potentially lead to sub-optimal predictions, especially when the target variable exhibits a strong residual spatial correlation and when sampling points are strongly spatially clustered (i.e., when the sampling density is dense in some areas and sparse in others). Note that we had strongly clustered data in our case study. There is a fundamental difference between the sampling design requirements for kriging and ML (Wadoux *et al.*, 2019). RK requires a sampling design that is optimized for estimating the trend, fitting the variogram and kriging interpolation (i.e., data must contain a good mix of small, medium and large lag distances, but also cover the geographic space well), whereas ML requires a sampling design that is optimized for coverage of the feature space (Mulder *et al.*, 2013; Ballabio *et al.*, 2016; Ma *et al.*, 2020). The data set used in this Chapter was not optimized for ML, and this consequently may have contributed to its poor relative performance. The sampling design was also not optimized for RK, but since there were many clustered measurements, the design did allow accurate estimation of the residual semivariogram, especially the nugget variance, which impact the krige value.

4.3.2.3 Quantification of prediction uncertainties

Accuracy plots for RF and RK built on LOOCV cross-validation for the three soil properties are shown in Fig. 4.3. The curves of the two models are close to the 1:1 line for $\alpha < 0.4$ for pH and clay and $\alpha < 0.3$ for SOC. This shows that at lower confidence levels, the PI performs as expected and validates the uncertainty assessment for this interval width (Arrouays *et al.*, 2014). However, for $\alpha > 0.4$ for pH and clay and $\alpha > 0.3$ for SOC, there is an increased deviation from the 1:1 line. The fact that the graphs are all above the 1:1 line, indicates that the predictions of uncertainty were overestimated by both models. In other words, the PIs are unnecessarily wide.

In general, the deviations from the 1:1 line are larger for RK than for RF, indicating that the overestimation of the local uncertainty of RK was higher than that of RF. For example, for pH, we find that for RK, 63% of the measurements is included in the 50% prediction interval, while it is 56% for RF. This indicates a better assessment of prediction uncertainty by RF as compared to RK, because for RF the observed percentage is closer to the expected percentage. Note that we could not test whether these differences are statistically significant because the validation data are not a probability sample from the study area. The same tendencies were observed for clay and SOC.

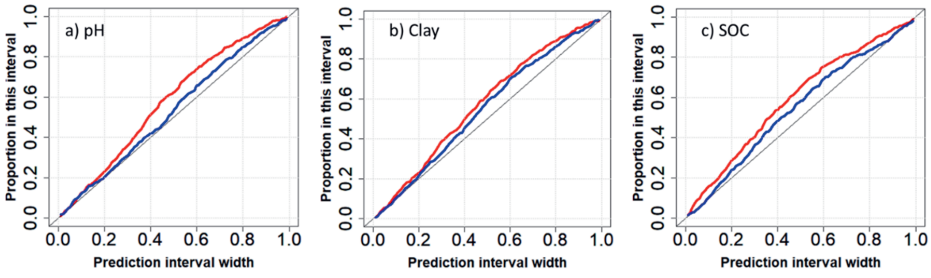


Fig. 4.3: Accuracy plots for a) pH, b) clay and c) SOC obtained using leave-one out cross-validation. Red lines represent regression kriging, blue lines random forest.

Although RF recorded slightly worse accuracy metrics in predicting soil properties compared to RK, its performance showed a better accuracy in terms of quantifying the prediction uncertainties. For instance, low performances for predicting clay values were registered for RF with lower RMSE and higher MEC values (Table 4.2), whereas accuracy plots revealed that RF delivered improved uncertainty predictions for this soil property (Fig. 4.3). These conflicting performances have been reported in previous studies that compare the performance of RK and RF (Vaysse and Lagacherie, 2017; Szatmári *et al.*, 2019). Consequently, standard validation metrics to compare the performance of prediction models should be accompanied by PICP and accuracy plots to also evaluate the ability of DSM models to assess the prediction uncertainties.

4.3.2.4. Spatial predictions

The maps of the spatial predictions of pH, clay and SOC predicted at 250 m resolution by RF and RK are presented in Fig. 4.4. Given the different models and different covariates selected, we observed relative absolute differences in predicted values in some parts of the study area between the two models of up to 0.8 for pH, 20% for clay and 20 g kg^{-1} for SOC. The differences are also illustrated in the scatter plots presented in Fig. 4.5.

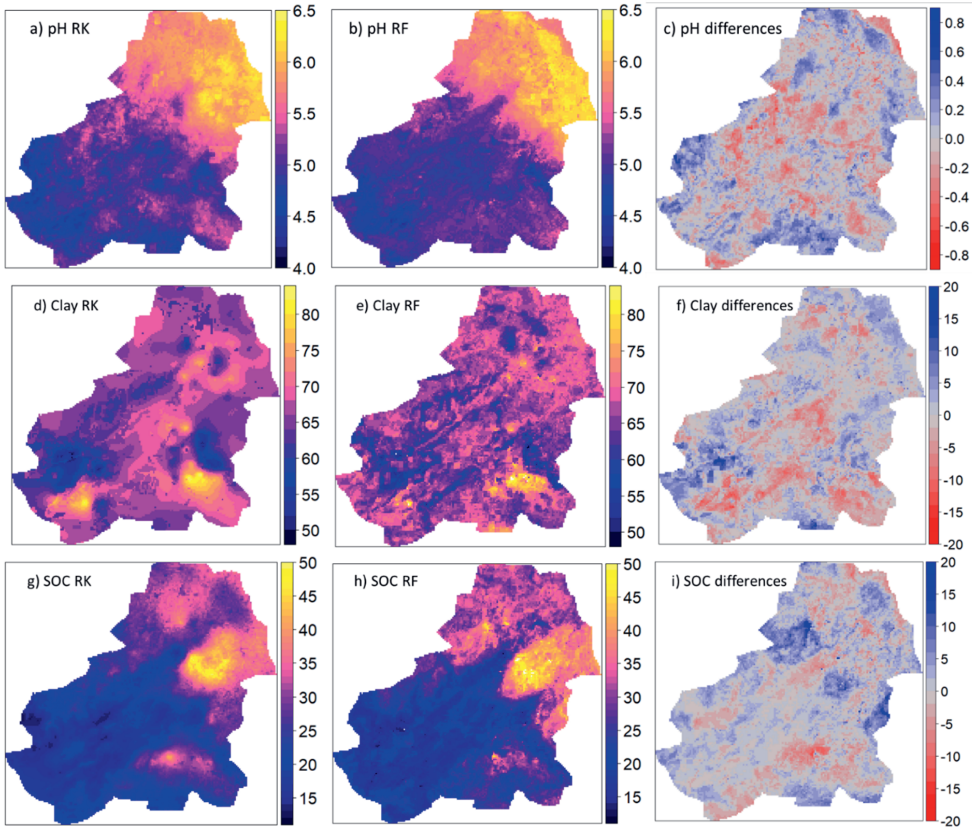


Fig. 4.4: Maps of RF and RK predictions and their differences: a) pH with RK, b) pH with RF, c) pH differences between RK and RF, d) clay with RK (%), e) clay with RF (%), f) clay differences between RK and RF (%), g) SOC with RK (g kg^{-1}), h) SOC with RF (g kg^{-1}), i) SOC differences between RK and RF (g kg^{-1}).

We observed significant differences between soil properties as exhibited by the residual variograms (**Chapter 3**, Fig. 3.3). pH and SOC were spatially structured, while clay exhibited less spatial structure and had a much smaller variogram range. This, together with the low amount of variance explained by environmental covariates for clay (**Chapter 3**), could explain the poor performance of the spatial models for clay. Predictions for pH and SOC were high in the northern parts of the study area, which is dominated by low altitude and warm climate. Unlike clay, which showed homogeneous spatial patterns, uniformly low SOC content was observed in the southern parts dominated by higher altitudes with low temperature and low vegetation covers. The spatial distribution of SOC is dominated by the influence of landscape attributes (Takoutsing *et al.*, 2018) and anthropogenic activities (Godswill *et al.*, 2016).

The spatial predictions of RF and RK were also compared using scatter density plots of RK versus RF predictions (Fig. 4.5). Despite the similarities in maps, the scatter around the 1:1 line attests the differences between the values predicted by RF and RK, with considerable differences in absolute values of up to 0.8, 20 % and 19 g kg^{-1} for pH, clay, and SOC respectively.

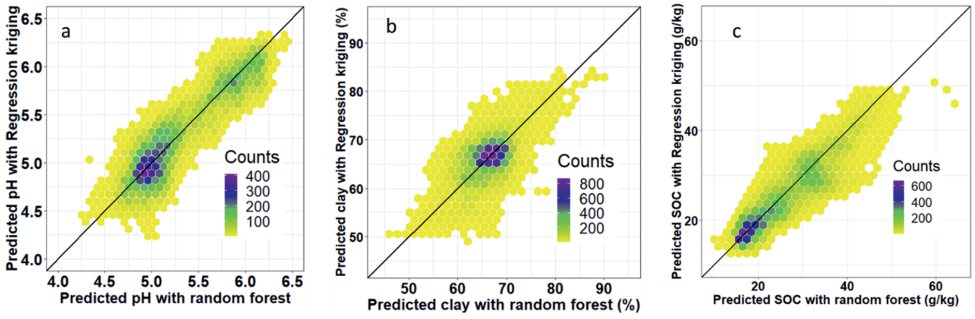


Fig. 4.5: Scatter density plots of RK versus RF predictions with the 1:1 line for: a) pH, b) clay and c) SOC.

The general appearance of the prediction uncertainty maps associated with each modelling approach differed notably (Fig. 4.6). The RK uncertainty maps vary much less across the study area and the largest prediction uncertainties are concentrated in areas not sampled. Thus, the prediction uncertainty, particularly for pH and clay provided by RK, depends mainly on the sampling configuration. For SOC, the spatial sampling design does not appear in the prediction uncertainty maps in the same way, because after back-transformation, these maps are not only influenced by the logSOC standard deviation but also by the logSOC prediction (Lark and Lapworth, 2012). Conversely, the RF uncertainty maps showed spatial patterns not related to the density of sampling locations but rather to the distribution of the explanatory variables. The prediction uncertainty maps for RK had lower range values than those for RF for the entire study area (Fig. 4.6). Assuming that the prediction uncertainty is correctly assessed (see Section 3.3.2) this attests the superiority of RK over RF in prediction performance and corroborates well with the cross-validation results (Table 4.2).

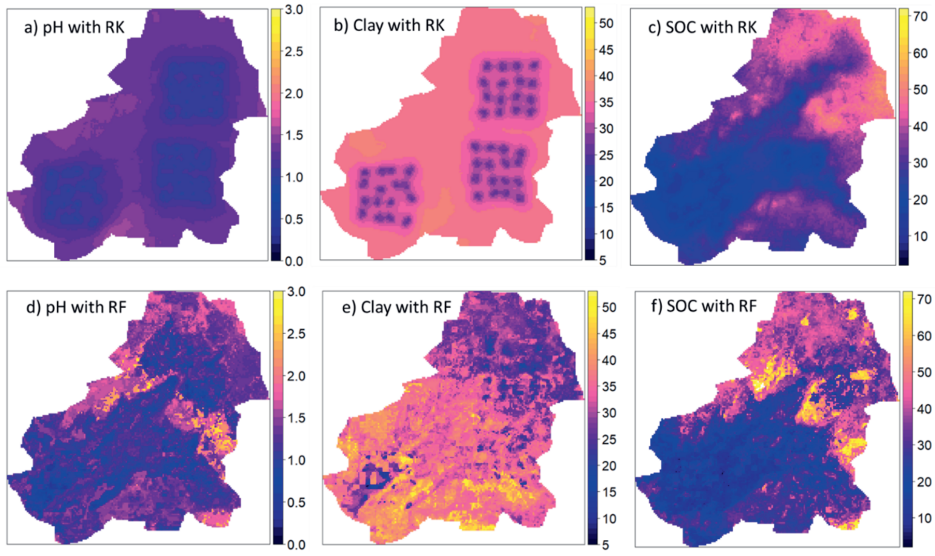


Fig. 4.6: Width of 90% prediction intervals: a) pH with RK, b) clay with RK (%), c) SOC with RK (g kg^{-1}), d) pH with RF, e) clay with RF (%), f) SOC with RF (g kg^{-1}).

4.3.3. Incorporation of measurement errors in random forest and regression kriging

The effect of accounting for measurement errors in RF and RK was assessed by comparing two Scenarios: Scenario 1, in which measurement errors were ignored, and Scenario 2, which accounts for measurement error variances.

4.3.3.1 Model selection and calibration

We selected covariates by applying RFE under Scenario 1 and used this same set of selected covariates in Scenario 2 (see Section 4.3.2.1). While the RF models by construction used the same covariates in Scenarios 1 and 2, the incorporation of measurement errors modified the covariate importance values for most of the covariates. As explained in Section 4.2.3.1, we used the RFE algorithm for model selection that rank-ordered the covariates and the ones with the least importance based on variable importance metrics (i.e., permutation-based mean square error reduction) iteratively removed. The final variable importance values were scaled from 0 to 100. We observed relative changes in the importance values of covariates between Scenario 1 and Scenario 2 of up to 22%, 8% and 30 % for pH, clay and logSOC, respectively (Table S4.2, Fig. 4.7). Scenario 2 showed a few covariates with much higher importance values.

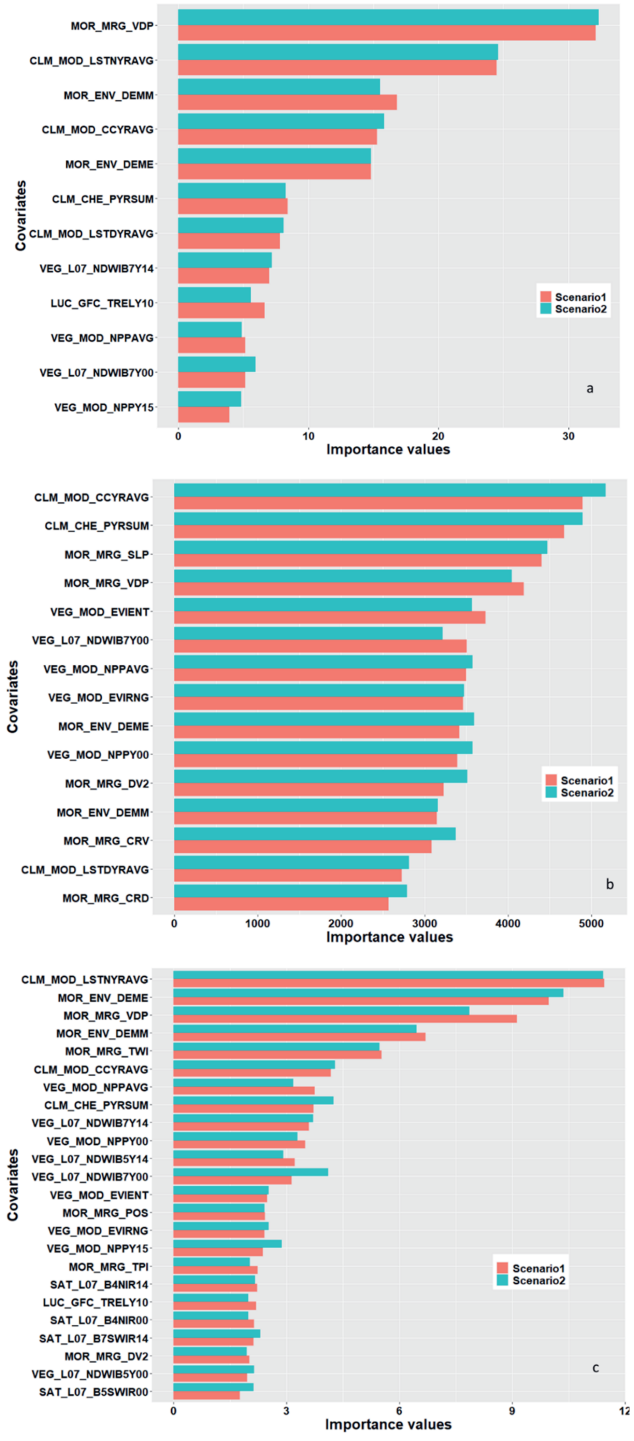


Fig. 4.7: Importance rankings of covariates for soil properties in RF models: a) pH; b) clay; c) logSOC. Red bars represent Scenario 1, green bars Scenario 2.

For RK, the inclusion of measurement errors modified the regression coefficient estimates, particularly for pH, where coefficient changes were up to 29% for some of the covariates (Table S4.3). Since pH had the largest number of covariates selected, it could be more sensitive to the incorporation of measurement error because of collinearity effects. To further avoid multicollinearity, we could have used PLRS instead of stepwise multiple linear regression. Previous studies have also demonstrated the effect of measurement errors on variogram parameters and regression coefficient estimates (Clark, 2010).

Although the RK residual variograms in both scenarios exhibited similar structures, we observed a decrease in the nugget and sill values for Scenario 2 (**Chapter 3**, Fig. 3.3). This could be attributed to the fact that in Scenario 2, the nugget represents only spatial variation at short distances and does not include measurement error variance (Chilès and Delfiner, 2012). In other words, part of the observed short-distance spatial variation of soil properties is explained by measurement errors, meaning that the spatial variation of the true (error-free) soil properties is lower than that of the measurements. We also observed an increase in the range parameter after the inclusion of measurement errors, particularly for clay.

4.3.3.2 Cross-validation and statistical evaluation of prediction uncertainty

The cross-validation metrics for RF and RK under the two scenarios are summarized in Table 4.2. While comparing the two scenarios, a similar performance was observed for RF and RK, attesting no significant change with the incorporation of measurement uncertainty. However, in all cases and as already noted in Section 4.3.2.2, RK outperformed RF with lower RMSE and higher MEC. The same trends were observed for both scenarios. As revealed by the prediction interval coverage probability plots based on LOOCV cross-validation, the curves for both scenarios deviate from the 1:1 line and show that both models tend to overestimate the prediction interval widths (Fig. 4.8). The deviations for RF were similar for both scenarios meaning that the quantification of prediction uncertainty was not much influenced by the measurement errors. For RK, the deviation from the 1:1-line was somewhat larger for Scenario 1 than for Scenario 2. For example, for pH, we found that for Scenario 1, 64% of the measurements were included in the 50% prediction interval, while it was only 57% for Scenario 2. This indicates that Scenario 2 represented the prediction uncertainty better than Scenario 1, although the difference was small.

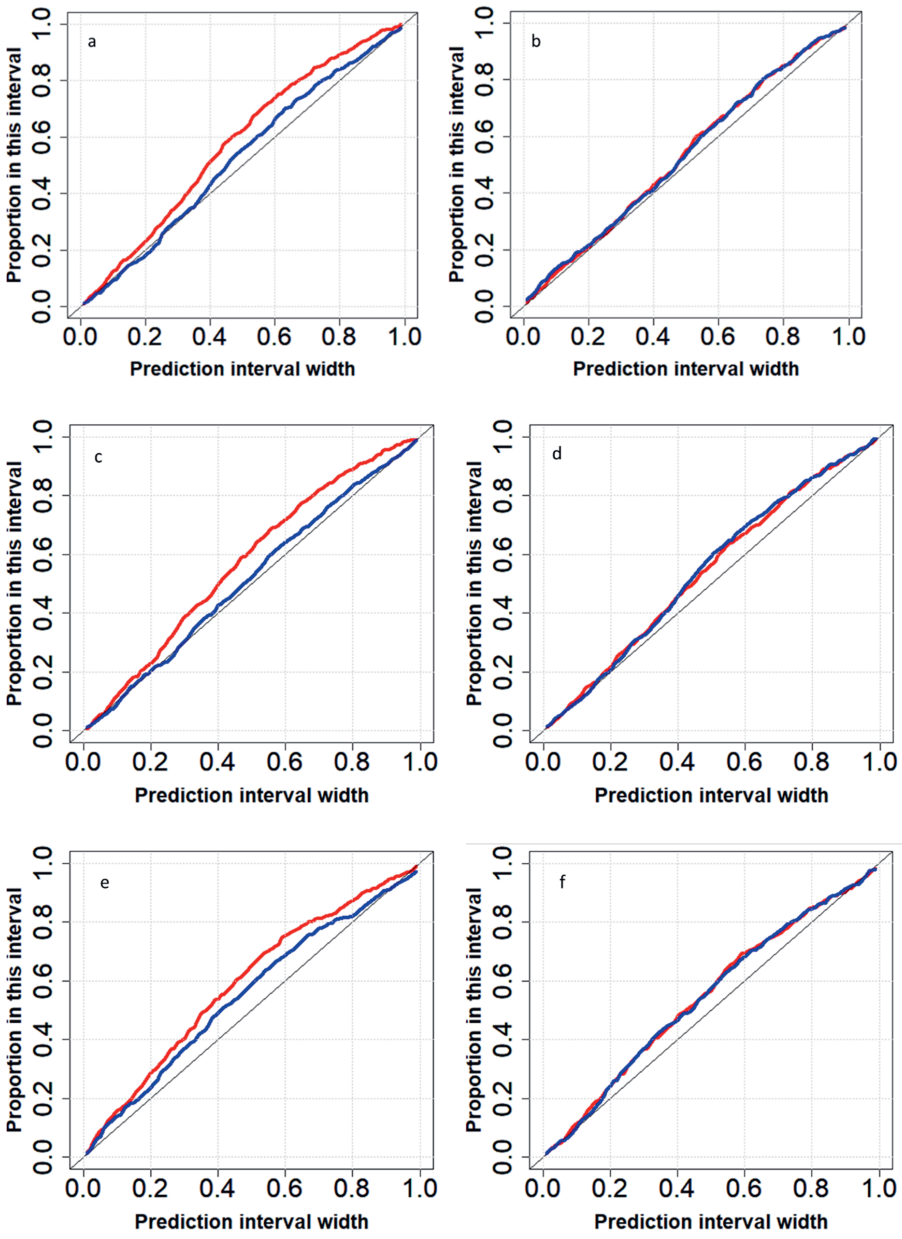


Fig. 4.8: Accuracy plots under the two scenarios obtained using leave-one out (LOOCV) cross-validation: a) pH with RK, b) pH with RF, c) clay with RK, d) clay with RF, e) SOC with RF, f) SOC with RF. Red lines represent Scenario 1, blue lines Scenario 2.

4.3.3.3 Spatial predictions

Broadly, the prediction surfaces of RF and RK under the two scenarios have similar spatial patterns. Therefore, only the associated map differences between the two scenarios are presented in the

supplementary information (Fig. S4.1). Both for RK and RF we observed relative differences in predicted values between the two scenarios, attesting the influence of measurement errors on the predicted values. However, the differences were small and therefore accounting for measurement uncertainty had only a small effect on the prediction maps in this Chapter. The map differences were also remarkably similar for RK and RF. Locations where RK had a smaller (greater) prediction for Scenario 1 than for Scenario 2, were almost the same as locations where RF had a smaller (greater) prediction for Scenario 1 than for Scenario 2. This is remarkable, because the models have a very different structure and used different covariates.

Scatter density plots of predictions between Scenario 1 and Scenario 2 for RF and RK are presented in Fig. 4.9 and show some scatter around the 1:1 line, which attests the effects of accounting for measurement errors. While comparing the two models, the predicted values between scenarios matched well for RK with values aligned close to the 1:1 line, while in contrast, significant scatter was observed for pH, clay, and SOC around the 1:1 line for RF. The large scatter for RF indicates that RF is more sensitive to accounting for measurement errors than RK.

Although no systematic differences were observed between the predictions generated by the two scenarios, relative absolute differences in predicted values in some areas of the study area between the two scenarios were up to 0.2, 4.2 % and 5.3 g kg^{-1} for pH, clay, and SOC for RF, and up to 0.1, 1.6% and 2 g kg^{-1} for pH, clay and SOC for RK. The differences in predicted values between the two scenarios are larger for RF than for RK, which confirms that RF was more sensitive to accounting for measurement errors than RK.

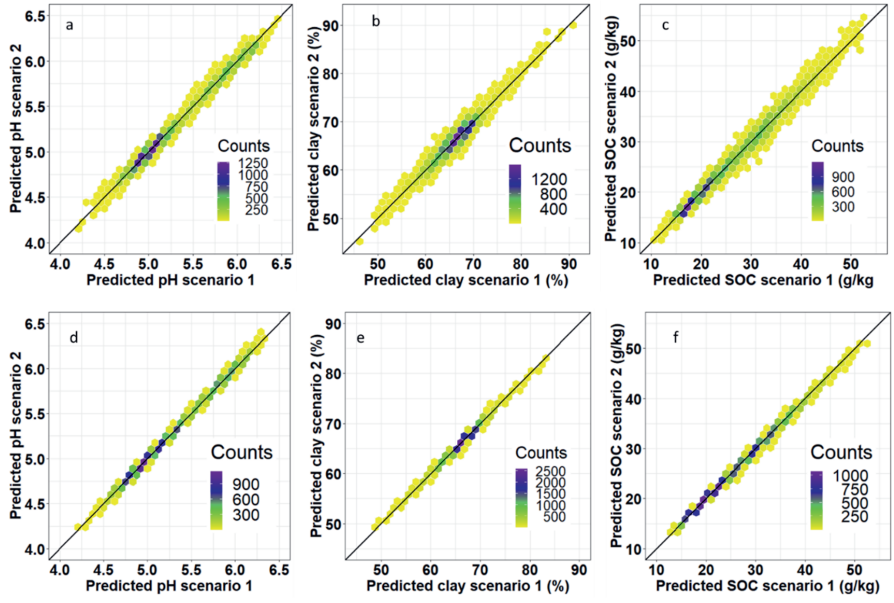


Fig. 4.9: Scatter density plots of predicted values for Scenario 1 and Scenario 2 for a) pH with RF, b) clay with RF, c) SOC with RF, d) pH with RK, e) clay with RK, f) SOC with RK. The solid black is the 1:1 line.

We observed small differences between the PI widths of the two scenarios for both models (Fig. S4.2). For instance, the mean SOC 90 % PI width for Scenario 1 was 30.83 g kg^{-1}, with the narrowest interval

being 2.6 g kg^{-1} and the widest interval 72.1 g/kg. The 90 % PI width for Scenario 2 was 30.5 g kg^{-1}, with a minimum of 3.9 g kg^{-1} and a maximum of 73.4 g kg^{-1}. Similar to the prediction case, relative differences were observed in PI widths between scenarios of up to 1.27 for pH, 15.5 % for clay and 34 g kg^{-1} for SOC.

For RK, we observed relative differences in prediction uncertainty maps in some areas of the study area between the two scenarios of up to 0.28 for pH, 9 % for clay and 16 g kg^{-1} for SOC (Fig. S4.3). The differences between the two scenarios may have been small in our case study because 90% of the data (i.e., the spectral data) had the same measurement error variance, leading to homogenous weights during the model fitting. Although the analytical data had much smaller measurement error variances than the spectral data, and therefore had much larger weights in Scenario 2, they represented only 10% of all data and had a spatial distribution similar to that of the spectral data (Fig. 4.1). In a case where the analytical and spectral data would have been of similar size, or located in different parts of the study area, we would expect larger differences between scenarios, both for predictions and prediction uncertainties.

4.3.4. Assessing the extrapolation potential of RF and RK

We performed LOOCV, LCOCV and LSOCV to evaluate the extrapolation potentials of RF and RK. The validation metrics for RF and RK under the two scenarios show that RF performed worse at extrapolation than RK (Table 4.2). As we move from LOOCV to LCOCV to LSOCV, we observe a decrease in model performance with increases in RMSE and decreases in MEC values, but the decrease in performance was much larger for RF than RK for all three soil properties (Table 4.2). For instance, for RF the RMSE values for pH in Scenario 1 increased from 0.237 to 0.307 to 0.605, while for RK they increased more moderately from 0.213 to 0.289 to 0.298. The same trends were observed for clay and logSOC. Negative values were observed for MEC under LCOCV and LSOCV for pH and clay. This attests that the RF models perform worse than using the mean of all measurements as predictor (McCuen *et al.*, 2006) and exhibit the low extrapolability potentials of RF as compared to RK.

The scatter plots of observed versus predicted for the three cross-validations are presented in Fig. 4.10. Since the plots exhibited similar trends for each of the three soil properties, only those of pH are shown. The scatter for LCOCV and particularly LSOCV is much larger for RF than for RK. This corroborates with the negative MEC values obtained in Table 4.2 and attests the superiority of RK over RF in terms of extrapolation potential. Generally, the ability of RF to capture complex nonlinear interactions between predictors would suggest that this technique is better able to extrapolate compared to RK. However, the flexibility of RF to fit non-linear relations also makes it more vulnerable to poor coverage of the feature space, due to the fact that a complex model is tuned towards the training data, and the predicted values are never outside the range of training data set (Fox *et al.*, 2020; Meyer and Pebesma, 2021). Consequently, this can lead to nonsensical predictions if applied outside the range of the training data set (Booker and Whitehead, 2018).

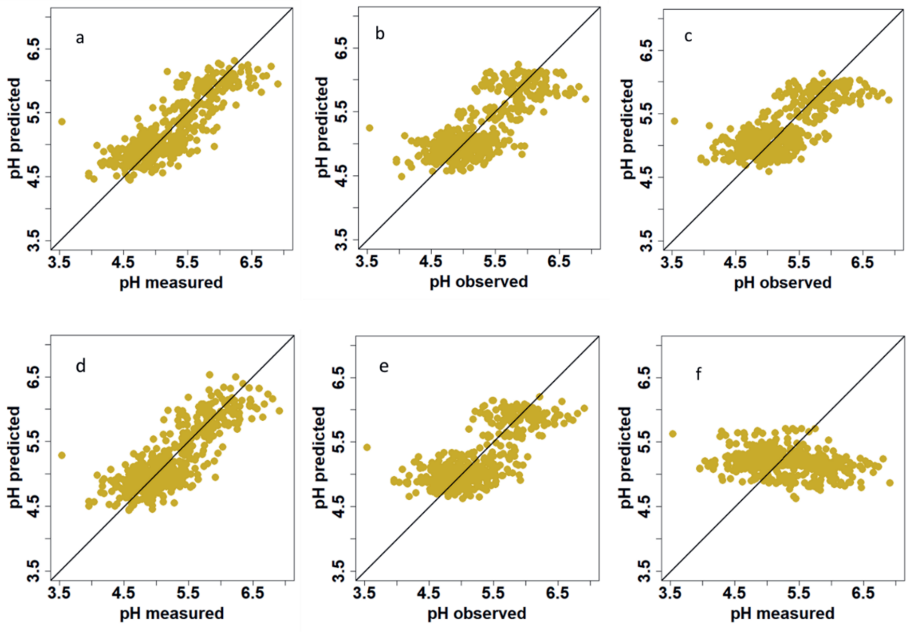


Fig. 4.10: Scatter plots for the three types of cross-validation for pH: a) LOOCV with RK, b) LCOCV with RK, c) LSOCV with RK, d) LOOCV with RF, e) LCOCV with RF, and f) LSOCV with RF.

4.4. General discussion

4.4.1. Comparing the performance of RF and RK

RK yielded more accurate predictions than RF (Table 4.2), while RF better quantified the prediction uncertainties (Fig. 4.3). There is currently no general consensus on how to explain the performance differences of spatial models. For instance, the lower prediction performance of RF in this Chapter may simply be because the sampling design was not suitable for ML algorithms, but it may also be very case-study dependent. Since no one-size-fits-all algorithm exists, the choice of the optimal spatial interpolation depends on the spatial data structure, the sampling design, the covariates, and the cross-validation techniques used (Guevara *et al.*, 2018). Consequently, without more case studies, it is difficult to draw conclusion with certainty which model approach is the most accurate in a given study. One of the possible extensions for modelers to overcome this challenge in comparing RK and RF is to combine RF with kriging by incorporating non-linear trends and use a hybrid model called 'random forest kriging' (Keskin and Grunwald, 2018; Szabó *et al.*, 2019). In the hybrid model, the deterministic component of spatial soil variation is modelled by RF, whereas the stochastic part of variation is modelled by kriging the RF residuals.

4.4.2. Accounting for measurement errors

Although the two models used the same covariates in both scenarios, we observed differences in variable importance values of RF models and the regression coefficient estimates of RK models after

accounting for measurement errors (Fig. 4.7, Tables S4.2 and S4.3). The effect of measurement errors on covariate importance values and regression coefficient estimates have not been fully investigated and to the best of our knowledge, it is unclear what caused the differences. However, previous studies have reported that measurement errors could affect machine learning by decreasing prediction performance (Nettleton *et al.*, 2010), increasing model complexity (Brodley and Friedl, 1999) and affecting feature selection and covariates importance values (Frénay *et al.*, 2014). In this Chapter, the insignificant influence of measurement errors on prediction performance could be because of the homogeneity of the measurement error variances. It should also be noted that feature importance is difficult to measure objectively when covariates are cross-correlated and their influence masked by other covariates (Wadoux *et al.*, 2020b). Models that use a large number of covariates appear to be more sensitive to measurement error. However, more studies are needed to evaluate if this is true more generally and whether there is a logical explanation for this effect.

Based on cross-validation (Table 4.2), the effects of measurement errors in general appeared not to be significant on the results obtained in this Chapter, but the relative differences between the two scenarios (Fig. S4.1) and the scatter plots (Fig. 4.9) acknowledges the effect of measurement errors, as well as indicates the higher sensitivity of RF to measurement errors compared to RK. A similar conclusion was reported by van der Westhuizen et al. (2022) which compared an error-filtered random forest with a regression kriging model and where the nonsignificant difference was attributed to the low variability of the measurement error variances. The magnitude and direction of the impact of measurement error are not easily predictable, and with the growing popularity of machine learning and the availability of big data, it is important to quantify and incorporate measurement errors to assess their effects on DSM accuracies (Heuvelink, 2018; Somarathna *et al.*, 2018), particularly in case measurement error variances are large and have large differences between measurements.

4.4.3. Assessing extrapolability potentials

Although we expected predictions outside the training space to lead to poor performance for both RF and RK, we did not expect RF to perform so much worse than RK. Thus, for RF models it is particularly important to avoid extrapolation in feature space, by computing the 'Area of applicability' of spatial prediction models (Meyer and Pebesma, 2021). Previous studies also warned that decision-tree based methods such as RF or Boosted Regression Trees can completely fail if applied in areas that have not been used for training (Hengl *et al.*, 2018a). The decrease in model performance as we moved from interpolation to extrapolation also stressed the importance of sampling design for interpolation and extrapolation that should be optimized enough to ensure that the feature space is well covered by the calibration data (Wadoux *et al.*, 2019).

One of the valuable insights of this Chapter, was to demonstrate that despite its advantages in the context of limited resources and poor accessibilities, cluster sampling design has some deficiencies, not only for RK where it is difficult to estimate the semivariogram and perform kriging, but also for RF because there is a higher risk of extrapolation in feature space. This made RF perform worse than RK in this Chapter. In spite of the practical advantages of cluster sampling, it is highly recommended to combine it with other sampling methods that cover the geographic and feature space well (Brus and Heuvelink, 2007; Wadoux *et al.*, 2019).

4.4.4. Limitations

The results of this study have some limitations, which should be considered in further studies. Firstly, the measurement uncertainties were not reliably estimated, particularly for the analytical data. Only few duplicate measurements were available to quantify the measurement error variances. Laboratories

should pay more attention to this problem and systematically quantify and publish the uncertainties of their measurements and benchmark them against standards to minimize systematic bias. Secondly, we used a fairly small dataset (n = 480) considering the surface area of the study area (1,053 km^2) in a single case study. Therefore, it is difficult to draw conclusions on the outcomes of a single study, and more studies are needed to detect consistent patterns in the effect of measurement error on DSM predictions and uncertainty. Thirdly, this Chapter assumed constant measurement error variances for each of the soil properties, while in many practical cases measurement errors are proportional to the measured values (Libohova *et al.*, 2019). Measurement errors proportional to the measured values could have yielded different results and would probably have led to a bigger impact of uncertainties on the DSM outputs.

4.5. Conclusions

Using analytical and spectral soil data, this Chapter assessed and compared the ability of RF and RK to predict soil properties and quantify prediction uncertainties while accounting for soil measurement errors. The potential of these models to extrapolate in geographical space were also evaluated. We found that both models produced comparable ranges of predicted values and similar maps of soil properties of interest, but that RK outperformed RF as shown by the validation metrics. Conversely, the prediction uncertainties were better quantified by the RF models as revealed by the accuracy plots. Overall, the results of this Chapter indicate that RF and RK were not much affected by measurement errors, and this is because most calibration data had the same measurement error variance. The potential to extrapolate was better for RK than RF, as RK was able to derive better predictions at unsampled sites based on validation metrics. This indicates that care should be taken when using machine learning algorithms for soil mapping in areas with strongly clustered data and large unsampled areas. The findings confirmed that there is no universal model that works best in all cases, and that model performance is case-dependent. Since, both RK and RF could perform better, it is advisable to compare performances of models using cross-validation modes that also evaluate uncertainty quantification and extrapolation potentials.

4.6. Supplementary information

Table S4.1: Description of environmental variables used as covariates in the study. The plus (+) and minus (–) signs indicate whether a covariate was selected for a soil property.

	Covariate codes	Descriptions	Sources	pH	Clay	logSOC
1	CLM_CHE_PYRSUM	Total annual precipitation	CHELSEA	+	+	+
2	CLM_MOD_CCYRAVG	Mean annual cloud cover	EarthEnv	+	+	+
3	CLM_MOD_LSTDYRAVG	Mean annual surface temperature	MODIS	+	+	-
4	CLM_MOD_LSTNYRAVG	Mean annual LST (nighttime) MODIS	MODIS	+	-	+
5	LUC_GFC_BARLY10	30 Meter Global Land Cover: Bare soil	ESA	-	-	-
6	LUC_GFC_TRELY10	30 Meter Global Land Cover: Tree cover	ESA	+	-	+
7	MOR_ENV_DEME	DEM-parameters: land surface elevation	SRTM	+	+	+
8	MOR_ENV_DEMM	DEM-parameters: Merged elevation	SRTM	+	+	+
9	MOR_MRG_CRD	DEM-parameters: Local downslope Curvature	SRTM	-	+	-
10	MOR_MRG_CRU	DEM-parameters: Local upslope Curvature	SRTM	-	-	-
11	MOR_MRG_CRV	DEM-parameters: Downslope Curvature	SRTM	-	+	-
12	MOR_MRG_DV2	DEM-parameters: Deviation from Mean Value (surface roughness)	SRTM	-	+	+
13	MOR_MRG_DVM	DEM-parameters: Deviation from Mean Value (surface roughness)	SRTM	-	-	-
14	MOR_MRG_MRN	DEM-parameters: Melton Ruggedness Number	SRTM	-	-	-
15	MOR_MRG_NEG	DEM-parameters: Negative Topographic Openness	SRTM	-	-	-
16	MOR_MRG_POS	DEM-parameters: Positive Topographic Openness	SRTM	-	-	+
17	MOR_MRG_SLP	DEM-parameters: Terrain slope	SRTM	-	+	-
18	MOR_MRG_TPI	DEM-parameters: Topographic Position Index	SRTM	-	-	+
19	MOR_MRG_TWI	DEM-parameters: SAGA Wetness Index	SRTM	-	-	+
20	MOR_MRG_VBF	DEM-parameters: Multiresolution Index of Valley Bottom Flatness	SRTM	-	-	-
21	MOR_MRG_VDP	DEM-parameters: Valley depth	SRTM	+	+	+
22	MOR_USG_F01	Landform class: Breaks/Foothills	SRTM	-	-	-
23	MOR_USG_F02	Landform class: Flat plains	USGS	-	-	-
24	MOR_USG_F04	Landform class: Hills	USGS	-	-	-
25	MOR_USG_F05	Landform class: Low Hills	SRTM	-	-	-
26	MOR_USG_F06	Landform class: Low mountains	USGS	-	-	-

27	MOR_USG_F07	Landform class: Smooth Plains	USGS		-	-	-
28	SAT_L07_B3RED00	Landsat Band 3 (red) for year 2000	Landsat		-	-	-
29	SAT_L07_B3RED14	Landsat Band 3 (red) for year 2014	Landsat		-	-	-
30	SAT_L07_B4NIR00	Landsat Band 4 (NIR) for year 2000	Landsat		-	-	+
31	SAT_L07_B4NIR14	Band 4 (NIR) for year 2014	Landsat		-	-	+
32	SAT_L07_B5SWIR00	Landsat Band 5 (SWIR) for year 2000	Landsat		-	-	+
33	SAT_L07_B5SWIR14	Landsat Band 5 (SWIR) for year 2014	Landsat		-	-	-
34	SAT_L07_B7SWIR00	Landsat Band 7 (SWIR) for year 2000	Landsat		-	-	-
35	SAT_L07_B7SWIR14	Landsat Band 7 (SWIR) for year 2014	Landsat		-	-	+
36	VEG_L07_NDWIB5Y00	Normalized Difference Water Index in 2010 with SW1 band	Global Change	Forest	-	-	+
37	VEG_L07_NDWIB5Y14	Normalized Difference Water Index in 2014 with SW1 band	Global Change	Forest	-	-	-
38	VEG_L07_NDWIB7Y00	Normalized Difference Water Index in 2010 with SW2 band	Global Change	Forest	+	+	+
39	VEG_L07_NDWIB7Y14	Normalized Difference Water Index in 2014 with SW2 band	Global Change	Forest	+		+
40	VEG_MOD_EVIENT	Entropy MODIS EVI	EarthEnv		-	+	+
41	VEG_MOD_EVIEVN	Evenness of MODIS EVI	EarthEnv		-	-	-
42	VEG_MOD_EVIMAX	Maximum MODIS EVI	EarthEnv		-	-	-
43	VEG_MOD_EVIRNG	Range MODIS EVI	EarthEnv		-	+	+
44	VEG_MOD_NPPAVG	Net Primary Productivity (2000-2015 average)	MODIS		+	+	+
45	VEG_MOD_NPPY00	Net Primary Productivity in 2000	MODIS		-	+	+
46	VEG_MOD_NPPY15	Net Primary Productivity in 2015	MODIS		+	-	-

Table S4.2: Estimated importance values of RF models for the environmental variables under Scenarios 1 and 2.

Covariates	Scenario 1	Scenario 2	Changes per covariate (%)
	pH		
MOR_MRG_VDP	32.12	32.34	-0.69
CLM_MOD_LSTNYRAVG	24.49	24.59	-0.42
CLM_MOD_CCYRAVG	15.29	15.84	-3.58
MOR_ENV_DEME	14.79	14.81	-0.17
MOR_ENV_DEMM	16.8	15.52	7.63
LUC_GFC_TRELY10	6.65	5.59	15.97
CLM_CHE_PYRSUM	8.39	8.23	1.92
VEG_MOD_NPPY15	3.94	4.84	-22.89
CLM_MOD_LSTDYRAVG	7.82	8.09	-3.47
VEG_L07_NDWIB7Y14	7.00	7.19	-2.78
VEG_MOD_NPPAVG	5.16	4.85	6.05
VEG_L07_NDWIB7Y00	5.16	5.92	-14.67
	Clay		
CLM_MOD_CCYRAVG	4896.1	5170.8	-5.61
MOR_MRG_VDP	4187.4	4048.1	3.33
MOR_MRG_SLP	4405.3	4474.2	-1.56
CLM_CHE_PYRSUM	4676.0	4893.8	-4.66
MOR_ENV_DEME	3415.9	3591.0	-5.13
MOR_ENV_DEMM	3146.6	3161.4	-0.47
MOR_MRG_CRV	3081.3	3375.2	-9.54
VEG_MOD_EVIRNG	3461.4	3473.1	-0.34
CLM_MOD_LSTDYRAVG	2726.3	2816.0	-3.29
VEG_MOD_NPPAVG	3499.5	3575.6	-2.17
VEG_MOD_EVIENT	3728.9	3567.4	4.33
VEG_MOD_NPPY00	3389.1	3574.8	-5.48
VEG_L07_NDWIB7Y00	3505.1	3219.2	8.16
MOR_MRG_DV2	3230.4	3511.8	-8.71
MOR_MRG_CRD	2571.1	2785.9	-8.35
	logSOC		
CLM_MOD_LSTNYRAVG	11.45	11.42	0.26
MOR_ENV_DEME	9.97	10.37	-4.01
MOR_MRG_VDP	9.13	7.86	13.91
MOR_ENV_DEMM	6.69	6.46	3.44
VEG_L07_NDWIB7Y14	3.60	3.71	-3.06
MOR_MRG_TWI	5.53	5.48	0.90
VEG_L07_NDWIB7Y00	3.14	4.11	-30.89
CLM_MOD_CCYRAVG	4.18	4.29	-2.63
VEG_MOD_NPPAVG	3.75	3.18	15.20
CLM_CHE_PYRSUM	3.72	4.25	-14.25
VEG_MOD_NPPY00	3.50	3.29	6.00
VEG_L07_NDWIB5Y14	3.22	2.91	9.63
VEG_MOD_NPPY15	2.37	2.87	-21.10
SAT_L07_B4NIR14	2.22	2.17	2.25
SAT_L07_B7SWIR14	2.13	2.31	-8.45
MOR_MRG_POS	2.43	2.41	0.82
SAT_L07_B4NIR00	2.14	1.98	7.48
LUC_GFC_TRELY10	2.19	1.98	9.59
MOR_MRG_DV2	2.01	1.94	3.48
VEG_MOD_EVIRNG	2.41	2.52	-4.56
MOR_MRG_TPI	2.23	2.03	8.97

VEG_MOD_EVIENT	2.48	2.53	-2.02
SAT_L07_B5SWIR00	1.76	2.13	-21.02
VEG_L07_NDWIB5Y00	1.95	2.14	-9.74

Table S4.3: Estimated regression coefficients for the environmental variables of RK models under Scenarios 1 and 2.

Covariates	Scenario 1	Scenario 2	Changes per covariate (%)
	pH		
Intercepts	7.27	7.6	-4.5
CLM_CHE_PYRSUM	-8.172E-04	-8.244E-04	-0.9
CLM_MOD_CCYRAVG	-1.446E-04	-1.804E-04	-24.8
MOR_MRG_CRU	8.370E-05	5.917E-05	29.3
MOR_MRG_TPI	-3.759E-04	-3.517E-04	6.4
MOR_MRG_VDP	-1.132E-04	-1.102E-04	2.7
MOR_USG_F02	-1.567E-03	-1.936E-03	-23.5
MOR_USG_F06	1.289E-03	1.413E-03	-9.6
SAT_L07_B4NIR00	1.845E-02	1.867E-02	-1.2
VEG_MOD_NPPY00	-4.133E-05	-4.664E-05	-12.9
	Clay		
Intercept	653.5	694.9	-6.3
CLM_CHE_PYRSUM	-8.691E-03	-8.585E-03	1.2
CLM_MOD_LSTDYRAVG	-1.918E-01	-2.058E-01	-7.3
MOR_USG_F02	-5.544E-02	-5.592E-02	-0.9
MOR_USG_F06	-2.668E-02	-2.758E-02	-3.4
	logSOC		
Intercept	3.547	3.568	-0.6
CLM_CHE_PYRSUM	-5.592E-04	-5.637E-04	-0.8
CLM_MOD_CCYRAVG	1.361E-04	1.344E-04	1.2
LUC_GFC_BARLY10	-4.213E-02	-4.316E-02	-2.5
MOR_MRG_VDP	-8.827E-05	-8.804E-05	0.3
MOR_USG_F04	1.340E-04	1.384E-04	-3.3

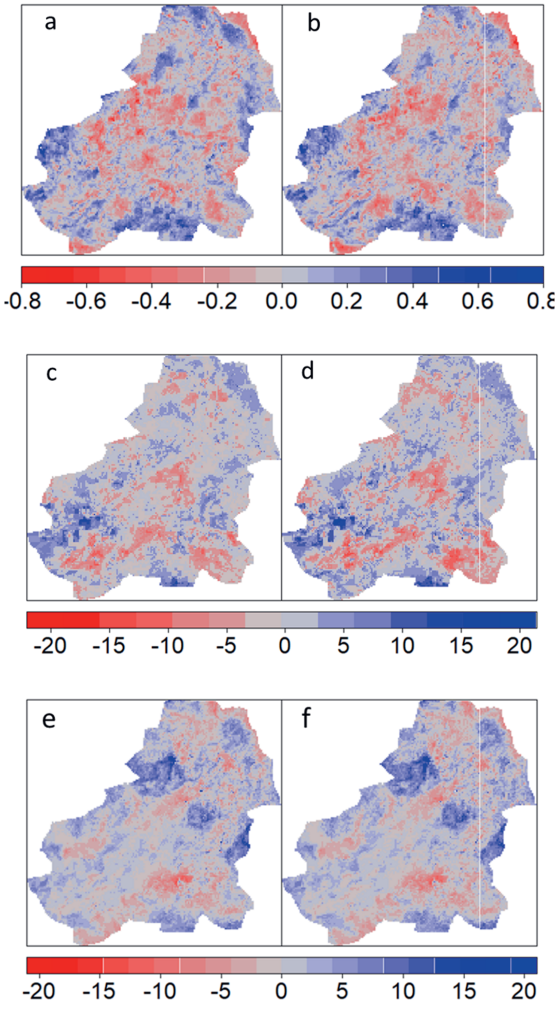


Fig. S4.1: Maps of prediction differences between Scenarios 1 and 2: a) pH map differences for RF, b) pH map differences for RK, c) clay map differences for RF, d) clay map differences for RK, e) SOC map differences for RF, f) SOC map differences for RK.

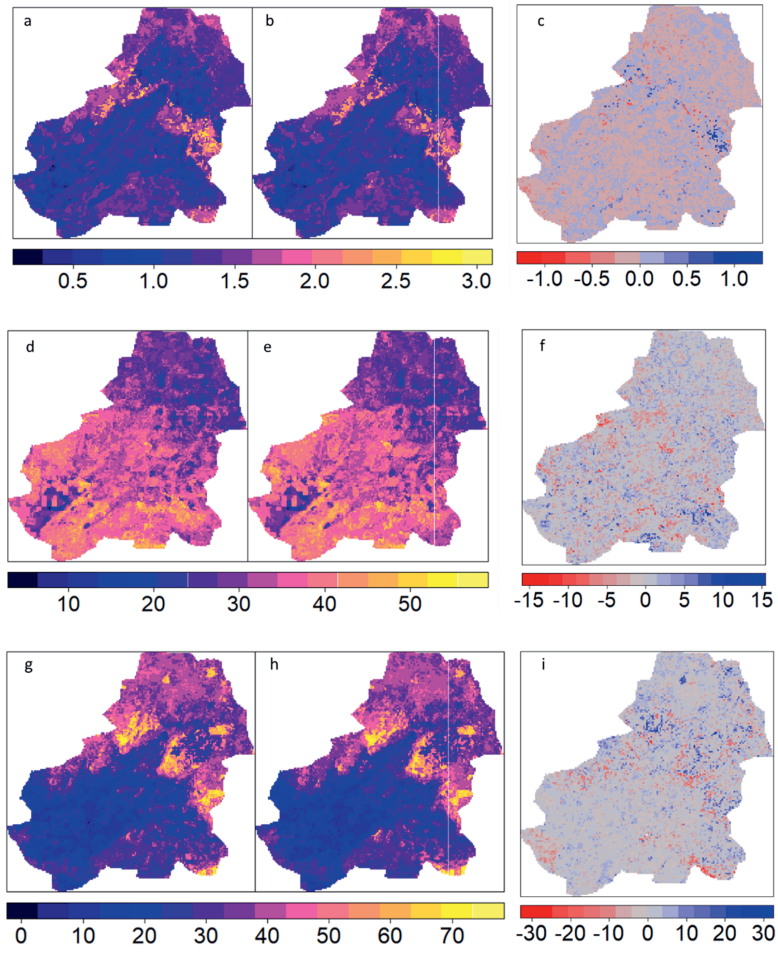


Fig. S4.2: Width of the RF 90% prediction intervals and differences: a) pH Scenario 1, b) pH Scenario 2, c) difference between pH Scenario 1 and pH Scenario 2, d) clay Scenario 1 (%), e) clay Scenario 2 (%), f) difference between clay Scenario 1 and clay Scenario 2 (%), g) SOC Scenario 1 (g kg^{-1}), h) SOC Scenario 2 (g kg^{-1}), i) difference between SOC Scenario 1 and SOC Scenario 2 (g kg^{-1}).

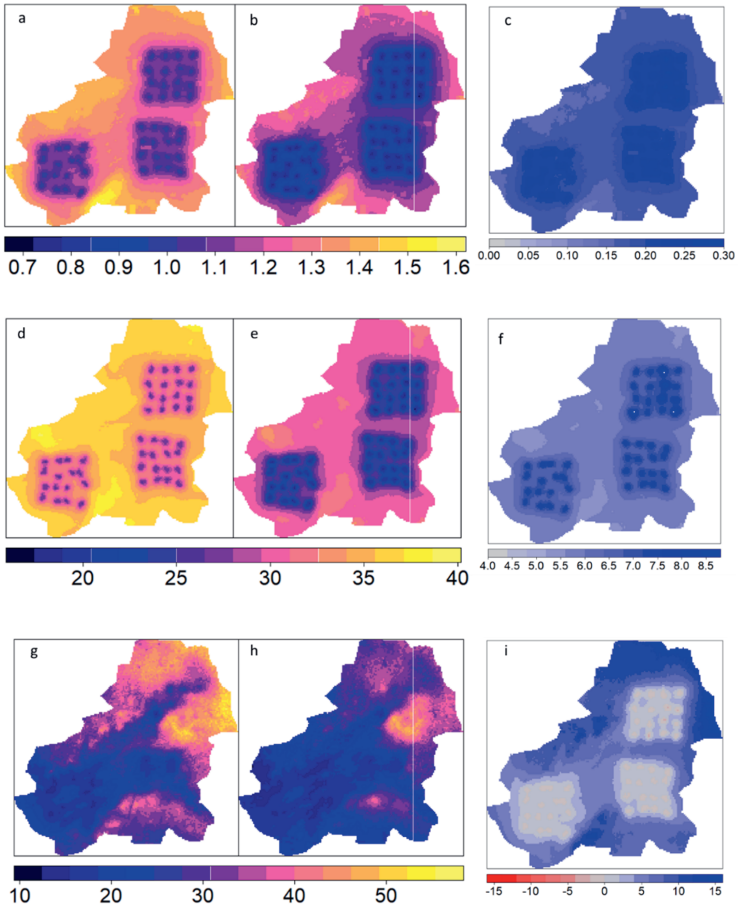


Fig. S4.3: Width of the RK 90% prediction intervals and differences: a) pH Scenario 1, b) pH Scenario 2, c) difference between pH Scenario 1 and pH Scenario 2, d) clay Scenario 1 (%), e) clay Scenario 2 (%), f) difference between clay Scenario 1 and clay Scenario 2 (%), g) SOC Scenario 1 (g kg^{-1}), h) SOC Scenario 2 (g kg^{-1}), i) difference between SOC Scenario 1 and SOC Scenario 2 (g kg^{-1}).

Chapter 5

Modelling and mapping maize yields and making fertilizer recommendations with uncertain soil information

Crop models hold the potential to enhance our understanding of crop responses to environmental conditions and agronomic management, however, the substantial uncertainty associated with model inputs can significantly influence the quality of the model outputs. This study aimed at quantifying the uncertainty in soil information and analyse how it propagates through a crop model to affect yield and fertilizer recommendation rates. Additional objectives were to analyse the uncertainty contributions of the individual soil inputs to model output uncertainty and discuss strategies to communicate uncertainty to end-users. For a case study in the highlands of Cameroon, the study used Monte Carlo simulation to analyse the propagation of soil input uncertainty through the Quantitative Evaluation of Fertility of Tropical Soils (QUEFTS) model. To effectively communicate uncertainty to end-users, a threshold probability map of yield gain was designed to serve as a visual tool.

The results showed that the impact of soil input uncertainty on model output uncertainty was significant and varied spatially. Large uncertainties in yield and fertilizer recommendation rates, with interquartile ranges larger than the median were observed in some parts of the study area. While comparing the results of a deterministic run (ignoring uncertainty in model soil inputs) with those of the Monte Carlo simulations, mean differences in predicted yield and fertilizer recommendation rates required to reach a target yield of 5 tons ha^{-1} in some parts of the study area were up to 1.0 tons ha^{-1} and up to 59, 42, and 20 kg ha^{-1} for N, P and K fertilizers, respectively. Accounting for soil input uncertainty leads to a systematic shift of the three fertilizers towards higher values. The spatial distribution of the uncertainty maps closely matched the spatial patterns of the soil input uncertainties, with high values where input uncertainty is high. Stochastic sensitivity analysis showed that pH is the main source of uncertainty for required K fertilizer (81.6%) and that soil organic carbon contributes to the uncertainty of required N fertilizer (97%), P fertilizer (25%) and K fertilizer (18%). Uncertainty in required P fertilizer mostly comes from soil extractable phosphorus (55%) and exchangeable potassium (20%). A threshold probability map designed using statistical predictions showed areas where fertilizer application will have more impact, and served as a visual tool that could enable farmers to swiftly make informed decisions about fertilizer application locations. The study highlighted meaningful relationships between the uncertainty of soil properties and the uncertainty associated with model outputs. It emphasized the necessity for future research to address correlations among uncertain soil inputs and consider additional sources of uncertainty for a more thorough and accurate quantification of model output uncertainty. The findings also underscore the importance of improving the accuracy of soil maps, which, in turn, positively influences the accuracy of QUEFTS model predictions. The findings of this chapter provide significant benefits to site-specific nutrient management proposed, by estimating the soil nutrient supply, recommending needed fertilizer rates and the corresponding theoretical yields. The findings also contribute to enhancing our understanding of the interplay between the accuracy of soil information and the reliability of model outputs, offering valuable insights for improved decision-making to achieve sustainable agricultural intensification.

A version of this chapter has been submitted as Takoutsing, B., Heuvelink, G.B.M., Aynekulu, E., Shepherd, K.D., 2024. Modelling and mapping maize yields and making fertilizer recommendations with uncertain soil information to the journal *Precision Agriculture*.

5.1. Introduction

Food security is one of the major challenges faced by populations in most sub-Saharan African countries (SSA) (Mesfin *et al.*, 2021) and agriculture remains the foremost food supplier in most of these countries. Substantial increase in agricultural production has been achieved over the past decades (FAOSTAT, 2020), but largely due to expansion into new land areas rather than increases in land productivity. The current yield gaps point to opportunities to increase food production through the efficient use of fertilizers (van Ittersum *et al.*, 2016). Obviously, the imbalance between current fertilizer application rates and the nutrient requirements for crops, leads to inefficient use of fertilizers and reduction in crop yields. A major challenge has been to recommend fertilizers that account for soil spatial variability (Rurinda *et al.*, 2020) and differences in landscape attributes (Takoutsing *et al.*, 2018).

Current fertilizer formulation in SSA has been conventionally promoted through blanket recommendations often developed based on a limited number of field-experimental data (Rurinda *et al.*, 2020). These recommendations are not able to provide the required optimal application rates because they ignore the often-high spatial variability in soil nutrient supply. This is bound to create imbalanced crop nutrition in heterogenous fields (Kihara *et al.*, 2016), leading to either over- or under-fertilization in different parts of the fertilized area. Site-specific fertilizer recommendations have been proposed to account for this variation by optimizing fertilizer use based on specific soil conditions and actual soil nutrient supply (Mesfin *et al.*, 2021). Current decision-making on fertilizer application rates is based primarily on yield responses, meanwhile farmers are eager to have detailed information on the nutrient status and fertilizer requirement specific to their fields (Breure *et al.*, 2022a). Though site-specific fertilization is much preferred over blanket fertilization, it can only achieve its objective if soil conditions are adequately known and uncertainty in soil properties does not prohibit deriving fertilizer recommendation. One way to provide fertilizer recommendations that is site-specific is the use of decision support tools such as crop models that combine soil nutrient supply and crop nutrient demand to recommend fertilizer application rates to achieve a target yield. While crop models have become common within agricultural research domains, they have traditionally been hampered by their complexity and high demand for input data that are seldomly available in SSA. In addition, some of these models are unable to predict nutrient-limited yield. The QUEFTS model can quantify the nutrient requirements of crops based on the target yield and nutrient uptake (Janssen *et al.*, 1990). The most prominent feature of QUEFTS is that it estimates soil N, P and K supply on the basis of soil data, and predicts N, P and K fertilizer rates to achieve target yields at specific locations (Smaling and Janssen, 1993). The model has been calibrated and validated for different crops in varying soils, climate, and management conditions in SSA (Ezui *et al.*, 2017) and other regions (Sattari *et al.*, 2014). Therefore, it can also be used to estimate the nutrient requirements of maize in Cameroon.

Most recent developments in crop modelling have acknowledged the need to quantify model uncertainties (Wallach and Thorburn, 2017). The sources of uncertainties are associated with errors in model structure, inputs, and parameters, and can overshade the spatiotemporal variability of simulated model outputs, thus limiting predictability (Ramirez-Villegas *et al.*, 2017; Chapagain *et al.*, 2022). Input uncertainty arises from uncertainty in climate (e.g. temperature), soil (e.g. soil properties), initial conditions and crop management practices, which are typical inputs required for most crop models (Ojeda *et al.*, 2021; Chapagain *et al.*, 2022). In particular, soil information embodies a substantial degree of error, because laboratory analyses are imperfect (van Leeuwen *et al.*, 2022) and most soil information is obtained from maps that suffer from prediction and interpolation errors. The effects of errors introduced due to soil sampling and chemical analysis procedures on fertilizer recommendations has been evaluated, with conclusions that large uncertainty exists in estimates of soil nutrient supply based on soil property measurements (Schut and Giller, 2020).

Uncertainty in soil information can be quantified by probability distributions (Heuvelink, 2014) and several methods for uncertainty propagation analysis have been developed and applied at various

scales. Monte Carlo analysis has often been used to compute output probability distributions by repeated model simulations with input variables randomly sampled from their probability distribution. Importantly, with the increased generation of spatially explicit gridded crop model simulations, failure to account for model output uncertainty may lead to poor decision-making by policy makers and stakeholders. Uncertainty in model outputs should therefore be communicated effectively so as to enable the end-users to draw valid conclusions and make sound decisions (Breure *et al.*, 2022b; Lark *et al.*, 2022). For a case study on maize (*Zea mays* L.) in the Western Highlands of Cameroon, this study aimed to: 1) quantify the uncertainty (probability distributions) in soil information used by the QUEFTS model; 2) analyse how uncertainty in soil information propagates through QUEFTS and affects yields and fertilizer recommendation rates; 3) compare the results of the uncertainty propagation analysis with a case where uncertainty in soil inputs is ignored; 4) analyse the uncertainty contributions of the individual soil inputs to model output uncertainty; 5) discuss strategies to communicate uncertainty in QUEFTS outputs to end-users and advise them on how uncertainty can be incorporated in their decision-making process.

5.2. Case study

5.2.1. Description of the study area

The study area is the west region of Cameroon that forms part of the Western Highlands and spans over 13,892 km² (Fig. 5.1). The climate is tropical humid with two seasons: a long, wet season of eight months from March to October, and a short, dry season of four months from November to February. The average annual temperature ranges between 20 °C and 28 °C, while annual average rainfall ranges from 1,200 mm to 2,300 mm. The area is characterized by accidented relief of massifs and mountains that consist of plains, undulating hills, and gentle sloping areas. Altitudes reach as high as 2,400 and as low as below 450 masl in valleys. The dominant soil types are Ferralsols and Nitisols of the World Reference Base system (IUSS, 2015). The area is largely an agrarian area subjugated by subsistence agricultural systems where high-skill farmers exploit virtually every strip of land available to grow a range of annual and perennial crops. Maize is the major staple crop, and farmers grow it either in association or rotation with other crops.

5.2.2. Field sampling using the Land Degradation Surveillance Framework

The Land Degradation Surveillance Framework (LDSF) was used to collect soil samples across the study area. The LDSF is a systematic methodology to conduct landscape-level assessments of soil and land health based on a consistent set of indicators and field protocols and uses the concept of sentinel sites (Vågen and Winowiecki, 2020). A sentinel site is a 100 km2 area stratified into 16 clusters of 1 km2 size, each containing 10 plots of size 1000 m2, while each plot is further subdivided into 4 subplots of size 100 m2. Four sentinel sites were randomly placed within the study area. The distribution of the 640 soil sampling plots is shown in Fig. 5.1. Topsoil samples (0 – 20 cm) were collected at the four subplots of each plot and thoroughly mixed to form a composite sample for each plot, resulting in 160 soil samples per site and 640 soil samples in total.

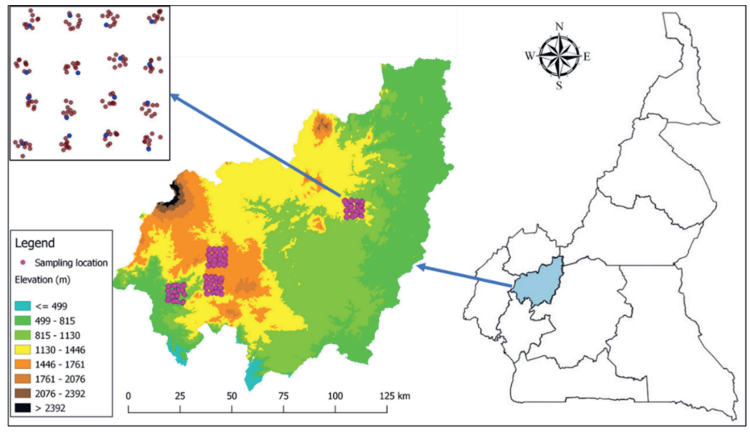


Fig. 5.1. Digital elevation map of the study area in Cameroon showing the soil sampling locations. Upper-left panel zooms in on one sentinel site. Red dots represent locations with spectral data, blue dots locations with analytical and spectral data.

5.2.3. Soil data

The soil samples were air dried, crushed by rolling pins and sieved through a 2-mm sieve. About 10 g of each soil sample was milled to pass a 75-µm sieve using a Retsch RM 200 mill (Retsch, Düsseldorf, Germany) and analysed using a high-throughput Bruker Tensor 27 Fourier Transform MIR spectrometer. MIR diffuse reflectance spectra were recorded at a waveband range of 601 to 4000 cm^{-1} with a resolution of 4 cm^{-1}. Ten per cent of the soil samples (n = 64) were randomly selected as references and also analysed using conventional wet chemistry methods for the determination of SOC (dry combustion), pH (1:1 solution in water), macro- and micro-elements including base cations (Melich-3 extraction), and texture (Laser diffraction particle size analysis). Calibration models were developed using the paired observations of MIR spectra and analytical laboratory measurements to predict the soil properties at locations where only spectra were available. SOC, pH, P and K were used as inputs for crop modelling, while Al, Fe, and Ca were used for pedotransfer functions as explained below. Prior to applying the MIRS, Olsen Phosphorus (POlsen) and exchangeable potassium (KExch) in mmol kg^{-1} were estimated from available Mehlich 3 data using transfer functions proposed by Breure *et al.* (2022a) and shown in Equations 5.1 and 5.2:

$$ln\_POlsen = 0.77 * ln(P_{M3}) + 0.62 * ln(Al_{M3}) + 0.13 * ln(Fe_{M3}) + 0.10 * ln(Ca_{M3}) - 0.19 * pH - 4.31 \quad (5.1)$$

$$KExch = 0.028 * K_{M3} + 0.015 \quad (5.2)$$

Note that the nutrients in Mehlich 3 (P, Al, Ca, and Fe) in mg kg^{-1} were log-transformed for application of the P transfer function (Equation 5.1). The predicted POlsen values were obtained by back-transformation following Lark and Lapworth (2012):

$$POlsen = exp(ln\_POlsen + 0.5 * \sigma^2_{pred}) \quad (5.3)$$

where ln_POlsen is the predicted value with the transfer function (Equation 5.1) and σ^2_{pred} denotes the associated prediction error variance, which can be obtained from the residual variance of the multiple linear regression model and variance covariance matrix of the estimation errors of the model coefficients.

5.3. Methodology

5.3.1. Spatial modelling of soil properties using random forest

The spatial modelling of soil properties has been described in detail in **Chapter 4**, and we only repeat the essentials here. Since soil properties are correlated with environmental variables (Minasny and McBratney, 2016), a set of 191 spatially distributed environmental variables was collected from different sources to represent the major soil forming processes and surface characteristics of the study area. Firstly, we carried out a correlation analysis to address multicollinearity of the 191 environmental layers. Only covariate layers with a pairwise correlation coefficient <= 0.75 with all the other covariates were retained for further analyses. In case two covariates were correlated above this threshold; we only retained the first one in alphabetical order for use in the modelling process. This first step reduced the number of covariates to 99 layers. Next, we performed a selection procedure using the Recursive Feature Elimination algorithm as implemented in the caret package (Kuhn, 2008) to remove the least important covariates. The Recursive Feature Elimination procedure is an iterative process that starts by fitting a model using all covariates, assesses its performance and ranks the covariates according to their importance (Gomes *et al.*, 2019). The least important covariate is removed, and the process is repeated, until only one covariate is left. From all evaluated models the one with the most favourable cross-validation statistic (RMSE or R^2) is selected. For each soil property, an optimal set of covariates was selected, which can differ between soil properties.

The fitted models were applied to the stack of retained environmental variables to predict the four soil properties required for further modelling, i.e. SOC, pH, POlsen and KExch across the study area at 250 m resolution. The performance of each predictive model was evaluated by calculating the Mean Error (*ME*), the Root Mean Squared Error (*RMSE*) and the model efficiency coefficient (MEC) (Janssen and Heuberger, 1995) using leave-cluster-out cross-validation (LCOCV). The LCOCV has previously been demonstrated to be a suitable method to evaluate the performance of prediction models in the case of clustered data (**Chapter 4**). We used quantile regression forest (QRF) (Meinshausen (2006) to obtain model predictions and the associated prediction uncertainty. QRF derives quantiles of the conditional probability distribution at each prediction location. For each soil property, we computed the mean, 0.05 quantile and 0.95 quantile, i.e. the lower and upper limits of a 90 % prediction interval. The 90% prediction interval coverage probability (PICP) was used to validate the quantified prediction uncertainty (Shrestha and Solomatine, 2006). Ideally, the PICP should be closed to 0.90. A PICP value substantially greater than 0.90 suggests that the uncertainty is underestimated, while a value substantially smaller than 0.90 indicates that it was overestimated.

5.3.2. The QUEFTS model

The QUEFTS model was originally developed by Janssen *et al.* (1990) to estimate yield responses based on nutrients present in the soil and those added through application of NPK fertilizers. Further improvements were made by Smaling and Janssen (1993) to estimate the nutrient requirements for crop yield. The model estimates expected yields given limitations in nitrogen, phosphorus and potassium uptake and takes into account the relationship between N, P, and K rather than the demand for individual

nutrient elements alone. The model can also be used to generate fertilizer recommendations given soil properties for target yields in such a way that N, P and K are not limiting for yield (Rurinda *et al.*, 2020). The QUEFTS model entails a four-step process (Fig. S5.1) that simulates the potential supply of nutrients, plant nutrient uptake, yield ranges and the final yield based on nutrient accumulation and dilution. The four steps are described in detail in Ravensbergen *et al.* (2021), and we only repeat the essentials here:

Step 1: First the potential soil supplies of N, P and K are calculated, by applying relationships between four chemical soil properties, namely pH, SOC, POlsen and Kexch, and the maximum quantity of these nutrients that can be taken up by maize if no other nutrients and no other growth factors are yield-limiting. In addition to the nutrient supply from the soil, nutrient supply from fertilizer application is obtained by accounting for the fertilizer recovery of applied fertilizers.

Step 2: In the second step the actual uptake of each of the three nutrients N, P and K is calculated as a function of the potential supply of that nutrient, taking into account the potential supplies of the other two nutrients. In QUEFTS the actual uptake of a nutrient is calculated twice for each nutrient, where each time only one of the other two nutrients is considered. For instance, the actual uptake of nitrogen is calculated once as a function of its own supply and the supply of phosphorus and once as a function of its own supply and the supply of potassium.

Step 3: QUEFTS converts the estimated uptake of N, P and K into maize yield. For each nutrient, the upper and lower bounds yields are calculated based on the actual uptake of each nutrient. The upper bound yield refers to the yield attainable when for instance N is maximally diluted in the plant. The lower bound yield refers to the yield that could be obtained when N is maximally accumulated in the plant. The actual yield lies somewhere in-between these yields. This leads to six combinations describing the uptake of one nutrient given maximum dilution or accumulation of another nutrient.

Step 4: In the fourth step, the yield estimates are calculated in pairs on the basis of the actual uptake of each nutrient and the yield ranges calculated in step 3. This will result in six paired estimations which are averaged to obtain a final yield estimate.

In case the calculated yield is below a required level, either because no fertilizer was applied or because too little was applied, then QUEFTS can calculate the required NPK application rates to achieve the required yield level. To compute the required application rates the relation between fertilizer application and yield as outlined in the four steps above is mathematically inversed. The QUEFTS model was run to compute the amount of NPK fertilizer for target yield of 5.0 tons ha^{-1} under two scenarios: the NPK application rates with all soil inputs assumed certain (Scenario 1), and the NPK application rate and associated yield with uncertain soil information (Scenario 2). The R version of QUEFTS based on Sattari *et al.* (2014) was used and run with the various input parameters: 1) soil properties as described in Section 3.1 (i.e., pH, SOC, KExch and POlsen); 2) QUEFTS model default values for nutrient recovery fractions (0.5 for N and K, 0.1 for P fertilizer); 3) QUEFTS model default values for maximum physiological efficiency for maize (IE borderline) (i.e., 70, 600, 120 kg biomass per kg N, P and K, respectively, and for minimum physiological efficiency (IE borderline) (i.e., 30, 200, and 30 kg biomass per kg N, P and K, respectively; 4) maximum crop season potential yield for the Western Highlands of Cameroon of 5.0 tons ha^{-1} ; and 5) average temperature for the study area of 25 ^{o}C.

5.3.3. Uncertainty propagation using the Monte Carlo method

Model output uncertainty is generally determined from three main sources, namely input uncertainty, model structure uncertainty, and model parameter uncertainty. In this study, we solely focussed on model input uncertainty, and only considered uncertainty in soil inputs. QUEFTS requires four soil properties (i.e., pH, SOC, POlsen and KExch) as inputs to derive fertilizer recommendation rates to

achieve a target yield of 5.0 tons ha^{-1}. Soil input uncertainties propagate through the model, resulting in uncertainties in yield and fertilizer recommendation rates.

To assess how the uncertainty in soil input variables affects QUEFTS outputs, we performed a Monte Carlo (MC) simulation approach. Attractive characteristics of the MC method are easy implementation, general applicability and that it yields the entire probability distribution of the model output (Heuvelink, 1998). The aim of the uncertainty analysis is to quantify the uncertainty of the yield and fertilizer recommendation rates as the result of uncertainties in the soil input variables. It can also be used to analyse the uncertainty contribution of each individual uncertainty source. The MC method repeatedly samples realizations from the probability distributions of the uncertain input variables and runs the model for all realizations. The results of the model simulations are then analysed to estimate the probability distribution of the outputs and quantify the uncertainty. The method as applied here thus consisted of the following steps:

Define the mathematical model (QUEFTS) that is to be simulated, including the inputs, parameters, and the targeted output variables.

1. Quantify the uncertainty of the soil inputs by probability distributions (Section 5.3.1) and draw a large number of realisations (i.e., 500, see below) from them using a pseudo-random number generator.
2. Run the QUEFTS model repeatedly (we used 500 Monte Carlo runs), each time using one of the simulated sets of inputs and store the model outputs. Running the MC analysis several times while changing the seed confirm that 500 runs were sufficient to reach stable results.
3. Construct an empirical probability distribution of the 500 output values. From this distribution summary statistics can be derived, such as the mean, standard deviation, and percentiles.

The frequency distributions of the Monte Carlo simulations of fertilizer recommendation rates and yield represent the propagation of input uncertainty to the model output uncertainty. In particular, the width of these distributions (in this study characterized by the difference between the 0.95 and 0.05 quantiles) signifies the uncertainty of the QUEFTS predicted yield and fertilizer recommendation rates.

5.3.4. Contributions of soil input variables to the model output total uncertainties

The contribution of individual soil input uncertainty to the overall uncertainty of QUEFTS outputs was analysed using a stochastic sensitivity analysis (Saltelli *et al.*, 2008) The uncertainty contribution was expressed as the percentage of the output variance accounted for by each uncertain input. If m is the number of uncertain inputs, then we need $m+1$ MC analyses to compute the uncertainty contributions. Initially, the total output uncertainty MC_{tot} is computed by stochastically varying all input variables considered (as explained in Section 3.3). The uncertainty associated with the first input variable x_1 is next obtained through another MC simulation MC_1 in which x_1 is set equal to its deterministic value, while the other input variables vary stochastically. Similarly, the other MC analyses MC_2, MC_3, . . ., MC_m are used to quantify the uncertainty contribution for the other uncertain inputs. The contribution of individual input variables to the total uncertainty of the model output, that is the stochastic sensitivity S_i for each uncertain input x_i is then computed as shown in Torres-Matallana *et al.* (2021):

$$S_i = 1 - \frac{var(MC_i)}{var(MC_{tot})} \tag{5.4}$$

The larger the index S_i, the high the contribution of the input uncertainty x_i.

5.3.5. Communicating uncertainty and its integration in decision-making

Scientists may be familiar with the concept of uncertainty and methods to quantify it (Brown and Heuvelink, 2005), but end-users are often less familiar and used to deterministic solutions. They often assume or require error-free model outputs to facilitate decision-making. One of the consequences of this desire for simplicity is that modellers do not pay much attention to the uncertainty of their outputs, and therefore do not communicate it to end-users (Verstegen *et al.*, 2012). In practice, achieving error-free model outputs is impossible, and end-users should be assisted to understand uncertainty and how to take decision based on uncertain data. Our interest in this study was to communicate the uncertainty of QUEFTS outputs to a range of end-users including policy makers, extension service agents, and staff of non-governmental organizations. These individuals may in turn be required to communicate uncertainty to the farmers or to the general public and integrate it in decision-making.

Quantification of uncertainty can be straightforward, but communicating uncertainty to a range of users of information and its integration in decision-making processes is less so (Verstegen *et al.*, 2012; Chagumaira *et al.*, 2021). The challenges may include the difficulty to grasp uncertainty for users without basic knowledge of statistics, the fact that uncertainty is input-dependent, the variation of uncertainty over space, and the lack of software that can integrate spatial modelling, uncertainty analysis and visualization (Verstegen *et al.*, 2012). In addition, the success of a method to communicate uncertainty may depend on the subject matter and on the background of the users of information (Milne *et al.*, 2015). Uncertainty can be communicated using an empirical probability distribution, a probability interval, or can be simply described using words, for example, on a verbal scale (Milne *et al.*, 2015; Spiegelhalter, 2017). It is important that the uncertainty statistics are communicated in an efficient way that is both informative and understandable to users with varied backgrounds. There should be ample explanation of the results and a visual representation of the uncertainty to facilitate communication. In this study, we used the 0.05 and 0.95 quantiles of the probability distribution of the spatially distributed yield and fertilizer recommendation rates as a 90 % prediction interval, which expresses the uncertainty about the true values. Furthermore, we opted to visualize uncertainty in the form of maps which show the mean as well as the upper and lower bounds of the prediction interval separately.

Agricultural business decisions, e.g., decisions on whether or not to apply fertilizer in a specific field, depend heavily on model-based recommendations, which are associated with uncertainty. Decisions based on misinterpreted or erroneous model outputs can be costly due to the irreversibility of such decisions (Verstegen *et al.*, 2012). Without understanding this risk, it is not possible for end-users to draw proper conclusions and making optimal decisions. Farmers are aware that fertilizer application improves yield, but fertilization comes at a cost, and they will want to be sufficiently certain that there will be a positive return on investment. Farmers could establish 'rules' which account for uncertainty. For example, they could state that they are only willing to apply fertilizer if the probability of obtaining a yield gain above a threshold (e.g., 2.0 tons ha^{-1}) is greater than 90 per cent. Using this example rule, we evaluated which part of the study area would be fertilized. We first derived the optimal fertilizer application for Scenario 1 (no uncertainty in soil inputs) and used these recommendation rates to derive the probability distribution of the yield gain at each location, by subtracting the simulated yield without fertilizer application from that obtained with fertilizer application. Note that this leads to a probability distribution of the yield gain because both the yield with and without fertilizer application are uncertain due to uncertainty about the soil properties. Next, we delineated the area where the expected yield gain (i.e., the mean of the probability distribution) is bigger than the threshold as well as the area where we are at least 90% certain that the yield gain is bigger than the threshold.

5.3.6. Data processing and statistical computing

All analysis and modelling in this study were conducted in the R statistical open-source software (R Core Team, 2021). The R packages used in this study included "ranger" (Wright and Ziegler, 2017) for fitting the random forest and quantile regression forest models; "raster" (Hijmans, 2021) for handling raster layers; and spplot (Pebesma and Bivand, 2005) and ggplot (Wickham, 2016) for plotting. All soil maps were produced in QGIS version 3.22 using the Albers equal area projection (EPSG:22832).

5.4. Results and discussion

5.4.1. Descriptive statistics of soil property data

The dataset used in this study combined both analytical and spectral soil data. For soil samples with both analytical and spectral data, only analytical data were retained. This resulted in 64 analytical and 546 spectral observations, making a total of 640 observations. Summary statistics of the soil properties are provided in Table S5.1. The SOC content ranged from 8.7 to 45.6 g kg^{-1} and had a mean of 25.5 g kg^{-1}. This range indicates high potential of crop responses and recovery of applied fertilizers and therefore, a high productivity of crops such as maize could be expected in the area. Soils are generally acidic with pH ranging from 4.2 to 6.3 and a mean of 5.2 and show moderate to severe limitations to crop production. For a large proportion of the soil samples (> 60%), POlsen values were below the critical value of 10 mg kg^{-1} for optimum maize yield (Ussiri *et al.*, 1998; Bai *et al.*, 2013). The deficiency could be due to inherent low P concentration in the parent material and P-fixation (Eijk *et al.*, 2006). The values of KExch were far below the threshold value of 2 mmol kg^{-1} required for the growth of major crops (Chilimba *et al.*, 1999). These findings indicate that P and K fertilizers are needed to enhance maize production in the study area.

5.4.2. Spatial prediction of soil properties and uncertainty quantification

The correlation analysis and variable selection procedure indicated that not all 191 covariates were useful in explaining the spatial variation of the soil properties of interest. The optimal number of covariates included in the RF predictive model for pH, SOC, POlsen and KExch were 9, 3, 22 and 57, respectively. Plots of recursive feature elimination showing model performance as a function of number of covariates are shown in Fig. S5.1. The LCOCV statistics used to evaluate the performance of each predictive model are presented in Table 1. The ME values were close to zero for all properties. The degree to which the spatial variation of soil properties was predicted from the available covariates varied substantially. Spatial variation in pH, POlsen and SOC was best described, with MEC values above 0.60, while variation in KExch was less described, with a MEC value of 0.41. The low MEC value for KExch is likely caused by the weak relation between the soil property and the covariates, which leads to a higher prediction uncertainty. Table 5.1 also summarizes the PICPs for the models. For all four soil properties, the PICP ranged between 0.89 and 0.92. This indicates that the prediction intervals obtained with QRF were a realistic representation of the various model prediction uncertainties, as these were close to 0.90.

Table 5.1: Leave-cluster-out cross-validation metrics for the random forest predictions of the four soil properties.

Soil property	ME	RMSE	MEC	PICP
pH	-0.001	0.157	0.832	0.91
SOC (g kg^{-1})	-0.020	5.620	0.618	0.89
POlsen (mg kg^{-1})	0.010	3.450	0.640	0.89
KExch (mmolc kg^{-1})	-0.001	0.130	0.410	0.92

For each soil property, we spatially predicted the mean value using RF, as well as the lower and the upper limits of the 90% PI with QRF. The results for SOC are presented in Fig. 5.2, while those of the three other soil properties are provided in Fig. S5.3. Predicted pH values ranged from 4.68 to 6.04, SOC values from 12.30 to 43.85 g kg^{-1}, POlsen from 1.36 to 11.81 mg kg^{-1} and KExch from 0.10 to 0.74 mmol kg^{-1}. It can be seen from the maps that lower values of pH and SOC are found in the south-western parts, while lower values of POlsen and KExch are found in the central and the northern parts of the study area. These two areas are dominated by Ferralsols which have a low nutrient retention capacity. High values of pH, SOC and KExch are found in a region in between these two areas. This region is a mountainous volcanic area dominated by fertile Nitisols which are well-drained with very favourable chemical and physical properties.

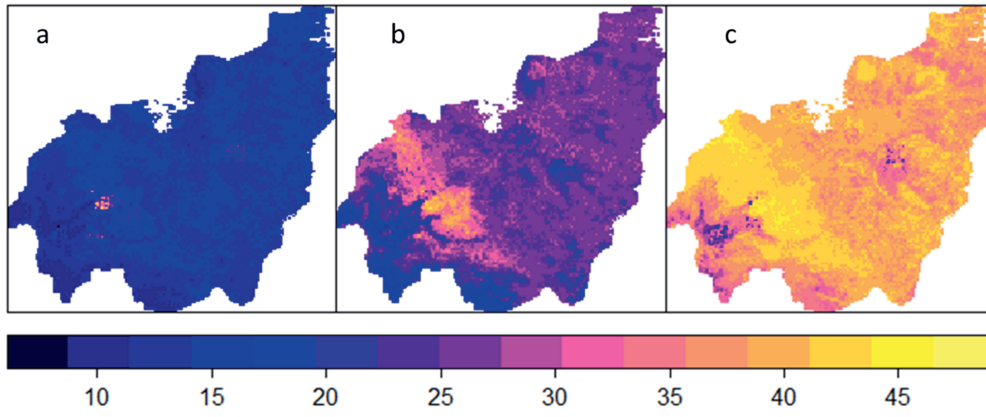


Fig. 5.2. Spatial distribution of SOC predictions and associated limits of the 90% prediction interval (g kg^{-1}): a) lower limit, b) mean values, c) upper limit.

5.4.3. QUEFTS model outputs ignoring uncertainty in soil inputs

Maps of the potential soil supply of the three macro-nutrients nitrogen, phosphorus and potassium estimated by QUEFTS under Scenario 1 are presented in Fig. S5.4. It can be seen from the maps that the south-western and central parts of the study area have low soil N and P supplies, whereas in between these areas there is a region with high N and P supply. The soil K supply map has a very different spatial pattern and shows high supply of K in the southern part of the study area. With no fertilizer applied, the predicted yield ranged from 1.44 to 4.94 with a mean of 3.26 tons ha^{-1} (Table 5.2).

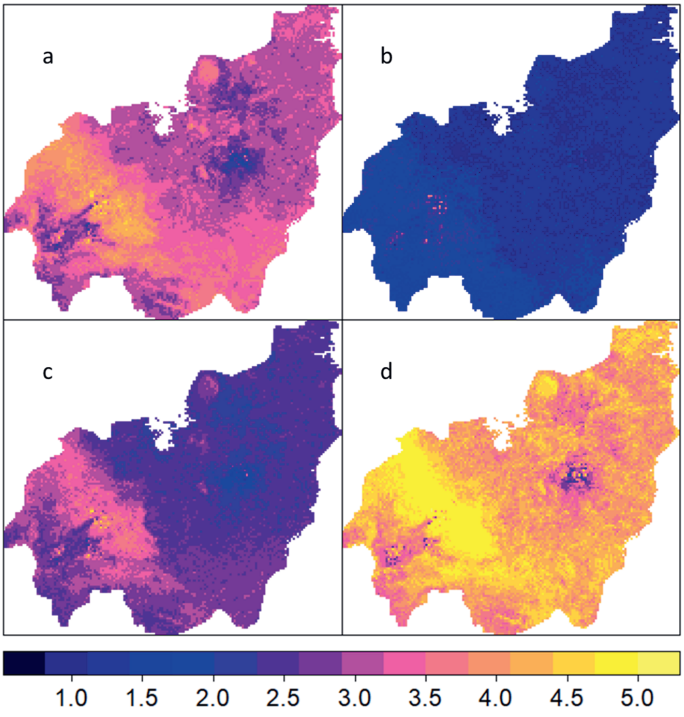


Fig. 5.3: Spatial distribution of yield as predicted by QUEFTS without fertilizer application: a) yield ignoring uncertainty in soil inputs, b) 0.05 quantile of 500 MC yield maps; c) mean of 500 MC yields; d) 0.95 quantile of 500 MC yields.

The spatial distribution of the yield as predicted by the QUEFTS model without fertilizer application is presented in Fig. 5.3a. Predicted yield values in some areas varied from low to high within a very short distance, e.g., in the central and south-west parts of the study area. The spatial distribution of the predicted yield agreed with those of the N and P soil supply maps and much less with the K soil supply map (Fig. 5.4). With the estimated soil nutrient supplies, the QUEFTS model was run to estimate the optimum NPK fertilizer recommendation rates required to achieve the target yield of 5.0 tons ha^{-1}, assuming the predicted soil properties to be certain (Scenario 1). Summary statistics of the predicted NPK recommendation rates are presented in Table 5.2. Fertilizer application allowed to reach the target yield across the entire study area. There were considerable yield gains across the area when NPK fertilizer were applied ranging from 0.06 to 3.56 with a mean of 1.74 tons ha^{-1}.

The spatial distribution of recommended NPK fertilizers to achieve a target yield of 5.0 tons ha^{-1} is presented in Fig. 5.4 and shows high variability across the study area. Summary statistics of the predicted yield gain and NPK recommendation rates are presented in Table 5.2. As expected, the maps are a mirror image of the soil nutrient supply maps shown in Fig. S5.4. Some parts of the study area have sufficient N and P supply from the soil and do not require N and P application to reach the target yield, whereas other parts have low soil supply and need a high fertilizer rate to reach the target yield. The spatial variation of the optimized NPK application is very large indicating that blanket fertilizer recommendation would not be a suitable policy to achieve the target yield of 5.0 tons ha^{-1} across the

study area. The maximum N application rate is higher than that of P and K (Fig. 5.4, Table 5.2), while the mean recommended application rate is highest for K. This is probably the result of spatial variation in soil nutrient supplies, which are influenced by other environmental factors, not accounted for by QUEFTS.

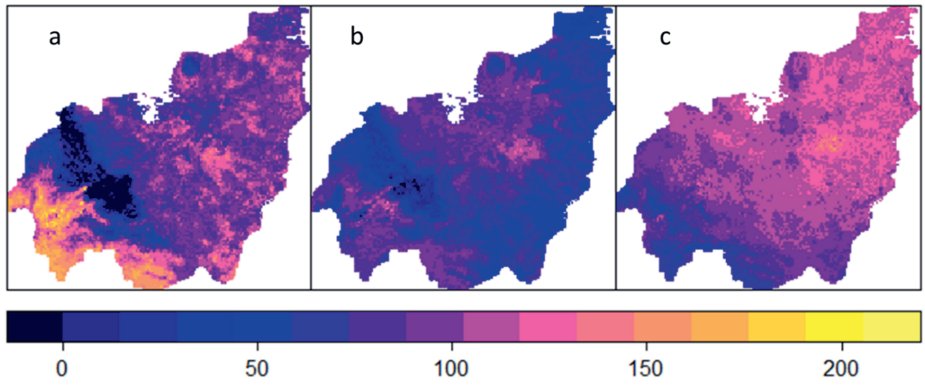


Fig. 5.4. Fertilizer recommendation rates required to obtain the targeted yield of 5.0 tons ha^{-1} as predicted by the QUEFTS model while ignoring uncertainty in soil inputs (in kg ha^{-1}): a) N fertilizer; b) P fertilizer; c) K fertilizer.

Table 5.2: Summary statistics of the QUEFTS model predicted yield without fertilizer application and NPK recommendation rates required to achieve a target yield of 5.0 tons ha^{-1} under Scenario 1 (ignoring soil input uncertainty).

	Tons ha^{-1}	kg ha^{-1}			
Statistic	Yield without NPK application (tons ha^{-1})	N Fertilizer (kg ha^{-1})	P fertilizer (kg ha^{-1})	K fertilizer (kg ha^{-1})	Yield gain (tons ha^{-1})
Maximum	4.94	205.8	143.2	157.0	3.56
75th percentile	3.47	107.2	84.4	115.2	2.00
Mean	3.26	89.8	69.1	102.6	1.74
25th percentile	3.02	77.4	56.9	93.8	1.52
Minimum	1.44	0.0	0.0	32.9	0.06

5.4.4. QUEFTS model outputs with uncertain soil inputs

We applied the same procedure as in Section 5.4.3 and used QUEFTS to calculate the yield without fertilizer application and predict the fertilizer recommendation rates required to achieve the target yield of 5.0 tons ha^{-1}, but now assuming the four soil inputs to be uncertain (Scenario 2). Each of the 500 MC runs produced a realisation of the yield and fertilizer recommendations, so that the frequency distribution of the 500 outputs approximates the probability distribution of the uncertain yield and fertilizer recommendation rates. We verified that 500 MC runs was sufficient to obtain stable results, by redoing the analysis using a different random seed. Differences were indeed small (results not shown). In a case without fertilizer application, we compared the yield map of the deterministic run with the

mean of the results of the 500 MC simulations and found substantial differences (Fig. 5.3). For some locations, differences in predicted yield between the deterministic run and the mean of the 500 MC runs were close to 1.0 tons ha^{-1}. We observed a spatial average of 3.26 tons ha^{-1} for the deterministic run (Fig. 5.3a) and 2.64 tons ha^{-1} for the mean of the MC runs (Fig. 5.3b). This shows that the yield is systematically overestimated when uncertainty in soil inputs is ignored. These systematic differences arise from the non-linear nature of the QUEFTS model, where yield increments are more pronounced for soil supply increments in the low range compared to the high range of soil nutrient supply (Dhakal and Lange, 2021).

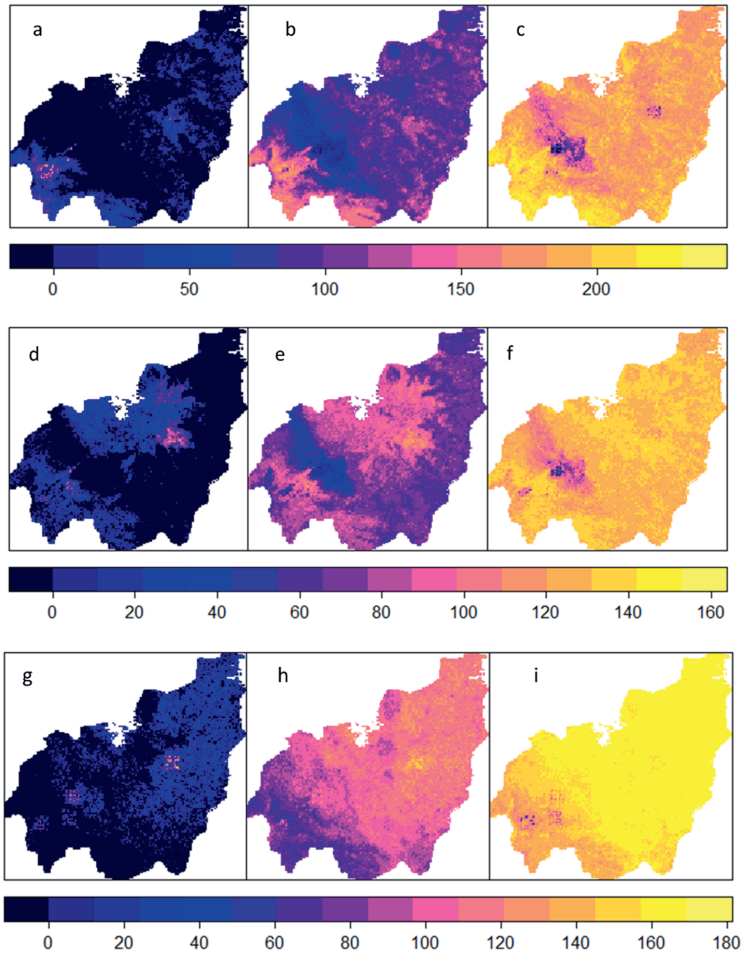


Fig. 5.5. Spatial distribution of fertilizer recommendation rates calculated using QUEFTS (kg ha^{-1}) under Scenario 2: lower limit, mean values, and upper limit for N (a, b, c); P (d, e, f); and K (g, h, i).

Maps of the mean and 0.05 and 0.95 quantiles obtained by propagating uncertainty in soil input variables through the QUEFTS model are presented in Fig. 5.5, while those of the 90% prediction interval widths are presented in Fig. S5.5. Uncertainty in soil inputs resulted in a large uncertainty of the NPK fertilizer

recommendation rates required to reach the target yield. As shown in Fig. 5.5, there is a large difference between the lower and upper limits of the prediction intervals. As a result, the 90% prediction interval widths are very large. High uncertainty of soil properties leads to high uncertainty of soil nutrient supplies, and this in turn means that we are highly uncertain about how much fertilizer should be applied to reach the target yield 5.0 tons ha^{-1}. The uncertainty maps of Fig. S5.5 show that the spatial distribution of uncertainty is not homogeneous and varies substantially across the study area. Relatively low values are found close to the sentinel sites, where soil input uncertainty is lower than elsewhere in the study area. The 90% prediction interval was generally wider for N and K fertilizers and narrower for P fertilizer, particularly in the western and northern parts of the study area. For instance, lower uncertainty for N was observed in areas where N fertilizer recommendation is low (see Fig. 5.4a). This corresponds to areas with fairly high soil supply of N. In these areas, we are less uncertain about the recommended N fertilization, because in many MC runs the soil supply is sufficient to reach the target yield, so that the N application rate is zero.

The spatial variation of the QUEFTS model outputs can also be displayed as spatial cumulative frequency distributions. In Scenario 2, this yields a cumulative distribution for each single MC run. We represented the 500 MC curves by curves of the mean and the 0.05 and 0.95 quantiles (Fig. 5.6). Incorporation of uncertainty leads to a systematic shift of the cumulative distributions of the three fertilizer applications to the right (higher values) while that of the yield shifted to the left (lower values) when uncertainty in soil inputs is accounted for. This indicate an overestimation of the yield and slight underestimation of the fertilizer recommendation when uncertainty in soil inputs is ignored. This can be termed the flaw of averages, which commonly occurs when input distributions are skewed, or thresholds exist (Savage, 2009). Though the spatial variation is not large, as shown by the steepness of the curves, large uncertainty was observed, as indicated by the wide prediction intervals for the four model outputs, in particular for N fertilizer recommendation. Thus, Fig. 5.6 nicely combines and allows to compare spatial variability and uncertainty in one figure (Heuvelink and Pebesma, 1999).

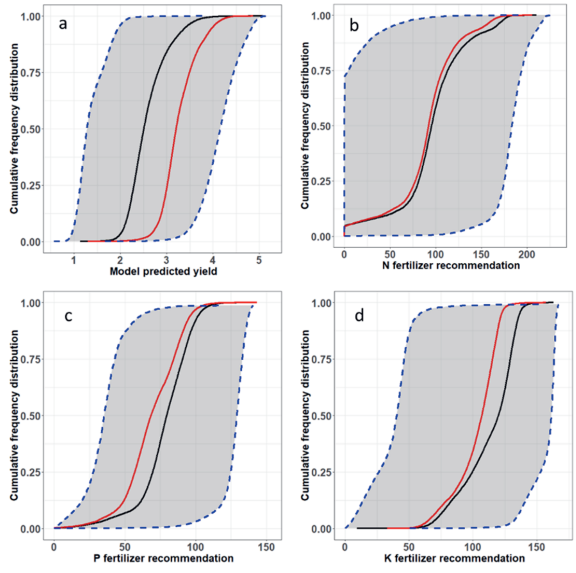


Fig. 5.6. Spatial cumulative frequency distribution of the yield (tons ha^{-1}) and the N, P and K fertilizer recommendations (kg ha^{-1} for cases with and without soil uncertainty: red lines represent deterministic run, black lines the mean of all MC runs, dashed blue lines represent the 0.05 and 0.95 quantiles. The grey area indicates uncertainty about the position of the 'true' cumulative distribution, while the steepness of the curves shows the degree of spatial variation.

5.4.5. Contribution of soil input variables to the total uncertainty of fertilizer recommendation rates

We decomposed the total uncertainty of each of the three fertilizer recommendation rates into the contribution of the four soil input variables using the stochastic sensitivity analysis as described in Section 3.4. An additional four MC simulations with 500 runs each were performed to estimate the stochastic sensitivity Si of the input variables pH, SOC, POlsen and KExch (Table 5.3). The relative contributions of input variables varied greatly among the four soil input variables.

Soil pH is the main source of uncertainty for K fertilizer (82%) while uncertainty in SOC is by far the most dominant for N fertilizer (97%), and a substantial source of uncertainty for P fertilizer (25%) and K fertilizer (18%). As expected, POlsen is the main source of uncertainty for P fertilizer (55%). The second-most important variable that contributes to the uncertainty of P fertilizer is KExch, with a stochastic sensitivity coefficient of about 21%. KExch is the least important uncertainty contributor for N and P fertilizers. From these results, we can infer that pH, SOC and POlsen were the dominant sources of uncertainty for the N, P, and K fertilizer recommendation rates. K fertilizer uncertainty is more influenced by pH uncertainty as compared to N and P. At low pH, the soil's ability to keep supplying potassium to plants is decreased, therefore potentially increasing the need for additional P fertilizer application. SOC uncertainty contributes strongly to N fertilizer uncertainty due to its direct relationship with organic matter and the strongly correlation between the two properties (**Chapter 2**).

Table 5.3: Relative contributions of soil input variables to the uncertainty of QUEFTS model outputs in terms of percentage of total variance.

	Stochastic sensitivity (S_i) of input variable (%)		
Soil inputs	N fertilizer	P fertilizer	K fertilizer
pH	0.1	0.0	81.6
SOC	96.7	24.8	18.2
POlsen	0.1	54.8	0.0
KExch	3.2	20.5	0.2

5.4.6. Communicating model output uncertainty to end-users and integration in decision-making processes

The uncertainty propagation analysis showed that the fertilizer recommendation rates derived by the QUEFTS model were highly uncertain due to uncertainty about soil inputs. The uncertainty was visualised by jointly plotting the lower and upper limits of prediction intervals (Figs. 5.2, 5.5, S5.3) and the prediction interval width maps (Fig. S5.5). The probability that the true value lies within a prediction interval might not be easily interpreted by a range of users, particularly farmers. Moreover, it is also challenging to include the reported uncertainty in decision-making on fertilizer recommendations. Section 5.3.5 explained how end-users can be supported by providing maps that show where it is sufficiently certain that fertilizer application will pay off.

We considered a yield gain of 2.0 tons ha^{-1} as the minimum threshold that will encourage farmers to apply fertilizers. Fig. 5.7 shows which parts of the study area are expected to have a yield gain above the threshold, as well as parts where the probability of a yield gain above the threshold is at least 90%. The map shown in Fig. 5.7 could be a simple and useful tool to communicate uncertainty to end-users for a rapid decision on whether to apply fertilizer in a specific plot or not. For instance, farmers are

advised to apply fertilizer in the green areas, but not to the orange areas. However, the decision to apply in the yellow areas is risky for the farmers because it is far from certain that the threshold will be reached. Farmers who are risk-aversive would probably not apply fertilizer in these areas, meanwhile those who can afford the risk can apply. The yellow areas could be targeted by farmers to reduce uncertainty in soil properties through additional soil sampling, hence also reducing uncertainty in yield gain.

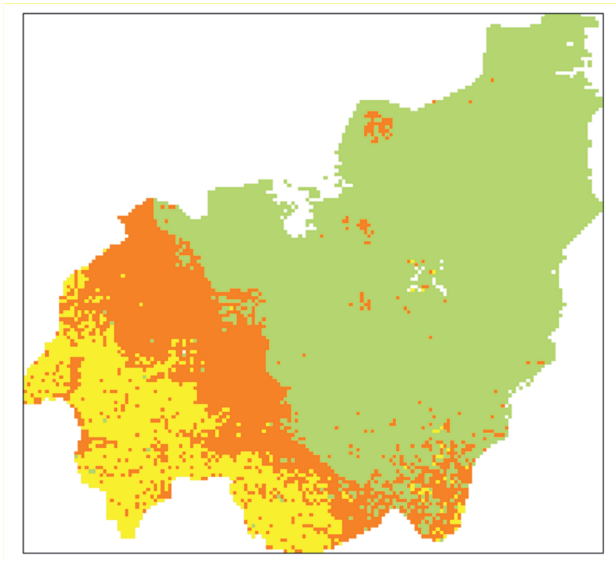


Fig. 5.7. Maps showing areas where yield gains are above the threshold of 2.0 tons ha^{-1}: green = probability that yield gains are above threshold ≥ 0.90, yellow = average of yield gain over 500 MC runs not certain to be > 2.0 tons ha^{-1}, orange = average of yield gain over 500 MC runs < 2.0 tons ha^{-1}.

5.5. General discussion

5.5.1 Uncertainty propagation analysis and implication for QUEFTS modelling

One of the key objectives of this study was to quantify the uncertainty in soil information and analyse how this uncertainty propagates through the QUEFTS model to affect yield and fertilizer recommendations. We illustrated the methodology by combining QRF with Monte Carlo simulation to quantify uncertainty in soil inputs and explore their impacts on the overall uncertainty of the model outputs. We also quantified the contributions of individual soil input uncertainty to the QUEFTS outputs. Then, we summarised the model output probability distributions in a map that shows where fertilizer application has almost certainly a large effect, thus providing a tool to communicate uncertainty to end-users and supporting the integration of uncertain information in decision-making. The methodology applied in a case study in the Western Highlands of Cameroon yielded satisfactory results in recommending fertilizer application rates required to achieve a target maize yield across the study area, along with associated uncertainty. To the best of our knowledge, this study represents the first analysis of the uncertainty in QUEFTS model outputs in Cameroon.

However, there is general consensus among modellers that uncertainty originating from inputs could be a major source of uncertainty in the uncertainty of crop model outputs, given that inputs constitute the most substantial data source (Dokoohaki *et al.*, 2021). However, soil input uncertainty is only one of many sources of uncertainty in crop models, and other sources of uncertainty might well be more important sources. But in this study, we focused only on the uncertainty in four soil inputs, which in itself is worthwhile but of course other uncertainty sources should also be considered in future research. Moreover, we found that just uncertainty in the four soil inputs already resulted in large uncertainty in yield and fertilizer recommendation, thus underlining the importance of this study and the importance of improving the accuracy of the soil maps that are used as input in the QUEFTS model. The results of this study showed a large contribution of soil input to the uncertainty of yield and fertilizer recommendation rates, warranting the need to further reduce crop modelling uncertainty by optimizing the accuracy of soil property maps. Improving soil map accuracy could be achieved by increasing data collection through extensive field sampling, using superior covariates, and employing advanced DSM models like ensemble approaches.

Not all soil inputs had equal contribution to the output uncertainty, and this varied greatly depending on the output variable considered (Table 5.3). Clearly, the overall output uncertainty and uncertainty contributions will vary from one case study to another. This is because the magnitude of the soil map uncertainty is heavily influenced by the soil spatial variability, the spatial sampling design and sampling density, and by the ability of the covariates to explain soil spatial variation (de Bruin *et al.*, 2022; Pusch *et al.*, 2023). In other case studies, it is conceivable that the model output may exhibit varying degrees of sensitivity to specific inputs, with the potential for greater sensitivity in one scenario compared to another. Therefore, the contribution of uncertainty is significant only when the input uncertainty is large, and the model demonstrates sensitivity to that specific input in a specific case study (Nol *et al.*, 2010). Understanding the apportioned uncertainty from different soil inputs helps recognize the key soil properties influencing model outcomes. For instance, certain soil properties account for over 90% of the uncertainty in fertilizer recommendations This knowledge guides efforts to map these inputs more accurately, focusing resources on components with the most significant impact on model uncertainty (Brown and Heuvelink, 2005).

Though we ignored other sources of uncertainty, we contend that comprehensive uncertainty quantification that account for all sources of uncertainty is vital for crop model validation and reproducibility. Other inputs, such as climate, crop parameters, and management practices are very uncertainty and could have major contributions to QUEFTS output uncertainty, The uncertainty in soil inputs was translated into large uncertainty in fertilizer recommendations (Fig. 5.5) to reach the target yield of 5 tons ha^{-1} across the entire study area. We think that aiming for a constant target yield for all locations may not be fair, given the high uncertainty in fertilizer recommendations. The results indicated that some parts of the study area do not require fertilizer to reach the targeted yield due to inherent favourable soil conditions (Fig 5.3). A higher yield could be aimed for such areas, while investment to increase yield in areas with very poor soil conditions should be limited.

5.5.2 Communicating model uncertainty and integration in decision-making processes

One of the objectives of this study was to provide a method for communicating QUEFTS model outputs to end-users and how it can be integrated in decision-making. We argue that instead of reporting a single mean value, i.e. result from a deterministic run, the entire probability distribution in model outputs should be recognized and reported, since it represents the uncertainty about the output, and how confident modellers are with the results of their predictions. The uncertainty in soil inputs resulted in large uncertainty in fertilizer recommendation rates and yield gains (Fig 5.6). We summarized the output probability distributions in a map that shows where fertilizer application has almost certainly a large effect on yield gains (Fig. 5.7). This map provided a simple visual aid to show the spatial distribution of

the uncertainty for a threshold of yield gains. By showing the areas where there is sufficient certainty that fertilization pays off, the map provides an intuitive way for end-users to integrate uncertainty in decision-making. This is under the assumption that the only source of uncertainty is the four soil properties considered in this study. In the real world, there are many other factors including climate, pest and diseases, and crop management practices that affect yield, so we are not certain that the yield gain will always be above the threshold at all locations across the study area. This mode of visualization for uncertain spatial data could be very suitable for end-users without profound knowledge of statistics. Communicating the impacts of uncertainty in this spatially explicit way is a contribution to the ongoing dialogue taking place between modellers, policy makers, and farmers on how uncertainty about model outputs could be communicated (Getson *et al.*, 2022).

A significant factor contributing to the gap between a modellers' comprehension of model output uncertainty and the end-users' utilization concerns the uncertainty communication method and tools. Ineffective communication strategies, especially concerning risk communication (Begho *et al.*, 2022), can result in heightened misinterpretation of uncertainty, potentially fostering disbelief in the impact of model outputs. This is especially critical in the context of fertilizer recommendations that entail substantial investments on the part of small-scale farmers with limited resources (Islam *et al.*, 2022). Drawing from experiences in other fields, particularly the health sector, where effective communication and visualization approaches for conveying uncertainty have been developed and used in critical situations such as the COVID-19 pandemic (McCabe *et al.*, 2021), it becomes evident that scientists often do not convey uncertainty through appropriate communication channels or presented in a format easily digestible by farmers for integration into decision-making (Spiegelhalter, 2017). This could also be the lack of skills on the part of scientists to design suitable uncertainty communication tools that cater to a diverse range of stakeholders. Therefore, efforts to enhance the capacities of both modellers and end-users could help overcome the challenge (Milne *et al.*, 2015). Well-formulated fertilizers can boost yields but pose risks for small-scale farmers. Our QUEFTS study showed that large uncertainty in fertilizer recommendations can lead to risk-averse decisions (Monjardino et al., 2015). Due to financial constraints, farmers tend to be risk-averse because fertilizer application is a high-risk, high-return agricultural business (Haile et al., 2020). In a single input case like fertilizer, a risk-averse farmer uses fewer inputs than a risk-neutral counterpart if the input increases output variability, assuming all other factors constant (Begho et al., 2022).

5.5.3 Limitations to our study and possible improvement

While the QUEFTS model demonstrated satisfactory performance based on the objectives of this study, it is essential to acknowledge certain limitations and challenges. These areas provide opportunities for further development and improvement in the model, as well as enhancing the results of uncertainty quantification and propagation analysis for crop modelling.

Non-consideration of other sources of uncertainty: The DSM and crop modelling processes encompass multiple error sources, contributing to overall output uncertainty. These errors stem from input data inaccuracies, model limitations, environmental variations, and challenges in scaling. Human errors, imprecise parameterization, and the dynamic nature of cropping system further add complexity to error propagation. While this study specifically addressed soil input uncertainties, neglecting other inputs and QUEFTS model uncertainties, it is crucial to recognize that overlooking an important uncertainty source may lead to an underestimation of the overall output uncertainty. To enhance future studies, careful consideration of all sources of uncertainties is essential for a comprehensive estimation of the accuracy of the model outputs.

Not-accounting for cross-correlation between the uncertainty of soil inputs: In our uncertainty propagation analyses, we did not account for interactions between the uncertainty of soil inputs, a factor we recognize as a limitation. We posit that incorporating such correlations would have provided a more

realistic model uncertainty, given the known strong correlations among some of the soil variables. Employing methods that simultaneously model multiple soil properties, such as multivariate random forest or regression kriging models could be a valuable approach to quantify the correlations between the uncertainties of multiple variables, and even simulate from the joint probability distribution (van der Westhuizen et al., 2023). If such correlation can be quantified, then it can also be incorporated in the MC analysis, and this can be explored in further research.

5.6. Conclusion

Our study proposed a methodological procedure for quantifying and propagating uncertainty in soil inputs to QUEFTS outputs, focusing on its impacts on yield and fertilizer recommendation rates in the Western Highlands of Cameroon. The results showed that uncertainty in soil inputs resulted in a large uncertainty in the NPK fertilizer recommendation rates required to reach the target yield of 5 tons ha^{-1}. There was an overestimation of the yield and underestimation of the fertilizer recommendations when uncertainty in soil inputs was ignored. Comparison of the results of a deterministic model run with the mean of the Monte Carlo simulation runs showed systematic differences up to 1.0 tons ha^{-1} for maize yield and up to 59, 42, and 20 kg ha^{-1} for N, P, and K fertilizers, respectively in some parts of the study area. Stochastic sensitivity analysis showed that pH was the main source of uncertainty for K fertilizer (81.6%) and that SOC contributed most to the uncertainty of N fertilizer (97%). Uncertainty in P fertilizer mostly came from uncertainty in POlsen (55%) and KExch (20%). A threshold probability map of yield gains designed using statistical predictions served as a visual tool that empowers farmers to swiftly make informed decisions about fertilizer application locations. The apportionment of the contribution of different soil inputs to model outputs facilitates prioritizing future efforts in crop modelling to reduce uncertainty around yield gains and enhance agricultural intensification. This approach developed in this study provides valuable spatial insights into crop nutrient requirement estimations with associated uncertainty, enabling farmers to adopt tailored fertilizer application based on specific soil conditions.

5.7. Supplementary information

Table S5.1: Descriptive statistics of the soil parameters (n = 640).

Soil properties	Min	Max	Mean	SD	Skewness	CV (%)
N (%)	0.6	3.3	1.8	0.6	0.5	0.34
SOC (g kg^{-1})	8.7	46.6	25.5	9.1	0.3	0.36
pH	4.2	6.3	5.3	0.4	0.7	0.07
m.3.Al (mg kg^{-1})	698	1653	1332	236	-1.0	0.18
m.3.Fe (mg kg^{-1})	64.7	349.8	166.1	67.5	0.6	0.41
ExCa (mg kg^{-1})	0.3	11.5	3.8	2.6	0.7	0.68
Clay (%)	44	93	69	9	0.1	0.13
POlsen (mg kg^{-1})	1.1	14.1	4.1	2.5	1.2	0.60
KExch (mmolc kg^{-1})	0.034	0.060	0.042	0.005	0.9	0.13

m3 = Mehlich 3 extractable; N = Nitrogen; SOC = soil organic carbon; ExCa = Exchangeable Ca; ExchK = Exchangeable K; SD = standard deviation; CV = coefficient of variation.

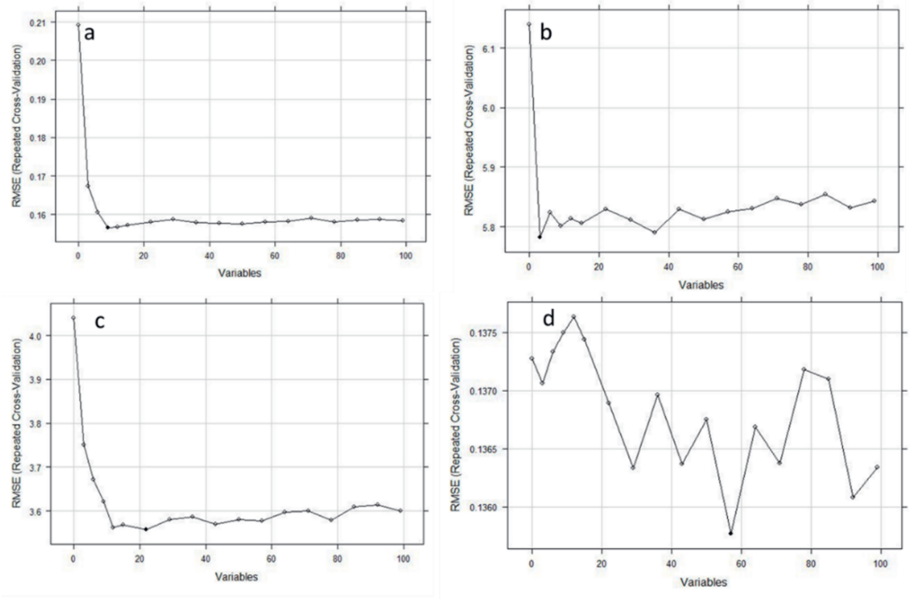


Fig. S5.1: Plot of Recursive feature elimination showing the number of selected covariates: a) pH, b) SOC, c) POlsen, and d) KExch.

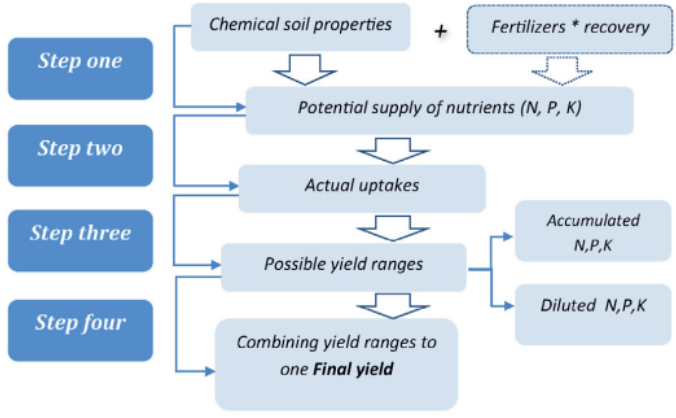


Fig. S5.2. QUEFTS simulation procedure (Sattari et al., 2014).

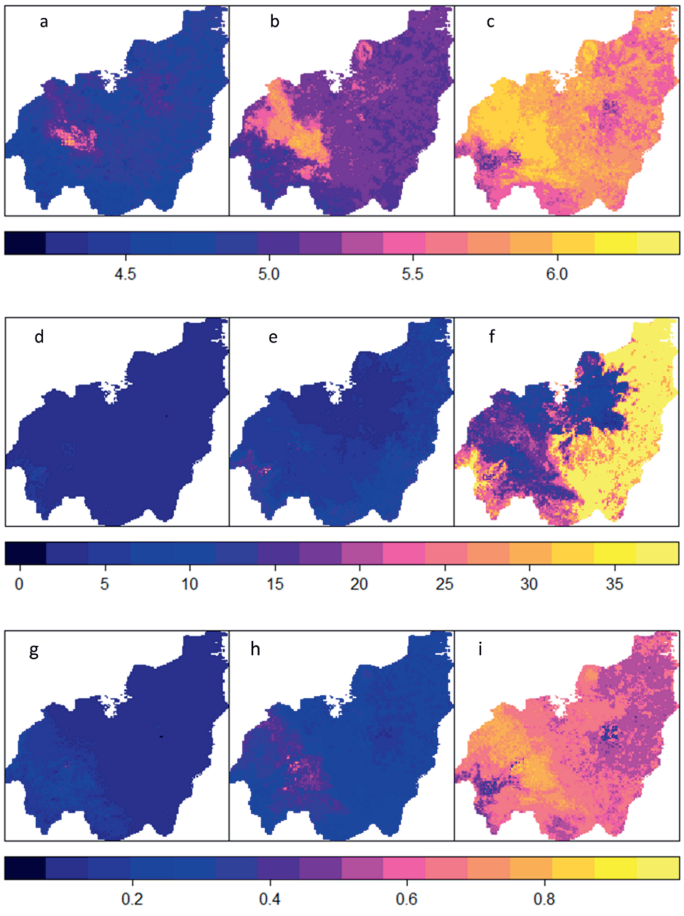


Fig. S5.3. Spatial distribution of soil properties and associated limits of the 90% prediction interval: lower limit, mean values, and upper limit for pH (a, b, c); POlsen (mg kg^{-1}) (d, e, f); and KExch (mmolc kg^{-1}) (g, h, i).

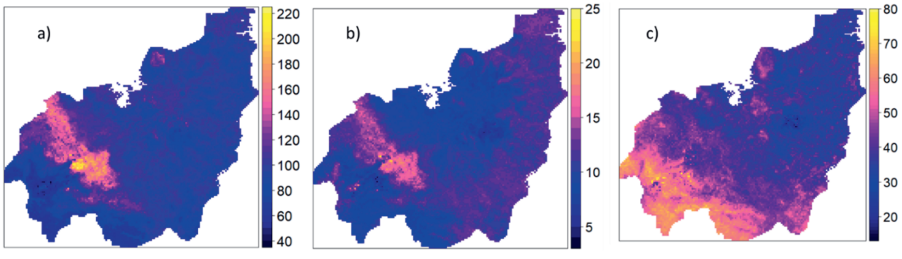


Fig. S5.4. Maps of soil nutrient supply as generated by the QUEFTS model while ignoring uncertainty in soil inputs (kg ha^{-1}): a) Nitrogen; b) Phosphorous; and c) Potassium.

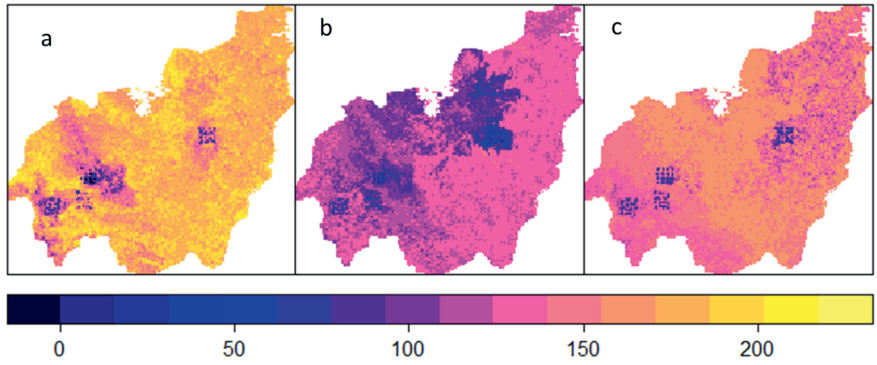


Fig. S5.5. Spatial distribution of the 90% prediction interval width for N fertilizer (a), P fertilizer (b), and K fertilizer (c) required to reach the target yield as obtained by propagating measurement uncertainty in soil input variables through the QUEFTS model.

Chapter 6

Synthesis

6.1. Introduction

The recent advances in proximal soil sensing, geospatial technologies, machine learning, and spatial statistical analyses are enabling exciting opportunities to efficiently generate soil information that is more consistent, detailed, and accurate, but also provides information about the associated uncertainties. The numerous modelling initiatives increasingly require soil information as data sources or data inputs to enable stakeholders to understand and account for soil variation both plot scales and at landscape level with associated uncertainty. This thesis contributed to these advancements by addressing four interconnected key objectives that aimed at enhancing and broadening our understanding of the application of DSM approaches to analyse the spatial patterns of soil properties, while accounting for uncertainty in soil measurements. Additionally, it involves using derived uncertain DSM maps as inputs in crop modelling to predict maize yield and fertilizer recommendations.

The four key objectives were to:

- Describe and quantify the spatial variation of soil properties using simple geostatistical methods (ordinary kriging).
- Quantify the errors in soil measurements and incorporate these into a state-of-the art geostatistical method (regression kriging) for spatial interpolation and compare results with a case in which measurement errors are ignored.
- Extend the calibration and prediction of DSM models using uncertain soil measurements from linear kriging with external drift, i.e. regression kriging (RK) to non-linear machine learning algorithms-based DSM models, i.e. random forest (RF).
- Quantify the uncertainty of QUEFTS model predictions considering uncertainty in soil input variables using Monte Carlo simulation.

To achieve these objectives and assess how uncertainty propagates through DSM and crop models and affects the quality of the model outputs, I applied various DSM approaches including ordinary kriging (OK), regression kriging (RK), random forest (RF), and quantile regression forests (QRF) using uncertain analytical and spectral soil measurements. I also conducted an uncertainty propagation analysis of the QUEFTS model using Monte Carlo simulation. A flowchart with a summary of the methodology that I used in this PhD research to achieve the defined objectives is presented in Fig. 6.1.

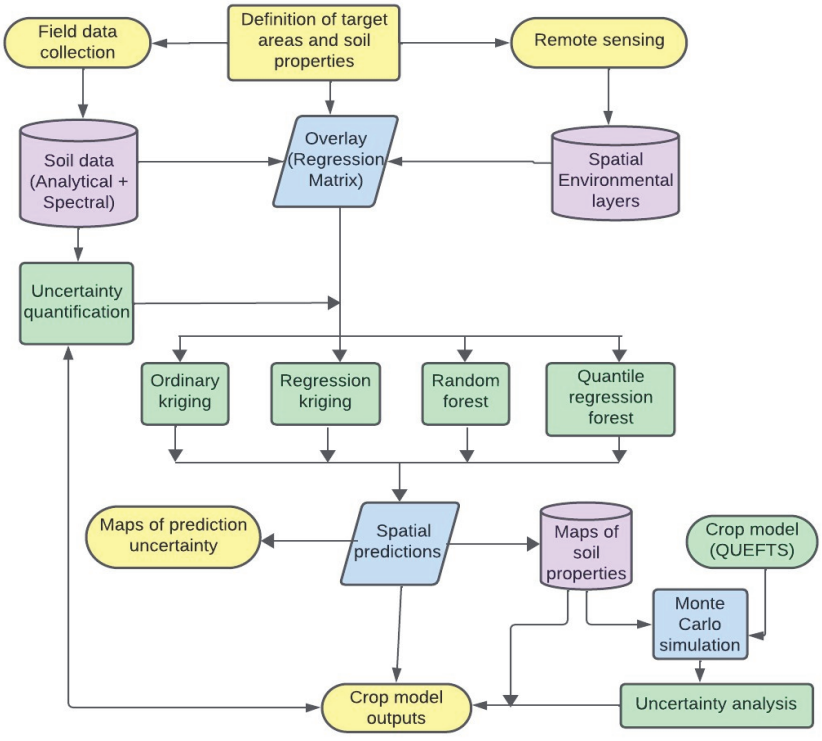


Fig. 6.1: Summary of the methodologies used in this PhD research.

This final chapter consists of four sections. In Section 6.2, I synthesize the results of the previous chapters, reflecting how the objectives of the PhD research have been attained, while in Section 6.3, I discuss the broader implications of the findings for sustainable agricultural intensification for Cameroon and SSA, as well as the methodological underpinnings. Finally, Section 6.4 presents the limitations and provides recommendations to enhance further research endeavours.

6.2. Summary of key findings

This PhD research demonstrated the successful integration of uncertain analytical and spectral soil data to generate digital maps of soil properties in a case study area in Cameroon, using various DSM techniques. The resultant soil maps served as inputs for modelling crop yield and formulating fertilizer recommendations. The key findings of the research are outlined below:

In Chapter 2, simple geostatistical approach is applied to characterise and map the spatial variation of selected soil physico-chemical properties among different land-use types including forest, grassland, fallow, cropland, and pasture in two sites: Bamendjou dominated by agricultural activities, and Koutaba dominated by livestock activities. A total of 320 topsoil samples (0-20 cm) were collected in the two sites and were analysed in for a range of soil

properties using diffuse reflectance mid-infrared spectroscopy. Variograms were used to quantify the spatial variation of soil properties, while ordinary kriging was applied to generate the respective maps of SOC, N, and clay content. Accuracy of the prediction performance of models was assessed using cross-validation. Soil properties differed considerably across the study area, with significant positive and negative correlation coefficients among many pairs of soil properties. The coefficients of variation (CV) helped in comparing the degree of variation of soil properties within the two sites. For Bamendjou, the most variable properties (CV >38%) were P, Mn, and Ca. Moderate variability (2.8% < CV < 38%) was observed for sand, SOC, N, K and Mg, while properties with very low variability (CV < 2.8%) were clay, soil pH, and Al. For Koutaba, the most variable properties were sand, Ca, K, Mg, Mn, P and Zn. Moderate variables were N, SOC, and silt, while clay, Al and pH were the least variables. Spherical variogram models were chosen as the best-fitted models for the investigated soil properties as attested by cross-validation. The spatial correlation ranges were significantly wider for SOC and N in Bamendjou than in Koutaba. Land use with more vegetation cover (forest, grassland, and fallow) exhibited the highest concentration of soil properties, attesting that land use types had significant impacts on spatial patterns and distribution of the soil properties. Well-defined pattern of higher concentrations of SOC and N were observed in the lowlands, valleys, and areas dominated by annual vegetation. Kriged maps provided detailed visualization of soil properties at landscape scale and helped to identify hot spots of land degradation and critical areas in need of specific land management practices to improve land productivity. These findings can be a helpful tool in achieving efficient site-specific land management interventions, that lead to better decisions aimed at enhancing the efficient use of agricultural inputs, such as fertilizers, in the context of limited resources. Although the spatial models could explain a large part of the spatial variation of soil properties, they may be improved by quantifying the measurement errors in soil observations and expanding the analysis with relevant covariates that represent the soil forming factors (**Chapter 3)**.

In Chapter 3, a geostatistical-based DSM approach (regression kriging) is applied to incorporate quantified measurement error variances of analytical and spectral soil data in the covariance structure of the spatial model, weigh measurements in accordance with their quantified measurement accuracies and assess the effects of measurement errors on the accuracies of the resulted DSM predictions. The method was applied on soil data analysed for pH, clay content and SOC using both conventional and mid infrared spectroscopy methods. Variogram parameters and regression coefficients were estimated using residual maximum likelihood under two scenarios: with and without taking measurement errors into account. Performance of the spatial models in the two scenarios was compared using common validation metrics obtained with three types of cross-validation, namely leave-one-out, leave-cluster-out and leave-sentinel-site-out cross-validation. Accounting for measurement errors had significant impacts on the estimated regression coefficients and influenced the variogram parameters by reducing the nugget and sill variance for the three soil properties. Validation metrics including mean error, root mean squared error and model efficiency coefficient, were quite similar in both scenarios, but the prediction uncertainties were more realistically quantified by the models that accounted for measurement errors. There were relatively small absolute differences in predicted values of soil properties of up to 0.1 for pH, 1.6% for clay and 2 g kg^{-1} for SOC between the two scenarios. While differences in prediction maps and cross-validation metrics of predictions did not differ significantly between the two scenarios, substantial differences were obtained in prediction error standard deviation maps and in the evaluation of the prediction uncertainty. Relative differences in standard deviations were observed in some areas of the study area between the two scenarios of up to 0.08 for pH, 2.7% for clay and 0.5 g kg^{-1} for SOC. The best modelling approach would therefore be the one that accounts for

measurement errors in soil observations. The overarching findings suggest that the additional investment in quantification of measurement error and its incorporation in the spatial models is a worthwhile endeavour. This is underscored by the observed improvement in the quantification of prediction uncertainties. Furthermore, these findings also emphasised the necessity of incorporating measurement errors in further studies not only with geostatistical-based DSM approach, but also by exploring the use of non-linear machine learning regression methods (**Chapter 4**) to improve uncertainty quantification. This becomes particularly crucial when spectral data are used as the main soil data source.

In Chapter 4, RK and RF are compared with respect to their ability to deliver accurate predictions and quantify prediction uncertainties, while accounting for errors in the soil measurements. The soil dataset was a combination of analytical and spectral data the case in **Chapter 3**. I also evaluated the sensitivity of the results of both DSM approaches to soil measurement errors, together with their spatial extrapolation potentials while mapping soil pH, clay content, and SOC. Both models produced comparable ranges and maps of predicted values for the three soil properties. Compared to RF, RK outperformed RF by presenting generally a higher model efficiency coefficient, lower root mean squared error values and better extrapolation performance. The improvement in root mean squared error was about 10, 12 and 2% while the improvement in the model efficiency coefficient was on average 5, 22 and 1% for pH, clay, and SOC, respectively. Overestimation of the local uncertainty observed for RK was larger than that of RF, indicating that prediction uncertainties were better quantified by the RF model. The effects of incorporating measurement errors appeared not significant both for predictions and prediction uncertainties. This was partly attributed to the fact that this chapter assumed constant measurement error variances for each of the soil property, and for analytical and spectral data used for calibration instead of considering measurement errors that are proportional to the measured values. Another important finding of this chapter was that model comparison should go beyond using only common validation metrics to evaluate prediction accuracy of DSM approaches but should also account for their ability to quantify prediction uncertainty at unsampled locations. The findings of this chapter show that soil maps are not error free, and if use as soil data inputs, the uncertainty can propagate to affect the results of further modelling. Hence, it is essential to evaluate the repercussions of uncertain soil information in other modelling processes, particularly in crop modelling (**Chapter 5**).

In Chapter 5, quantified the uncertainty of QUEFTS model predicted yield and fertilizer recommendation rates is quantified considering the uncertainty in soil inputs using Monte Carlo simulation. It also determined the contributions of individual soil input uncertainties to the overall model output uncertainties, discussed strategies to communicate uncertainty of QUEFTS outputs to end-users and advised on how uncertainty can be incorporated in the decision-making process. The study demonstrated that the impacts of soil input uncertainty on model output uncertainty were significant and varied spatially. Notably, the results showed considerable uncertainties in yield and fertilizer recommendations, with interquartile ranges larger than the median in some parts of the study area. While comparing the results of the deterministic run with those of the Monte Carlo simulations, mean differences in predicted yield and fertilizer recommendation rates in some areas of the study area were up to 1.0 tons ha^{-1} and up to 59, 42, and 20 kg ha^{-1} for Nitrogen, Phosphorous and Potassium fertilizers, respectively. Accounting for soil input uncertainty leads to a systematic shift of the three fertilizers towards higher values. The spatial distribution of the uncertainty maps closely matched the spatial patterns of the soil input uncertainties, with high values where input uncertainty is high. Stochastic sensitivity analysis showed that pH is the main source of uncertainty for K fertilizer (82%) and that SOC contributes to the uncertainty of N fertilizer

(97%), P fertilizer (25%) and K fertilizer (18%). Uncertainty in P fertilizer mostly comes from soil extractable phosphorus (55%) and exchangeable potassium (20%). A threshold probability map designed using statistical predictions served as a visual aid that could enable farmers to swiftly make informed decisions about fertilizer application locations. The study highlighted meaningful relationships between the uncertainty of soil properties and the uncertainty associated with model outputs. It emphasized the necessity for future research to address correlations among uncertain soil inputs and consider additional sources of uncertainty for a more thorough and accurate quantification of model output uncertainty. In general, the findings of the research underscore the importance of improving the accuracy of soil measurements and maps, which, in turn, positively influences the accuracy of DSM and QUEFTS model predictions. The findings of this chapter provide significant benefits for site-specific nutrient management proposed in **Chapter 2.** These benefits include estimating the soil nutrient supply, recommending required fertilizer application rates, and determining corresponding theoretical yields. This PhD research also contribute to enhancing our understanding of the interplay between the accuracy of soil information and the reliability of DSM model outputs (**Chapter 3** and **Chapter 4**), offering valuable insights of the uncertainty in soil inputs for improved decision-making to achieve sustainable agricultural intensification (**Chapter 5**).

6.3. Implications and practical applications of the findings

The DSM and crop modelling approaches proposed in this PhD research hold the potential for wide application across various domains and areas. Their implementation can enhance the accuracy of DSM and crop model predictions. Therefore, it is recommended to use these approaches to enhance the laboratory procedures for soil samples analysis by incorporating the systematic quantification of measurement errors, leverage proximal soil sensing techniques, or improve the accuracy of DSM and crop models outputs, especially when dealing with limited and uncertain soil data. These methodological approaches can be more efficiently and effectively applied due to recent advancements in digital information and computational technologies, making it highly adaptable and versatile. Based on the findings of the research, implications and recommendations are provided as follows:

Soil information for sustainable land management

Agriculture is the backbone of most countries in SSA and employs more than 60% of its population. Intensification of crop production, which is mainly rainfed-based, has not been able to match the population growth (Giller *et al.*, 2021). Meanwhile it stands as the only viable approach for small-scale farmers to increase yield per unit area though agricultural intensification (Kansiime *et al.*, 2022). The low productivity of small holdings is attributed to declining soil fertility, limited use of external input, unfavourable agricultural policies, as well as the effects of climate variability (Vanlauwe *et al.*, 2015). The findings of this PhD research have reiterated the fundamental role of accurate soil information in supporting land management decisions, as an important step in the way towards agricultural intensification.

Maps of soil properties represent an important basis for the evaluation of the quality of land resources, so as to take appropriate decisions for productivity improvement (AbdelRahman and Metwaly, 2023; von Fromm *et al.*, 2024). The understanding of the spatial variability of soil properties and the phenomena that influence their variation in space is crucial for the implementation of practices that prevent degradation and foster restoration (Winowiecki *et al.*, 2018). This is particularly important in SSA where there is great diversity of climate, geology,

landscapes, and soil forming processes within a relatively small area. In this context, accurate soil information plays a key role in the restoration of degraded lands, which has received increased international attention in recent years and many countries in SSA have committed to restore million hectares of lands. Some of the initiatives for large-scale restoration include the Bonn Challenge (www.bonnchallenge.org) and related regional initiatives such as the AFR100 (www.afr100.org), the UN Decade on Ecosystem Restoration (2021–2030) (Gnacadja and Vidal, 2023), and the concept of land degradation neutrality (Feng *et al.*, 2022). Soil maps are an important source to identify areas affected by degradation and optimize the resources available to restore them. The results of this PhD research have shown that an important prerequisite in achieving these restoration initiatives/targets is the knowledge of the spatial distribution of soil properties (**Chapter 2**). The generated soil maps act as visual tools to identify areas in need of restoration and the suitable specific practice that can be implemented to restore productivity potentials.

These findings support those of previous studies in Chad and Kenya that demonstrated the usefulness of spatial soil information in determining the potential of land restoration and the level of investment needed in heterogenous landscapes (UNEP, 2012; Winowiecki *et al.*, 2018; Takoutsing *et al.*, 2023). Other studies have also successfully used soil information to evaluate land suitability for agricultural investments and land use plannings (AbdelRahman *et al.*, 2022; Dornik *et al.*, 2022). With the recent developments in digital technology, GIS, remote sensing, machine learning and spatial modelling, there is a greater awareness by land managers and policy makers of the need for detailed soil information to be incorporated into decision support systems (Bulmer *et al.*, 2019). Such spatial land management tools are a vital component of land management, not just for policy makers, but also for farmers, researchers and technicians who directly make use of the information. Policy makers may implement appropriate policies depending on soil quality for specific areas, while technicians can apply effective site-specific practices. The government can utilize soil maps in combination with crop models to make more efficient management decisions on fertilizer import, distribution, and recommendations that are based on nutrient requirements. In addition, government should implement policies that incentivize farmers to invest in maintaining and improving soil health as the foundation of sustainable and regenerative food systems (Wolde-Meskel *et al.*, 2022).

This PhD research demonstrated the critical role of assessing the spatial variability of soil properties in identifying targeted areas for implementing practices that mitigate land degradation and enhance productivity. In resource-constrained contexts, such as the targeted landscapes of this research, the generated soil information becomes a valuable tool for optimizing the impact of land management and restoration projects by strategically orienting activities to areas where the impact will be most significant, and the efficiency of available resources enhanced. But many initiatives utilizing spatial soil information to guide restoration activities or assessing land suitability for priority crops often overlook the inherent uncertainty in the data. This oversight can have serious consequences, leading to irreversible economic and environmental outcomes for many stakeholders. This PhD research addressed this knowledge gap by providing insights into how to manage uncertainty in both linear kriging and non-linear machine learning algorithms-based DSM models (**Chapter 3** and **Chapter 4**), as well as in crop modelling (**Chapter 5**). Applying these methods enables a realistic estimation of prediction accuracy, empowering land managers with more reliable information for informed decision-making. Furthermore, the research underscores that crop modelling, crucial for soil nutrient management and land productivity, has often neglected the uncertainty associated with soil information used for predicting yields and formulating fertilizer recommendations. Ignoring this uncertainty in policy design and small-scale farmer decision-making can have

severe repercussions. **Chapter 5** emphasized the substantial uncertainty in crop model outputs resulting from uncertainties in soil inputs. By raising awareness about this issue, the research advocates for a shift in perspective among stakeholders, urging investments in obtaining accurate soil information to enhance land productivity and advance sustainable agricultural intensification.

Quantification of soil laboratory measurement errors

Errors in soil measurements used as data source propagate in further applications, such as pedotransfer functions (Padarian *et al.*, 2018), spectral models (Semella *et al.*, 2022) and digital soil mapping (Heuvelink, 2018). For the case of spectroscopy techniques, the last decades have witnessed the establishment of visible-to-near infrared (VNIR) and mid-infrared (MIR) soil spectroscopy as a rapid and cost-effective method of soil analysis (Towett *et al.*, 2015; Ng *et al.*, 2022). The spectral techniques allow for a larger number of samples and as a result thus a better description of soil variability. The findings of this research have demonstrated the usefulness of spectroscopy techniques in generating soil information through MIR spectroscopy. Prediction of soil properties from reflectance data requires the development of multivariate calibrations, i.e., regression models that relate the measured spectral signal to the soil data obtained using conventional laboratory methods (Viscarra Rossel *et al.*, 2022).

Previous studies have reported accurate predictions for soil properties, such as soil organic carbon, soil nitrogen, clay, and carbonate contents, and concluded that spectroscopic models can be sufficiently accurate to partially replace standard laboratory methods in soil analysis (McBratney *et al.*, 2006; Poppiel *et al.*, 2022). The authors of these papers argued that the common assumption that spectral predictions cannot be better than the reference method used for the calibration is not strictly true, since multivariate modelling with many predictor variables is theoretically capable of compensating for noise in laboratory reference data. However, in most of these studies dealing with soil properties prediction from spectra, no clear statement about input data uncertainties is made. The reported prediction errors only refer to the model building procedure, while uncertainties from analytical measurements used to train the models are neglected.

The results of this PhD research show that there is an issue with the assumption that spectral predictions can be better than the reference method, because error metrics for spectroscopic models are computed against the analytical reference data under the assumption that these data are essentially error-free. This is incorrect because during the estimation of the error, it is the deviation from the analytical reference data that is actually measured, which includes the inherent uncertainty in the analytical data. Despite the success of MIR spectroscopy in predicting soil properties, the findings of this research showed that ignoring uncertainty in analytical data leads to underestimation of the spectral error (**Chapter 3**). However, there is a problem with attributing errors or poor performance of spectral models to laboratory analytical error because the magnitude of these errors is usually unknown. This is why one of the objectives of this research was to assess the uncertainty inherent in laboratory reference data as one error source in spectroscopic modelling. In this thesis I was able to distinguish between true and apparent spectral prediction errors, with the latter arising, for example, from inaccurately measured reference values.

I applied a simple method to quantify the soil measurement error variances using a repeatability procedure (van Leeuwen *et al.*, 2022). Measurement error variance is derived from comparison of measurements on the same soil sample multiple times under the same conditions, such as the same laboratory, same analyst, same time, and same equipment. The procedure has also

been used previously to re-analyse a set of soil samples so as to calculate the reproducibility error of reference measurements during the development of the European LUCAS soil database (Stevens *et al.*, 2013). Furthermore, variation in a laboratory over time and from one to another can be large but are also rarely considered. Actually, there is a general tendency to attribute the uncertainty in soil property estimates using spectroscopic models to the deficits in the accuracy and reproducibility of the respective spectral measurements, e.g., by instrumental characteristics, sample type and preparation, and also from error sources in the model building procedure (Ellinger *et al.*, 2019).

This PhD research also demonstrated that a complete evaluation of spectroscopic modelling results requires a separate error assessment of the analytical reference measurements, in addition to an assessment of the variance of the spectral model. This is in line with previous studies that analysed the impact of spectral variation and repeated measurements on the validation of spectroscopic models and found that an error component could be attributed to the combined uncertainty introduced by variation in spectral measurements and reference values (Ellinger *et al.*, 2019). In addition to the quantification of measurement errors in analytical data, this research quantified the measurement errors in spectral data using the residual variance of the partial least square regression models that were used to predict the soil properties. This is a simple approach to assess spectral prediction measurement errors that combined both error in analytical data and that of the spectral model.

The methodological approach developed in this PhD research can be applied to improve the use of soil spectroscopy to generate soil information while being able to quantify separately the different sources of uncertainty. By knowing the various sources, measures could be taken to reduce their impacts on the final production and give more credibility to the model outputs and help users who need to invest in improving soil analysis and soil maps. This PhD research suggests an enhancement in budget allocation for laboratory analysis of soil samples within research and development projects implemented in the study area, Cameroon, and SSA. Specifically, projects aiming to generate soil information for various applications are encouraged to allocate additional resources to ensure thorough laboratory analysis, thereby improving the quality and reliability of the soil data generated. This will help in implementing the methods developed in this thesis by taking and analysing replicate samples so that measurement errors can be quantified. Also, laboratories, including those that measure soil properties using spectra, should routinely quantify and communicate the uncertainty of their measurements (analytical as well as spectral) in such a way that a user can decide whether to invest more in improving the quality of the measurements. Note that this is only possible if the measurement uncertainty is well quantified and communicated.

Furthermore, the widespread application of soil spectroscopy has recently received new impetus through the establishment of large spectral libraries and the introduction of portable MIR instruments (Zhong *et al.*, 2021; Shepherd *et al.*, 2022; Moloney *et al.*, 2023). This PhD research is a contribution to the ongoing debate on the accuracy of soil data generated not only using conventional laboratory methods, but also proximal soil sensing, and stressed the need to address the various sources of uncertainty and assess the implications for end-users. This could form the basis for the development of standardized protocols for soil sample analysis in laboratories and improve the accuracy of data in soil spectral libraries (Shepherd *et al.*, 2022; GLOSOLAN, 2023). Measurement errors in laboratory analytical methods that occurs across laboratory should also be considered in future research (Shepherd et al., 2022).

Soil maps as an important data input for spatialising crop models

Generally, the most common source of soil data used as input for crop modelling has been obtained from yield response experimental trials, often estimated by one representative soil profile (Lagacherie *et al.*, 2022). This often leads to prediction of crop model outputs at plot level, which may not be useful to many farmers due to local soil spatial variability, and also not the priority for extension agents and policy makers who are more interested in information at landscape and national level respectively. Accessing soil data has been the main challenge for the spatialisation of crop modelling as limited data lead to extrapolation of input data, which increases uncertainty. Uncertainty on soil input will propagate to the outputs of the crop models (Wallach and Thorburn, 2017; Mukhtar *et al.*, 2019). Modellers might be aware of this uncertainty and the harms it causes on the model outputs, but they rarely make mention of input errors while reporting the accuracy of the model outputs. Previous studies have used soil maps as inputs for the spatialisation of the QUEFTS model to predict outputs beyond the plot scale (Leenaars *et al.*, 2017; Breure *et al.*, 2022a).

Through this PhD research, I showed that DSM approaches such as OK, RK and RF if well implemented can be alternative sources for providing soil information useful for crop modelling spatialisation. DSM approaches not only provide spatial soil information, but they are also able to ex-ante estimate the uncertainty associated with the information (Heuvelink, 2014). This research demonstrated that for the case study in Cameroon, estimation of soil inputs is one of the most critical sources of uncertainty that may affect the spatialization of crop models. This research significantly contributes to recognizing the role of DSM maps and the associated uncertainty in crop modelling. It highlights the potential for various crop models and studies in diverse regions globally to leverage the spatialization of model inputs and outputs. Additionally, it emphasizes the importance of quantifying uncertainties associated with crop model outputs to enhance decision-making at the landscape level. The methodology employed in this PhD research to quantify the uncertainty of crop model outputs, by accounting for the inherent uncertainties associated with soil inputs, has been notably effective. By quantifying the contributions of individual soil input uncertainty sources to the overall uncertainty of the model outputs, as well as quantifying the level of accuracy of the model outputs, end-users could easily identify which soil property constitutes the weakest link and which measures can be taken to increase accuracy.

The findings derived from this PhD research have the potential to empower farmers by enabling them to juxtapose their actual yields with the model-predicted yields, and to fertilizer providers to understand the impact of soil properties on fertilizer formulation and recommendations. This comparative analysis could serve as a valuable tool for them in making informed decisions aimed at optimizing crop yield. Such decisions may encompass strategic choices like adopting high-yield fertilization practices, implementing balanced fertilization strategies, and incorporating soil-test-based fertilization approaches, all of which contribute to the overarching goal of enhancing agricultural productivity. Indeed, using a crop model for the estimation of fertilizer application rates, while considering the inherent uncertainties associated with soil inputs, provides a nuanced approach. This methodology enables the adjustment of soil input assessments based on their respective levels of uncertainty and the ensuing impacts on model outputs. By incorporating this adaptive framework, farmers can tailor their soil input evaluations, taking into account the uncertainties, thereby enhancing the accuracy and reliability of fertilizer application recommendations (Schils *et al.*, 2022).

Communicating uncertainty to end-users of DSM and crop model outputs

The findings of this research have shown that most model outputs for the case study in Cameroon had substantial uncertainty, which has consequences on the decisions based on this inaccurate information. This case study likely represents a broader trend, suggesting that encountering uncertainties is not an isolated incident, but rather a common occurrence in general. The findings from this case study therefore serve as a potential indicator that significant uncertainties are prevalent in similar contexts, emphasizing the importance of acknowledging and addressing these uncertainties in crop modelling in general. The potential impacts of poor decisions based on inaccurate information can be huge and irreversible, particularly for small-holder farmers with limited resources. Communication of uncertainties therefore aimed at a range of stakeholders, as well as others involved in policy making and is of high importance. In addition, communicating uncertainty is a matter of good scientific practice, accountability, credibility, and openness towards the general public (Wardekker *et al.*, 2008). Failure to do so may create a false sense of certainty among users, which can lead to distrust when that certainty proves to be overstated (Simpkin and Armstrong, 2019).

Several methods have been developed to communicate and present uncertainty in various scientific domains, including soil science, health, engineering, and many others, but not all these methods are easy to understand by non-technical audiences, and they can also unexpectedly lead to misinterpretation (Wardekker *et al.*, 2008). The success of a method to properly communicate uncertainty may depend on the subject matter and on the background knowledge of the users (Milne *et al.*, 2015; Chagumaira *et al.*, 2021; Chagumaira *et al.*, 2022). In soil mapping, communicating uncertainty to end-users has most frequently been done through parameters of the probability distribution of the model outputs, such as the median and interquartile ranges (Heuvelink and Webster, 2022). The presentation of the uncertainty of the output of various modelling processes quantified in this research is an example of such illustration. However, communicating the probability that the true value lies within a prediction interval might not be easily interpreted by all end-users for whom the results of a research are intended. This is because these users, particularly farmers in the context of this research, might not have basic statistical knowledge to easily understand information on uncertainty presented in the form of prediction intervals (Spiegelhalter, 2017).

This research has made a significant contribution to innovating uncertainty communication by employing statistical predictions that transform the probability distribution of QUEFTS model outputs into maps depicting the degree of certainty of an effect of fertilizer application. In **Chapter 5**, I delineated areas where it was sufficiently certain that the anticipated maize yield gain surpasses a predetermined threshold following fertilizer application. Such maps serve as a valuable visual aid for farmers, allowing them to make informed decisions about whether to apply fertilizer to a specific plot or not. Empowered with this information, farmers can promptly act when the yield gain either exceeds or falls below the threshold in a particular farm, based on the insights provided by the probability maps. This approach not only facilitates swift decision-making at the farm level but also addresses a longstanding need expressed by numerous land-users and extension agents in various local communities across SSA.

By using this approach, farmers can avoid or minimise loss in input investment and put themselves in the safer side of risk taking, because the map tells them where it is sufficiently certain that fertilizer application will lead to substantial yield gain. This uncertainty communication approach is also useful for extension agents in their mission to support farmers in identifying areas that require action to increase maize production through fertilizer applications, while minimising the risk of wasting the inputs and reducing return in investment. Policy makers could also use the same approach to delineate areas where they are certain that

fertilizer application would lead to agronomic, as well as determining the type and the quantity of fertilisers required. This helps to define policies with evidence-based and robust technical arguments that can enhance their implementation, while at the same time increasing the level of trust between farmers, modellers, and extension agents. This can only be achieved if all stakeholders involved in the modelling processes quantify and communicate uncertainty of their results in such a way that the uncertainty of the final products can easily be assessed, including assessment of the contribution of all sources of uncertainty.

6.4. Limitations of the research and recommendations for future research

While this PhD research has yielded satisfactory results based on the objectives, it is important to acknowledge certain limitations and identify areas that warrant additional attention for further improvements. These limitations may include aspects such as data constraints, or methodological considerations that could impact the generalizability of the findings. Additionally, attention should be directed towards exploring avenues for refinement or expansion of the research framework, potentially incorporating emerging technologies or alternative methodologies to enhance the overall robustness of the research. Addressing these considerations will not only contribute to the comprehensive understanding of the research outcomes but also set the stage for future advancements and refinements in the field of DSM and crop modelling.

Soil spatial sampling design

Designing an effective soil sampling strategy is a crucial step in optimizing the outcomes of DSM models. The selection of sampling locations, distribution, and density significantly influences the accuracy and reliability of the derived soil information (Wadoux *et al.*, 2019). A well-thought-out sampling design that considers spatial variability of soil properties and environmental covariates, ensuring representative data points across the targeted area. By optimizing the sampling design, DSM efforts can enhance precision and provide a more robust foundation for creating accurate soil maps and models. The cluster soil sampling design I used in this research was not initially intended for spatial mapping, but rather designed to provide a biophysical baseline, and a monitoring and evaluation framework for assessing processes of land degradation and the effectiveness of rehabilitation measures (Vågen & Winowiecki, 2013). Consequently, the design was not optimised to properly assess spatial correlation over large distances (Brus et al., 2011) and this was a limitation for this research. Most initiatives in Africa geared towards collecting soil samples favour cluster sampling due to accessibility challenges in difficult terrain and limited resources allocated to field survey. Examples of such initiatives include the Africa Soil Information Service (AfSIS) (http://africasoils.net/) and Soils4Africa (https://www.soils4africa-h2020.eu/). Cluster sampling is however prone to biases and large prediction errors at unsampled locations when applied to cover large areas. As demonstrated in **Chapter 4**, it also complicates the evaluation of map accuracy using cross-validation.

The sampling design used for this PhD research likely did not meet the soil mapping requirements in terms of feature space coverage, as a key element to sampling optimisation for spatial prediction using machine learning. This soil sampling design was not initially intended for geostatistical mapping, but rather to provide a biophysical baseline, and a monitoring and evaluation framework for assessing processes of land degradation and the

effectiveness of rehabilitation measures (Vågen and Winowiecki, 2013). The DSM output accuracy is influenced by the sample size and spatial locations of the sampling units that are used to calibrate the model (Wadoux *et al.*, 2019). This is even more important for kriging methods where the soil property is mapped using a known model of spatial variation, i.e., the variogram, which requires that paired comparisons are made at all distances. In addition, the kriging variance is larger in areas that are under-sampled (**Chapter 3**), indicating that sampling units should be spread evenly throughout the targeted area (Delmelle and Goovaerts, 2009).

This PhD research used the sentinel site cluster sampling design (Vågen and Winowiecki, 2020) to produce results both with RK and RF for the set objectives, though we acknowledge its limitations. However, optimising sampling designs is not a straightforward problem, and the optimal design depends on the mapping method used. There is no single best sampling design for all DSM approaches (Brus, 2019). For instance, the spatial coverage design is recommended for kriging as it ensures a fairly uniform spread of the measurements in geographic space (Brus and Heuvelink, 2007). In the case of machine learning with covariates, feature space coverage sampling or conditioned Latin Hypercube sampling is recommended (Brus, 2019). Based on the findings of this research, future sampling designs could be improved by combining cluster sampling with other sampling methods to account for variation, both at short and larger distances, in such a way that the sampling design cover the feature space well (Brus and Heuvelink, 2007; Wadoux *et al.*, 2019). Nevertheless, the suggested integrated sampling approach may come with associated cost implications, potentially rendering it financially challenging in SSA. However, the importance of implementing such a strategy is underscored by the necessity to align with advancements in spatial modelling. While cost-effectiveness is a valid consideration, the benefits of embracing this combined sampling scheme, in terms of improved accuracy and relevance in the context of evolving spatial modelling techniques, make it a worthwhile investment for staying abreast of developments in the field of soil mapping. Balancing cost constraints with the potential gains in accuracy and model performance becomes imperative for achieving a judicious and forward-thinking approach to spatial modelling endeavours.

While statistical optimization of soil sampling designs is valuable, practical considerations, particularly the safety of field staff, must take precedence. In the African context, it is important to recognize that certain regions may be inaccessible due to political and military conflicts, adding a layer of complexity to soil sampling endeavours. In conflict-affected areas, the feasibility and security of implementing any proposed soil sampling strategy become paramount. Striking a balance between statistical rigor and real-world challenges imposed by conflict-related constraints is essential. Ensuring the well-being of field staff and adapting methodologies to navigate safety concerns becomes a critical aspect of any soil sampling initiative in such environments.

Use of a small soil dataset considering the large study area

In this thesis I used a fairly small dataset (n = 480, **Chapter 3** and **Chapter 4**; n = 640, **Chapter 5**) considering the surface area of the study areas. This is a limitation because it is widely recognised that a larger number of soil samples could capture more spatial heterogeneity, thus improving map accuracy when using DSM. In this PhD research, I combined analytical with spectral data to increase the sample size, but the number of soil samples was still constrained by limited resources allocated to fieldwork and laboratory analysis. As well-known from the literature (Viscarra Rossel *et al.*, 2016a) and confirmed in this PhD thesis, laboratory analysis costs could be significantly reduced using soil spectroscopic techniques

(Viscarra Rossel and Bouma, 2016; Chen *et al.*, 2020). In light of the advantages demonstrated in this research by the soil spectroscopic techniques, development projects with soil sampling objectives can afford extensive soil sampling density and analysis to provide a better understanding of soil spatial variation for DSM practices (Somarathna *et al.*, 2018). As a recommendation, it is required that the sample data should be sufficient to represent the soil – environment relationships across the study area for reliable model calibration or construction. From probability theory and geostatistical analysis, the soil sample data should be collected through a well-defined field sampling design, which typically aims to allocate soil samples in geographical space and/or environmental covariate space (feature space) in such a way that the designed sample set is representative of the targeted area (Zhu *et al.*, 2015).

Inadequate estimation of uncertainties in soil measurements in the laboratory

The method used to quantify uncertainty in soil measurements (**Chapter 3**), particularly for analytical data, had some limitations and as such, the estimates may not be as good as expected. I used soil samples from another research with similar environmental and soil conditions to perform duplicate laboratory measurements, while also the number of duplicate soil samples was limited. With this limitation, I therefore assumed constant measurement error variances for each of the soil properties. This is probably one of the reasons behind the non-significant differences in the DSM outputs between the case where uncertainty was accounted for and the case where it was ignored (**Chapter 3** and **Chapter 4**). It might well be that the methodology applied to quantify soil measurement errors in this research could have yielded different results, such as bigger impacts on the uncertainty of DSM outputs, if measurement errors were taken to be proportional to the measured values in the laboratory rather than considering constant measurement errors. Methodology development for the quantification of measurement errors in future research, particularly in analytical data, can further be improved if laboratories pay more attention and systematically quantify and publish the uncertainties associated with their measurements and benchmark these against standards, to minimise bias. This of course has time and financial implications as more samples will need to be collected from the field, which translates into an increase in the cost of fieldwork and analysis in the laboratory.

Soil forming factors and covariates

The DSM framework is based on the idea that unknown soil properties can be represented by environmental factors (covariates). Therefore, the covariate layers should be selected to meet all SCORPAN soil forming factors (i.e., soil, climate, organisms, topography, parent materials, age and space) (McBratney *et al.*, 2003) and to represent the complex processes behind the soil attribute of interest (Broeg *et al.*, 2023). Then, the quantitative relationship between a measured soil property and relevant covariates is capitalised to predict values over the area of interest and to create soil property maps. In this research, I used up to 191 resampled environmental variables at 250 m resolution as covariates for the implementation of DSM approaches. Though the spatial resolution of the covariates might appear coarse, and not able to express well the variation of soil properties at small scales, the findings also revealed that not all SCORPAN soil forming factors were well represented in the selected set of covariates used to predict each of the soil property. It is also possible that during covariate selection, some of the important factors of the SCORPAN model were inadvertently removed. Taking this into account, covariate selection in further research should carefully ensure that the set of covariates selected to predict a soil property contains all soil forming factors. In addition, the

acquisition of covariates should strike a balance between the various soil forming factors to better ensure that a larger portion of the spatial variability is captured across the study area.

Non-consideration of other sources of uncertainty

The DSM and crop modelling processes involve multiple sources of uncertainty that collectively contribute to the uncertainty in model outputs. These uncertainties may arise from inaccuracies in input data, limitations in model algorithms, environmental variability, and uncertainties associated with scaling from point-based measurements to larger spatial extents. Additionally, human error in data collection, imprecise model parameterization, and the dynamic nature of agricultural systems further contribute to the complexity of uncertainty quantification and propagation. Identifying and addressing these sources of uncertainty is essential for a comprehensive understanding of the uncertainty in DSM and crop modelling outputs. In **Chapter 3** and **Chapter 4** of this research, I focused on the uncertainty in soil measurements in the laboratory, while in **Chapter 5**, only soil input uncertainties were considered, ignoring uncertainty in other inputs, and ignoring QUEFTS model structural and parameter uncertainties. In a broader context, it is also crucial to broaden the incorporation of uncertainty to cover other soil properties, beyond those investigated in this thesis, that play significant roles in agricultural production. This expanded perspective should consider properties such micronutrients, water holding capacity, and erodibility for a more comprehensive understanding. Similarly, it is essential to incorporate a temporal component of soil properties, a consideration that is gaining attention in DSM work. The exploration of how uncertainty can be effectively managed within this temporal framework is also of paramount importance. In a practical sense, uncertainty in model outputs is the combination of uncertainties from various sources. Ignoring one of the sources without prior investigation that its contribution is insignificant may lead to underestimation of the model output uncertainty. To enhance future studies, it is crucial for both modellers and end-users to conscientiously account for all sources of uncertainty to evaluate and understand their combined effects on decision outcomes, so as to direct efforts to reducing the errors where it matters most (Glynn *et al.*, 2022). Taking measures to improve the overall uncertainties of model outputs is essential from both a modelling and practical application standpoint. This approach ensures a more comprehensive understanding of the variability in model predictions, contributing to more robust and reliable.

Cross-correlation between soil inputs variables

In **Chapter 5**, I ignored the interactions between various soil inputs during the uncertainty propagation analysis. I considered this as a limitation because in many circumstances, most soil properties are cross-correlated (**Chapter 2**). As such, their uncertainty will also be cross-correlated. Thus, when multiple inputs are uncertain as in the case of soil inputs used in crop modelling in **Chapter 5**, cross-correlations between them will influence the degree of uncertainty in the model outputs. Such analysis was beyond the scope of this research, but I believe that including cross-correlation would have yielded a more realistic crop model uncertainty estimation. This observation underscores the complexity of uncertainty analysis. The interaction between different sources of uncertainty may lead to compounded effects, potentially influencing the model outputs more significantly than any single source of uncertainty. Furthermore, the unaccounted uncertainty in the relationship between soil nutrient supply capacity and soil properties adds an additional layer of complexity to the overall uncertainty framework. Future studies could benefit from a more comprehensive exploration of these interactions and considerations, allowing for a nuanced understanding of the intricate

factors influencing model outcomes in the context of soil nutrient dynamics. A viable solution could involve adopting a methodology capable of modelling multiple correlated soil properties.

In summary, this PhD research underscored the importance of accurate soil information on enhancing land management interventions and promoting sustainable agricultural intensification in not only in Cameroon, but in West and Central Africa countries in general. It also showed that it is important to quantify and communicate uncertainty, as well as account for it in decision making. However, the ultimate goal of this PhD research should not be solely to generate the most accurate soil information, but also to recognize that despite our efforts and available resources, complete elimination of uncertainty is unattainable. Therefore, it is essential to adapt our methods to effectively account for uncertainty.

The DSM approaches developed in this research have practical implications for initiatives led by Center for International Forestry Research - World Agroforestry (CIFOR-ICRAF) and other development organizations West and Central Africa, with a direct bearing on land management, agricultural production, and food security. The generated soil maps serve as valuable tools for identifying land degradation hotspots, determining suitable areas for specific practices like tree planting, soil, and water conservation practices, and evaluating landscape potential for meeting land restoration objectives - all critical elements in ensuring long-term agricultural intensification and food security. Key challenges faced by development projects in West and Central Africa, including insufficient knowledge to design context-specific land restoration strategies and challenges in quantifying uncertainty in available information, directly impact the success of initiatives aimed at improving sustainable agricultural intensification and food security. The extensive knowledge I gained in PSS, spatial analysis, pedometric, and DSM approaches throughout this PhD research not only addressed these challenges but also strengthened technical and knowledge sharing collaboration between CIFOR-ICRAF, Wageningen University and Research, and the International Soil Reference and Information Centre (ISRIC) in their endeavours to provide soil information useful in addressing global environmental and societal challenges. Despite delving into fundamental and theoretical aspects of statistical methods and spatial modelling, the research ultimately produced results and products that are highly relevant to applied research and concrete projects, emphasizing the practical significance of the research outcomes for achieving sustainable agricultural intensification in West and Central Africa.

References

AbdelRahman, M.A.E., Metwaly, M.M., 2023. Digital soil characteristics mapping for aiding site-specific management practices in the West Nile Delta, Egypt. Discover Sustainability 4, 47.

AbdelRahman, M.A.E., Saleh, A.M., Arafat, S.M., 2022. Assessment of land suitability using a soil-indicator-based approach in a geomatics environment. Scientific Reports 12, 18113.

Amirinejad, A.A., Kamble, K., Aggarwal, P., Chakraborty, D., Pradhan, S., Mittal, R.B., 2011. Assessment and mapping of spatial variation of soil physical health in a farm. Geoderma 160, 292-303.

Angelini, M.E., Kempen, B., Heuvelink, G.B.M., Temme, A.J.A.M., Ransom, M.D., 2020. Extrapolation of a structural equation model for digital soil mapping. Geoderma 367, 114226.

Arrouays, D., Grundy, M.G., Hartemink, A.E., Hempel, J.W., Heuvelink, G.B.M., Hong, S.Y., Lagacherie, P., Lelyk, G., McBratney, A.B., McKenzie, N.J., Mendonca-Santos, M.d.L., Minasny, B., Montanarella, L., Odeh, I.O.A., Sanchez, P.A., Thompson, J.A., Zhang, G.-L., 2014. Chapter Three - GlobalSoilMap: Toward a Fine-Resolution Global Grid of Soil Properties. In: Sparks, D.L. (Ed.), Advances in Agronomy. Academic Press, pp. 93-134.

Arrouays, D., Lagacherie, P., Hartemink, A.E., 2017. Digital soil mapping across the globe. Geoderma Regional 9, 1-4.

Arrouays, D., Poggio, L., Salazar Guerrero, O.A., Mulder, V.L., 2020. Digital soil mapping and GlobalSoilMap. Main advances and ways forward. Geoderma Regional 21, e00265.

Awal, R., Safeeq, M., Abbas, F., Fares, S., Deb, S.K., Ahmad, A., Fares, A., 2019. Soil Physical Properties Spatial Variability under Long-Term No-Tillage Corn. Agronomy 9, 750.

Bai, Z., Li, H., Yang, X., Zhou, B., Shi, X., Wang, B., Li, D., Shen, J., Chen, Q., Qin, W., Oenema, O., Zhang, F., 2013. The critical soil P levels for crop yield, soil fertility and environmental safety in different soil types. Plant and Soil 372, 27-37.

Ballabio, C., Panagos, P., Monatanarella, L., 2016. Mapping topsoil physical properties at European scale using the LUCAS database. Geoderma 261, 110-123.

Batjes, N.H., Ribeiro, E., van Oostrum, A., Leenaars, J., Hengl, T., Mendes de Jesus, J., 2017. WoSIS: providing standardised soil profile data for the world. Earth Syst. Sci. Data 9, 1-14.

Begho, T., Eory, V., Glenk, K., 2022. Demystifying risk attitudes and fertilizer use: A review focusing on the behavioral factors associated with agricultural nitrogen emissions in South Asia. Frontiers in Sustainable Food Systems 6.

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

Berhongaray, G., Alvarez, R., De Paepe, J., Caride, C., Cantet, R., 2013. Land use effects on soil carbon in the Argentine Pampas. Geoderma 192, 97-110.

Biau, G., Scornet, E., 2016. A random forest guided tour. TEST 25, 197-227.

Bishop, T.F.A., Horta, A., Karunaratne, S.B., 2015. Validation of digital soil maps at different spatial supports. Geoderma 241-242, 238-249.

Booker, D.J., Whitehead, A.L., 2018. Inside or Outside: Quantifying Extrapolation Across River Networks. Water Resources Research 54, 6983-7003.

Bouma, J., 1997. Precision agriculture: introduction to the spatial and temporal variability of environmental quality. Ciba Foundation symposium 210, 5-13; discussion 14-17, 68-78.

Breiman, L., 2001. Random Forests. Machine Learning 45, 5-32.

Breure, M.S., Kempen, B., Hoffland, E., 2022a. Spatial predictions of maize yields using QUEFTS – A comparison of methods. Geoderma 425, 116018.

Breure, T.S., Haefele, S.M., Hannam, J.A., Corstanje, R., Webster, R., Moreno-Rojas, S., Milne, A.E., 2022b. A loss function to evaluate agricultural decision-making under uncertainty: a case study of soil spectroscopy. Precision Agriculture.

Brevik, E.C., Calzolari, C., Miller, B.A., Pereira, P., Kabala, C., Baumgarten, A., Jordán, A., 2016. Soil mapping, classification, and pedologic modeling: History and future directions. Geoderma 264, 256-274.

Brevik, E.C., Hartemink, A.E., 2010. Early soil knowledge and the birth and development of soil science. CATENA 83, 23-33.

Brevik, E.C., Steffan, J.J., Rodrigo-Comino, J., Neubert, D., Burgess, L.C., Cerdà, A., 2019. Connecting the public with soil to improve human health. European Journal of Soil Science 70, 898-910.

Brodley, C.E., Friedl, M.A., 1999. Identifying mislabeled training data. J. Artif. Int. Res. 11, 131–167.

Broeg, T., Blaschek, M., Seitz, S., Taghizadeh-Mehrjardi, R., Zepp, S., Scholten, T., 2023. Transferability of Covariates to Predict Soil Organic Carbon in Cropland Soils. Remote Sensing 15, 876.

Brown, J.D., Heuvelink, G.B.M., 2005. Assessing Uncertainty Propagation through Physically Based Models of Soil Water Flow and Solute Transport. Encyclopedia of Hydrological Sciences.

Brus, D.J., 2019. Sampling for digital soil mapping: A tutorial supported by R scripts. Geoderma 338, 464-480.

Brus, D.J., Bogaert, P., Heuvelink, G.B.M., 2008. Bayesian Maximum Entropy prediction of soil categories using a traditional soil map as soft information. European Journal of Soil Science 59, 166-177.
Brus, D.J., Heuvelink, G.B.M., 2007. Optimization of sample patterns for universal kriging of environmental variables. Geoderma 138, 86-95.
Brus, D.J., Kempen, B., Heuvelink, G.B.M., 2011. Sampling for validation of digital soil maps. European Journal of Soil Science 62, 394-407.
Buenemann, M., Coetzee, M.E., Kutuahupira, J., Maynard, J.J., Herrick, J.E., 2023. Errors in soil maps: The need for better on-site estimates and soil map predictions. PLoS One 18, e0270176.
Bulmer, C., Paré, D., Domke, G.M., 2019. Chapter 14 - A new era of digital soil mapping across forested landscapes. In: Busse, M., Giardina, C.P., Morris, D.M., Page-Dumroese, D.S. (Eds.), Developments in Soil Science. Elsevier, pp. 345-371.
Bünemann, E.K., Bongiorno, G., Bai, Z., Creamer, R.E., De Deyn, G., de Goede, R., Fleskens, L., Geissen, V., Kuyper, T.W., Mäder, P., Pulleman, M., Sukkel, W., van Groenigen, J.W., Brussaard, L., 2018. Soil quality – A critical review. Soil Biology and Biochemistry 120, 105-125.
Cambardella, C.A., Moorman, T.B., Parkin, T.B., Karlen, D.L., Novak, J.M., Turco, R.F., Konopka, A.E., 1994. Field-Scale Variability of Soil Properties in Central Iowa Soils. Soil Science Society of America Journal 58.
Chagumaira, C., Chimungu, J.G., Gashu, D., Nalivata, P.C., Broadley, M.R., Milne, A.E., Lark, R.M., 2021. Communicating uncertainties in spatial predictions of grain micronutrient concentration. Geosci. Commun. 4, 245-265.
Chagumaira, C., Nalivata, P.C., Chimungu, J.G., Gashu, D., Broadley, M.R., Milne, A.E., Lark, R.M., 2022. Stakeholder interpretation of probabilistic representations of uncertainty in spatial information: an example on the nutritional quality of staple crops. International Journal of Geographical Information Science 36, 2446-2472.
Chapagain, R., Remenyi, T.A., Harris, R.M.B., Mohammed, C.L., Huth, N., Wallach, D., Rezaei, E.E., Ojeda, J.J., 2022. Decomposing crop model uncertainty: A systematic review. Field Crops Research 279, 108448.
Chartin, C., Stevens, A., Goidts, E., Krüger, I., Carnol, M., van Wesemael, B., 2017. Mapping Soil Organic Carbon stocks and estimating uncertainties at the regional scale following a legacy sampling strategy (Southern Belgium, Wallonia). Geoderma Regional 9, 73-86.
Chen, L., Gong, J., Fu, B., Huang, Z., Huang, Y., Gui, L., 2007. Effect of land use conversion on soil organic carbon sequestration in the loess hilly area, loess plateau of China. Ecol Res 22, 641-648.
Chen, S., Arrouays, D., Leatitia Mulder, V., Poggio, L., Minasny, B., Roudier, P., Libohova, Z., Lagacherie, P., Shi, Z., Hannam, J., Meersmans, J., Richer-de-Forges, A.C., Walter, C., 2022a. Digital mapping of GlobalSoilMap soil properties at a broad scale: A review. Geoderma 409, 115567.
Chen, S., Xu, D., Li, S., Ji, W., Yang, M., Zhou, Y., Hu, B., Xu, H., Shi, Z., 2020. Monitoring soil organic carbon in alpine soils using in situ vis-NIR spectroscopy and a multilayer perceptron. Land Degradation & Development 31, 1026-1038.
Chen, S., Zhang, G., Zhu, P., Wang, C., Wan, Y., 2022b. Impact of slope position on soil erodibility indicators in rolling hill regions of northeast China. CATENA 217, 106475.
Chilès, J.P., Delfiner, P., 2012. Geostatistics: Modeling Spatial Uncertainty. John Wiley & Sons, Inc. .
Chilimba, A.D.C., Mughogho, S.K., Wendt, J., 1999. Mehlich 3 or modified olsen for soil testing in Malawi. Communications in Soil Science and Plant Analysis 30, 1231-1250.
Chytrý, M., Danihelka, J., Ermakov, N., Hájek, M., Hájková, P., Kočí, M., Kubešová, S., Lustyk, P., Otýpková, Z., Popov, D., Roleček, J., Řezníčková, M., Šmarda, P., Valachovič, M., 2007. Plant species richness in continental southern Siberia: effects of pH and climate in the context of the species pool hypothesis. Global Ecology and Biogeography 16, 668-678.
Clark, I., 2010. Statistics or geostatistics? Sampling error or nugget effect? Journal of the Southern African Institute of Mining and Metallurgy 110, 307-312.
Côté, L., Brown, S., Paré, D., Fyles, J., Bauhus, J., 2000. Dynamics of carbon and nitrogen mineralization in relation to stand type, stand age and soil texture in the boreal mixedwood. Soil Biology and Biochemistry 32, 1079-1090.
Dangal, S.R.S., Sanderman, J., Wills, S., Ramirez-Lopez, L., 2019. Accurate and Precise Prediction of Soil Properties from a Large Mid-Infrared Spectral Library. Soil Systems 3, 11.
de Bruin, S., Brus, D.J., Heuvelink, G.B.M., van Ebbenhorst Tengbergen, T., Wadoux, A.M.J.C., 2022. Dealing with clustered samples for assessing map accuracy by cross-validation. Ecological Informatics 69, 101665.
Delmelle, E.M., Goovaerts, P., 2009. Second-phase sampling designs for non-stationary spatial variables. Geoderma 153, 205-216.
Dhakal, C., Lange, K., 2021. Crop yield response functions in nutrient application: A review. Agronomy Journal 113, 5222-5234.
Di Virgilio, N., Monti, A., Venturi, G., 2007. Spatial variability of switchgrass (Panicum virgatum L.) yield as related to soil parameters in a small field. Field Crops Research 101, 232-239.
Dokoohaki, H., Kivi, M.S., Martinez-Feria, R., Miguez, F.E., Hoogenboom, G., 2021. A comprehensive

uncertainty quantification of large-scale process-based crop modeling frameworks. Environmental Research Letters 16, 084010.
Dormann, C.F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., Marquéz, J.R.G., Gruber, B., Lafourcade, B., Leitão, P.J., Münkemüller, T., McClean, C., Osborne, P.E., Reineking, B., Schröder, B., Skidmore, A.K., Zurell, D., Lautenbach, S., 2013. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. Ecography 36, 27-46.
Dornik, A., Cheţan, M.A., Drăguţ, L., Iliuţă, A., Dicu, D.D., 2022. Importance of the mapping unit on the land suitability assessment for agriculture. Computers and Electronics in Agriculture 201, 107305.
Durr, C., Constantin, J., Wagner, M.H., Navier, H., Demilly, D., Goertz, S., Nesi, N., 2016. Virtual modeling based on deep phenotyping provides complementary data to field experiments to predict plant emergence in oilseed rape genotypes. European Journal of Agronomy 79, 90-99.
Ebentier, D.L., Hanley, K.T., Cao, Y., Badgley, B.D., Boehm, A.B., Ervin, J.S., Goodwin, K.D., Gourmelon, M., Griffith, J.F., Holden, P.A., Kelty, C.A., Lozach, S., McGee, C., Peed, L.A., Raith, M., Ryu, H., Sadowsky, M.J., Scott, E.A., Domingo, J.S., Schriewer, A., Sinigalliano, C.D., Shanks, O.C., Van De Werfhorst, L.C., Wang, D., Wuertz, S., Jay, J.A., 2013. Evaluation of the repeatability and reproducibility of a suite of qPCR-based microbial source tracking methods. Water Research 47, 6839-6848.
Eijk, D., Janssen, B., Oenema, O., 2006. Initial and residual effects of fertilizer phosphorus on soil phosphorus and maize yields on phosphorus fixing soils. A case study in South-West Kenya. Agriculture, Ecosystems & Environment 116, 104-120.
Ellinger, M., Merbach, I., Werban, U., Ließ, M., 2019. Error propagation in spectrometric functions of soil organic carbon. SOIL 5, 275-288.
Ewert, F., Rötter, R.P., Bindi, M., Webber, H., Trnka, M., Kersebaum, K.C., Olesen, J.E., van Ittersum, M.K., Janssen, S., Rivington, M., Semenov, M.A., Wallach, D., Porter, J.R., Stewart, D., Verhagen, J., Gaiser, T., Palosuo, T., Tao, F., Nendel, C., Roggero, P.P., Bartošová, L., Asseng, S., 2015. Crop modelling for integrated assessment of risk to food production from climate change. Environmental Modelling & Software 72, 287-303.
Ezui, K.S., Franke, A.C., Ahiabor, B.D.K., Tetteh, F.M., Sogbedji, J., Janssen, B.H., Mando, A., Giller, K.E., 2017. Understanding cassava yield response to soil and fertilizer nutrient supply in West Africa. Plant and Soil 420, 331-347.
Faivre, R., Leenhardt, D., Voltz, M., Benoît, M., Papy, F., Dedieu, G., Wallach, D., 2004. Spatialising crop models. Agronomie 24, 205-217.
FAOSTAT, 2020. Statistical Database. Food and Agriculture Organization of the United Nations, Rome.
Feng, S., Zhao, W., Zhan, T., Yan, Y., Pereira, P., 2022. Land degradation neutrality: A review of progress and perspectives. Ecological Indicators 144, 109530.
Forkuor, G., Hounkpatin, O.K.L., Welp, G., Thiel, M., 2017. High Resolution Mapping of Soil Properties Using Remote Sensing Variables in South-Western Burkina Faso: A Comparison of Machine Learning and Multiple Linear Regression Models. PLOS ONE 12, e0170478.
Fouedjio, F., 2020. Exact Conditioning of Regression Random Forest for Spatial Prediction. Artificial Intelligence in Geosciences 1, 11-23.
Fox, E.W., Ver Hoef, J.M., Olsen, A.R., 2020. Comparing spatial regression to random forests for large environmental data sets. PLOS ONE 15, e0229509.
Frénay, B., Doquire, G., Verleysen, M., 2014. Estimating mutual information for feature selection in the presence of label noise. Computational Statistics & Data Analysis 71, 832-848.
Fu, W., Tunney, H., Zhang, C., 2010. Spatial variation of soil nutrients in a dairy farm and its implications for site-specific fertilizer application. Soil and Tillage Research 106, 185-193.
Gao, X., Song, P.X.K., 2010. Composite likelihood Bayesian Information Criteria for model selection in high-dimensional data. Journal of the American Statistical Association 105, 1531-1540.
Getson, J.M., Church, S.P., Radulski, B.G., Sjöstrand, A.E., Lu, J., Prokopy, L.S., 2022. Understanding scientists' communication challenges at the intersection of climate and agriculture. PLOS ONE 17, e0269927.
Giller, K.E., Delaune, T., Silva, J.V., van Wijk, M., Hammond, J., Descheemaeker, K., van de Ven, G., Schut, A.G.T., Taulya, G., Chikowo, R., Andersson, J.A., 2021. Small farms and development in sub-Saharan Africa: Farming for food, for income or for lack of better options? Food Security 13, 1431-1454.
Giller, K.E., Witter, E., Corbeels, M., Tittonell, P., 2009. Conservation agriculture and smallholder farming in Africa: The heretics' view. Field Crops Research 114, 23-34.
Ginaldi, F., Bajocco, S., Bregaglio, S., Cappelli, G., 2019. Spatializing Crop Models for Sustainable Agriculture. In: Farooq, M., Pisante, M. (Eds.), Innovations in Sustainable Agriculture. Springer International Publishing, Cham, pp. 599-619.
GLOSOLAN, 2023. The Global Soil Laboratory Network, https://www.fao.org/global-soil-partnership/glosolan/en.
Glynn, P.D., Rhodes, C.R., Chiavacci, S.J., Helgeson, J.F., Shapiro, C.D., Straub, C.L., 2022. Value of Information and Decision Pathways: Concepts and Case Studies. Frontiers in Environmental Science 10.
Gnacadja, L., Vidal, A., 2023. How can science help to implement the UN Decade on Ecosystem Restoration

2021–2030? Philosophical Transactions of the Royal Society B: Biological Sciences 378, 20210066.
Godswill, A.A., Bernard, P.K.Y., Aaron, S.T., 2016. Spatial variability of selected physico-chemical properties of soils under vegetable cultivation in urban and peri-urban wetland gardens of Bamenda municipality, Cameroon. African Journal of Agricultural Research 11, 74-86.
Gomes, L.C., Faria, R.M., de Souza, E., Veloso, G.V., Schaefer, C.E.G.R., Filho, E.I.F., 2019. Modelling and mapping soil organic carbon stocks in Brazil. Geoderma 340, 337-350.
Goovaerts, P., 1997. Geostatistics for natural resources evaluation. Oxford University Press, USA.
Goovaerts, P., 2001. Geostatistical modelling of uncertainty in soil science. Geoderma 103, 3-26.
Goovaerts, P., Journel, A.G., 1995. Integrating soil map information in modelling the spatial variation of continuous soil properties. European Journal of Soil Science 46, 397-414.
Grego, C.R., Vieira, S.R., Lourenção, A.L., 2006. Spatial distribution of Pseudaletia sequax Franclemlont in triticale under no-till management. Scientia Agricola 63, 321-327.
Grundy, M.J., Rossel, R.A.V., Searle, R.D., Wilson, P.L., Chen, C., Gregory, L.J., 2015. Soil and Landscape Grid of Australia. Soil Research 53, 835-844.
Guan, F., Tang, X., Fan, S., Zhao, J., Peng, C., 2015. Changes in soil carbon and nitrogen stocks followed the conversion from secondary forest to Chinese fir and Moso bamboo plantations. CATENA 133, 455-460.
Guedes Filho, O., Vieira, S.R., Chiba, M.K., Nagumo, C.H., Dechen, S.C.F., 2010. Spatial and temporal variability of crop yield and some Rhodic Hapludox properties under no-tillage. Revista Brasileira de Ciência do Solo 34, 1-14.
Guevara, M., Olmedo, G.F., Stell, E., Yigini, Y., Aguilar Duarte, Y., Arellano Hernández, C., Arévalo, G.E., Arroyo-Cruz, C.E., Bolivar, A., Bunning, S., Bustamante Cañas, N., Cruz-Gaistardo, C.O., Davila, F., Dell Acqua, M., Encina, A., Figueredo Tacona, H., Fontes, F., Hernández Herrera, J.A., Ibelles Navarro, A.R., Loayza, V., Manueles, A.M., Mendoza Jara, F., Olivera, C., Osorio Hermosilla, R., Pereira, G., Prieto, P., Ramos, I.A., Rey Brina, J.C., Rivera, R., Rodríguez-Rodríguez, J., Roopnarine, R., Rosales Ibarra, A., Rosales Riveiro, K.A., Schulz, G.A., Spence, A., Vasques, G.M., Vargas, R.R., Vargas, R., 2018. No silver bullet for digital soil mapping: country-specific soil organic carbon estimates across Latin America. SOIL 4, 173-193.
Han, G., Hao, X., Zhao, M., Wang, M., Ellert, B.H., Willms, W., Wang, M., 2008. Effect of grazing intensity on carbon and nitrogen in soil and vegetation in a meadow steppe in Inner Mongolia. Agriculture, Ecosystems & Environment 125, 21-32.
Hanchuan, P., Fuhui, L., Ding, C., 2005. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on Pattern Analysis and Machine Intelligence 27, 1226-1238.
Hansen, M.C., Potapov, P.V., Moore, R., Hancher, M., Turubanova, S.A., Tyukavina, A., Thau, D., Stehman, S.V., Goetz, S.J., Loveland, T.R., Kommareddy, A., Egorov, A., Chini, L., Justice, C.O., Townshend, J.R.G., 2013. High-resolution global maps of 21st-century forest cover change. Science 342, 850-853.
Haregeweyn, N., Poesen, J., Deckers, J., Nyssen, J., Haile, M., Govers, G., Verstraeten, G., Moeyersons, J., 2008. Sediment-bound nutrient export from micro-dam catchments in Northern Ethiopia. Land Degradation & Development 19, 136-152.
Hateffard, F., Steinbuch, L., Heuvelink, G.B.M., 2024. Evaluating the extrapolation potential of random forest digital soil mapping. Geoderma 441, 116740.
Hendriks, C.M.J., Stoorvogel, J.J., Álvarez-Martínez, J.M., Claessens, L., Pérez-Silos, I., Barquín, J., 2021. Introducing a mechanistic model in digital soil mapping to predict soil organic matter stocks in the Cantabrian region (Spain). European Journal of Soil Science 72, 704– 719.
Hengl, T., Heuvelink, G.B.M., Kempen, B., Leenaars, J.G.B., Walsh, M.G., Shepherd, K.D., Sila, A., MacMillan, R.A., Mendes de Jesus, J., Tamene, L., Tondoh, J.E., 2015. Mapping Soil Properties of Africa at 250 m Resolution: Random Forests Significantly Improve Current Predictions. PLOS ONE 10, e0125814.
Hengl, T., Heuvelink, G.B.M., Rossiter, D.G., 2007. About regression-kriging: From equations to case studies. Computers & Geosciences 33, 1301-1315.
Hengl, T., Miller, M.A.E., Križan, J., Shepherd, K.D., Sila, A., Kilibarda, M., Antonijević, O., Glušica, L., Dobermann, A., Haefele, S.M., McGrath, S.P., Acquah, G.E., Collinson, J., Parente, L., Sheykhmousa, M., Saito, K., Johnson, J.-M., Chamberlin, J., Silatsa, F.B.T., Yemefack, M., Wendt, J., MacMillan, R.A., Wheeler, I., Crouch, J., 2021. African soil properties and nutrients mapped at 30 m spatial resolution using two-scale ensemble machine learning. Scientific Reports 11, 6130.
Hengl, T., Nussbaum, M., Wright, M.N., Heuvelink, G.B.M., Gräler, B., 2018a. Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. PeerJ 6, e5518.
Hengl, T., Walsh, M.G., Sanderman, J., Wheeler, I., Harrison, S.P., Prentice, I.C., 2018b. Global mapping of potential natural vegetation: an assessment of machine learning algorithms for estimating land potential. PeerJ 6, e5457-e5457.
Heuvelink, G.B.M., 1998. Uncertainty analysis in environmental modelling under a change of spatial scale. In: Finke, P.A., Bouma, J., Hoosbeek, M.R. (Eds.), Soil and Water Quality at Different Scales: Proceedings of the Workshop "Soil and Water Quality at Different Scales" held 7–9 August 1996,

Wageningen, The Netherlands. Springer Netherlands, Dordrecht, pp. 255-264.
Heuvelink, G.B.M., 2014. Uncertainty quantification of GlobalSoilMap products. In: D. Arrouays, N. McKenzie, J. Hempel, A. Richer de Forges, McBratney, A. (Eds.), GlobalSoilMap. Basis of the Global Spatial Soil Information System, pp. 335 - 340.
Heuvelink, G.B.M., 2018. Uncertainty and Uncertainty Propagation in Soil Mapping and Modelling. In: McBratney, A.B., Minasny, B., Stockmann, U. (Eds.), Pedometrics. Springer International Publishing, Cham, pp. 439-461.
Heuvelink, G.B.M., Brown, J.D., van Loon, E.E., 2007. A probabilistic framework for representing and simulating uncertain environmental variables. International Journal of Geographical Information Science 21, 497-513.
Heuvelink, G.B.M., Pebesma, E.J., 1999. Spatial aggregation and soil process modelling. Geoderma 89, 47-65.
Heuvelink, G.B.M., Webster, R., 2001. Modelling soil variation: past, present, and future. Geoderma 100, 269-301.
Heuvelink, G.B.M., Webster, R., 2022. Spatial statistics and soil mapping: A blossoming partnership under pressure. Spatial Statistics, 100639.
Hibbert, D.B., 2007. Systematic errors in analytical measurement results. Journal of Chromatography A 1158, 25-32.
Hijmans, R.J., 2021. raster: Geographic Data Analysis and Modeling. R package version 3.5-2
Hounkpatin, K.O.L., Schmidt, K., Stumpf, F., Forkuor, G., Behrens, T., Scholten, T., Amelung, W., Welp, G., 2018. Predicting reference soil groups using legacy data: A data pruning and Random Forest approach for tropical environment (Dano catchment, Burkina Faso). Sci Rep 8, 9959.
Islam, M.S., Bell, R.W., Miah, M.A.M., Alam, M.J., 2022. Farmers' fertilizer use gaps relative to government recommendations in the saline coastal zone of the Ganges Delta. Agronomy for Sustainable Development 42, 59.
IUSS, W.G.W., 2006. World reference base for soil resources 2006. FAO, Rome.
IUSS, W.G.W., 2015. World reference base for soil resources 2014. FAO, Rome.
Janssen, B.H., Guiking, F.C.T., van der Eijk, D., Smaling, E.M.A., Wolf, J., van Reuler, H., 1990. A system for quantitative evaluation of the fertility of tropical soils (QUEFTS). Geoderma 46, 299-318.
Janssen, P.H.M., Heuberger, P.S.C., 1995. Calibration of process-oriented models. Ecological Modelling 83, 55-66.
Jenny, H., 1941. Factors of Soil Formation, a System of Quantitative Pedology. Agronomy Journal 33, 857-858.
Jin, Y., Ge, Y., Wang, J., Heuvelink, G.B.M., Wang, L., 2018. Geographically Weighted Area-to-Point Regression Kriging for Spatial Downscaling in Remote Sensing. Remote Sensing 10, 579.
Jones, J.W., Antle, J.M., Basso, B., Boote, K.J., Conant, R.T., Foster, I., Godfray, H.C.J., Herrero, M., Howitt, R.E., Janssen, S., Keating, B.A., Munoz-Carpena, R., Porter, C.H., Rosenzweig, C., Wheeler, T.R., 2017a. Brief history of agricultural systems modeling. Agricultural Systems 155, 240-254.
Jones, J.W., Antle, J.M., Basso, B., Boote, K.J., Conant, R.T., Foster, I., Godfray, H.C.J., Herrero, M., Howitt, R.E., Janssen, S., Keating, B.A., Munoz-Carpena, R., Porter, C.H., Rosenzweig, C., Wheeler, T.R., 2017b. Toward a new generation of agricultural system data, models, and knowledge products: State of agricultural systems science. Agricultural Systems 155, 269-288.
Jones, J.W., Hoogenboom, G., Porter, C.H., Boote, K.J., Batchelor, W.D., Hunt, L.A., Wilkens, P.W., Singh, U., Gijsman, A.J., Ritchie, J.T., 2003. The DSSAT cropping system model. European Journal of Agronomy 18, 235-265.
Kadiyala, M.D.M., Nedumaran, S., Singh, P., S, C., Irshad, M.A., Bantilan, M.C.S., 2015. An integrated crop model and GIS decision support system for assisting agronomic decision making under climate change. Science of The Total Environment 521-522, 123-134.
Kahle, M., Kleber, M., Jahn, R., 2002. Predicting carbon content in illitic clay fractions from surface area, cation exchange capacity and dithionite-extractable iron. European Journal of Soil Science 53, 639-644.
Kansiime, M.K., Njunge, R., Okuku, I., Baars, E., Alokit, C., Duah, S., Gakuo, S., Karanja, L., McHana, A., Mibei, H., Musebe, R., Romney, D., Rware, H., Silvestri, S., Sones, D., Watiti, J., 2022. Bringing sustainable agricultural intensification practices and technologies to scale through campaign-based extension approaches: lessons from Africa Soil Health Consortium. International Journal of Agricultural Sustainability 20, 743-757.
Keating, B.A., Carberry, P.S., Hammer, G.L., Probert, M.E., Robertson, M.J., Holzworth, D., Huth, N.I., Hargreaves, J.N.G., Meinke, H., Hochman, Z., McLean, G., Verburg, K., Snow, V., Dimes, J.P., Silburn, M., Wang, E., Brown, S., Bristow, K.L., Asseng, S., Chapman, S., McCown, R.L., Freebairn, D.M., Smith, C.J., 2003. An overview of APSIM, a model designed for farming systems simulation. European Journal of Agronomy 18, 267-288.
Keesstra, S.D., Bouma, J., Wallinga, J., Tittonell, P., Smith, P., Cerdà, A., Montanarella, L., Quinton, J.N., Pachepsky, Y., van der Putten, W.H., Bardgett, R.D., Moolenaar, S., Mol, G., Jansen, B., Fresco, L.O., 2016. The significance of soils and soil science towards realization of the United Nations Sustainable

Development Goals. SOIL 2, 111-128.
Kempen, B., Brus, D.J., Stoorvogel, J.J., Heuvelink, G.B.M., de Vries, F., 2012. Efficiency Comparison of Conventional and Digital Soil Mapping for Updating Soil Maps. Soil Science Society of America Journal 76, 2097-2115.
Kephe, P.N., Ayisi, K.K., Petja, B.M., 2021. Challenges and opportunities in crop simulation modelling under seasonal and projected climate change scenarios for crop production in South Africa. Agriculture & Food Security 10, 10.
Keskin, H., Grunwald, S., 2018. Regression kriging as a workhorse in the digital soil mapper's toolbox. Geoderma 326, 22-41.
Kihara, J., Nziguheba, G., Zingore, S., Coulibaly, A., Esilaba, A., Kabambe, V., Njoroge, S., Palm, C., Huising, J., 2016. Understanding variability in crop response to fertilizer and amendments in sub-Saharan Africa. Agriculture, Ecosystems & Environment 229, 1-12.
Knotters, M., Brus, D.J., Oude Voshaar, J.H., 1995. A comparison of kriging, co-kriging and kriging combined with regression for spatial interpolation of horizon depth with censored observations. Geoderma 67, 227-246.
Kopittke, P.M., Menzies, N.W., Wang, P., McKenna, B.A., Lombi, E., 2019. Soil and the intensification of agriculture for global food security. Environment International 132, 105078.
Kuhn, M., 2008. Building Predictive Models in R Using the caret Package. Journal of Statistical Software 28, 1 - 26.
Kumar, Y., Subramanyam, G., Mujumdar, M., 2023. Chapter 15 - Spatializing crop models for sustainable agriculture and environment. In: Farooq, M., Gogoi, N., Pisante, M. (Eds.), Sustainable Agriculture and the Environment. Academic Press, pp. 421-436.
Lagacherie, P., Arrouays, D., Bourennane, H., Gomez, C., Nkuba-Kasanda, L., 2020. Analysing the impact of soil spatial sampling on the performances of Digital Soil Mapping models and their evaluation: A numerical experiment on Quantile Random Forest using clay contents obtained from Vis-NIR-SWIR hyperspectral imagery. Geoderma 375, 114503.
Lagacherie, P., Buis, S., Constantin, J., Dharumarajan, S., Ruiz, L., Sekhar, M., 2022. Evaluating the impact of using digital soil mapping products as input for spatializing a crop model: The case of drainage and maize yield simulated by STICS in the Berambadi catchment (India). Geoderma 406, 115503.
Lamichhane, S., Kumar, L., Wilson, B., 2019. Digital soil mapping algorithms and covariates for soil organic carbon mapping and their implications: A review. Geoderma 352, 395-413.
Lark, R.M., Chagumaira, C., Milne, A.E., 2022. Decisions, uncertainty and spatial information. Spatial Statistics 50, 100619.
Lark, R.M., Cullis, B.R., 2004. Model-based analysis using REML for inference from systematically sampled data on soil. European Journal of Soil Science 55, 799-813.
Lark, R.M., Cullis, B.R., Welham, S.J., 2006. On spatial prediction of soil properties in the presence of a spatial trend: the empirical best linear unbiased predictor (E-BLUP) with REML. European Journal of Soil Science 57, 787-799.
Lark, R.M., Lapworth, D.J., 2012. Quality measures for soil surveys by lognormal kriging. Geoderma 173-174, 231-240.
Laslet, G.M., McBratney, A.B., 1990. Estimation and implications of instrumental drift, random measurement error and nugget variance of soil attributes—a case study for soil pH. European Journal of Soil Science 41, 451-471.
Laurent, A.G., 1963. The Lognormal Distribution and the Translation Method: Description and Estimation Problems. Journal of the American Statistical Association 58, 231-235.
Leenaars, J.G.B., Gonzalez, M.R., Kempen, B., 2017. Extrapolation of fertilizer nutrient recommendations for major food crops in West Africa; a proof of concept (with dataset). Project report for IFDC, USAID - West Africa Fertilizer Program, Accra. . ISRIC - World Soil Information, Wageningen, the Netherlands, Accra.
Leenaars, J.G.B., Oostrum, A.v., Gonzalez, M.R., 2014. Africa Soil Profiles Database, Version 1.2. A compilation of georeferenced and standardised legacy soil profile data for Sub-Saharan Africa (with dataset).
Lei, Z., Yu, D., Zhou, F., Zhang, Y., Yu, D., Zhou, Y., Han, Y., 2019. Changes in soil organic carbon and its influencing factors in the growth of Pinus sylvestris var. mongolica plantation in Horqin Sandy Land, Northeast China. Scientific Reports 9, 16453.
Li, C., Ashlock, J.C., Wang, X., 2019. Quantifying repeatability reproducibility sources of error and capacity of a measurement: demonstrated using laboratory soil plasticity tests. Advances in Civil Engineering 2019, 4539549.
Li, C., Hao, X., Zhao, M., Han, G., Willms, W.D., 2008. Influence of historic sheep grazing on vegetation and soil properties of a Desert Steppe in Inner Mongolia. Agriculture, Ecosystems & Environment 128, 109-116.
Li, T., Hasegawa, T., Yin, X., Zhu, Y., Boote, K., Adam, M., Bregaglio, S., Buis, S., Confalonieri, R., Fumoto, T., Gaydon, D., Marcaida III, M., Nakagawa, H., Oriol, P., Ruane, A.C., Ruget, F., Singh, B.-., Singh, U., Tang, L., Tao, F., Wilkens, P., Yoshida, H., Zhang, Z., Bouman, B., 2015. Uncertainties in predicting

rice yield by current crop models under a wide range of climatic conditions. Global Change Biology 21, 1328-1341.
Liang, H., Hu, K., Batchelor, W.D., Qi, Z., Li, B., 2016. An integrated soil-crop system model for water and nitrogen management in North China. Scientific Reports 6, 25755.
Libohova, Z., Seybold, C., Adhikari, K., Wills, S., Beaudette, D., Peaslee, S., Lindbo, D., Owens, P.R., 2019. The anatomy of uncertainty for soil pH measurements and predictions: Implications for modellers and practitioners. European Journal of Soil Science 70, 185-199.
Liu, J., Li, Y.P., Huang, G.H., Zhuang, X.W., Fu, H.Y., 2017. Assessment of uncertainty effects on crop planning and irrigation water supply using a Monte Carlo simulation based dual-interval stochastic programming method. Journal of Cleaner Production 149, 945-967.
Liu, Y., Lv, J., Zhang, B., Bi, J., 2013. Spatial multi-scale variability of soil nutrients in relation to environmental factors in a typical agricultural region, Eastern China. Science of The Total Environment 450–451, 108-119.
Loiseau, T., Arrouays, D., Richer-de-Forges, A.C., Lagacherie, P., Ducommun, C., Minasny, B., 2021. Density of soil observations in digital soil mapping: A study in the Mayenne region, France. Geoderma Regional 24, e00358.
Lopes, M.E., 2019. Estimating the algorithmic variance of randomized ensembles via the bootstrap. The Annals of Statistics 47, 1088-1112, 1025.
Loveland, P., Webb, J., 2003. Is there a critical level of organic matter in the agricultural soils of temperate regions: a review. Soil and Tillage Research 70, 1-18.
Ma, T., Brus, D.J., Zhu, A.X., Zhang, L., Scholten, T., 2020. Comparison of conditioned Latin hypercube and feature space coverage sampling for predicting soil classes using simulation from soil maps. Geoderma 370, 114366.
Makungwe, M., Chabala, L.M., Chishala, B.H., Lark, R.M., 2021. Performance of linear mixed models and random forests for spatial prediction of soil pH. Geoderma 397, 115079.
Maleki, S., Khormali, F., Mohammadi, J., Bogaert, P., Bagheri Bodaghabadi, M., 2020. Effect of the accuracy of topographic data on improving digital soil mapping predictions with limited soil data: An application to the Iranian loess plateau. CATENA 195, 104810.
Malone, B.P., Kidd, D.B., Minasny, B., McBratney, A.B., 2015. Taking account of uncertainties in digital land suitability assessment. PeerJ 3, e1366.
Malone, B.P., McBratney, A.B., Minasny, B., 2011. Empirical estimates of uncertainty for mapping continuous depth functions of soil attributes. Geoderma 160, 614-626.
Mandal, D., Chandrakala, M., Alam, N.M., Roy, T., Mandal, U., 2021. Assessment of soil quality and productivity in different phases of soil erosion with the focus on land degradation neutrality in tropical humid region of India. CATENA 204, 105440.
Marchant, B., Saby, N., Lark, R., Bellamy, P., Jolivet, C., Arrouays, D., 2010. Robust analysis of soil properties at the national scale: cadmium content of French soils. European journal of soil science 61, 144-152.
Mariano, C., Mónica, B., 2021. A random forest-based algorithm for data-intensive spatial interpolation in crop yield mapping. Computers and Electronics in Agriculture 184, 106094.
McBratney, A., Field, D.J., Koch, A., 2014. The dimensions of soil security. Geoderma 213, 203-213.
McBratney, A.B., Mendonça Santos, M.L., Minasny, B., 2003. On digital soil mapping. Geoderma 117, 3-52.
McBratney, A.B., Minasny, B., Viscarra Rossel, R., 2006. Spectral soil analysis and inference systems: A powerful combination for solving the soil data crisis. Geoderma 136, 272-278.
McCabe, R., Kont, M.D., Schmit, N., Whittaker, C., Løchen, A., Walker, P.G.T., Ghani, A.C., Ferguson, N.M., White, P.J., Donnelly, C.A., Watson, O.J., 2021. Communicating uncertainty in epidemic models. Epidemics 37, 100520.
McCuen, R.H., Knight, Z., Cutter, A.G., 2006. Evaluation of the Nash–Sutcliffe Efficiency Index. Journal of Hydrologic Engineering 11, 597-602.
McGrath, D.A., Smith, C.K., Gholz, H.L., Oliveira, F.d.A., 2001. Effects of Land-Use Change on Soil Nutrient Dynamics in Amazônia. Ecosystems 4, 625-645.
McKenzie, B.M., 2012. Spatial and temporal variability in soil physical conditions for root growth: interactions between soil, water and root growth for sustainable agriculture. Agrociencia 16, 208-213.
McLauchlan, K.K., Hobbie, S.E., Post, W.M., 2006. Conversion from agriculture to grassland builds soil organic matter on decadal timescales. Ecological applications : a publication of the Ecological Society of America 16, 143-153.
Meersmans, J., Martin, M., De Ridder, F., Lacarce, E., Wetterlind, J., De Baets, S., Le Bas, C., Louis, B., Orton, T., Bispo, A., Arrouays, D., 2012. A novel soil organic C model using climate, soil type and management data at the national scale in France. Agronomy for Sustainable Development 32, 873-888.
Meersmans, J., van Wesemael, B., Goidts, E., Van Molle, M., De Baets, S., De Ridder, F., 2011. Spatial analysis of soil organic carbon evolution in Belgian croplands and grasslands, 1960–2006. Global

Change Biology 17, 466-479.
Meinshausen, N., 2006. Quantile Regression Forests. Journal of Machine Learning Research 7, 983-999.
Mesfin, S., Haile, M., Gebresamuel, G., Zenebe, A., Gebre, A., 2021. Establishment and validation of site specific fertilizer recommendation for increased barley (Hordeum spp.) yield, northern Ethiopia. Heliyon 7, e07758-e07758.
Meyer, H., Pebesma, E., 2021. Predicting into unknown space? Estimating the area of applicability of spatial prediction models. Methods in Ecology and Evolution 12, 1620-1633.
Meyer, H., Reudenbach, C., Hengl, T., Katurji, M., Nauss, T., 2018. Improving performance of spatio-temporal machine learning models using forward feature selection and target-oriented validation. Environmental Modelling & Software 101, 1-9.
Miller, M.A.E., Shepherd, K.D., Kisitu, B., Collinson, J., 2021. iSDAsoil: The first continent-scale soil property map at 30 m resolution provides a soil information revolution for Africa. PLOS Biology 19, e3001441.
Milne, A.E., Glendining, M.J., Lark, R.M., Perryman, S.A., Gordon, T., Whitmore, A.P., 2015. Communicating the uncertainty in estimated greenhouse gas emissions from agriculture. J Environ Manage 160, 139-153.
Minasny, B., McBratney, A.B., 2016. Digital soil mapping: A brief history and some lessons. Geoderma 264, 301-311.
Minasny, B., McBratney, A.B., Malone, B.P., Wheeler, I., 2013. Digital mapping of soil carbon. Adv. Agron 118, 1-47.
Moges, A., Holden, N., 2008. Soil Fertility in Relation to Slope Position and Agricultural Land Use: A Case Study of Umbulo Catchment in Southern Ethiopia. Environmental Management 42, 753-763.
Moloney, J.P., Malone, B.P., Karunaratne, S., Stockmann, U., 2023. Leveraging large soil spectral libraries for sensor-agnostic field condition predictions of several agronomically important soil properties. Geoderma 439, 116651.
Mueller, T.G., Hartsock, N.J., Stombaugh, T.S., Shearer, S.A., Cornelius, P.L., Barnhisel, R.I., 2003. Soil Electrical Conductivity Map Variability in Limestone Soils Overlain by Loess. Agronomy Journal 95, 496-507.
Mukhtar, H., Wunderlich, R.F., Petway, J.R., Lin, Y.-P., 2019. Uncertainty Propagation in Crop Adaptation Responses to Climate Change: A Modelling Perspective. Environments 6, 115.
Mulder, V.L., de Bruin, S., Schaepman, M.E., 2013. Representing major soil variability at regional scale by constrained Latin Hypercube Sampling of remote sensing data. International Journal of Applied Earth Observation and Geoinformation 21, 301-310.
Musumba, M., Grabowski, P., Palm, C. and Snapp, S., 2017. Guide for the sustainable intensification assessment framework. . Kansas State University, Kansas, USA.
Nanos, N., Grigoratos, T., Rodríguez Martín, J., Samara, C., 2015. Scale-dependent correlations between soil heavy metals and As around four coal-fired power plants of northern Greece. Stoch Environ Res Risk Assess 29, 1531-1543.
Nash, J.E., Sutcliffe, J.V., 1970. River flow forecasting through conceptual models part I — A discussion of principles. Journal of Hydrology 10, 282-290.
Nassiri Mahallati, M., 2020. Chapter 9 - Advances in modeling saffron growth and development at different scales. In: Koocheki, A., Khajeh-Hosseini, M. (Eds.), Saffron. Woodhead Publishing, pp. 139-167.
Naul, B., Bloom, J.S., Pérez, F., van der Walt, S., 2018. A recurrent neural network for classification of unevenly sampled variable stars. Nature Astronomy 2, 151-155.
Nave, L.E., Vance, E.D., Swanston, C.W., Curtis, P.S., 2009. Impacts of elevated N inputs on north temperate forest soil C storage, C/N, and net N-mineralization. Geoderma 153, 231-240.
Nduwamungu, C., Ziadi, N., Parent, L.-É., Tremblay, G.F., Thuriès, L., 2009. Opportunities for, and limitations of, near infrared reflectance spectroscopy applications in soil analysis: A review. Canadian Journal of Soil Science 89, 531-541.
Neba, A., 1999. Modern Geography of the Republic of Cameroon. Neba Publishers, Bamenda.
Nelson, M.A., Bishop, T.F.A., Triantafilis, J., Odeh, I.O.A., 2011. An error budget for different sources of error in digital soil mapping. European Journal of Soil Science 62, 417-430.
Nettleton, D.F., Orriols-Puig, A., Fornells, A., 2010. A study of the effect of different types of noise on the precision of supervised learning techniques. Artificial Intelligence Review 33, 275-306.
Neyestani, M., Sarmadian, F., Jafari, A., Keshavarzi, A., Sharififar, A., 2021. Digital mapping of soil classes using spatial extrapolation with imbalanced data. Geoderma Regional 26, e00422.
Ng, W., Minasny, B., Jeon, S.H., McBratney, A., 2022. Mid-infrared spectroscopy for accurate measurement of an extensive set of soil properties for assessing soil functions. Soil Security 6, 100043.
Nielsen, D.R., Bouma, J., 1985. Soil spatial variability. Pudoc, Wageningen, The Netherlands.
Nikou, M., Tziachris, P., 2022. Prediction and Uncertainty Capabilities of Quantile Regression Forests in Estimating Spatial Distribution of Soil Organic Matter. ISPRS International Journal of Geo-Information 11, 130.
Nocita, M., Stevens, A., van Wesemael, B., Aitkenhead, M., Bachmann, M., Barthès, B., Ben Dor, E., Brown, D.J., Clairotte, M., Csorba, A., Dardenne, P., Demattê, J.A.M., Genot, V., Guerrero, C., Knadel, M., Montanarella, L., Noon, C., Ramirez-Lopez, L., Robertson, J., Sakai, H., Soriano-Disla, J.M., Shepherd,

K.D., Stenberg, B., Towett, E.K., Vargas, R., Wetterlind, J., 2015. Chapter Four - Soil Spectroscopy: An Alternative to Wet Chemistry for Soil Monitoring. In: Sparks, D.L. (Ed.), Advances in Agronomy. Academic Press, pp. 139-159.
Nol, L., Heuvelink, G.B.M., Veldkamp, A., de Vries, W., Kros, J., 2010. Uncertainty propagation analysis of an N2O emission model at the plot and landscape scale. Geoderma 159, 9-23.
Nourzadeh, M., Hashemy, S.M., Rodriguez Martin, J.A., Bahrami, H.A., Moshashaei, S., 2012. Using fuzzy clustering algorithms to describe the distribution of trace elements in arable calcareous soils in northwest Iran. Archives of Agronomy and Soil Science 59, 435-448.
Nussbaum, M., Spiess, K., Baltensweiler, A., Grob, U., Keller, A., Greiner, L., Schaepman, M.E., Papritz, A., 2018. Evaluation of digital soil mapping approaches with large sets of environmental covariates. SOIL 4, 1-22.
Odeh, I.O.A., McBratney, A.B., Chittleborough, D.J., 1995. Further results on prediction of soil properties from terrain attributes: heterotopic cokriging and regression-kriging. Geoderma 67, 215-226.
Odumo, B.O., Carbonell, G., Angeyo, H.K., Patel, J.P., Torrijos, M., Rodriguez Martin, J.A., 2014. Impact of gold mining associated with mercury contamination in soil, biota sediments and tailings in Kenya. Environmental science and pollution research international 21, 12426-12435.
Ojeda, J.J., Rezaei, E.E., Kamali, B., McPhee, J., Meinke, H., Siebert, S., Webb, M.A., Ara, I., Mulcahy, F., Ewert, F., 2021. Impact of crop management and environment on the spatio-temporal variance of potato yield at regional scale. Field Crops Research 270, 108213.
Oliver, M., Webster, R., 2014. A tutorial guide to geostatistics: Computing and modelling variograms and kriging. CATENA 113, 56-69.
Oteng-Darko, P., Yeboah, S., Addy, S., Amponsah, S.K., Danquah, E.O., 2013. Crop modeling: A tool for agricultural research – A review.
Özgöz, E., 2009. Long Term Conventional Tillage Effect on Spatial Variability of Some Soil Physical Properties. Journal of Sustainable Agriculture 33, 142-160.
Özgöz, E., Akbş, F., Çetin, M., Erşahin, S., Günal, H., 2007. Spatial variability of soil physical properties as affected by different tillage systems. New Zealand Journal of Crop and Horticultural Science 35, 1-13.
Padarian, J., Minasny, B., McBratney, A.B., 2019. Using deep learning for digital soil mapping. SOIL 5, 79-89.
Padarian, J., Morris, J., Minasny, B., McBratney, A.B., 2018. Pedotransfer Functions and Soil Inference Systems. In: McBratney, A.B., Minasny, B., Stockmann, U. (Eds.), Pedometrics. Springer International Publishing, Cham, pp. 195-220.
Parfitt, J.M.B., Timm, L.C., Pauletto, E.A., Sousa, R.O.d., Castilhos, D.D., Ávila, C.L.d., Reckziegel, N.L., 2009. Spatial variability of the chemical, physical and biological properties in lowland cultivated with irrigated rice. Revista Brasileira de Ciência do Solo 33, 819-830.
Pebesma, E., Bivand, R., 2005. S Classes and Methods for Spatial Data: the sp Package.
Perreault, S., El Alem, A., Chokmani, K., Cambouris, A.N., 2022. Development of Pedotransfer Functions to Predict Soil Physical Properties in Southern Quebec (Canada). Agronomy 12, 526.
Piikki, K., Wetterlind, J., Söderström, M., Stenberg, B., 2021. Perspectives on validation in digital soil mapping of continuous attributes—A review. Soil Use and Management 37, 7-21.
Pino, V., McBratney, A., O'Brien, E., Singh, K., Pozza, L., 2022. Citizen science & soil connectivity: Where are we? Soil Security 9, 100073.
Poggio, L., de Sousa, L.M., Batjes, N.H., Heuvelink, G.B.M., Kempen, B., Ribeiro, E., Rossiter, D., 2021. SoilGrids 2.0: producing soil information for the globe with quantified spatial uncertainty. SOIL 7, 217-240.
Poggio, L., Gimona, A., Spezia, L., Brewer, M.J., 2016. Example of Bayesian Uncertainty for Digital Soil Mapping. In: Zhang, G.-L., Brus, D., Liu, F., Song, X.-D., Lagacherie, P. (Eds.), Digital Soil Mapping Across Paradigms, Scales and Boundaries. Springer Singapore, Singapore, pp. 181-193.
Poppiel, R.R., Paiva, A.F.d.S., Demattê, J.A.M., 2022. Bridging the gap between soil spectroscopy and traditional laboratory: Insights for routine implementation. Geoderma 425, 116029.
Pouladi, N., Møller, A.B., Tabatabai, S., Greve, M.H., 2019. Mapping soil organic matter contents at field level with Cubist, Random Forest and kriging. Geoderma 342, 85-92.
Pozza, L.E., Field, D.J., 2020. The science of Soil Security and Food Security. Soil Security 1, 100002.
Probst, P., Boulesteix, A.-L., 2017. To tune or not to tune the number of trees in random forest. J. Mach. Learn. Res. 18, 6673–6690.
Pusch, M., Samuel-Rosa, A., Sergio Graziano Magalhães, P., Rios do Amaral, L., 2023. Covariates in sample planning optimization for digital soil fertility mapping in agricultural areas. Geoderma 429, 116252.
Qi, Y.-C., Dong, Y.-S., Liu, J.-Y., Domroes, M., Geng, Y.-B., Liu, L.-X., Liu, X.-R., Yang, X.-h., 2007. Effect of the conversion of grassland to spring wheat field on the CO2 emission characteristics in Inner Mongolia, China. Soil and Tillage Research 94, 310-320.
R Core Team, 2016. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
R Core Team, 2021. R: A language and environment for statistical computing. R Foundation for Statistical

Computing. Vienna, Austria. URL https://www.R-project.org/.
Rabus, B., Eineder, M., Roth, A., Bamler, R., 2003. The shuttle radar topography mission—a new class of digital elevation models acquired by spaceborne radar. ISPRS Journal of Photogrammetry and Remote Sensing 57, 241-262.
Ramankutty, N., Mehrabi, Z., Waha, K., Jarvis, L., Kremen, C., Herrero, M., Rieseberg, L.H., 2018. Trends in Global Agricultural Land Use: Implications for Environmental Health and Food Security. Annual Review of Plant Biology 69, 789-815.
Ramirez-Villegas, J., Koehler, A.-K., Challinor, A.J., 2017. Assessing uncertainty and complexity in regional-scale crop model simulations. European Journal of Agronomy 88, 84-95.
Ran, H., Kang, S., Hu, X., Yao, N., Li, S., Wang, W., Galdos, M.V., Challinor, A.J., 2022. A framework to quantify uncertainty of crop model parameters and its application in arid Northwest China. Agricultural and Forest Meteorology 316, 108844.
Ravensbergen, A.P.P., Chamberlin, J., Craufurd, P., Shehu, B.M., Hijbeek, R., 2021. Adapting the QUEFTS model to predict attainable yields when training data are characterized by imperfect management. Field Crops Research 266, 108126.
Rees, R.M., Griffiths, B.S., McVittie, A., 2018. Sustainable Intensification of Agriculture: Impacts on Sustainable Soil Management. In: Ginzky, H., Dooley, E., Heuser, I.L., Kasimbazi, E., Markus, T., Qin, T. (Eds.), International Yearbook of Soil Law and Policy 2017. Springer International Publishing, Cham, pp. 7-16.
Reza, S.K., Nayak, D.C., Chattopadhyay, T., Mukhopadhyay, S., Singh, S.K., Srinivasan, R., 2016. Spatial distribution of soil physical properties of alluvial soils: a geostatistical approach. Archives of Agronomy and Soil Science 62, 972-981.
Reza, S.K., Sarkar, D.K., Baruah, U., Das, T.H., 2010. Evaluation and comparison of ordinary kriging and inverse distance weighting methods for prediction of spatial variability of some chemical parameters of Dhalai district, Tripura.
Roberts, D.R., Bahn, V., Ciuti, S., Boyce, M.S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J.J., Schröder, B., Thuiller, W., Warton, D.I., Wintle, B.A., Hartig, F., Dormann, C.F., 2017. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography 40, 913-929.
Robinson, N.J., Benke, K.K., Norng, S., 2015. Identification and interpretation of sources of uncertainty in soils change in a global systems-based modelling process. Soil Research 53, 592-604.
Rodríguez Martín, J., Carbonell Martín, G., López Arias, M., Grau Corbí, J., 2009. Mercury content in topsoils, and geostatistical methods to identify anthropogenic input in the Ebro basin (Spain). Spanish Journal of Agricultural Research 7, 107-118.
Rodriguez Martin, J., Nanos, N., Grigoratos, T., Carbonell, G., Samara, C., 2014. Local deposition of mercury in topsoils around coal-fired power plants: is it always true? Environ Sci Pollut Res, 1-10.
Rodríguez Martín, J.A., Álvaro-Fuentes, J., Gonzalo, J., Gil, C., Ramos-Miras, J.J., Grau Corbí, J.M., Boluda, R., 2016. Assessment of the soil organic carbon stock in Spain. Geoderma 264, Part A, 117-125.
Rodríguez Martín, J.A., Gutiérrez, C., Escuer, M., García-González, M.T., Campos-Herrera, R., Águila, N., 2014. Effect of mine tailing on the spatial variability of soil nematodes from lead pollution in La Union (Spain). Science of The Total Environment 473–474, 518-529.
Rossiter, D.G., Liu, J., Carlisle, S., Zhu, A.X., 2015. Can citizen science assist digital soil mapping? Geoderma 259-260, 71-80.
Rurinda, J., Zingore, S., Jibrin, J.M., Balemi, T., Masuki, K., Andersson, J.A., Pampolino, M.F., Mohammed, I., Mutegi, J., Kamara, A.Y., Vanlauwe, B., Craufurd, P.Q., 2020. Science-based decision support for formulating crop fertilizer recommendations in sub-Saharan Africa. Agricultural Systems 180, 102790.
Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M., Tarantola, S., 2008. Global sensitivity analysis: the primer. John Wiley & Sons.
Samuel-Rosa, A., Heuvelink, G.B.M., Vasques, G.M., Anjos, L.H.C., 2015. Do more detailed environmental covariates deliver more accurate soil maps? Geoderma 243-244, 214-227.
Sanchez, P.A., Ahamed, S., Carré, F., Hartemink, A.E., Hempel, J., Huising, J., Lagacherie, P., McBratney, A.B., McKenzie, N.J., Mendonça-Santos, M.d.L., Minasny, B., Montanarella, L., Okoth, P., Palm, C.A., Sachs, J.D., Shepherd, K.D., Vågen, T.-G., Vanlauwe, B., Walsh, M.G., Winowiecki, L.A., Zhang, G.-L., 2009. Digital Soil Map of the World. Science 325, 680-681.
Sattari, S.Z., van Ittersum, M.K., Bouwman, A.F., Smit, A.L., Janssen, B.H., 2014. Crop yield response to soil fertility and N, P, K inputs in different environments: Testing and improving the QUEFTS model. Field Crops Research 157, 35-46.
Savage, S.L., 2009. The flaw of averages : why we underestimate risk in the face of uncertainty. John Wiley & Sons Hoboken, Hoboken.
Sayre, R., Dangermond, J., Frye, C., Vaughan, R., Aniello, P., Breyer, S.P., Cribbs, D., Hopkins, D., Nauman, R., Derrenbacher, W., Wright, D.J., Brown, C., Convis, C., Smith, J.H., Benson, L., VanSistine, D.P., Warner, H., Cress, J.J., Danielson, J.J., Hamann, S.L., Cecere, T., Reddy, A.D., Burton, D., Grosse, A., True, D., Metzger, M., Hartmann, J., Moosdorf, N., Durr, H., Paganini, M., Defourny, P., Arino, O., Maynard, S., Anderson, M., Comer, P., 2014. A new map of global ecological land units—An

ecophysiographic stratification approach. Association of American Geographers.
Schils, R.L.M., van Voorn, G.A.K., Grassini, P., van Ittersum, M.K., 2022. Uncertainty is more than a number or colour: Involving experts in uncertainty assessments of yield gaps. Agricultural Systems 195, 103311.
Schmidinger, J., Heuvelink, G.B.M., 2023. Validation of uncertainty predictions in digital soil mapping. Geoderma 437, 116585.
Schulte, R.P.O., Creamer, R.E., Donnellan, T., Farrelly, N., Fealy, R., O'Donoghue, C., O'hUallachain, D., 2014. Functional land management: A framework for managing soil-based ecosystem services for the sustainable intensification of agriculture. Environmental Science & Policy 38, 45-58.
Schut, A.G.T., Giller, K.E., 2020. Soil-based, field-specific fertilizer recommendations are a pipe-dream. Geoderma 380, 114680.
Semella, S., Hutengs, C., Seidel, M., Ulrich, M., Schneider, B., Ortner, M., Thiele-Bruhn, S., Ludwig, B., Vohland, M., 2022. Accuracy and Reproducibility of Laboratory Diffuse Reflectance Measurements with Portable VNIR and MIR Spectrometers for Predictive Soil Organic Carbon Modeling. Sensors (Basel) 22.
Sheahan, M., Barrett, C.B., 2017. Ten striking facts about agricultural input use in Sub-Saharan Africa. Food Policy 67, 12-25.
Shehu, B.M., Merckx, R., Jibrin, J.M., Kamara, A.Y., Rurinda, J., 2018. Quantifying Variability in Maize Yield Response to Nutrient Applications in the Northern Nigerian Savanna. Agronomy 8, 18.
Shepherd, K.D., Ferguson, R., Hoover, D., van Egmond, F., Sanderman, J., Ge, Y., 2022. A global soil spectral calibration library and estimation service. Soil Security 7, 100061.
Shepherd, K.D., Shepherd, G., Walsh, M.G., 2015. Land health surveillance and response: A framework for evidence-informed land management. Agricultural Systems 132, 93-106.
Shepherd, K.D., Walsh, M.G., 2002. Development of Reflectance Spectral Libraries for Characterization of Soil Properties. Soil Sci. Soc. Am. J. 66, 988-998.
Shiva Shankar, Y., Khan, M.L., Qureshi, A., 2023. Chapter 14 - Spatial applications of crop models in the Indian context and sustainability. In: Farooq, M., Gogoi, N., Pisante, M. (Eds.), Sustainable Agriculture and the Environment. Academic Press, pp. 395-420.
Shrestha, D.L., Solomatine, D.P., 2006. Machine learning approaches for estimation of prediction interval for the model output. Neural Networks 19, 225-235.
Shukla, A.K., Behera, S.K., Lenka, N.K., Tiwari, P.K., Prakash, C., Malik, R.S., Sinha, N.K., Singh, V.K., Patra, A.K., Chaudhary, S.K., 2016. Spatial variability of soil micronutrients in the intensively cultivated Trans-Gangetic Plains of India. Soil and Tillage Research 163, 282-289.
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.
Simpkin, A.L., Armstrong, K.A., 2019. Communicating Uncertainty: a Narrative Review and Framework for Future Research. J Gen Intern Med 34, 2586-2591.
Smaling, E.M.A., Janssen, B.H., 1993. Calibration of quefts, a model predicting nutrient uptake and yields from chemical soil fertility indices. Geoderma 59, 21-44.
Somarathna, P.D.S.N., Minasny, B., Malone, B.P., Stockmann, U., McBratney, A.B., 2018. Accounting for the measurement error of spectroscopically inferred soil carbon data for improved precision of spatial predictions. Science of The Total Environment 631-632, 377-389.
Spiegelhalter, D., 2017. Risk and Uncertainty Communication. Annual Review of Statistics and Its Application 4, 31-60.
Stevens, A., Nocita, M., Tóth, G., Montanarella, L., van Wesemael, B., 2013. Prediction of Soil Organic Carbon at the European Scale by Visible and Near InfraRed Reflectance Spectroscopy. PLOS ONE 8, e66409.
Stone, J.R., Gilliam, J.W., Cassel, D.K., Daniels, R.B., Nelson, L.A., Kleiss, H.J., 1985. Effect of Erosion and Landscape Position on the Productivity of Piedmont Soils. Soil Science Society of America Journal 49.
Stoorvogel, J.J., Kooistra, L., Bouma, J., 2015. Managing soil variability at different spatial scales as a basis for precision agriculture. In: Lal R, Stewart BA (Eds.), Soil-Specific Farming: Precision Agriculture, pp. 37-72.
Styc, Q., Lagacherie, P., 2021. Uncertainty assessment of soil available water capacity using error propagation: A test in Languedoc-Roussillon. Geoderma 391, 114968.
Szabó, B., Szatmári, G., Takács, K., Laborczi, A., Makó, A., Rajkai, K., Pásztor, L., 2019. Mapping soil hydraulic properties using random-forest-based pedotransfer functions and geostatistics. Hydrol. Earth Syst. Sci. 23, 2615-2635.
Szatmári, G., Pásztor, L., 2019. Comparison of various uncertainty modelling approaches based on geostatistics and machine learning algorithms. Geoderma 337, 1329-1340.
Szatmári, G., Pásztor, L., Heuvelink, G.B.M., 2021. Estimating soil organic carbon stock change at multiple scales using machine learning and multivariate geostatistics. Geoderma 403, 115356.
Szatmári, G., Pirkó, B., Koós, S., Laborczi, A., Bakacsi, Z., Szabó, J., Pásztor, L., 2019. Spatio-temporal assessment of topsoil organic carbon stock change in Hungary. Soil and Tillage Research 195, 104410.
Takoutsing, B., Asaah, E., Yuh, R., Tchoundjeu, Z., Degrande, A., Kouodiekong, L., 2013. Impact of Organic

Soil Amendments on the Physical Characteristics and Yield Components of Potato (Solanum tuberosum L.) in the Highlands of Cameroon. Journal of Agricultural Science and Technology 3, 257-266.

Takoutsing, B., Weber, J., Aynekulu, E., Rodríguez Martín, J.A., Shepherd, K., Sila, A., Tchoundjeu, Z., Diby, L., 2016. Assessment of soil health indicators for sustainable production of maize in smallholder farming systems in the highlands of Cameroon. Geoderma 276, 64-73.

Takoutsing, B., Weber, J., Tchoundjeu, Z., Shepherd, K., 2015. Soil chemical properties dynamics as affected by land use change in the humid forest zone of Cameroon. Agroforest Syst, 1-14.

Takoutsing, B., Weber, J.C., Rodríguez Martín, J.A., Shepherd, K., Aynekulu, E., Sila, A., 2018. An assessment of the variation of soil properties with landscape attributes in the highlands of Cameroon. Land Degradation & Development 29, 2496-2505.

Takoutsing, B., Winowiecki, L.A., Bargués-Tobella, A., Vågen, T.-G., 2023. Determination of land restoration potentials in the semi-arid areas of Chad using systematic monitoring and mapping techniques. Agroforest Syst 97, 1289-1305.

Taylor, J.A., Jacob, F., Galleguillos, M., Prévot, L., Guix, N., Lagacherie, P., 2013. The utility of remotely-sensed vegetative and terrain covariates at different spatial resolutions in modelling soil and watertable depth (for digital soil mapping). Geoderma 193-194, 83-93.

ten Berge, H.F.M., Hijbeek, R., van Loon, M.P., Rurinda, J., Tesfaye, K., Zingore, S., Craufurd, P., van Heerwaarden, J., Brentrup, F., Schröder, J.J., Boogaard, H.L., de Groot, H.L.E., van Ittersum, M.K., 2019. Maize crop nutrient input requirements for food security in sub-Saharan Africa. Global Food Security 23, 9-21.

Terhoeven-Urselmans, T., Vagen, T.G., Spaargaren, O., Shepherd, K.D., 2010. Prediction of soil fertility properties from a globally distributed soil mid-infrared spectral library. Soil Science Society of America Journal 74, 1792-1799.

Tesfahunegn, G.B., Tamene, L., Vlek, P.L.G., 2011. Catchment-scale spatial variability of soil properties and implications on site-specific soil management in northern Ethiopia. Soil and Tillage Research 117, 124-139.

Tittonell, P., Dis, R.V., Vanlauwe, B., Shepherd, K.D., 2015. Managing Soil Heterogeneity in Smallholder African Landscapes Requires a New Form of Precision Agriculture. In: Lal, R., Stewart, B.A. (Eds.), Soil Specific Farming: Precision Agriculture. Advances in Soil Science. CRC Press. http://www.crcnetbase.com/doi/abs/10.120/b18759-9.

Torres-Matallana, J.A., Leopold, U., Heuvelink, G.B.M., 2021. Multivariate autoregressive modelling and conditional simulation for temporal uncertainty analysis of an urban water system in Luxembourg. Hydrol. Earth Syst. Sci. 25, 193-216.

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, 1375-1385.

Trap, J., Blanchart, E., 2023. Intensifying the soil ecological functions for sustainable agriculture: Acting with stakeholders. Current Research in Environmental Sustainability 5, 100225.

Tsegaye, T., Hill, R.L., 1998. Intensive tillage effects on spatial variability of soil test, plant growth, and nutrient uptake measurements. Soil Science 163, 155-165.

UNEP, 2012. Land Health Surveillance: An Evidence-Based Approach to Land Ecosystem Management. Illustrated with a Case Study in the West Africa Sahel. United Nations Environment Programme, Nairobi, http://wedocs.unep.org/handle/20.500.11822/8571.

Ussiri, D.A., Mnkeni, P.N.S., MacKenzie, A.F., Semoka, J.M.R., 1998. Soil test calibration studies for formulation of phosphorus fertilizer recommendations for maize in Morogoro district, Tanzania. I. evaluation of soil test methods. Communications in Soil Science and Plant Analysis 29, 2801-2813.

Vågen, T.-G., Davey, F., Shepherd, K., 2012. Land Health Surveillance: Mapping Soil Carbon in Kenyan Rangelands. In: Nair, P.K.R., Garrity, D. (Eds.), Agroforestry - The Future of Global Land Use. Springer Netherlands, pp. 455-462.

Vågen, T.-G., Shepherd, K.D., Walsh, M.G., 2006. Sensing landscape level change in soil fertility following deforestation and conversion in the highlands of Madagascar using Vis-NIR spectroscopy. Geoderma 133, 281-294.

Vågen, T.-G., Winowiecki, A., 2020. The Land Degradation Surveillance Framework Field Guide. ICRAF, Nairobi.

Vågen, T.-G., Winowiecki, L.A., 2013. Mapping of soil organic carbon stocks for spatially explicit assessments of climate change mitigation potential. Environmental Research Letters 8, 015011.

Vågen, T.-G., Winowiecki, L.A., Abegaz, A., Hadgu, K.M., 2013. Landsat-based approaches for mapping of land degradation prevalence and soil functional properties in Ethiopia. Remote Sensing of Environment 134, 266-275.

Vågen, T.-G., Winowiecki, L.A., Tondoh, J.E., Desta, L.T., Gumbricht, T., 2016. Mapping of soil properties and land degradation risk in Africa using MODIS reflectance. Geoderma 263, 216-225.

Vågen, T.G., Winowiecki, L.A., Neely, C., Chesterman, S., Bourne, M., 2018. Spatial assessments of soil organic carbon for stakeholder decision-making – a case study from Kenya. SOIL 4, 259-266.

van der Westhuizen, S., Heuvelink, G.B.M., Hofmeyr, D.P., Poggio, L., 2022. Measurement error-filtered machine learning in digital soil mapping. Spatial Statistics 47, 100572.
van Diepen, C.A., Wolf, J., van Keulen, H., Rappoldt, C., 1989. WOFOST: a simulation model of crop production. Soil Use and Management 5, 16-24.
van Ittersum, M.K., Cassman, K.G., Grassini, P., Wolf, J., Tittonell, P., Hochman, Z., 2013. Yield gap analysis with local to global relevance—A review. Field Crops Research 143, 4-17.
van Ittersum, M.K., van Bussel, L.G.J., Wolf, J., Grassini, P., van Wart, J., Guilpart, N., Claessens, L., de Groot, H., Wiebe, K., Mason-D'Croz, D., Yang, H., Boogaard, H., van Oort, P.A.J., van Loon, M.P., Saito, K., Adimo, O., Adjei-Nsiah, S., Agali, A., Bala, A., Chikowo, R., Kaizzi, K., Kouressy, M., Makoi, J.H.J.R., Ouattara, K., Tesfaye, K., Cassman, K.G., 2016. Can sub-Saharan Africa feed itself? Proceedings of the National Academy of Sciences 113, 14964-14969.
van Leeuwen, C.C.E., Mulder, V.L., Batjes, N.H., Heuvelink, G.B.M., 2021. Statistical modelling of measurement error in wet chemistry soil data. European Journal of Soil Science 73.
van Leeuwen, C.C.E., Mulder, V.L., Batjes, N.H., Heuvelink, G.B.M., 2022. Statistical modelling of measurement error in wet chemistry soil data. European Journal of Soil Science 73, e13137.
Van Looy, K., Bouma, J., Herbst, M., Koestel, J., Minasny, B., Mishra, U., Montzka, C., Nemes, A., Pachepsky, Y.A., Padarian, J., Schaap, M.G., Tóth, B., Verhoef, A., Vanderborght, J., van der Ploeg, M.J., Weihermüller, L., Zacharias, S., Zhang, Y., Vereecken, H., 2017. Pedotransfer Functions in Earth System Science: Challenges and Perspectives. Reviews of Geophysics 55, 1199-1256.
Vanino, S., Pirelli, T., Di Bene, C., Bøe, F., Castanheira, N., Chenu, C., Cornu, S., Feiza, V., Fornara, D., Heller, O., Kasparinskis, R., Keesstra, S., Lasorella, M.V., Madenoğlu, S., Meurer, K.H.E., O'Sullivan, L., Peter, N., Piccini, C., Siebielec, G., Smreczak, B., Thorsøe, M.H., Farina, R., 2023. Barriers and opportunities of soil knowledge to address soil challenges: Stakeholders' perspectives across Europe. Journal of Environmental Management 325, 116581.
Vanlauwe, B., Six, J., Sanginga, N., Adesina, A.A., 2015. Soil fertility decline at the base of rural poverty in sub-Saharan Africa. Nat Plants 1, 15101.
Vasques, G.M., Grunwald, S., Sickman, J.O., 2008. Comparison of multivariate methods for inferential modeling of soil carbon using visible/near-infrared spectra. Geoderma 146, 14-25.
Vaysse, K., Lagacherie, P., 2017. Using quantile regression forest to estimate uncertainty of digital soil mapping products. Geoderma 291, 55-64.
Vejre, H., Callesen, I., Vesterdal, L., Raulund-Rasmussen, K., 2003. Carbon and Nitrogen in Danish Forest Soils—Contents and Distribution Determined by Soil Order. Soil Science Society of America Journal 67, 335-343.
Veronesi, F., Schillaci, C., 2019. Comparison between geostatistical and machine learning models as predictors of topsoil organic carbon with a focus on local uncertainty estimation. Ecological Indicators 101, 1032-1044.
Verstegen, J.A., Karssenberg, D., van der Hilst, F., Faaij, A., 2012. Spatio-temporal uncertainty in Spatial Decision Support Systems: A case study of changing land availability for bioenergy crops in Mozambique. Computers, Environment and Urban Systems 36, 30-42.
Viscarra Rossel, R.A., Behrens, T., Ben-Dor, E., Brown, D.J., Demattê, J.A.M., Shepherd, K.D., Shi, Z., Stenberg, B., Stevens, A., Adamchuk, V., Aïchi, H., Barthès, B.G., Bartholomeus, H.M., Bayer, A.D., Bernoux, M., Böttcher, K., Brodský, L., Du, C.W., Chappell, A., Fouad, Y., Genot, V., Gomez, C., Grunwald, S., Gubler, A., Guerrero, C., Hedley, C.B., Knadel, M., Morrás, H.J.M., Nocita, M., Ramirez-Lopez, L., Roudier, P., Campos, E.M.R., Sanborn, P., Sellitto, V.M., Sudduth, K.A., Rawlins, B.G., Walter, C., Winowiecki, L.A., Hong, S.Y., Ji, W., 2016a. A global spectral library to characterize the world's soil. Earth-Science Reviews 155, 198-230.
Viscarra Rossel, R.A., Behrens, T., Ben-Dor, E., Chabrillat, S., Demattê, J.A.M., Ge, Y., Gomez, C., Guerrero, C., Peng, Y., Ramirez-Lopez, L., Shi, Z., Stenberg, B., Webster, R., Winowiecki, L., Shen, Z., 2022. Diffuse reflectance spectroscopy for estimating soil properties: A technology for the 21st century. European Journal of Soil Science 73, e13271.
Viscarra Rossel, R.A., Bouma, J., 2016. Soil sensing: A new paradigm for agriculture. Agricultural Systems 148, 71-74.
Viscarra Rossel, R.A., Brus, D.J., Lobsey, C., Shi, Z., McLachlan, G., 2016b. Baseline estimates of soil organic carbon by proximal sensing: Comparing design-based, model-assisted and model-based inference. Geoderma 265, 152-163.
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, 59-75.
Viscarra Rossel, R.A., Brus, D.J., 2018. The cost-efficiency and reliability of two methods for soil organic C accounting. Land Degradation & Development 29, 506-520.
von Fromm, S.F., Doetterl, S., Butler, B.M., Aynekulu, E., Berhe, A.A., Haefele, S.M., McGrath, Steve P., Shepherd, K.D., Six, J., Tamene, L., Tondoh, E.J., Vågen, T.-G., Winowiecki, L.A., Trumbore, Susan E., Hoyt, A.M., 2024. Controls on timescales of soil organic carbon persistence across sub-Saharan Africa. Global Change Biology 30, e17089.

Vrebos, D., Jones, A., Lugato, E., O'Sullivan, L., Schulte, R., Staes, J., Meire, P., 2021. Spatial evaluation and trade-off analysis of soil functions through Bayesian networks. European Journal of Soil Science 72, 1575-1589.

Wadoux, A., Brus, D.J., Heuvelink, G.B.M., 2018. Accounting for non-stationary variance in geostatistical mapping of soil properties. Geoderma 324, 138-147.

Wadoux, A.M.J.C., 2019. Using deep learning for multivariate mapping of soil with quantified uncertainty. Geoderma 351, 59-70.

Wadoux, A.M.J.C., Brus, D.J., Heuvelink, G.B.M., 2019. Sampling design optimization for soil mapping with random forest. Geoderma 355, 113913.

Wadoux, A.M.J.C., Heuvelink, G.B.M., de Bruin, S., Brus, D.J., 2021. Spatial cross-validation is not the right way to evaluate map accuracy. Ecological Modelling 457, 109692.

Wadoux, A.M.J.C., Minasny, B., McBratney, A.B., 2020a. Machine learning for digital soil mapping: Applications, challenges and suggested solutions. Earth-Science Reviews 210, 103359.

Wadoux, A.M.J.C., Samuel-Rosa, A., Poggio, L., Mulder, V.L., 2020b. A note on knowledge discovery and machine learning in digital soil mapping. European Journal of Soil Science 71, 133-136.

Wallach, D., Thorburn, P.J., 2017. Estimating uncertainty in crop model predictions: Current situation and future prospects. European Journal of Agronomy 88, A1-A7.

Wardekker, J.A., van der Sluijs, J.P., Janssen, P.H.M., Kloprogge, P., Petersen, A.C., 2008. Uncertainty communication in environmental assessments: views from the Dutch science-policy interface. Environmental Science & Policy 11, 627-641.

Waswa, B.S., Vlek, P.L.G., Tamene, L.D., Okoth, P., Mbakaya, D., Zingore, S., 2013. Evaluating indicators of land degradation in smallholder farming systems of western Kenya. Geoderma 195–196, 192-200.

Webster, R., Oliver, M.A., 2001. Geostatistics for Environmental Scientists. . John Wiley & Sons Ltd., Chichester, England.

Webster, R., Oliver, M.A., 2007. Geostatistics for Environmental Scientists, 2nd Edition. John Wiley & Sons, Inc.

Wezel, A., Soboksa, G., McClelland, S., Delespesse, F., Boissau, A., 2015. The blurred boundaries of ecological, sustainable, and agroecological intensification: a review. Agronomy for Sustainable Development 35, 1283-1295.

Wickham, H., 2016. Programming with ggplot2. In: Wickham, H. (Ed.), ggplot2: Elegant Graphics for Data Analysis. Springer International Publishing, Cham, pp. 241-253.

Wilson, A.M., Jetz, W., 2016. Remotely sensed high-resolution global cloud dynamics for predicting ecosystem and biodiversity distributions. PLOS Biology 14, e1002415.

Winowiecki, L., Vågen, T.-G., Huising, J., 2016. Effects of land cover on ecosystem services in Tanzania: A spatial assessment of soil organic carbon. Geoderma 263, 274-283.

Winowiecki, L.A., Vågen, T.-G., Kinnaird, M.F., O'Brien, T.G., 2018. Application of systematic monitoring and mapping techniques: Assessing land restoration potential in semi-arid lands of Kenya. Geoderma 327, 107-118.

Wolde-Meskel, E., Amanu, T., Aynekulu, E., Winowiecki, L., 2022. Including soil organic carbon into nationally determined contributions: Insights from Ethiopia. AICCRA Policy Brief. Accelerating Impacts of CGIAR Climate Research for Africa (AICCRA). https://hdl.handle.net/10568/126514.

Wright, M.N., Ziegler, A., 2017. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. Journal of Statistical Software 77, 17.

Yang, L., Brus, D.J., Zhu, A.X., Li, X., Shi, J., 2018. Accounting for access costs in validation of soil maps: A comparison of design-based sampling strategies. Geoderma 315, 160-169.

Yong-Zhong, S., Yu-Lin, L., Jian-Yuan, C., Wen-Zhi, Z., 2005. Influences of continuous grazing and livestock exclusion on soil properties in a degraded sandy grassland, Inner Mongolia, northern China. CATENA 59, 267-278.

Zanter, K., 2019. Landsat 4-7 Surface Reflectance Code (LEDAPS) Product Guide, landsat-normalized-difference-vegetation-index. www.usgs.gov/land-resources/nli/landsat/landsat-normalized-difference-vegetation-index.

Zhao, C., Jia, X., Shao, M.a., Zhang, X., 2020. Using pedo-transfer functions to estimate dry soil layers along an 860-km long transect on China's Loess Plateau. Geoderma 369, 114320.

Zhong, L., Guo, X., Xu, Z., Ding, M., 2021. Soil properties: Their prediction and feature extraction from the LUCAS spectral library using deep convolutional neural networks. Geoderma 402, 115366.

Zhu, A.X., Liu, J., Du, F., Zhang, S.J., Qin, C.Z., Burt, J., Behrens, T., Scholten, T., 2015. Predictive soil mapping with limited sample data. European Journal of Soil Science 66, 535-547.

Zuber, S.M., Veum, K.S., Myers, R.L., Kitchen, N.R., Anderson, S.H., 2020. Role of inherent soil characteristics in assessing soil health across Missouri. Agricultural & Environmental Letters 5, e20021.

Summary

The surge in the demand for accurate soil information has become increasingly notable, aligning with global challenges related to sustainable agriculture, land management, and environmental conservation. This has spurred the adoption of innovative approaches, with proximal soil sensing (PSS) methods such as spectral data emerging as a prominent technique for generating accurate and timely soil data. While PSS has proven successful in delivering soil data, an aspect that has been relatively overlooked is the quantification of errors in soil measurements used for digital soil mapping (DSM) and crop model calibration and prediction. The calibration of PSS models relies on wet chemistry measurements, which are not error-free. The measurement error in analytical data combines with that of the PSS model, and eventually propagates to the DSM and crop model to affect the final predictions. Unfortunately, measures for error associated with compilations of such analytical and spectral datasets are seldom provided, leaving modellers with little knowledge about the quality or uncertainty of the data. Ignoring measurement errors in soil data used for model training may result in suboptimal models and systematic overestimation of prediction accuracy, impacting end-user decisions related to land management interventions. Practical solutions for quantifying measurement errors in analytical and spectral data have been proposed, offering a pathway for laboratories to report measurements along with their associated uncertainties, and enhancing the accuracy and reliability of the reported data.

Despite the widespread application of DSM products, high-resolution digital soil data availability in sub-Saharan Africa (SSA), particularly at local scales, remains limited. Various approaches, such as geostatistics and machine learning, are used to model and predict soil properties together with the associated uncertainty. However, there is a scarcity of studies in SSA that have systematically evaluated and compared the performance of DSM approaches such regression kriging (RK) and random forest (RF). Such evaluations are crucial for understanding the accuracy of these models in predicting soil properties and the effectiveness in modelling prediction uncertainty. RK leverages both the information from auxiliary variables through regression analysis and the spatial correlation captured by kriging to provide improved predictions for the variable of interest at unsampled locations. RK is particularly useful when dealing with spatial datasets where data points are unevenly distributed, and it helps account for both the global trend and local variations in the spatial structure of the variable. On the other hand, RF is an ensemble learning algorithm that uses a collection of decision trees to make predictions. Each decision tree is trained on a different subset of the data, and the predictions of all the trees are averaged to produce the final prediction. RF can handle complex relationships between soil attributes and spatial variability. It is robust against overfitting, and its ability to capture non-linear patterns makes it well-suited for modelling the intricate nature of soil properties across landscapes.

Conducting large-scale soil sampling in Sub-Saharan Africa (SSA) faces numerous challenges, given the complex and remote terrain that hampers easy access to various parts of the targeted landscape and the limited resources allocated for field sampling. The prevalent use of cluster sampling designs in this context offers several advantages, but it may lead to the under-representation of certain areas in the DSM modelling process. To address these challenges, alternative approaches, such as extrapolation of DSM models have been proposed. However, the issue of spatial extrapolation, especially in regions with no training data, poses a significant challenge, as extrapolation in geographic space can adversely impact prediction accuracy, and this concern has not been thoroughly investigated, particularly in SSA, despite the widespread use of sparsely distributed and highly clustered soil data.

Opinions converge on the fact that increase in food production to feed the anticipated increase in population cannot be attained without an increased use of fertilizers, particularly in SSA where soil fertility is often low. This, therefore, presents a scenario where the information needed for agricultural decision-making at all levels from farm management to national policy are also increasing and methods of supplying such information in relatively short timeframes are crucially needed. While traditional agronomic research based on field experiments has been a reliable information source, its limitations are becoming evident as the need for data surpasses the capacity of these methods. Field experiments often yield plot-specific results from trials conducted at specific points in time and space, making it challenging to extrapolate findings to other plots or areas.

Furthermore, agricultural landscapes in SSA are commonly managed based on uniform agronomic recommendations, despite the substantial inherent spatial variation in soil nutrients. This approach raises the risk of both over- and under-application of fertilizers, resulting in undesirable environmental impacts and increased production costs. While site-specific fertilizer recommendations have been proposed as a solution, existing strategies still heavily rely on results from experimental plots. Recognizing these limitations, there is a pressing need for tools that can integrate new data and research findings and provide stakeholders with actionable information for decision-making. Consequently, the development and effective utilization of tools, such as crop models, are essential for projecting cropping systems under various scenarios, conditions, and scales. However, model outputs are inevitably associated with uncertainty from various sources that require thorough analysis and quantification.

The main objective of this thesis was to apply and extend state-of-the-art DSM approaches to analyse and quantify the spatial patterns of soil properties while accounting for uncertainty in soil measurements. Additionally, it involved using derived uncertain DSM maps as inputs in crop modelling to derive maize yield and fertilizer recommendations. The impact of realistic quantification of accuracies of soil measurements and soil maps could improve the performance of DSM and crop models and support the development of strategies and policies that enhance sustainable agricultural intensification. The objective was approached based on: 1) description and quantification of the spatial variation of soil properties using simple geostatistical methods (**Chapter 2**); 2) quantification of errors in soil measurements and incorporation of the associated uncertainty into a state-of-the art geostatistical method (regression kriging) for spatial interpolation and comparison with a case in which measurement errors are ignored (**Chapter 3**); 3) Extension of the calibration and prediction of DSM models using uncertain soil measurements from linear kriging with external drift, i.e. regression kriging (RK) to non-linear machine learning algorithms-based DSM models, i.e. random forest (RF) (**Chapter 4**); and 4) quantification of uncertainty of QUEFTS model predictions considering uncertainty in soil input variables using Monte Carlo simulation (**Chapter 5**). **Chapters 2, 3 and 4** focused on analysis at sub-area, while **Chapter 5** was carried out at a regional scale that covered the entire study area.

In **Chapter 2**, I used the land health surveillance concept in combination with simple geostatistical approaches to describe selected soil properties among land use types and characterize their spatial variability. A total of 320 soil samples were collected in two sites (Bamendjou and Koutaba) with contrasting landscape attributes and dominant land uses (agricultural vs pasture) and were analysed in the laboratory for granulometric fraction, soil organic carbon (SOC), nitrogen (N), soil reaction (pH), phosphorous (P), calcium (Ca), potassium (K), magnesium (Mg), aluminium (Al) and zinc (Zn) using diffuse reflectance mid-infrared spectroscopy (MIRS). I used variogram models to quantify the spatial variation of soil properties, while ordinary kriging was applied to generate the respective maps of SOC, N, and

clay content. I assessed the accuracy of the prediction performance of models using cross-validation, and also evaluated the implications of spatial relationships for land management interventions and restoration. The results showed that soil properties differed considerably across the area, with significant positive and negative correlation coefficients among many pairs of soil properties. The coefficients of variation (CV) helped in comparing the degree of variation of soil properties relative to the mean within the two sites. For Bamendjou, the most variable properties (CV >38%) were P, Mn, and Ca. Moderate variability (2.8% < CV < 38%) was observed for sand, SOC, N, K and Mg, while properties with very low variability (CV < 2.8%) were clay, soil pH, and Al. For Koutaba, the most variable properties were sand, Ca, K, Mg, Mn, P and Zn. Moderate variables were N, SOC, and silt, while clay, Al and pH were the least variables. Spherical variogram models were chosen as the best-fitted models for the investigated soil properties as attested by cross-validation. The spatial correlation ranges were significantly larger for SOC and N in Bamendjou than in Koutaba. Land use with more vegetation cover (forest, grassland, and fallow) exhibited the highest concentration of soil properties, attesting that land use types had significant impacts on spatial patterns and distribution of the soil properties. Well-defined patterns of higher concentrations of SOC and N were observed in the lowlands, valleys, and areas dominated by annual vegetation. Kriged maps provided a detailed visualization of soil properties at the landscape scale and helped to identify 'hotspots' of land degradation and critical areas in need of specific land management practices to improve land productivity. These findings can be a helpful tool in achieving efficient site-specific land management interventions, that lead to better decisions aimed at enhancing the efficient use of agricultural inputs, such as fertilizers, in the context of limited resources. Although the spatial models could explain a large part of the spatial variation of soil properties, they may be improved by quantifying the measurement errors in soil observations, and expanding the analysis with relevant covariates, as was the case in **Chapter 3.**

In **Chapter 3**, I applied a geostatistical-based DSM approach - RK to incorporate quantified measurement error variances of analytical and spectral soil data in the covariance structure of the spatial model, weigh measurements in accordance with their quantified measurement accuracies and assess the effects of measurement errors on the accuracies of the resulted DSM predictions. I applied the method on soil data analysed for pH, clay and SOC using conventional and mid infrared spectroscopy methods. Variogram parameters and regression coefficients were estimated using residual maximum likelihood under two scenarios: with and without taking measurement errors into account. I compared the performance of the spatial models in the two scenarios using common validation metrics obtained with three types of cross-validation, namely leave-one-out, leave-cluster-out and leave-sentinel-site-out cross-validation. The results showed that accounting for measurement errors had significant impacts on the estimated regression coefficients and influenced the variogram parameters by reducing the nugget and sill variance for the three soil properties. Validation metrics including mean error, root mean squared error and model efficiency coefficient, were quite similar in both scenarios, but the prediction uncertainties were more realistically quantified by the models that accounted for measurement errors. There were relatively small absolute differences in predicted values of soil properties of up to 0.1 for pH, 1.6% for clay and 2 g kg^{-1} for SOC between the two scenarios. While differences in prediction maps and cross-validation metrics of predictions did not differ significantly between the two scenarios, substantial differences were obtained in prediction error standard deviation maps and in the evaluation of the prediction uncertainty. Relative differences in standard deviations were observed in some areas of the study area between the two scenarios of up to 0.08 for pH, 2.7% for clay and 0.5 g kg^{-1} for SOC. The best modelling approach would therefore be the one that accounts for measurement errors in soil observations. The findings overall indicated that the additional investment in quantification of

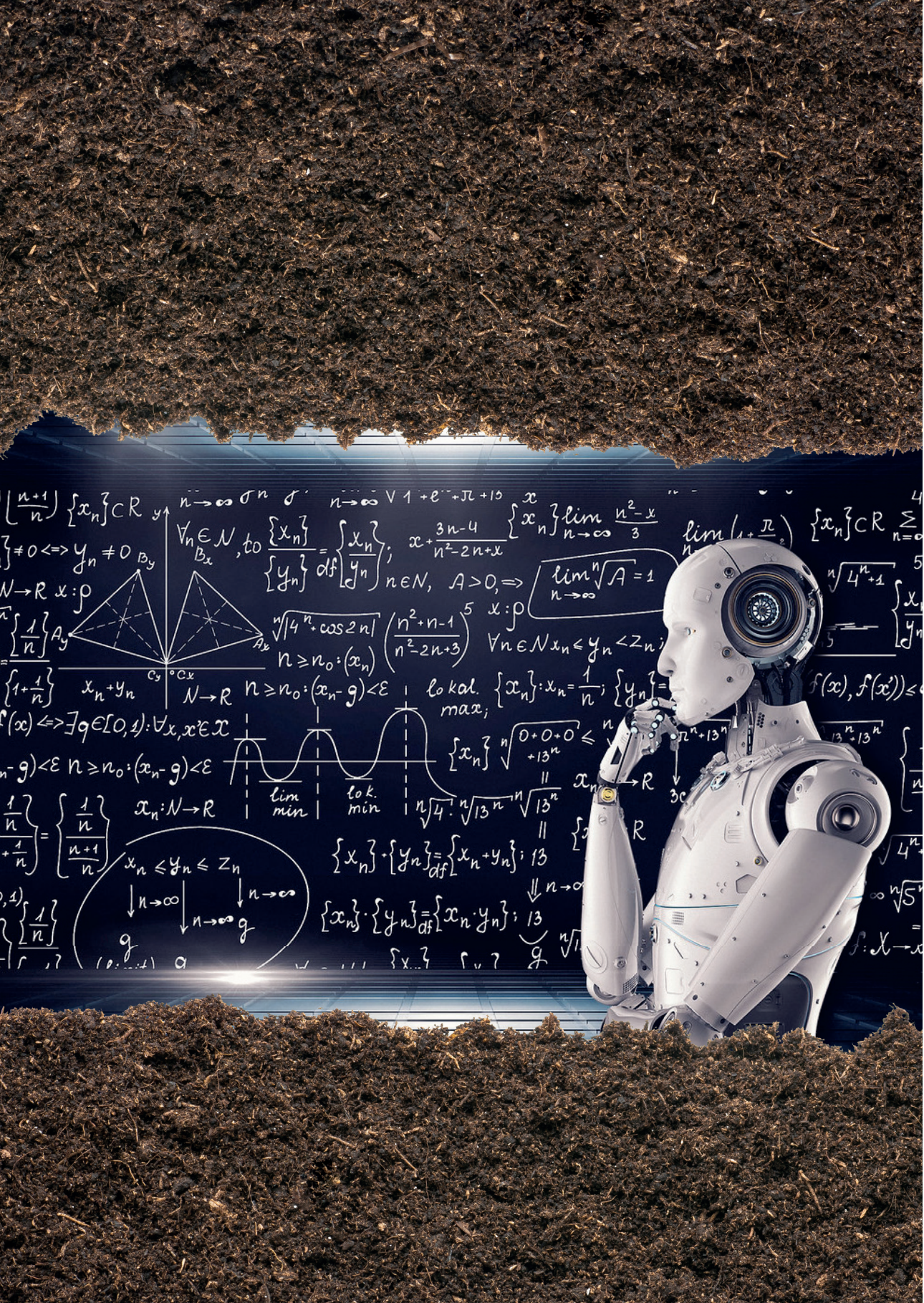
measurement error and the incorporation in the spatial models is worth the effort, as shown by the improvement in the quantification of the prediction uncertainties. These findings also emphasised the need of incorporating measurement errors in further studies not only in a geostatistical-based DSM approach, but also when using non-linear machine learning regression methods (**Chapter 4**) to improve uncertainty quantification, particularly when spectral data are used as the main soil data source.

In **Chapter 4**, I compared a geostatistical-based DSM approach - RK - and non-linear machine learning regression methods - RF – with respect to their ability to deliver accurate predictions and quantify prediction uncertainties, while accounting for errors in soil measurements. I also evaluated the sensitivity of the results of both DSM approaches to soil measurement errors, together with their spatial extrapolation potentials while mapping soil pH, clay, and SOC in the Cameroon study area. The results showed that both models produced comparable ranges and maps of predicted values for the three soil properties. Compared to RF, RK outperformed RF by presenting generally a higher model efficiency coefficient, lower root mean squared error values and better extrapolation performance. The improvement in root mean squared error was about 10, 12 and 2% while the improvement in the model efficiency coefficient was on average 5, 22 and 1% for pH, clay, and SOC, respectively. Overestimation of the local uncertainty observed for RK was larger than that of RF, indicating that prediction uncertainties were better quantified by the RF model. The effects of incorporating measurement errors appeared not significant for predictions and prediction uncertainties. This was partly attributed to the fact that this chapter assumed constant measurement error variances for each of the soil properties, and for analytical and spectral data used for calibration instead of considering measurement errors that are proportional to the measured values. An important finding of this chapter was that model comparison should go beyond using only common validation metrics to evaluate prediction accuracy of DSM approaches but should also account for their ability to quantify prediction uncertainty at unsampled locations. The findings of this chapter show that soil maps are not error free, and if use as soil data inputs, the uncertainty can propagate to affect the results of further modelling. Hence, it is essential to evaluate the repercussions of uncertain soil information in other modelling processes, particularly in crop modelling (**Chapter 5)**.

In **Chapter 5**, since no model has perfect input information, I quantified the uncertainty of QUEFTS model predicted yield and fertilizer recommendations considering uncertainty in soil inputs, using Monte Carlo simulation. I also determined the contributions of individual soil input uncertainties to the overall model output uncertainties, discussed strategies to communicate uncertainty of QUEFTS outputs to end-users and advised on how uncertainty can be incorporated in the decision-making process. The results showed that the impact of soil input uncertainty on model output uncertainty was significant and varied spatially. Large uncertainties in yield and fertilizer recommendation rates, with interquartile ranges larger than the median were observed in some parts of the study area. While comparing the results of a deterministic run (ignoring uncertainty in model soil inputs) with those of the Monte Carlo simulations, mean differences in predicted yield and fertilizer recommendation rates required to reach a target yield of 5 tons ha^{-1} in some parts of the study area were up to 1.0 tons ha^{-1} and up to 59, 42, and 20 kg ha^{-1} for N, P and K fertilizers, respectively. Accounting for soil input uncertainty leads to a systematic shift of the three fertilizers towards higher values. The spatial distribution of the uncertainty maps closely matched the spatial patterns of the soil input uncertainties, with high values where input uncertainty is high. Stochastic sensitivity analysis showed that pH is the main source of uncertainty for required K fertilizer (81.6%) and that soil organic carbon contributes to the uncertainty of required N fertilizer (97%), P fertilizer (25%) and K fertilizer (18%). Uncertainty in required P fertilizer mostly comes from soil extractable

phosphorus (55%) and exchangeable potassium (20%). A threshold probability map designed using statistical predictions served as a visual tool that could enable farmers to swiftly make informed decisions about fertilizer application locations. The study highlighted meaningful relationships between the uncertainty of soil properties and the uncertainty associated with model outputs. It emphasized the necessity for future research to address correlations among uncertain soil inputs and consider additional sources of uncertainty for a more thorough and accurate quantification of model output uncertainty. The findings underscore the importance of improving the accuracy of soil maps, which, in turn, positively influences the accuracy of QUEFTS model predictions. The findings of this chapter provide significant benefits to site-specific nutrient management proposed in **Chapter 2**, by estimating the soil nutrient supply, recommending needed fertilizer rates and the corresponding theoretical yields. The findings also contribute to enhancing our understanding of the interplay between the accuracy of soil information and the reliability of model outputs (**Chapter 3** and **Chapter 4**), offering valuable insights for improved decision-making to achieve sustainable agricultural intensification (**Chapter 5**).

The thesis synthesis is given in **Chapter 6**. It discusses the main findings of this thesis, my personal implications and recommendations for stakeholders and policy makers, while acknowledging the limitations of this research. Ultimately, the significance of quantifying uncertainty in soil information varies across stakeholders and purposes, underscoring the need for a nuanced understanding at different levels. Policy makers should formulate and implement tailored agricultural policies that consider the soil quality of specific areas. Simultaneously, technicians are encouraged to adopt effective site-specific practices that account for the unique soil and environmental conditions of each location. The government can leverage DSM-derived soil maps in conjunction with crop models to make more informed and efficient management decisions related to fertilizer formulation/import, distribution, and recommendations based on crop nutrient requirements. Extension agents and farmers can make use of threshold probability maps designed using statistical predictions as a practical visual aid, to swiftly make informed decisions about fertilizer application locations and rates. Considering the uncertainty contributions of soil input data to the overall model output uncertainty, modellers are encouraged to standardize the soil sampling design and analytical protocols to enhance soil measurement accuracy. This approach also serves as inspiration for other scientists to leverage available data through DSM approaches and develop more suitable models capable of comprehensively accounting for all sources of uncertainty.



Acknowledgements

Embarking on the journey of completing the PhD research is a monumental task, fraught with many challenges and obstacles that test one's resilience, determination, and perseverance. The characteristics of my journey were such that I could not have completed it alone. From the initial stages of research conceptualization to the final stages of thesis completion, many individuals and institutions have played integral roles in shaping the trajectory of this academic pursuit. I take this opportunity to give credit to all those that have helped and supported me along the way, each within their own capacity and in unique ways.

I start by thanking my supervisors who closely monitored the research progress and nurture me through the whole process. I am profoundly grateful to my esteemed supervisors and co-supervisors Prof Gerard Heuvelink, Prof Jetse Stoorvogel, Dr Ermias Betemariam, and Dr Keith Shepherd, whose unwavering support, invaluable guidance, and profound expertise have been instrumental in shaping the trajectory of my doctoral journey. Their commitment to excellence, profound insights, and tireless dedication to my academic and personal development have enriched this research endeavour beyond measure. I am deeply indebted to them for their encouragement and constructive feedback, which have not only shaped the direction of this thesis but also profoundly impacted my growth as a scholar and researcher. Particularly to Gerard, my daily supervisor, I feel immensely fortunate to have had the privilege of working alongside you throughout this journey. You were always patient with me and available to guide looking for solutions to my preoccupations. I have always admired your ability to explain complex geostatistics concepts in a manner that is accessible to all levels of understanding. Your guidance, support, and mentorship have been invaluable to me, and I am deeply grateful for the trust you placed in me by providing the opportunity to pursue my PhD at Wageningen University; my dream university since my childhood captivated by its esteemed reputation, renowned faculty, and vibrant academic community. Your belief in my capabilities has been a source of inspiration and motivation, propelling me forward during both the triumphs and challenges encountered along the way. Thank you for your untiring support and for being an integral part of this transformative experience.

I extend my heartfelt appreciation to the management of the World Agroforestry (ICRAF) for their generous provision of resources, access to state-of-the-art facilities, and the invaluable datasets that have facilitated the completion of this research. I am deeply grateful for the support and cooperation of the Soil and Land Health Research Theme and the Spatial Data Science & Applied Learning Laboratory, whose collaboration and assistance have been indispensable throughout this journey. Their willingness to share their expertise, data, and resources has significantly enriched the depth and scope of this PhD research. I am immensely grateful to the staff of the ICRAF Soil-Plant Spectral Laboratory for their invaluable assistance in analysing the soil samples used in the PhD research. Their expertise and meticulous attention to detail have played a crucial role in ensuring the accuracy and reliability of the analytical methods used in the laboratory. Their dedication to providing high-quality soil analysis services as well as the associated measurement errors has been instrumental in advancing the understanding of the quality of measurements. I extend my sincerest appreciation for their professionalism and support throughout this process.

I express my sincere gratitude to ISRIC - World Soil Information (International Soil Reference and Information Centre) for their invaluable contribution to this research endeavour. Their provision of spatial data, expertise, and resources has been instrumental in advancing the scope and quality of this study. The provision of the environmental layers used as covariates throughout the PhD research has been indispensable. Without their support and collaboration,

assembling such a comprehensive dataset that accurately represents the soil forming factors and surface characteristics of the study area would have been an arduous task. Their commitment to promoting soil research and their dedication to facilitating access to essential soil information have been pivotal in shaping the outcomes of this research project. I am deeply grateful for their collaboration and support.

To my dear classmates and esteemed colleagues of the Soil Geography and Landscape group, whose camaraderie, intellectual discourse, and unwavering encouragement have been a constant source of inspiration and motivation, I extend my deepest gratitude. Your support, friendship, and shared experiences have provided solace and friendship during the inevitable challenges and successes of this PhD research. I am profoundly thankful for our collaborative efforts, stimulating discussions, and the bonds of friendship forged along this arduous yet rewarding path.

I am profoundly grateful to my beloved family for their unyielding love, unwavering support, and enduring patience throughout my PhD pursuit. Their steadfast encouragement, sacrifices, and firm belief in my abilities have been a beacon of strength and resilience during times of uncertainty and doubt. I express profound appreciation to my parents, whose endless love, sacrifices, and steadfast confidence in my capabilities have served as a continual wellspring of inspiration and encouragement throughout my journey. I am forever indebted to them for their staunch support and encouragement. To my siblings and extended family members, whose untiring support, encouragement, and understanding have provided solace and reassurance during the peaks and valleys of this PhD research, I extend my heartfelt appreciation. Your unwavering faith in my abilities and persistent support have been a constant source of strength and motivation.

Last but certainly not least, I owe an immense debt of gratitude to my beloved wife, Lucienne, and our cherished children, Romaric, Yann, Rhoda-Ann, Rodrigue, and Janny. Throughout the ups and downs of my PhD journey, your unwavering love, steadfast support, and boundless patience have been my guiding lights. Your understanding, encouragement, and sacrifices have sustained me through the long day and night hours of research, the challenges of academia, and the demands of balancing work and family life. I am endlessly grateful for your presence, your sacrifices, and your unwavering belief in me.

List of publications

Fonkeng, E.E., Chevallier, T., Sauvadet, M., Enock, S., Rakotondrazafy, N., Chapuis-Lardy, L., **Takoutsing, B**., Tabi Fritz, O., Harmand, J.-M., 2024. Dynamics of soil organic carbon pools following conversion of savannah to cocoa agroforestry systems in the Centre region of Cameroon. Geoderma Regional 36, e00758.

Takoutsing, B., Winowiecki, L.A., Bargués-Tobella, A., Vågen, T.-G., 2022. Determination of land restoration potentials in the semi-arid areas of Chad using systematic monitoring and mapping techniques. Agroforestry Systems.

Takoutsing, B., Heuvelink, G.B.M., 2022. Comparing the prediction performance, uncertainty quantification and extrapolation potential of regression kriging and random forest while accounting for soil measurement errors. Geoderma 428, 116192.

Takoutsing, B., Heuvelink, G.B.M., Stoorvogel, J.J., Shepherd, K.D., Aynekulu, E., 2022. Accounting for analytical and proximal soil sensing errors in digital soil mapping. European Journal of Soil Science 73, e13226.

Lohbeck, M., Albers, P., Boels, L.E., Bongers, F., Morel, S., Sinclair, F., **Takoutsing, B**., Vågen, T.-G., Winowiecki, L.A., Smith-Dumont, E., 2020. Drivers of farmer-managed natural regeneration in the Sahel. Lessons for restoration. Scientific Reports 10, 15038.

Takoutsing, B., Heuvelink, G.B.M., Stoorvogel, J.J., Shepherd, K.D., Aynekulu, E., 2019. Accounting for measurement errors in soil observations to improve the accuracies of digital soil mapping outputs (oral presentation). Wageningen conference, Wageningen, the Netherlands.

Takoutsing, B., Heuvelink, G.B.M., Shepherd, K.D., Aynekulu, E., 2019. Accounting for measurement errors in SOC estimates introduced by proximal sensing methods (oral presentation). 5th Global Science Conference on Climate Smart Agriculture, Bali, Indonesia.

Takoutsing, B., Weber, J.C., Rodríguez Martín, J.A., Shepherd, K., Aynekulu, E., Sila, A., 2018. An assessment of the variation of soil properties with landscape attributes in the highlands of Cameroon. Land Degradation & Development 29, 2496-2505.

Takoutsing, B., Rodríguez Martín, J.A., Weber, C.J. Shepherd, K., Sila, A., Tondoh J., 2017. "Landscape approach to assess key soil functional properties in the highlands of Cameroon: Repercussions of spatial relationships for land management interventions." Journal of Geochemical Exploration 178: 35-44.

Takoutsing, B., Weber, J., Aynekulu, E., Rodríguez Martín, J.A., Shepherd, K., Sila, A., Tchoundjeu, Z., Diby, L., 2016. Assessment of soil health indicators for sustainable production of maize in smallholder farming systems in the highlands of Cameroon. Geoderma 276, 64-73.

About the author

Bertin Takoutsing was born and raised in a rural community located in the West Region of Cameroon. His passion for agriculture was cultivated at a young age, spending cherished moments with grandparents tending to crop and nurturing the land. These formative experiences instilled a deep appreciation for the intricacies of soil health and agricultural practices, sparking a lifelong curiosity and commitment to understanding and enhancing soil ecosystems. Throughout secondary school, his fascination with the natural world continued to blossom, fuelling a desire to delve deeper into the science behind agricultural systems. This pursuit led to the completion of a Bachelor of Engineering in Agricultural and Bioresources from the University of Nigeria, Nsukka in 1999, where foundational knowledge in soil science, crop production, and sustainable agricultural practices was acquired. Building upon this academic foundation, Bertin pursued a Master's degree in Soil and Water Management in the same university in 2013, delving into the complexities of soil-water interactions, nutrient cycling, soil variability, and land management strategies. This period of study provided invaluable insights into the vital role of soil health in ecosystem functioning and agricultural sustainability, inspiring a deeper exploration into the realm of soil spatial variation and mapping. Motivated by a desire to leverage cutting-edge technology and advanced analytical techniques to address pressing soil management related challenges, Bertin embarked on a Ph.D. journey focused on pedometrics and digital soil mapping at the Soil Geography and Landscape group of the Wageningen University and Research in 2019. This interdisciplinary research endeavour combined expertise in soil science, soil health, geospatial analysis, machine learning, and spatial data science to develop innovative approaches for assessing soil information and mapping soil variability at high resolutions. Through this academic journey, Bertin has been driven by a passion for advancing our understanding of soil health dynamics and enhancing soil management practices to promote sustainable agriculture and environmental stewardship. This thesis represents the culmination of years of dedicated research, aimed at contributing novel insights and practical solutions to the field of soil science, digital soil mapping, and agricultural sustainability. During the course of his PhD journey, Bertin made significant contributions to the academic community through his prolific publication record, mentorship of Master's students, and active participation in various conferences and symposiums. In conjunction with his doctoral studies, Bertin has seamlessly balanced his academic pursuits with his role as an Associate Soil Health scientist at the World Agroforestry (ICRAF), a position he has held steadfastly for the past eight years. Through this role, Bertin has applied his expertise and research findings to address real-world challenges in soil health and agroforestry, contributing valuable insights to sustainable land management practices. His tenure at ICRAF has not only enhanced his professional growth but has also allowed him to make tangible contributions to improving land management to support sustainable agricultural systems and fostering environmental resilience on a global scale.

Contact email address:

Work email: b.takoutsing@cifor-icraf.org;
Private email: btakoutsing@gmail.com;
Address: Soil and Land Health Research Theme, World Agroforestry (ICRAF), P.O. Box, 16317, Yaoundé, Cameroon.

PE&RC Training and Education Statement

With the training and education activities listed below the PhD candidate has complied with the requirements set by the C.T. de Wit Graduate School for Production Ecology and Resource Conservation (PE&RC) which comprises of a minimum total of 32 ECTS (= 22 weeks of activities)

Review/project proposal (9 ECTS)

- Digital soil mapping using uncertain spectral soil observations to support sustainable agricultural intensification in West and Central Africa

Post-graduate courses (7.4 ECTS)

- Machine learning for spatial data; PE&RC (2019)
- Hands on digital soil mapping; ISRIC (2019)
- Introduction to R for statistical analysis; PE&RC (2019)
- Uncertainty propagation in spatial environmental modelling; PE&RC (2020)
- Predicting soil properties using mid infrared spectroscopy; ICRAF (2020)
- Remote sensing data, acquisition, processing, and analysis; ICRAF (2021)
- Tidy data transformation and visualization with R; PE&RC (2021)

Laboratory training and working visits (3 ECTS)

- Introduction to soil spectroscopy; Global soil partnerships and FAO (2021)
- A future of soil spectral inference; Global soil partnerships and FAO (2021)
- Characterization of soil properties using French national Vis-NIR and MIR spectral libraries; Global soil partnerships and FAO (2021)
- Application of infrared spectroscopy on soil and plant materials; the Research Institute for Development, France (2022)

Invited review of journal manuscripts (3 ECTS)

- Land degradation and development: effect of land use dynamics on physico- chemical soil properties and soil carbon stock in the northern Ethiopia
- Science of the total environment: forms and dynamics of soil potassium in acid soil in the Wolaita zone of southern Ethiopia
- Geoderma: variability of soil properties as influenced by land use and land management practices in the upper Blue Nile basin of Ethiopia

Deficiency, refresh, brush-up courses (2 ECTS)

- Using R for digital soil mapping; reading and practical exercises (2019-2021)

Competence, skills and career-oriented activities (2.1 ECTS)

- Introduction to latex; PE&RC (2019)
- Rmarkdown; VLAG (2019)
- Reviewing scientific publication; WGS (2021)
- Intensive writing week; Wageningen in'to Languages (2024)
- Communicating research results through social media; ICRAF (2024)

Scientific integrity/ethics in science activity (0.3 ECTS)

- Ethics & STICs: scientific integrity, research ethics & information ethics for ICTs; France Université Numérique (2024)

PE&RC Annual meetings, seminars and PE&RC weekend/retreat (1.5 ECTS)

- PE&RC Weekend for first years (2019)
- PE&RC Last year retreat (2024)

Discussion groups/local seminars or scientific meetings (6.3 ECTS)

- International symposium on plant-soil feedback linkages between root traits and soil biota Wageningen, the Netherlands (2019)
- Participate in PE&RC discussion group plant-soil interaction (2019-2020)
- Global soil biodiversity initiative (2020-2021)
- International soil modelling consortium (2020-2021)
- Restoration for resilience champions virtual meet-up (2021)
- Digital soil mapping with uncertain data University of Ndjamena, Republic of Chad (2021)
- International network on soil pollution (2022-2023)

International symposia, workshops, and conferences (11.3 ECTS)

- Wageningen soil conference; oral presentation; Wageningen, the Netherlands
- Global climate-smart agriculture conference; oral presentation; Bali, Indonesia

Societally relevant exposure (2 ECTS)

- Championing healthy soils and ecosystems through data blog (2020)
- Systematic monitoring and mapping techniques for land restoration blog (2022)

BSc/MSc thesis supervision (6 ECTS)

- Influence of soil properties and pedological requirements on the spatial distribution of *irvingia gabonensis* in the centre region of Cameroon
- Effect of spatial variation of soil properties on the yields of priority crop in the semi-arid area of Chad
- Assessment of the effects of agricultural practices on land degradation with gender perspectives in the western highland of Cameroon
- Gender-based assessment of the impacts of agricultural practices on land restoration in the dryland areas of Cameroon

Colophon

Cover design: Bertin Takoutsing
Printed by: ProefschriftMaken Utrecht