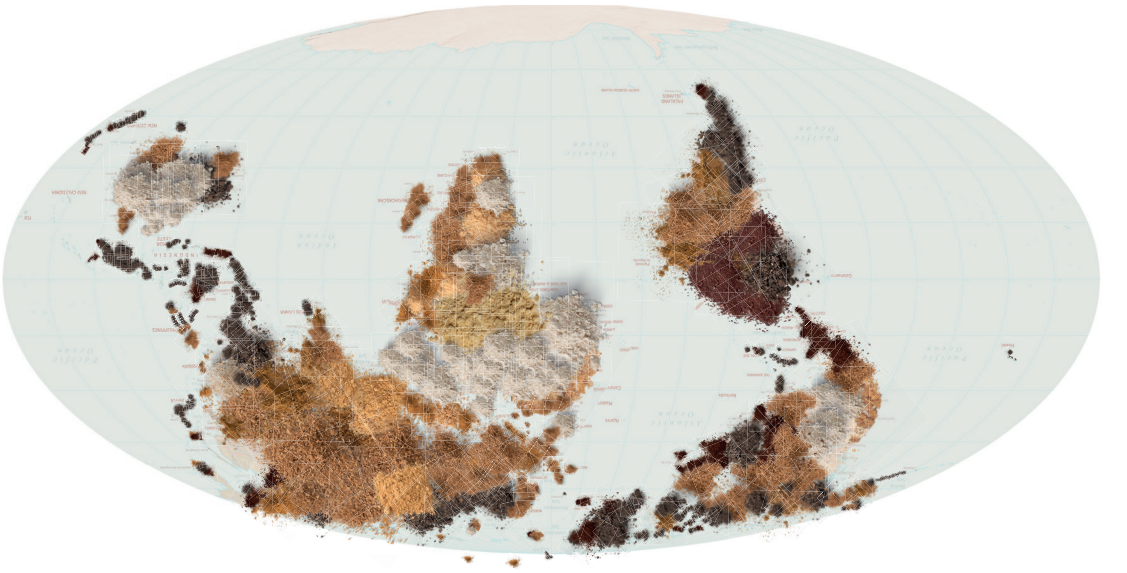
C.S. VAN DER WESTHUIZEN

ENHANCEMENT OF THE USE OF MACHINE LEARNING IN DIGITAL SOIL MAPPING



Propositions

1. Variable importance statistics used in isolation are worthless for soil scientists.

(this thesis)

2. Too much emphasis is placed on a single statistic, i.e., the mean-square-error, to convey map quality.

(this thesis)

3. Machine learning is abused for publications in scientific research.

4. PhD programmes in South Africa do not produce well-rounded academics.

5. Promotion and celebration of sports icons to boost a country's morale is misguided.

6. The mainstream approach to teaching statistics at school level does little to cultivate critical thinking.

Propositions belonging to the thesis, entitled

Enhancement of the use of machine learning in digital soil mapping

Stephan van der Westhuizen

Wageningen, 14 June 2024

Enhancement of the use of machine learning in digital soil mapping

C.S. van der Westhuizen

Thesis committee

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

Co-promotors
Dr David P. Hofmeyr
Senior Lecturer Mathematical Statistics
Lancaster University, United Kingdom

Dr Laura Poggio
Digital Soil Mapping and Remote Sensing Expert
ISRIC - World Soil Information, Wageningen

Other members
Prof. Dr I. Athanasiadis, Wageningen University & Research
Prof. Dr I. Fabris-Rotelli, University of Pretoria, South Africa
Prof. Dr E.J. Pebesma, University of Münster, Germany
Dr A.M.J.-C. Wadoux, French National Institute for Agriculture, Food, and Environment (INRAE), Paris, France

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

Enhancement of the use of machine learning in digital soil mapping

C.S. van der Westhuizen

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 Friday 14 June 2024
at 11 a.m. in the Omnia Auditorium.

C.S. van der Westhuizen
Enhancement of the use of machine learning in digital soil mapping,
174 pages.

PhD thesis, Wageningen University, Wageningen, the Netherlands (2024)
With references, with summaries in English and Afrikaans

ISBN: 978-94-6469-929-6
DOI: 10.18174/656077

Contents

Chapter 1

Introduction

"*Then you don't remember a world without robots. There was a time when humanity faced the universe alone and without a friend. Now he has creatures to help him; stronger creatures than himself, more faithful, more useful, and absolutely devoted to him.*"

Isaac Asimov - I, Robot

1.1 Background

1.1.1 Implementing machine learning

Artificial intelligence and machine learning have greatly influenced human society. Since the inception of the Information Age in the mid-20th century, sectors such as business, medicine, government and science have increasingly relied on machine learning for tasks of optimisation, automation, and prediction. In the realm of business, call centers utilise machine learning for intelligent call routing and interactive voice response, a game-changer to reduce waiting times. Healthcare institutions use machine learning to make predictions and offer proactive preventive care recommendations. Governments benefit from increased innovation, enabling the design of more effective policies that enhance communication and engage with citizens. The scientific community has also experienced notable advantages through the application of machine learning, employing it for diverse tasks ranging from reviewing literature to predicting environmental trends. Human dependence on machine learning has certainly reached a point that evokes echoes of science fiction novelist Isaac Asimov's classic work, "I, Robot" (Asimov, 1950). The reality is that machine learning has undeniably brought about positive societal changes.

Machine learning models are commonly characterised as being "black-box" in nature. This is due to the difficulty of understanding the process by which the model produces predictions. In this thesis I have adopted the definition in Belle & Papantonis (2021), which classifies a machine learning model as a model that is not simulateable by a human, lacks decomposability, and is algorithmically nontransparent. Specifically, a model is simulateable when it is simple and compact. Decomposability refers to the ability to break down a model into parts, such as inputs and parameters, and then to be able to explain these parts. Lastly, algorithmic tranparency denotes the ability to understand the procedure that the model goes through to generate its output (i.e., make predictions). This definition includes well known models such as random forests (Breiman, 2001), support vector machines (Cortes & Vapnik, 1995), and multi-layered artificial neural networks (Anderson, 1995). I mainly used random forests in this thesis, and so various descriptions of this class of models are given in several chapters of this thesis, depending on the problem at hand.

It is also important to note that machine learning models are data-driven. In other words, these are statistical models that execute specific regression or classification tasks by identifying patterns in data, and in such do not rely on explicit instructions or assumptions (Bishop, 2006). This stands in contrast to traditional statistical models like multiple linear regression or geostatistical models, whose formulations are based on assumptions about the data generating process (Webster & Oliver, 2007). For instance, a linear regression model requires the residuals to be normally distributed, if it is to be used for inferential purposes. It is for this reason, and the fact that machine learning can well capture complex

relationships between inputs and outputs, and make more accurate predictions, that has led to its widespread adoption in numerous scientific fields, including pedometrics, a branch of soil science primarily concerned with employing statistical and mathematical methods to analyse soil data. Unfortunately, this surge in usage has also resulted in the misuse or overuse of machine learning, including in environmental research. This might occur for example when the increase in prediction accuracy is minimal compared to that achieved by traditional statistical models. In such cases, it introduces a questionable trade-off between the complexity and the interpretability of the model.

The misuse of machine learning is sometimes because of a lack of understanding of key concepts in data analysis and modelling. Common pitfalls are related to data sufficiency, data cleaning, data leakage, model selection and evaluation, model optimisation, and model interpretability (Zhu et al., 2023). Data sufficiency is not only concerned with whether the required data set is large enough or whether the data quality is adequate, but also whether the data can address the research question at hand (Zhu et al., 2023). Data cleaning is often the most time consuming component of a data analysis, but it is a critical step to ensure good model development. Some typical data cleaning activities include the selection of the best variables (i.e., inputs), imputation of missing values, standardisation of data, and splitting the data into sets that are used to develop the model (training phase) and to evaluate its performance (testing phase). Data leakage is often overlooked and can lead to severely biased results (Zhu et al., 2023). This happens when a model gains information from unseen data during the training phase. For instance, a model might get an unfair advantage if variable selection is performed before data-splitting (Hastie et al., 2008). In the model selection phase it is important to include different types of machine learning models, for example tree-based or feature-learning methods or even a traditional statistical model. Furthermore, in the model selection phase, it is also important to know which models are appropriate and how they can be adapted to account for a specific challenge. For instance, when soil data are contaminated with errors, a model needs to be adapted to take the errors into account. With model evaluation, it is important to use a variety of metrics (e.g., mean error, mean square error, mean absolute error, model efficiency coefficient). Model optimisation is concerned with hyper-parameter optimisation as well as using an appropriate validation method (e.g., cross-validation). Finally, model interpretability holds significance as it is crucial to understand the extent to which the internal mechanisms of a model can be explained in human terms, and how well the model explanation can describe causal relationships (Zhu et al., 2023).

A theme that I address in this thesis pertains to the responsible and justified use of machine learning, incorporating sound "statistical thinking." The enhancement of the use of machine learning does not pertain to increasing its usage by "throwing things at the wall and see what sticks", but rather to employ it responsibly, ensuring optimal output for the task at hand.

1.1.2 Digital soil mapping and spatial prediction

The applications of this thesis are within digital soil mapping (DSM)[1], which is typically characterised as the computer-aided production of digital maps of soil types and properties. This is done by the use of mathematical and statistical models that combine information from soil observations with information contained in explanatory environmental variables. The framework of DSM, formalised by McBratney et al. (2003), is based on the *clorpt* model (Jenny, 1941) which expresses soil formation as a function of climate, organisms, relief, parent material and time. Note that McBratney et al. (2003) adapted the *clorpt* model to also include spatial position to what is known as the *scorpan* model. Spatial autocorrelation in the soil property is therefore commonly utilised, and it is convenient to associate the values of the soil property at all locations over a region of interest with the realisation of a stochastic process.

In the simplest scenario a stochastic process is often characterised by a constant mean plus a zero-mean, spatially correlated Gaussian residual process. In the presence of additional information made available through a set of covariates, we can model more general scenarios than those having constant mean by representing the process of interest through a spatially varying mean, which depends on the covariates, and the residual Gaussian process which captures spatial autocorrelation on top of the varying mean. This formulation is popular in the geostatistical literature and has motivated the use of methods such as regression kriging (Goovaerts, 1999). However, in many real-world situations the spatial variation which is naturally occurring in the covariates also provides substantial information about the spatial variation in the process of interest, in that its spatial autocorrelation can actually be captured through the covariates, provided a sufficiently complex model is used to represent the relation between covariates and the soil property of interest. In such cases, the spatial autocorrelation in the remaining residuals, i.e., those left after the covariates allow to capture much of this spatial autocorrelation in the process of interest, is sufficiently low that it can be ignored without too much degradation in model accuracy.

1.1.3 Why map soil?

Soil makes life on land possible. It not only supplies us with food, biomass, and raw materials but also plays a key role in regulating water, carbon, and nutrient cycles (Weil & Brady, 2017; Banwart et al., 2014; Barnes, 2015). It hosts over 25 percent of the biodiversity on the planet (FAO, 2020), and serves as the cornerstone of the food chains that support both human sustenance and above-ground biodiversity (Barnes, 2015). Soils are also a source of physical and cultural heritage, and act as a platform for man-made structures. With soils representing the largest terrestrial carbon pool on Earth (Poggio et al., 2021), they play a crucial role in mitigating climate change. It is for these reasons,

[1] A division of pedometrics, `http://pedometrics.org/`

and the fact that the process of soil formation takes thousands of years to produce just a few centimeters, that preserving this essential resource is vital.

The importance of soil is recognised by governments and policy makers. The European Union (EU), for example, has adopted in 2021 the EU Soil Strategy[2] that envisions that by 2050 all EU soil ecosystems are in healthy condition. This new vision is also anchored in the Climate Adaptation Strategy[3]. In Africa, initiatives like the Soils4Africa project[4], led by ISRIC[5], aim to establish an open-access soil information system, complete with a methodology for consistently monitoring soil conditions throughout the continent. It is important to map soil as it provides valuable information about the spatial distribution of soil properties that is also used in fields such as agronomy, ecology, water- and land management, and climate studies.

Soil properties that are typically mapped include soil organic carbon concentration and corresponding stocks, pH and soil texture (Wadoux et al., 2020a). It is also important to map nutrients such as nitrogen, potassium and phosphorus, as well as soil conditions such as soil thickness. In this thesis I use machine learning to address certain challenges when mapping soil organic carbon, nitrogen, clay and soil thickness for different regions across the African, European and North-American continents.

1.1.4 Opportunities for machine learning in digital soil mapping

Practitioners in DSM are faced with various challenges (Wadoux et al., 2021), and with many of these challenges it is not clear how to address the problem when a machine learning model is used. In this section I review four challenges, but the reader should note that there are many others, and in Chapter 6 I will provide an overview of these.

One significant challenge in DSM is that soil data are often contaminated with measurement errors. A measurement error is the difference between the actual and recorded value of a soil property. When we have error-contaminated soil observations we would like to take the measurement errors into account and make predictions that reflect the underlying error-free process of interest. With geostatistical models, measurement errors can be accounted for with methods such as measurement error-filtered kriging as introduced by Cressie (1993), or by simply adding variances of the measurement errors to the diagonal of the variance-covariance matrix of the kriging system (Delhomme, 1978). Although accounting for measurement errors in statistical models is a well-established domain (Buonaccorsi, 2010; Schennach, 2016), in the DSM literature, however, little work has been done on the incorporation of measurement errors in soil data with machine

[2] https://environment.ec.europa.eu/topics/soil-and-land/soil-strategy_en

[3] https://climate.ec.europa.eu/eu-action/adaptation-climate-change/eu-adaptation-strategy_en

[4] https://www.soils4africa-h2020.eu/

[5] https://www.isric.org/about

learning. In addition, research is required to investigate a sound statistical framework that allows machine learning models to account for measurement errors. This could be useful not just to DSM, but also other research domains that face problems concerning measurement errors.

It is often the case that there is more than one soil property of interest in a DSM project. Calibrating a different model for each is not just cumbersome, but if each soil property is modelled independently, thereby ignoring the covariance structure between the soil properties, it may lead to inconsistent predictions (Heuvelink et al., 2016). Multivariate mapping refers to the simultaneous prediction of several soil properties with a single statistical model. In the geostatistical literature multivariate modelling of soil properties is possible with methods such as co-kriging (Goovaerts, 1999), or regression co-kriging in the case the trend is not assumed constant. The use of machine learning models to map more than one soil property is less common, which is surprising especially since models such as random forests have multivariate extensions. Yet to the best of my knowledge the multivariate counterpart of the random forest has not been used to map several soil properties simultaneously.

Data on soil thickness are often right-censored, which means that the true soil thickness is not known and that it is greater than the sampling depth. Many DSM studies have attempted to model soil thickness, but in many of these cases, the censored data are treated as if they were true measurements which then might severely underestimate the actual soil thickness. A machine learning model that can account for right-censored data is a random survival forest model, but studies that have employed this model only used it to map the probability that soil thickness exceeds certain depths (Chen et al., 2019). In Malone & Searle (2020a) the authors stated that the results for mapping soil thickness with this model produced poor results, and so they modelled soil thickness with regular random forest models. Therefore, there is no current literature in DSM in which machine learning is used to map soil thickness itself while also accounting for the censored nature of the data. In addition, since current literature suggest that the random survival forest model performs poorly when modelling soil thickness, alternative methods might also be required.

It is often required in DSM to be able to provide insight and explain the outputs of a prediction model. However, recall that machine learning is generally regarded as a black-box which makes interpreting such a model difficult. Explainable machine learning (XML) is a rapidly growing field which focuses on methods to understand the predictions made by machine learning (Biecek & Burzykowski, 2021). In spatial prediction, XML methods can give interpretations on the overall behaviour of the model over the entire region of interest (global explanations) and/or interpretations for predictions for specific locations (local explanations). XML methods can either be model-specific or model-agnostic, which means that the method might either depend on the architecture of a certain machine

learning model, or is independent of it. Useful XML methods are model-agnostic and provide intuitive visualisations on how a certain covariate might influence the predictions of a soil property. However, these methods usually require covariates to be uncorrelated, and only one or two covariates can be visualised at a time. Therefore, a method is needed that is not only not hindered by the requirement of uncorrelated covariates, but also is able to visualise more than two covariates at a time.

1.2 Objectives and research questions

The overarching aim of this thesis is to show how the use of machine learning can be enhanced when faced with certain challenges. To be clear, the aim is not to provide an exhaustive list of solutions, but rather to illustrate how machine learning can be adequately implemented to tackle these challenges. I have identified four objectives, each corresponding to the four opportunities discussed in the preceding section. The following objectives incorporate research questions that I explore in this thesis.

1. Machine learning for measurement error-contaminated soil data.
    - Can measurement errors in a soil property be effectively accounted for within a machine learning model?
    - Does the prediction accuracy increase when measurement errors are accounted for?
2. Multivariate random forest as a viable option for multivariate mapping of more than one soil property.
    - Is the covariance structure between soil properties maintained when modelled with a multivariate random forest model?
    - Does the multivariate random forest model produce maps with higher accuracy compared to when soil properties are mapped independently with separate univariate random forest models?
3. Accounting for right-censored soil thickness data with machine learning
    - Can a random survival forest be used to map soil thickness, and is it a viable option?
    - Is there an alternative machine learning method that can model and account for right-censored soil thickness data, and how would this method compare to the random survival forest model?
4. Visually explain machine learning predictions with respect to multiple covariates.
    - Is a biplot a viable option for explaining machine learning predictions?

- Are there situations in which biplots reveal trends that are not detected by other XML methods?

1.3 Thesis outline

This thesis comprises six chapters, including this introductory chapter. Chapter 2 introduces a novel statistical framework designed to account for measurement errors in a soil property with a machine learning model. The framework is tested with different machine learning models in a synthetic simulation study and in a case study of error-contaminated clay data from Namibia. In Chapter 3 the use of a multivariate random forest model is studied and explored in a case study of European soil organic carbon and nitrogen data. Chapter 4 deals with machine learning models to account for right-censored soil thickness data. The models are compared in a synthetic simulation study in which various parameters such as proportion of censored data, and different types of censoring are investigated. The models are also compared in two case studies, one from the United States of America, and one from Switzerland. Chapter 5 presents a biplot methodology to explain machine learning predictions. This method is also compared to other popular XML visualisation methods. Chapter 6 gives the conclusion of the thesis.

1.4 A note on mathematical notation

Chapters 2-5 introduce distinct statistical challenges across various domains. It is important for the reader to be aware that the mathematical notation employed in this thesis might vary from one chapter to the next. However, within each chapter, the notation is clearly explained, ensuring a seamless reading experience without disruption to the thesis's flow.

Chapter 2

Measurement error-filtered machine learning in digital soil mapping

This chapter is based on:

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**, p. e100572. doi: https://doi.org/10.1016/j.spasta.2021.100572.

2.1 Introduction

Soil maps contain vital information for fields such as ecology and agronomy and can be produced with statistical models in digital soil mapping (DSM) (McBratney et al., 2003; Goovaerts, 1999). Nelson et al. (2011) address four sources of error in the DSM error budget. These are: (1) Model error, which refers to covariates that do not fully explain the variation in the target quantity and error in the estimation of the model parameters; (2) Measurement error, the difference between the actual and recorded value of a soil property; (3) Positional (location) error, uncertainty about the locations of the soil samples and; (4) Covariate error, which represents errors in the environmental covariates. With the latter two less pronounced (Nelson et al., 2011), it is becoming clearer that as model error is in decline, due to more covariate data that are available and due to more complex models that are used, taking measurement errors into account will play an increasingly important role in the future of DSM.

It is often the case that models in DSM ignore measurement errors. This is important to note, because the degree of measurement errors in soil properties is also increasing. We see this for example when soil samples are obtained with low-cost techniques such as infrared spectroscopy instead of more expensive but also more accurate wet chemistry techniques (Nocita et al., 2015). Soil information is also more frequently obtained from citizen science (Rossiter et al., 2015), and these data are expected to be less accurate, because they mostly come from non-domain experts. Therefore, with a growing demand for more detailed soil maps which require higher sampling densities, the need to account for measurement errors is further amplified.

When we have measurement error-contaminated soil observations we would like to take the measurement errors into account and make predictions that reflect the underlying error-free process of interest. This can be achieved, in principle, when the measurement errors have mean zero and their measurement error variance (MEV) is known, in that, we can attach more weight to observations with lower MEVs compared to observations with larger MEVs. In the geostatistical literature, Delhomme (1978) was the first to introduce such an approach with a kriging method, which essentially adds the MEVs to the diagonal of the variance-covariance matrix of the kriging system. A similar approach was developed by Cressie (1993), called measurement error-filtered kriging, or simply filtered kriging (FK), which incorporates the MEVs directly in the variogram of the kriging system. Christensen (2011) also developed heterogeneous filtered kriging (HFK) for situations when the MEVs is not constant across sampling locations. Somarathna et al. (2018) proposed an approach by including the MEVs in the covariance structure of the random effect of a linear mixed model. In hydrology, similar approaches were developed by De Marsily (1986) and Mazzetti & Todini (2009) called kriging with external drift for uncertain data, or KEDUD.

Machine learning models have gained tremendous traction in DSM in the last two decades, and there are many examples of publications where machine learning is utilised in DSM (see Wadoux et al. (2020a) for an overview of various machine learning models that have been implemented in DSM). The reasons for an increased utilization of machine learning models in DSM is because machine learning models are data-driven, i.e. they are not as dependent on assumptions as traditional statistical models, and they can be easily implemented with diverse and large data sets.

Accounting for measurement errors in the response in statistical models is a well-established domain (Buonaccorsi, 2010; Schennach, 2016). In the DSM literature, however, little work has been done on the incorporation of measurement errors into machine learning models (Wadoux et al., 2020a). At the time of writing this paper, and to the best of our knowledge, we found only two examples of accounting for measurement errors with machine learning in DSM. In Wadoux et al. (2019) and Hengl et al. (2018) the authors used weights to account for measurement errors. Wadoux et al. (2019) calculated weights as a function of the variance of the measurements predicted from an infrared spectroscopy model relative to the variance of the measurements used to calibrate the spectroscopy model. These weights were then used in a loss function which the authors used to calibrate a convolutional neural network. In Hengl et al. (2018) the authors used weights, calculated as the inverse of the measurement error standard deviations, in a random forest (RF) model to account for measurement errors.

In this paper we expand on the idea of using weights to account for measurement errors associated with the soil property of interest, by using a maximum likelihood approach (Buonaccorsi, 2010). We present a two-stage maximum likelihood framework, called measurement error-filtered machine learning, that aims to filter out measurement errors through the incorporation of, assumed known, MEVs into weights for a given machine learning model (Buonaccorsi, 2010). In this likelihood framework the optimal weights are well defined, but require knowledge of the residual variance of the measurement error-free model. In Section 2.3.3 we describe the two-stage approach which allows us to appropriately estimate these optimal weights. We illustrate our proposed framework with RF (Breiman, 2001) and projection pursuit regression (PPR) (Friedman & Stuetzle, 1981) models. These models were deliberately chosen to not only fit into our framework, but also to represent two completely different areas of machine learning for the purpose of regression, i.e. RF as ensemble learning with trees, and projection pursuit regression as a feature-learning approach (similar to a simple neural network).

The article is structured in the following way: in Section 2.2 we discuss a general framework of spatial prediction with measurement error-free soil observations. In Section 2.3 we adapt the general framework to allow for measurement error-contaminated observations, and present our proposed methodology of measurement error-filtered machine learning. In Section 2.4 we provide a simulation study to compare error-filtered machine learning to

the conventional machine learning models, and we benchmark this against the comparison of a regression kriging (RK) model that filters out measurement errors with a residual maximum likelihood approach to a regular RK model. A real-world case study is then presented in Section 2.5 where we map topsoil clay content from error-contaminated observations in Namibia. In Section 2.6 we provide a general discussion, and in Section 2.7 we present a summary of our conclusions.

2.2 Spatial prediction of measurement error-free measurements

2.2.1 Spatial modelling framework

Formally, suppose that we have a spatial process of interest on a geographical region $\mathcal{D}$, say $\{Y(\mathbf{s}), \mathbf{s} \in \mathcal{D}\}$ which is characterised by a mean $\{m(\mathbf{s}), \mathbf{s} \in \mathcal{D}\}$, and a certain spatially varying residual process, $\{\varepsilon(\mathbf{s}), \mathbf{s} \in \mathcal{D}\}$, so that we have

$$Y(\mathbf{s}) = m(\mathbf{s}) + \varepsilon(\mathbf{s}); \forall \mathbf{s} \in \mathcal{D}. \tag{2.1}$$

Without additional information it is difficult to separate the mean from the residual, since the spatial autocorrelation in the residual process, $\varepsilon(\mathbf{s})$, can mimic a spatially varying mean. Additional information in the form of covariates can mitigate this challenge, in that $Y(\mathbf{s})$, conditional on the values of the covariates, can be described through

$$Y(\mathbf{s}) = g(\mathbf{x}(\mathbf{s})|\theta_g) + \eta(\mathbf{s}), \tag{2.2}$$

where $\mathbf{x}(\mathbf{s})$ is a p-dimensional vector of covariates at location $\mathbf{s}$. Here $g(\cdot)$ is a regression function which is assumed to be linear in the covariates, θ_g describes its parameters, and $\eta(\mathbf{s})$ is assumed to be a zero-mean second-order stationary Gaussian process. We typically use a geostatistical model such as RK to describe Eq. (2.2). See Section 2.2.2 for further details.

Suppose now that the covariates $\{\mathbf{x}(\mathbf{s}), \mathbf{s} \in \mathcal{D}\}$ are related (in a potentially complex way) to the mean of the process of interest and are also able to capture the spatial autocorrelation in the residual process well, then it is reasonable to expect that there exists some function $f(\cdot)$ for which

$$Y(\mathbf{s}) = f(\mathbf{x}(\mathbf{s})|\theta_f) + \xi(\mathbf{s}), \tag{2.3}$$

where the *excess* residual $\{\xi(\mathbf{s}), \mathbf{s} \in \mathcal{D}\}$ is close to Gaussian white noise. This formulation has motivated the application of highly flexible regression methods (i.e. machine learning models) which ignore the dependence between the residuals, which is weak or even absent in the formulation above, to problems in spatial prediction. See Section 2.2.3 for more information.

In the remainder of this section we focus on a scenario where we have error-free, partial information of the realisation of $\{Y(\mathbf{s}), \mathbf{s} \in \mathcal{D}\}$ in the form of measurements $y(\mathbf{s}_i)$, where $\mathbf{s}_i$ are the sample locations for $i = 1, \ldots, n$, and where we have complete information of $\{\mathbf{x}(\mathbf{s}), \mathbf{s} \in \mathcal{D}\}$ in the form of maps of environmental covariates. Depending on the relation between $Y(\mathbf{s})$ and $\mathbf{x}(\mathbf{s})$, we will use the observation pairs $\{(y(\mathbf{s}_1), \mathbf{x}(\mathbf{s}_1)), \ldots, (y(\mathbf{s}_n), \mathbf{x}(\mathbf{s}_n))\}$ to model Eq. (2.1) either with Eq. (2.2) or Eq. (2.3). Then in Section 2.3 we consider the case where our information of $\{Y(\mathbf{s}), \mathbf{s} \in \mathcal{D}\}$ is contaminated with measurement errors.

The primary objective of spatial prediction in DSM is to obtain a model and an associated error-free *prediction function*, $\hat{y}(\cdot)$, which can be used to predict the realisation of $\{Y(\mathbf{s}), \mathbf{s} \in \mathcal{D}\}$ at various *prediction locations* in $\mathcal{D}$. Suppose that we use $\mathbf{s}_0$ to represent such a prediction location, then we minimise the *expected loss*, given by

$$E[L(\hat{y}(\mathbf{s}_0), y(\mathbf{s}_0))], \tag{2.4}$$

where $L(\cdot, \cdot)$ is a chosen loss function.

2.2.2 Regression kriging

Suppose that we wish to model a spatial response variable with a RK model such as described in Eq. (2.2). A state-of-the-art approach to estimate the required parameters for the RK predictor is to use the empirical best linear unbiased predictor (EBLUP) with REML (Stein, 1999; Lark et al., 2006). REML estimates the variogram model parameters with maximum likelihood that is independent of the trend parameters. The variogram parameters are then used to derive the covariances which are used to obtain the EBLUP. For a detailed outline of this approach we refer to Stein (1999); Harville (1977), but we also provide an overview in Web Appendix A. We just note that the RK predictor is given by

$$\hat{y}_{RK}(\mathbf{s}_0) = \mathbf{x}(\mathbf{s}_0)^T \hat{\boldsymbol{\beta}} + \hat{\mathbf{c}}(\mathbf{s}_0)^T \hat{\mathbf{C}}^{-1} \boldsymbol{\eta}, \tag{2.5}$$

where $\mathbf{x}(\mathbf{s}_0)$ is a $p + 1$ vector of a 1 and the covariate values at prediction location $\mathbf{s}_0$, $\hat{\boldsymbol{\beta}}$ are the trend parameters estimated with generalised least squares (GLS), the i-th element of $\hat{\mathbf{c}}(\mathbf{s}_0)$ is the estimated covariance of $Y(\mathbf{s}_i)$ with $Y(\mathbf{s}_0)$, $\hat{\mathbf{C}}$ is the estimated variance-covariance matrix, and $\boldsymbol{\eta}$ is a vector of n regression residuals at the sample locations. Note that we retain the use of RK in the subscript in Eq. (2.5) so that it can be distinguished from filtered regression kriging (FRK), a RK model that filters out measurement errors which will be discussed in Section 2.3.2.

2.2.3 Machine learning

When the covariates are spatially autocorrelated and also highly correlated to the residual process $\{\varepsilon(\mathbf{s}), \mathbf{s} \in \mathcal{D}\}$, then the relation in Eq. (2.1) can be described with Eq. (2.3). Therefore, with a machine learning model one is often able to make accurate predictions

of the process of interest without explicitly incorporating the spatial autocorrelation that it inherits from the residual process. Following from Eq. (2.3) the explicit structure of θ_f will depend on the specific machine learning model that is used (for example, see sections on the RF and PPR models below), but in general we can estimate θ_f with maximum likelihood. Therefore, for a given set of observation locations $\mathbf{s}_1, \ldots, \mathbf{s}_n$, the log-likelihood of the n observations $y(\mathbf{s}_1), \ldots, y(\mathbf{s}_n)$ associated with Eq. (2.3) is given as

$$\mathcal{L}(\mathbf{X}, \mathbf{y}, \theta_f) = -\frac{n}{2}\log\left(\sigma_\xi^2\right) - \frac{n}{2}\log\left(2\pi\right) - \frac{1}{2}\sum_{i=1}^{n}\frac{(y(\mathbf{s}_i) - f(\mathbf{x}(\mathbf{s}_i)|\theta_f))^2}{\sigma_\xi^2}, \tag{2.6}$$

where the $\mathbf{x}(\mathbf{s}_i)$ are the covariate values at sample locations $\mathbf{s}_i$, σ_ξ^2 is the common variance of i.i.d. Gaussian random variables $\xi(\mathbf{s}_i)$, and $\mathbf{X}$ is a $n \times p$ data matrix. The non-constant part with respect to θ_f is then proportional to

$$-\sum_{i=1}^{n} w(\mathbf{s}_i)(y(\mathbf{s}_i) - f(\mathbf{x}(\mathbf{s}_i)|\theta_f))^2, \tag{2.7}$$

where we can interpret $w(\mathbf{s}_i)$ as a weight, given by

$$w(\mathbf{s}_i) = \frac{1}{\sigma_\xi^2}. \tag{2.8}$$

Note that in the current formulation the weights are equal. We introduce these weights here so that in Section 2.3.3 we can adjust them to account for measurement errors by incorporating the measurement error variances, which will differ over different $\mathbf{s}_i$. The maximum likelihood estimation of θ_f can be obtained by maximising Eq. (2.7). Since the weights are all equal the optimisation can be done without first estimating σ_ξ^2.

Random forests

A RF model is a modification of bootstrap aggregation or bagging (Breiman, 2001). Bagging is a method that fits a certain model, such as regression trees, many times over bootstrapped training samples in order to improve the model's accuracy and to reduce its variance (Hastie et al., 2008). RF provides an improvement over bagging regression trees by decorrelating the trees, in that at each split only a subset of m_{try} of the p covariates is considered which leads to trees that are less correlated in their predictions.

If we consider that $f(\cdot)$ in Eq. (2.3) is a RF model then $\boldsymbol{\theta}_{RF}$ would denote the split variables, split points and terminal-node values which define the trees. In addition to $\boldsymbol{\theta}_{RF}$, one also needs to find the optimal number of covariates to subset at each split, m_{try}, the number of trees to grow in each forest, and the minimal node size which sets the depth of the trees in the forest. These hyper-parameters can be found, for example with cross-validation, so that the expected loss in Eq. (2.4) is minimized.

Projection pursuit regression

If we consider that $f(\cdot)$ in Eq. (2.3) is a PPR model, then $f(\cdot)$ is estimated with an additive model which is fitted in a forward step-wise manner (Friedman & Stuetzle, 1981), in that

$$f(\mathbf{x}(\mathbf{s}_i)) = \sum_{m=1}^{M} h_m(\boldsymbol{\varphi}_m^T \mathbf{x}(\mathbf{s}_i)). \tag{2.9}$$

The function, $h_m(\cdot)$, is a smoothing function and the $\boldsymbol{\varphi}_m$, for $m = 1, \ldots, M$, are unit p-vectors of unknown parameters. The $h_m(\boldsymbol{\varphi}_m^T(\cdot))$ is called a ridge function since it varies only in the direction defined by the vector $\boldsymbol{\varphi}_m$, for $m = 1, \ldots, M$, where M is the number of added pairs of $(h_m, \boldsymbol{\varphi}_m)$. The variable $\boldsymbol{\varphi}_m^T \mathbf{x}(\mathbf{s}_i)$ is the projection of $\mathbf{x}(\mathbf{s}_i)$ onto the unit vector $\boldsymbol{\varphi}_m$. The two steps of estimating h_m and $\boldsymbol{\varphi}_m$ are done with adding one pair $(h_m, \boldsymbol{\varphi}_m)$ at a time in a forward fashion. We can denote these parameters, h_m and $\boldsymbol{\varphi}_m$, as $\boldsymbol{\theta}_{PPR}$. The number of pairs, M, can be selected with cross-validation so that the expected loss in Eq. (2.4) is minimised. For more details concerning PPR we invite the reader to refer to Friedman & Stuetzle (1981); Hastie et al. (2008).

2.3 Spatial prediction of measurement error contaminated measurements

2.3.1 Adapted spatial modelling framework

Suppose that measurement errors, $\delta(\mathbf{s}_i)$, assumed to be independent over the sample locations, $\mathbf{s}_i$, are zero-mean Gaussian random variables, each with known but possibly different variance $\sigma_\delta^2(\mathbf{s}_i)$. Let $\mathbf{V}$ be the $n \times n$ variance-covariance matrix of the vector of n measurement errors, which has the $\sigma_\delta^2(\mathbf{s}_i)$ on the diagonal and zeros on the off-diagonals. We then have

$$Z(\mathbf{s}_i) = Y(\mathbf{s}_i) + \delta(\mathbf{s}_i), \tag{2.10}$$

where $Z(\mathbf{s}_i)$ is a measurement error-contaminated random variable which we observe (Buonaccorsi, 2010). By using Eq. (2.10) we can write Eq. (2.2) and Eq. (2.3) at the sample locations $\mathbf{s}_i$ as

$$Z(\mathbf{s}_i) = g(\mathbf{x}(\mathbf{s}_i)|\theta_g) + \eta(\mathbf{s}_i) + \delta(\mathbf{s}_i), \tag{2.11}$$

and

$$Z(\mathbf{s}_i) = f(\mathbf{x}(\mathbf{s}_i)|\theta_f) + \xi(\mathbf{s}_i) + \delta(\mathbf{s}_i). \tag{2.12}$$

We further assume that the $\delta(\mathbf{s}_i)$ are independent of $\eta(\mathbf{s}_i)$ and $\xi(\mathbf{s}_i)$.

We now use observation pairs $\{(z(\mathbf{s}_1), \mathbf{x}(\mathbf{s}_1)), \ldots, (z(\mathbf{s}_n), \mathbf{x}(\mathbf{s}_n))\}$ to fit the models in Sections 2.3.2 and 2.3.3, along with the known values of $\sigma_\delta^2(\mathbf{s}_i)$ to take the measurement errors associated with $y(\mathbf{s}_i)$ into account. Our aim is still to minimise the expected loss in Eq. (2.4).

2.3.2 Filtered regression kriging

With RK measurement errors can be filtered out by adding the assumed known MEVs to the diagonal of $\mathbf{C}$ (Delhomme, 1978; Somarathna et al., 2018). In this paper we will refer to this as FRK. We can estimate the required variogram parameters with REML by simply replacing $\mathbf{C}$ with the variance-covariance matrix of the measurement error-contaminated process $Z(\mathbf{s})$, that is $\mathbf{C}_z = \mathbf{C} + \mathbf{V}$. We can then use GLS to estimate the regression parameters, $\hat{\boldsymbol{\beta}}_z$. The FRK predictor is then given by

$$\hat{y}_{FRK}(\mathbf{s}_0) = \mathbf{x}(\mathbf{s}_0)^T\hat{\boldsymbol{\beta}}_z + \hat{\mathbf{c}}(\mathbf{s_0})^T\hat{\mathbf{C}}_z^{-1}\boldsymbol{\eta}_z, \tag{2.13}$$

where the i-th element of $\boldsymbol{\eta}_z$ is $(z(\mathbf{s}_i) - \hat{y}(\mathbf{s}_i))$. The predictor in Eq. (2.13) will filter out the measurement errors by giving more weight to observations with lower values for $\sigma^2_\delta(\mathbf{s}_i)$ compared to observations with higher values for $\sigma^2_\delta(\mathbf{s}_i)$. A more detailed outline of FRK is supplied in Web Appendix B.

2.3.3 Measurement error-filtered machine learning

In order to obtain a prediction function $\hat{y}(\cdot)$ for a machine learning model with observation pairs $\{z(\mathbf{s}_i), \mathbf{x}(\mathbf{s}_i))\}$, for $i = 1, \ldots, n$, we need to filter out the measurement errors associated with $y(\mathbf{s}_i)$. This can be achieved by incorporating $\sigma^2_\delta(\mathbf{s}_i)$ into the weights first introduced in Eq. (2.7) and Eq. (2.8) (Buonaccorsi, 2010). The log-likelihood of $p(z(\mathbf{s}_i)|\mathbf{x}(\mathbf{s}_i), \theta_f)$ is now given by

$$\begin{aligned}\mathcal{L}(\mathbf{X}, \mathbf{z}, \theta_f) = &-\frac{1}{2}\sum_{i=1}^{n}\log(\sigma^2_\xi + \sigma^2_\delta(\mathbf{s}_i)) - \frac{n}{2}\log(2\pi) \\ &- \frac{1}{2}\sum_{i=1}^{n}\frac{(z(\mathbf{s}_i) - f(\mathbf{x}(\mathbf{s}_i)|\theta_f))^2}{\sigma^2_\xi + \sigma^2_\delta(\mathbf{s}_i)}.\end{aligned} \tag{2.14}$$

Notice that where previously we were able to estimate θ_f without knowledge of σ^2_ξ, since the solution was based only on minimising the sum-of-squared residuals, now this is no longer the case since the denominators in the second term in the log-likelihood are not all equal. We therefore follow an alternating procedure in which first a pilot estimate for the parameters in θ_f is obtained, say $\bar{\theta}_f$, by calibrating the model in a standard way on the observed pairs $\{(z(\mathbf{s}_1), \mathbf{x}(\mathbf{s}_1)), \ldots, (z(\mathbf{s}_n), \mathbf{x}(\mathbf{s}_n))\}$, similar to the approach discussed in Section 2.2.3. We then use this to estimate σ^2_ξ by using the conditional log-likelihood in which θ_f is kept fixed at $\bar{\theta}_f$, i.e., by minimising over σ^2_ξ in the objective

$$\sum_{i=1}^{n}\log(\sigma^2_\xi + \sigma^2_\delta(\mathbf{s}_i)) + \sum_{i=1}^{n}\frac{\left(z(\mathbf{s}_i) - f(\mathbf{x}(\mathbf{s}_i)|\bar{\theta}_f)\right)^2}{\sigma^2_\xi + \sigma^2_\delta(\mathbf{s}_i)}. \tag{2.15}$$

Since this is a univariate problem it can easily be solved. With this estimate for the residual variance, say $\hat{\sigma}^2_\xi$, we can then estimate θ_f by minimising the weighted sum of the squared residuals, i.e.,

$$\sum_{i=1}^{n} w(\mathbf{s}_i) \left(z(\mathbf{s}_i) - f(\mathbf{x}(\mathbf{s}_i)|\theta_f)\right)^2, \tag{2.16}$$

where $w(\mathbf{s}_i) = \left(\hat{\sigma}^2_\xi + \sigma^2_\delta(\mathbf{s}_i)\right)^{-1}$ for each $i = 1, ..., n$. Since $\hat{\sigma}^2_\xi$ is constant, measurements with larger values for $\sigma^2_\delta(\mathbf{s}_i)$ receive less weight and measurements with smaller values for $\sigma^2_\delta(s_i)$ receive more weight when θ_f is estimated. The resulting measurement error-filtered model's residual variance can then be estimated again by minimising Eq. (2.15), but replacing $\bar{\theta}_f$ with the new estimate for θ_f obtained as above. These steps can be repeated until convergence, each time using the updated pair of estimates for θ_f and σ^2_ξ in order to further refine the model. It is important to note that many models used in machine learning do not have convex objectives, and hence there is no guarantee that such an approach will always converge. Our experience with the chosen models is that in the case of RF the model stabilises after only a single iteration, suggesting that convergence tends to occur in practice despite the lack of convexity. However, in the case of PPR no such convergence seems to occur in practice. We therefore perform only a single iteration of this approach in our applications.

2.4 Synthetic simulation study

Synthetic data simulation experiments were performed in the `R` programming language (R Core Team, 2020) to evaluate the performance of the measurement error-filtered models, FRK, filtered PPR (FPPR) and filtered RF (FRF), relative to the regular models, RK, PPR and RF, i.e. where the measurement errors are ignored and not taken into account. The performance of these models was evaluated under different MEV scenarios regulated by three simulation parameters, $\{\mu_{\sigma^2_\delta}, \nu_{\sigma^2_\delta}, \breve{\nu}_{\sigma^2_\delta}\}$, similar to the simulation study performed in Christensen (2011). The simulation parameters will be used to simulate two sets of MEVs, that is, a "true" set and a set of "specified" MEVs.

The first two parameters are the average size of the true MEVs ($\mu_{\sigma^2_\delta}$) and the coefficient of variation of the true MEVs ($\nu_{\sigma^2_\delta}$), where the latter regulates the heterogeneity of the MEVs over the observations. Since the true MEVs will usually not be known exactly in a real-world analysis, we introduce an additional set of specified MEVs which will be used in the models. The third simulation parameter, $\breve{\nu}_{\sigma^2_\delta}$, will therefore regulate the uncertainty of the specified MEVs relative to the true MEVs. The methodology for the simulated experiments for a given set of simulation parameters is summarised in Figure 2.2.

In each experiment we simulated a synthetic data set on unit square discretised into a 40×40 grid comprising 1600 response values. The response values $y(\mathbf{s}_l)$ were generated from a Gaussian process with a covariance structure governed by an isotropic spherical variogram model with $\boldsymbol{\theta}_\gamma = \{c_0 = 0.1, c = 0.9, a = 0.2\}$, where a range of $a = 0.2$ corresponds to 8 cells in the grid. The covariate values were calculated using

$$x_1(\mathbf{s}_l) = \begin{cases} 2y(\mathbf{s}_l) + r(\mathbf{s}_l), & y(\mathbf{s}_l) \geq 0.5, \\ -2y(\mathbf{s}_l) + r(\mathbf{s}_l), & y(\mathbf{s}_l) < 0.5, \end{cases} \tag{2.17}$$

$$x_2(\mathbf{s}_l) = -\frac{1}{2}y(\mathbf{s}_l) + r(\mathbf{s}_l), \tag{2.18}$$

$$x_3(\mathbf{s}_l) = x_1(\mathbf{s}_l)x_2(\mathbf{s}_l), \tag{2.19}$$

for $l = 1, \ldots, N = 1600$, and where the residual values $r(\mathbf{s}_l)$, which are i.i.d., were generated from a standard Gaussian random variable. We simulated the synthetic data in this way to ensure that the simulation parameters, which regulate the MEV scenarios, are directly comparable to the response values, $y(\mathbf{s}_l)$. In addition, the purpose of this analysis is not to compare the performance between all types of models defining the relationship between the response variable and covariates, but rather to compare the performance of each measurement error-filtered model with that model's version that ignores the measurement errors. For this reason it is satisfactory to just have a few covariates to enable us to fit the various models. The covariates in Eqs. (2.17), (2.18) and (2.19), on average, describe about 40% of the variation in $y(\mathbf{s}_l)$. In Figure 2.1 we show the distribution of the generated response values and visualise the relationships between the response and the covariates for one experiment.

A sample of observations is obtained by randomly selecting 10% of the 1600 response variable values. We contaminated these observations with measurement errors by using Eq. (2.10). We used the approach implemented in Lingwall & Christensen (2007), and also used in Christensen (2011), to generate values for $\sigma_\delta^2(\mathbf{s}_i)$, in that at each sample location, $\mathbf{s}_i$, $\sigma_\delta^2(\mathbf{s}_i)$ was a random draw from a lognormal distribution with mean $\mu_{\sigma_\delta^2}$ and coefficient of variation $\nu_{\sigma_\delta^2}$. The specified MEVs, $\breve{\sigma}_\delta^2(\mathbf{s}_i)$, was a draw from a lognormal distribution with mean $\sigma_\delta^2(\mathbf{s}_i)$ and coefficient of variation $\breve{\nu}_{\sigma_\delta^2}$. We also consider a special case where the measurement errors, $\delta(\mathbf{s}_i)$, were drawn from a zero-mean Gaussian distribution with constant variance, σ_δ^2. In the simulation study this constant MEV, σ_δ^2, was set to $\mu_{\sigma_\delta^2}$ for a given experiment, and in order to determine $\breve{\sigma}_\delta^2$ for a given experiment, we calculated the mean of the values for $\breve{\sigma}_\delta^2(\mathbf{s}_i)$. The results for this case is discussed in Web Appendix C.

The models (RK, FRK, PPR, FPPR, RF and FRF) were fitted on $z(\mathbf{s}_i)$ along with the generated covariate values, by using the methodology discussed in Sections 2.3.2 and 2.3.3. In the case of the machine learning models we used the default value available in the `ppr` function in `R` for the number of added pairs for the PPR models, that is $M = 1$, and we used $m_{try} = 1$ for the RF models. For the RF models, we implemented

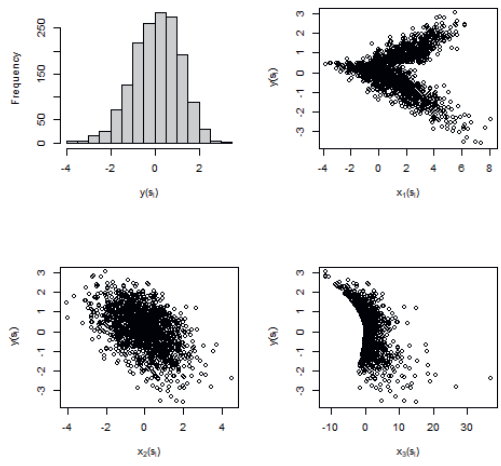


Figure 2.1: Distribution of response variable values and relationship with the covariates, as illustrated for one of the experiments.

the `ranger` package (Wright & Ziegler, 2017) to fit the RF models, and specifically the `case.weights` argument within the `ranger` function to incorporate the weights. This argument puts the weights on the sample observations in order to be selected for the bootstrap samples that are used to build the trees. In Web Appendix D we explain the similarity of our proposed methodology discussed in Section 2.3.3 as implemented through the `case.weights` argument in the `ranger` function. The effect of incorporating the weights in this way is however negligibly small. Note that for the measurement error-filtered models (FRK, FPPR and FRF) we used $\breve{\sigma}^2_\delta(\mathbf{s}_i)$ (or $\breve{\sigma}^2_\delta$ in the case of the constant MEV) in the models, and not $\sigma^2_\delta(\mathbf{s}_i)$ (or σ^2_δ). The latter is used to generate the measurement errors, but the former is considered as what is observed by the researcher, and hence used in the models.

Predictions were generated and the performance of the models was assessed with the mean square error (MSE). That is, in the case of generating predictions with RK, PPR and RF, i.e. where we ignore the measurement errors, we used

$$\frac{1}{N}\sum_{l=1}^{N}(y(\mathbf{s}_l) - \hat{z}(\mathbf{s}_l))^2, \tag{2.20}$$

for $l = 1, \ldots, N = 1600$, where $\hat{z}(\mathbf{s}_l)$ are the predictions of the region of interest of the model that ignores the measurement errors. For the measurement error-filtered models

FRK, FPPR and FRF we used

$$\frac{1}{N}\sum_{l=1}^{N}(y(\mathbf{s}_l)-\hat{y}(\mathbf{s}_l))^2, \tag{2.21}$$

for $l = 1, \ldots, N = 1600$, where the $\hat{y}(\mathbf{s}_l)$ are the measurement error-filtered model's predictions of the region of interest.

The MEV scenarios that we considered were all combinations of $\mu_{\sigma_\delta^2} = \{0.1, 0.25, 0.5, 1.0, 1.5\}$, $\nu_{\sigma_\delta^2} = \{0.1, 0.5, 1.0, 1.5\}$ and $\breve{\nu}_{\sigma_\delta^2} = \{0.0001, 1.0, 1.5\}$. Note that when $\breve{\nu}_{\sigma_\delta^2} = 0.0001$ the specified MEVs are considered as very reliable. Each experiment with a unique set of $\{\mu_{\sigma_\delta^2}, \nu_{\sigma_\delta^2}, \breve{\nu}_{\sigma_\delta^2}\}$ was repeated 300 times, which resulted in a total of 18000 experiments. Since the sill, $(c_0 + c)$, was set equal to 1, we can use the values of $\mu_{\sigma_\delta^2}$ in the simulation study to identify values for $\left(\frac{\mu_{\sigma_\delta^2}}{c_0+c}\right)$ where we expect the performance of each of the measurement error-filtered models to be superior to each of the respective regular model's performance.

The results for the spatially variable MEVs, $\sigma_\delta^2(\mathbf{s}_i)$, are shown in Table 2.1. The table values are the MSE values for the measurement error-filtered models, FRK, FPPR and FRF, relative to those of RK, PPR and RF. In each table we also show the ratios with the smallest MSE of the measurement error-filtered model in blue. The second- and third smallest values are shown in orange and red, respectively.

In Table 2.1, as $\mu_{\sigma_\delta^2}$ increases from 0.10 to 1.50 we observe the increasing advantage of filtering out measurement errors with FRK. This trend is seen for any given pair $\{\nu_{\sigma_\delta^2}, \breve{\nu}_{\sigma_\delta^2}\}$, but it is most notable when $\nu_{\sigma_\delta^2} = 1.50$ and $\breve{\nu}_{\sigma_\delta^2} = 0.0001$ with the RK ratio decreasing from 0.96 to 0.53. The improvement of FRK to RK is the smallest when $\nu_{\sigma_\delta^2} = 0.10$ and $\breve{\nu}_{\sigma_\delta^2} = 1.50$ with the ratio decreasing from 0.99 to 0.70. In general, we see larger improvements in FRK relative to RK for larger values of $\mu_{\sigma_\delta^2}$ for a given set of values for $\nu_{\sigma_\delta^2}$ and $\breve{\nu}_{\sigma_\delta^2}$. We also see larger improvements for larger values of $\nu_{\sigma_\delta^2}$ given $\mu_{\sigma_\delta^2}$ and $\breve{\nu}_{\sigma_\delta^2}$ and larger improvements for smaller values of $\breve{\nu}_{\sigma_\delta^2}$ given $\mu_{\sigma_\delta^2}$ and $\nu_{\sigma_\delta^2}$. This general trend can also be observed for the machine learning models.

For the measurement error-filtered machine learning models we only start to see notable improvements when $\nu_{\sigma_\delta^2} \geq 0.50$ and $\breve{\nu}_{\sigma_\delta^2} \leq 1.00$ (in the case of FPPR this is only at $\breve{\nu}_{\sigma_\delta^2} = 0.0001$). For any given value of $\breve{\nu}_{\sigma_\delta^2}$, and for increasing values in $\mu_{\sigma_\delta^2}$, the largest improvements are seen when $\nu_{\sigma_\delta^2} = 1.50$. We note that the effect of increasing $\mu_{\sigma_\delta^2}$ gradually disappears as the values for $\nu_{\sigma_\delta^2}$ become smaller. For example, when $\breve{\nu}_{\sigma_\delta^2} = 0.0001$, we see that, for $\nu_{\sigma_\delta^2} = 1.50$, notable improvements in FPPR and FRF relative to PPR and RF are already observed at $\mu_{\sigma_\delta^2} \geq 0.25$ in the case of FRF and at $\mu_{\sigma_\delta^2} \geq 0.50$ in the case of FPPR, but when $\nu_{\sigma_\delta^2} = 0.50$ notable improvements are only seen at $\mu_{\sigma_\delta^2} \geq 1.00$. When $\nu_{\sigma_\delta^2} = 0.10$ we do not observe improvements in FPPR and FRF with the ratios being mostly constant and close to 1.00. This is because the MEVs are still somewhat

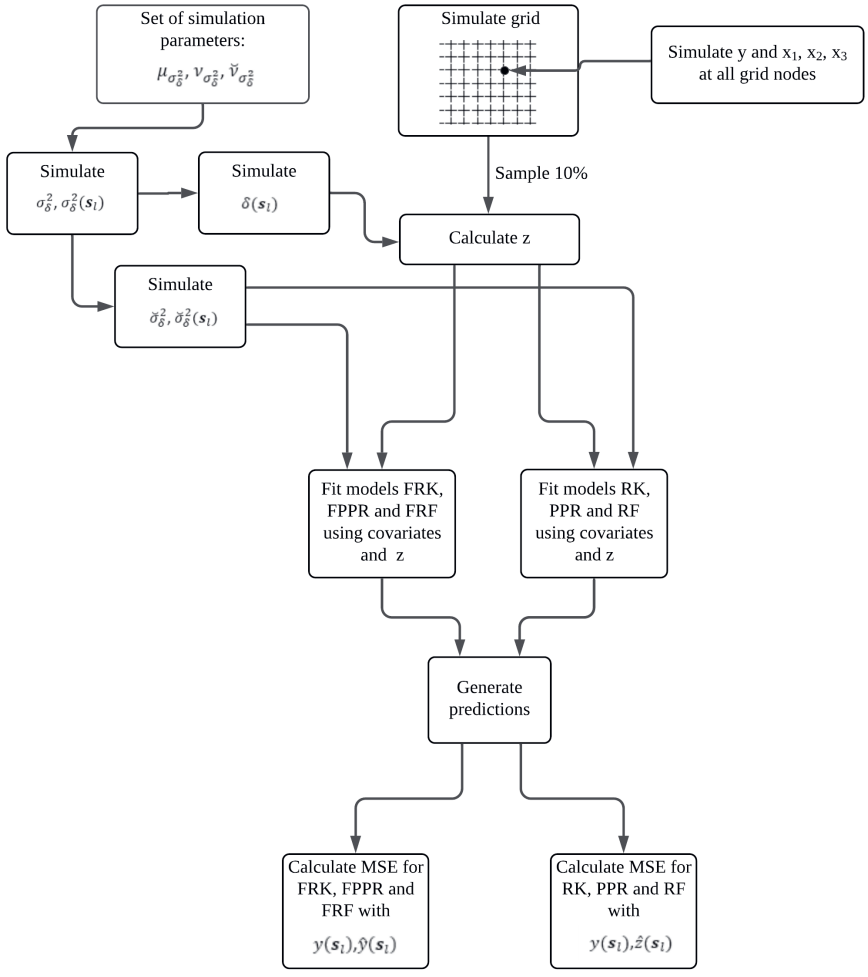


Figure 2.2: Flow diagram of methodology for the simulated experiments to evaluate the performance of measurement error-filtered models relative to regular models.

homogeneous with the standard deviation being only 10% of $\mu_{\sigma_\delta^2}$ for the values of $\sigma_\delta^2(\mathbf{s}_i)$. We even see FPPR and FRF performing worse when $\nu_{\sigma_\delta^2} = 0.10$, with the PPR and RF ratios increasing to 1.01 when $\breve{\nu}_{\sigma_\delta^2} = 1.00$, and the PPR ratio increasing to 1.03 and the RF ratio increasing to 1.02 when $\breve{\nu}_{\sigma_\delta^2} = 1.50$.

It is evident that the degree at which we observe an improvement of FRK relative to RK is much higher compared to the PPR and RF models. For example, when $\nu_{\sigma_\delta^2} = 1.50$ and $\breve{\nu}_{\sigma_\delta^2} = 0.0001$, we note that the RK ratio decreases from 0.96 to 0.53 while the PPR ratio decreases from 0.99 to 0.94, and the RF ratio from 0.99 to 0.88. It therefore appears that FRK is able to take more advantage of incorporating the MEVs to filter out the measurement errors than FPPR and FRF.

The effect of larger values for $\breve{\nu}_{\sigma_\delta^2}$ on the performance of FRK, FPPR and FRF becomes smaller as $\mu_{\sigma_\delta^2}$ decreases. For a given value of $\nu_{\sigma_\delta^2}$, if we look for example at the ratios where $\mu_{\sigma_\delta^2} = 0.10$ and compare it to the ratios where $\mu_{\sigma_\delta^2} = 1.50$ we note larger changes in the results between different values for $\breve{\nu}_{\sigma_\delta^2}$ when $\mu_{\sigma_\delta^2} = 1.50$. There are a few exceptions when $\nu_{\sigma_\delta^2} \leq 0.50$ where different values for $\breve{\nu}_{\sigma_\delta^2}$ does not lead to different results for different values in $\mu_{\sigma_\delta^2}$. For example, we see that when $\nu_{\sigma_\delta^2} = 0.50$ and $\breve{\nu}_{\sigma_\delta^2} \geq 1.00$ the PPR and RF ratios remain more or less constant and close to 1.00 regardless of $\mu_{\sigma_\delta^2}$.

The negative effect of larger values for $\breve{\nu}_{\sigma_\delta^2}$ is also greater on the performance of FRK relative to RK compared to FPPR to PPR and FRF to RF (this becomes more notable for larger values of $\mu_{\sigma_\delta^2}$ and $\nu_{\sigma_\delta^2}$). For example, we see that when both $\mu_{\sigma_\delta^2}$ and $\nu_{\sigma_\delta^2}$ are equal to 1.50, the RK ratio increases from 0.53 to 0.60 as $\breve{\nu}_{\sigma_\delta^2}$ increases from 0.0001 to 1.50, while the PPR ratio increases from 0.94 to 0.96 and the RF ratio from 0.88 to 0.91. However, this could just be due to the RK models starting from a much lower ratio, and therefore having more to lose when there is a decrease in the accuracy of the specified MEVs.

Table 2.1: Results of model performance based on the case where the values for $\delta(\mathbf{s}_i)$ were drawn from a multivariate, zero-mean Gaussian distribution with variance-covariance matrix, $\mathbf{V}$, with $\sigma^2_\delta(\mathbf{s}_i)$ on the diagonal. The table values are MSE values for FRK, FPPR and FRF relative to the MSE values of RK, PPR and RF, respectively. The results are shown for different values of the mean, $\mu_{\sigma^2_\delta}$, and for the coefficient of variation, $\nu_{\sigma^2_\delta}$ of the MEVs, and for different values of $\breve{\nu}_{\sigma^2_\delta}$, which regulates the variability of $\breve{\sigma}^2_\delta$ and $\breve{\sigma}^2_\delta(\mathbf{s}_i)$ as estimates of the "true" MEVs, σ^2_δ and $\sigma^2_\delta(\mathbf{s}_i)$. For each simulation parameter set the ratio including the measurement error-filtered model with the smallest MSE is shown in blue, the second smallest in orange, and the third smallest in red.

$\breve{\nu}_{\sigma^2_\delta}$		0.0001			1.00			1.50		
$\mu_{\sigma^2_\delta}$	$\nu_{\sigma^2_\delta}$	RK	PPR	RF	RK	PPR	RF	RK	PPR	RF
0.10	0.10	0.98	0.99	1.00	0.99	1.00	1.00	0.99	1.00	1.00
0.25	0.10	0.93	0.99	1.00	0.95	1.00	1.00	0.96	1.00	1.00
0.50	0.10	0.84	0.99	1.00	0.88	1.01	1.01	0.90	1.01	1.01
1.00	0.10	0.69	0.99	1.00	0.75	1.01	1.01	0.80	1.02	1.01
1.50	0.10	0.59	1.00	1.00	0.67	1.01	1.01	0.70	1.03	1.02
0.10	0.50	0.98	0.99	0.99	0.98	0.99	1.00	0.99	0.99	1.00
0.25	0.50	0.92	0.99	0.99	0.94	0.99	1.00	0.95	1.00	1.00
0.50	0.50	0.82	0.99	0.99	0.86	1.00	0.99	0.89	1.00	1.00
1.00	0.50	0.67	0.98	0.97	0.74	1.01	0.99	0.77	1.01	1.00
1.50	0.50	0.58	0.98	0.97	0.64	1.00	0.98	0.68	1.01	1.00
0.10	1.00	0.97	0.99	0.99	0.98	0.99	1.00	0.98	0.99	1.00
0.25	1.00	0.90	0.99	0.99	0.92	0.99	0.99	0.94	0.99	1.00
0.50	1.00	0.80	0.98	0.97	0.84	0.99	0.98	0.85	0.99	0.98
1.00	1.00	0.66	0.98	0.94	0.70	0.99	0.96	0.72	0.99	0.97
1.50	1.00	0.56	0.96	0.91	0.60	0.97	0.94	0.63	0.98	0.95
0.10	1.50	0.96	0.99	0.99	0.97	0.98	0.99	0.98	0.99	0.99
0.25	1.50	0.89	0.99	0.98	0.91	0.99	0.99	0.92	0.99	0.99
0.50	1.50	0.78	0.97	0.96	0.81	0.98	0.97	0.84	0.99	0.97
1.00	1.50	0.64	0.96	0.92	0.67	0.97	0.93	0.69	0.97	0.94
1.50	1.50	0.53	0.94	0.88	0.57	0.95	0.91	0.60	0.96	0.91

The FRK model outperforms FPPR and FRF for all MEV scenarios. This is partly due to the nature of the assumed model for the response variable and synthetic data, but we would like to reiterate that the purpose of this analysis was not to compare the performance between FRK, FPPR and FRF. However, it is interesting to note that when $\nu_{\sigma^2_\delta} \leq 0.50$, for any given value for $\breve{\nu}_{\sigma^2_\delta}$, FRF outperformed FPPR when $\mu_{\sigma^2_\delta} \leq 0.25$, while for larger values, that is, when $\mu_{\sigma^2_\delta} \geq 0.50$ it is the opposite. There are exceptions, for example when $\nu_{\sigma^2_\delta} = 1.50$, where FRF outperforms FPPR when $\mu_{\sigma^2_\delta} = 0.50$.

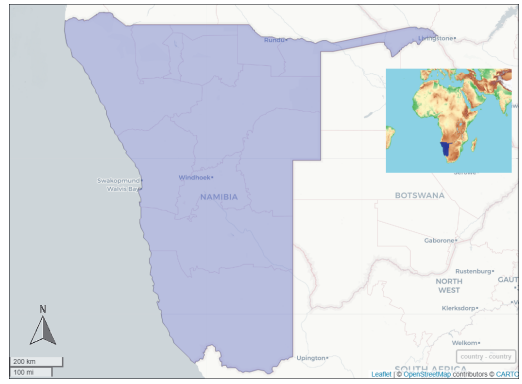


Figure 2.3: Region of interest, Namibia, used for the real-world case study.

2.5 Real-world case study

To illustrate the measurement error-filtered models in a real-world application we consider a data set of uncertain topsoil clay content of Namibia. Namibia (see Figure 2.3) is located in South-Western Africa and has a surface area of 823 680 km^2. The climate is mostly arid and rainfall is mostly limited to the northern regions of the country mainly as summer storms, from September to February.

The soil data were derived from the African Soil Profiles (AfSP) database (Leenaars et al., 2014) (laboratory analyses of 53 soil samples) and the Land Potential Knowledge System (LandPKS) database (Herrick et al., 2016) (310 field measurements). The original LandPKS data set included texture classes at three layers, 0-1cm, 1-10cm, and 10-20cm. We used an internal *TT2tri* function, available in Hengl (2021), to derive the mean and standard deviation for all clay measurements, by assuming a uniform distribution within each soil texture class.

The AfSP data also consisted of a reliability class which ranked from 1 (most reliable) to 4 (least reliable). Expert judgement was used to estimate the interquartile range of topsoil clay content for each reliability class. These estimates were 2% for class 1, 4% for class 2, 8% for class 3 and 16% for class 4. The MEVs were determined by dividing the interquartile range by 1.34 and then taking the square.

The soil property of interest, topsoil (0-20cm) clay content (measured in %), is an important soil attribute in land management and land use, because it affects the water holding capacity and hydraulic properties of soil (Jabro, 1992), and soil fertility (Prasad & Power, 1997). Clay refers to particles smaller than 0.002 mm. The clay observations are shown in Figure 2.5a. It can be seen that the sampled locations are well dispersed throughout the country, but with fewer to no observations in the southern, north-western and coastal regions. Clay observations range from 1.10% to 55.50%, with a mean of 13.55%

and a standard deviation of 9.89%. In the upper margin of Figure 2.4 we note that clay content is skewed to the right with most observations less than 15%. The spatial variation was modelled with a spherical variogram model and estimated with REML as discussed in Section 2.3.2. The variogram parameters were estimated as being equal to $\boldsymbol{\theta}_\gamma = \{58.34, 36.00, 10012.50\}$.

Figure 2.4 indicates a weak to moderate linear relationship between clay content and clay uncertainty. The estimated correlation coefficient between clay content and clay uncertainty is 0.46. The correlation is mostly due to the relationship between clay content and clay uncertainty in the LandPKS field measurement data. On the other hand, there does not appear to be a relationship between clay content and uncertainty in the laboratory measurements of the AfSP data.

The spatial distribution of clay uncertainty is presented in Figure 2.5b, in which the uncertainty is shown as MEVs. These variances range from 2.20% to 55.80%, with a mean of 7.76%. On the right margin of Figure 2.4 we note that clay uncertainty is skewed to the right with most observations having a MEV less than $12\%^2$. The coefficient of variation of the MEVs, which is equal to 0.72, indicates that the heterogeneity of the clay uncertainty over the region of interest is not very large. The ratio of the mean of the MEVs to the variance of the clay observations was estimated to be equal to

$$\frac{\hat{\mu}_{\sigma_\delta^2}}{\hat{c}_0 + \hat{c}} = 0.08. \tag{2.22}$$

We can use the coefficient of variation of the MEVs and Eq. (2.22) along with the results in Table 2.1 to get an expectation of how the measurement error-filtered models will perform relative to models that ignore measurement errors. We see that we fall in a region of ($\mu_{\sigma_\delta^2} \approx 0.1; \nu_{\sigma_\delta^2} = \{0.5; 1\}$) in Table 2.1. Thus, we expect a small improvement for the FRK and FPPR models and a very small improvement, if any, for the FRF model. The performance of the measurement error-filtered models will also depend on the reliability of the MEVs, but if the MEVs in this analysis deviate at most by 150% from the "true" MEVs, then the reliability of the MEVs should not have a large effect on the results.

In this analysis we compare the predictions of RK, FRK, PPR, FPPR, RF and FRF. We had 109 covariates (at $250m \times 250m$), provided by ISRIC, which included maps of various vegetation indices, a digital elevation model, precipitation and surface temperatures, largely the same as those used in SoilGrids (Poggio et al., 2021). The measurement error-filtered models were fitted as discussed in Sections 2.3.2 and 2.3.3. To select and assess the models we created 100 independent 80%/20% training and validation sets. In each training set we performed 5-fold cross-validation to select the hyper-parameters of the machine learning models, and to select the best subset of covariates for the RK models. Note that we used a measurement error-adjusted MSE for model selection (see discussion below). The models were then assessed by generating predictions for the validation sets. To evaluate model performance, we used the mean error (ME), MSE and concordance

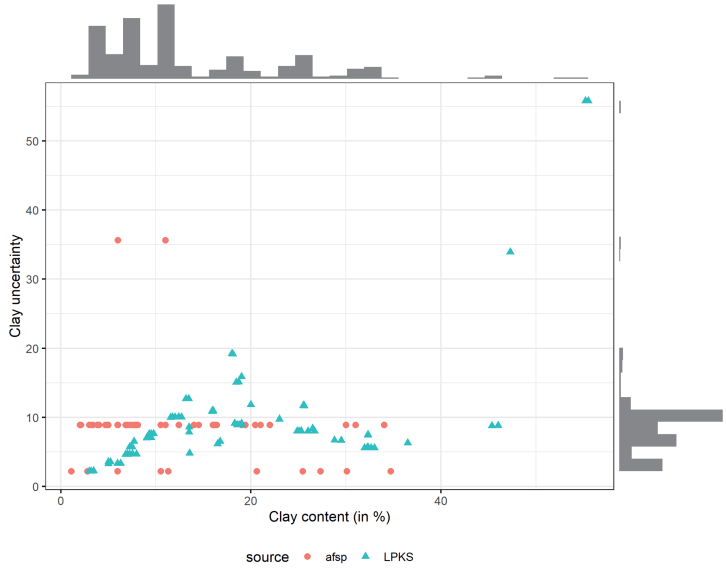


Figure 2.4: Scatter plot of clay uncertainty with clay content. Margins show histograms depicting the marginal distributions.

correlation coefficient (CCC) (Lawrence & Lin, 1989). The final results for model performance were then obtained by taking the average of the 100 independent validation sets.

In this case study the true clay contents at validation locations are unknown and we only know the measurement error-contaminated observations. It is therefore necessary to adjust the validation statistics by taking into account the measurement errors associated with the validation data. The adjusted ME, MSE and CCC were used in this case study and the derivation of these are discussed in Web Appendix E.

The overall prediction results are shown in Tables 2.2 and 2.3. Note that we took the square root after we averaged the MSE values and present the root MSE (RMSE). As previously mentioned we expect only small improvements in the measurement error-filtered models relative to the models that do not incorporate MEVs, mostly due to low variability of the MEVs and the low ratio of the means of the MEVs to the variance of the underlying spatial process.

In terms of the ME, which measures overall bias, we can see that FPPR and FRF lead to less biased predictions than PPR and RF, respectively. We also note that all of the machine learning models slightly overestimates clay content while RK and FRK underestimates it. The ME of RK is the closest to zero while the ME for FRK is equal to -0.12. In terms of RMSE, only FRK and FPPR show an improvement relative to their

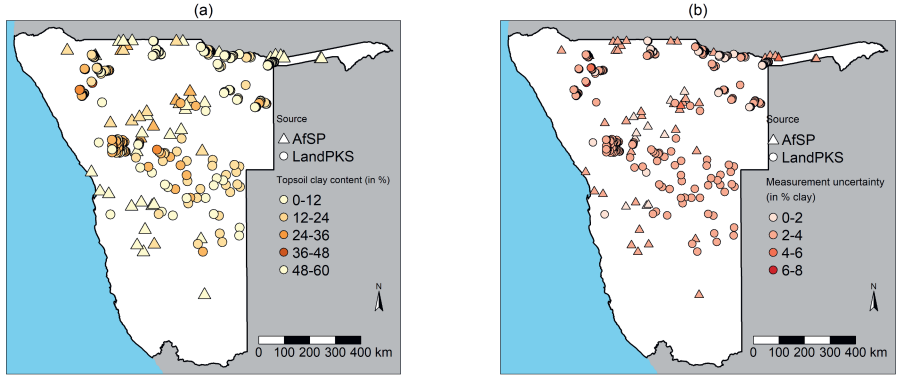


Figure 2.5: (a) Topsoil clay content observations from AfSP and LandPKS. (b) Measurement error standard deviation for topsoil clay content from AfSP and LandPKS.

conventional counterparts, with the RMSE decreasing from 8.84 to 8.82 for RK models, and from 11.21 to 10.91 for the PPR models. These are very small differences that may be due to chance affects and effectively mean that measurement error-filtered models do not improve prediction performance. FPPR showed the largest improvement relative to its conventional counterpart compared to the other measurement error-filtered models, but also here the improvement was only marginal. The RF models did not show an improvement, but did perform overall best with the smallest RMSE value of 8.03 for the RF model, and second smallest RMSE value of 8.09 for FRF. In terms of the association between the clay predictions and observations we observe no differences in the values for the CCCs for the RK and PPR models. The CCC for the FRF model is slightly lower compared to the CCC for the RF model.

Table 2.2: Topsoil clay content prediction results. Model performance results for the Namibia real-world case study. For each model the ME, RMSE and CCC are shown.

	RK	FRK	PPR	FPPR	RF	FRF
ME	-0.01	-0.12	0.48	0.24	0.20	0.11
RMSE	8.84	8.82	11.21	10.91	8.03	8.09
CCC	0.32	0.32	0.30	0.30	0.45	0.43

We also generated predictions for the entire region of interest by using the optimal model parameters obtained during model selection. The prediction maps are shown in Figure 2.6, and the five-point summary statistics of the predictions of each model is provided in Table 2.3. We also show the prediction differences between each error-filtered model and its conventional counterpart in Figure 2.7.

Table 2.3: Topsoil clay content prediction results. Five-point summary statistics for prediction maps of all six models.

	RK	FRK	PPR	FPPR	RF	FRF
min	-22.66	-23.42	5.97	4.96	4.47	4.03
q1	10.28	9.98	8.19	7.50	11.09	11.06
q2	12.79	12.64	8.79	8.11	13.01	12.97
q3	17.65	17.48	10.16	12.06	15.74	15.42
max	48.12	47.00	46.92	32.52	29.46	24.98

On a national scale it is difficult to observe any notable differences between the RK and FRK maps. When we observe the RK and FRK summary statistics in Table 2.3 and the map of differences in Figure 2.7(a), we observe overall lower predictions for FRK than RK. The lower predictions of clay content for FRK occur mostly in the northern and south-eastern regions. Local differences where FRK produced higher predictions for clay content compared to RK can be observed inland, parallel to the coastline.

The PPR model generally produced higher clay content predictions, especially for the top 25%. This can be seen in the five-point summary statistics where the third quartile and maximum for PPR are 10.16 and 46.92, while for the FPPR model, these are 12.06 and 32.52. In Figures 2.6(c) and (d), we can also see this when we compare the central and north-western regions of Namibia. In Figure 2.7(b) we observe greater local differences between PPR and FPPR throughout the country, but especially in the far southern, central, and north-western regions. In the southern region we observe mostly higher PPR clay content predictions, while in the central region we mostly observe higher predictions for FPPR. In the northern regions the local differences are more abrupt, and along the coast in the northern region we again observe higher clay predictions for FPPR.

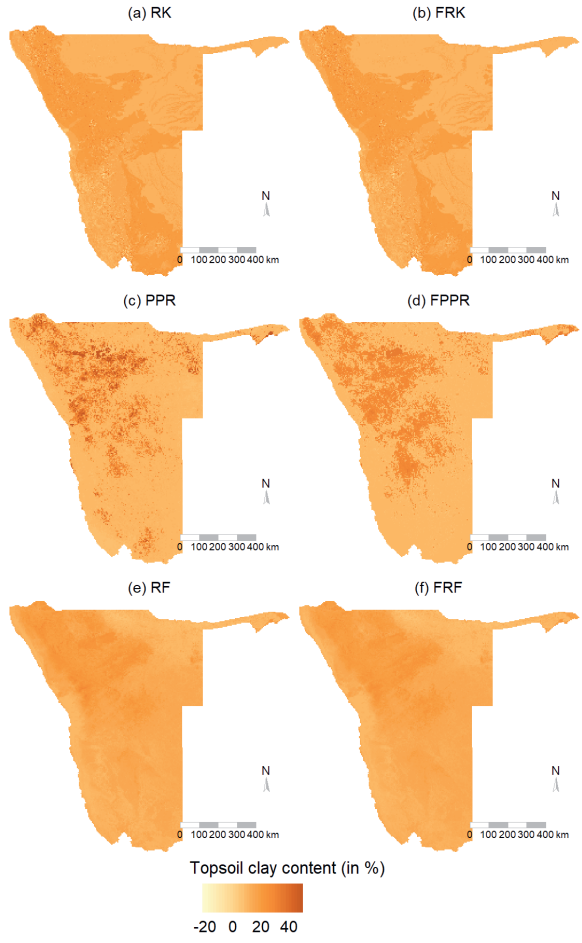


Figure 2.6: Topsoil clay content predictions for (a) RK; (b) FRK; (c) PPR; (d) FPPR; (e) RF; (f) FRF.

As mentioned earlier, RF and FRF performed best. This can be seen in Figures 2.6(e) and (f) where it is also difficult to note any notable differences between RF and FRF. As noted with the other measurement error-filtered models, FRF also produced overall lower predictions for clay content compared to its conventional counterpart, RF. In the five-point summary statistics we also observe very similar results, except for the maximum that is much lower for FRF than for RF. In Figure 2.7(c) we see that RF produced higher predictions for clay content along the coast in the northern and central regions, while FRF mostly produced higher values to the east. We note that the differences between the PPR models are much more abrupt compared to the RK and RF models. We also

observe larger local differences between PPR and FPPR compared to the RK and RF models.

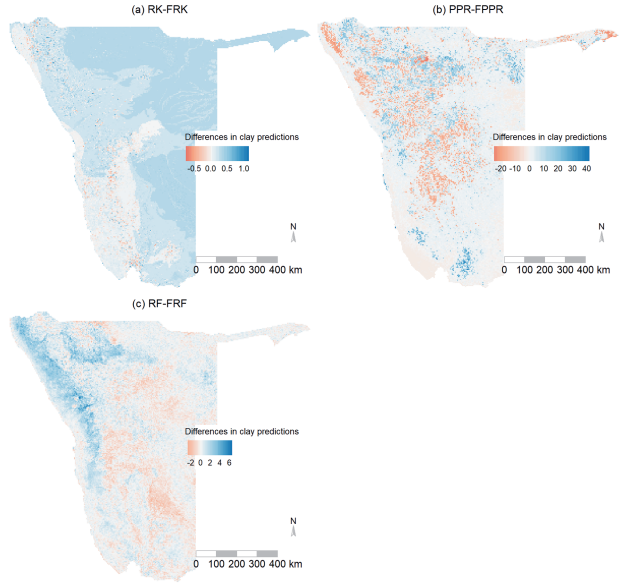


Figure 2.7: Prediction differences between each of the model pairs. (a) RK-FRK; (b) PPR-FPPR; (c) RF-FRF.

2.6 General discussion

2.6.1 Findings of this research

We introduced a novel, formalised framework for dealing with measurement errors in DSM when using a machine learning model. We acknowledge that both Wadoux et al. (2019) and Hengl et al. (2018) already used machine learning models to account for measurement errors in DSM. Wadoux et al. (2019) used weights to account for measurement errors with a convolutional neural network, while Hengl et al. (2018) used weights in a RF model. We formalised the use of weights to account for measurement errors with a machine learning model with a maximum likelihood framework (Buonaccorsi, 2010). As opposed to the weights used in Wadoux et al. (2019) and in and Hengl et al. (2018), the weights used in our proposed framework will minimise the loss function in Eq. (2.14), and allow for more efficient estimates of the model parameters, θ_f. Our framework is a two-stage approach in which first an initial estimate of the model parameters, $\bar{\theta}_f$, is obtained, which is then used to estimate the residual variance, σ^2_ξ, by minimising Eq. (2.15). The estimate of the residual variance can then be used to find a new estimate of the model parameters, θ_f

by minimising Eq. (2.16). One can alternate between these two steps until satisfactory estimates are obtained.

In addition to the proposed framework, this is also the first study (to the best of our knowledge) where a comprehensive simulation study was performed to investigate how RK and machine learning models perform under various MEV scenarios. A similar simulation study was performed in Christensen (2011), but here the author compared FK, HFK and HFK with variance-stabilising transformations.

From the synthetic simulation study it was seen that the average MEV size as well as the relative variability of the MEVs, had a significant effect on the results of the error-filtered models. This was also the case in Christensen (2011). The effect of increasing the average size of the MEVs was greater for FRK compared to FPPR and FRF. For FPPR and FRF there was also a stronger interaction between the average size and the relative variability of MEVs. That is, when the MEVs were less variable, we did not observe the same improvement of FRK over RK for an increasing size in the MEVs, compared to FPPR relative to PPR, and compared to FRF relative to RF. We only observed an increased improvement in FPPR and FRF for larger MEVs when the MEVs were relatively more variable. This is due to the weights that are incorporated in Eq. (2.16) which remain relatively constant when the MEVs are less variable.

The reliability of the MEVs had a minimal effect on the prediction results, especially for the machine learning models. This is because we compared error-filtered models to the conventional models that do not filter out measurement errors. Therefore, it does not matter to an extent (i.e. values of $\breve{\nu}_{\sigma^2_\delta}$ up to 150%) if the MEVs were reliable or not so reliable. Even if the specified MEVs were less reliable, for example, suppose that the true MEVs are small, the specified MEVs would also overall be considered small. In Christensen (2011) it was noted that the effect of changing $\breve{\nu}_{\sigma^2_\delta}$ was more notable on the prediction results, but the author compared HFK to FK, two error-filtered models, and not an error-filtered model to a conventional model as done in our paper.

It is also important to note that error-filtered machine learning models may perform worse than the conventional machine learning models when the average (relative to the variance of the underlying spatial process) and relative variability of the MEVs are small, and when the reliability of the MEVs is low. Researchers should therefore not incorporate weights to counter for measurement errors before establishing that it will be beneficial to do so. In DSM we often have poor information about the MEVs, and this may have been indeed the case for the Namibia case study. It is therefore important that soil data producers not only report their soil measurements, but also the uncertainty of the measurements (Laslett & McBratney, 1990; Van Leeuwen et al., 2021).

In the synthetic case study we looked at a scenario where the response values were Gaussian realisations which may take on negative values. In addition to the simulation study in Section 2.4, we also performed a simulation study where the response values were also

Gaussian realisations, but the negative values were set to 0.01. It is a typical scenario in DSM to have a response variable to be non-negative. The results and discussion are shown in Web Appendix F.

2.6.2 Limitations and future research

In our proposed methodology in Section 2.2.3 we assumed a constant variance for $\xi(\mathbf{s}_i)$ in Eq. (2.3). This is not realistic, because machine learning models will usually perform better in different parts of the study area (or in different parts of the feature space). Further research is therefore needed to investigate how to adapt the proposed framework to accommodate for a non-constant variance, that is $\sigma^2_\xi(\mathbf{s}_i)$. In the case of the RF model, a possible solution would be to use the variance derived from quantile regression forest (Meinshausen, 2006). In addition, the RF and PPR models used in this study do not have convex objectives, and therefore the two-stage alternating procedure in our proposed framework may not converge. In case of the RF model this did not appear to be much of a problem. Even though we did see improvements in FPPR relative to PPR in the simulation study and in the case study, further research will be required to investigate how often PPR would converge, and how the framework may be adapted to allow for similar models.

It was seen in the Namibia case study that the error-filtered models generally produced lower predictions compared to the conventional models. This was also noted in Hengl et al. (2018) where the predictions for the weighted RF were lower than the predictions of the unweighted RF model. In our case study this was especially evident for the RK models, as seen in Figure 2.7a where we observe lower predictions for FRK (i.e. blue regions). This could be due to the correlation that exists between the soil observations and the MEVs, as shown in Figure 2.4. A possible way to deal with this is to decorrelate the MEVs and the soil observations before they are entered into the models. Christensen (2011) addressed this problem by using variance-stabilising transformations to decorrelate the measurements and the MEVs.

In the Namibia study, where we mapped clay content, we did not observe any notable improvements in the error-filtered machine learning models. The main contributing factors to the under-performance of the error-filtered models, were the small MEVs (relative to the variance of the underlying spatial process) and the low variability of the MEVs. It will therefore be important to investigate these characteristics of the MEVs for other soil properties and other spatial domains, in order to obtain a knowledge pool of the performance of error-filtered machine learning models with various soil properties and spatial domains.

No notable global differences were observed for prediction accuracy between the error-filtered models and the conventional models in the Namibia case study. A similar conclusion was made by Somarathna et al. (2018) where the authors compared a linear mixed

model which accounts for measurement errors to a linear mixed model that ignores the measurement errors when mapping soil organic carbon. In our paper, we did observe greater local differences, especially for the PPR and RF models. It would therefore seem that the error-filtered machine learning models might deem some covariates more, or less important. Further research might therefore be necessary to understand which covariates are deemed important between the error-filtered and the conventional machine learning model, and why.

It is important to note that positional error can be translated into measurement error. When a perfect measurement at a location, say $\mathbf{s} + \mathbf{a}$, is taken to represent a soil property at location $\mathbf{s}$, with $\mathbf{a}$ a possible positional error, then short-distance spatial variation means that the soil property at $\mathbf{s}$ may differ from that at $\mathbf{s} + \mathbf{a}$. Therefore, even though the measurement at location $\mathbf{s} + \mathbf{a}$ is error-free, it still represents the soil property value at $\mathbf{s}$ with error. In light of this we could have augmented the MEV, but also accounting for positional error was outside the scope of this paper.

Accounting for measurement errors may also lead to a decrease in the prediction variance of a model. In Somarathna et al. (2018) the authors noted a significant decrease in prediction uncertainty when accounting for measurement errors in a linear mixed model. In this paper we did not investigate the effect of accounting for measurement errors on the prediction variances. This is because the prediction variances of machine learning models are not readily available, unless additional methods such as described in Hengl et al. (2018) are used.

Finally, the way we computed the validation metrics in a case where validation data have measurement errors, is sub-optimal. This is because ideally one would like to give a more accurate measurement not only a larger weight in prediction, but also in validation. Further research may therefore be required to investigate ways of improving validation with error-contaminated validation data.

2.7 Conclusions

In this paper we introduced a two-stage maximum likelihood framework to deal with measurement errors in DSM when using machine learning models. In this framework the MEVs are incorporated as weights into the log-likelihood function so that measurements with larger MEVs receive less weight when the machine learning model is calibrated. We illustrated our framework with the PPR and RF models, and we named these models FPPR and FRF. The performance of FPPR and FRF was also compared to that of FRK, a regression kriging model that incorporates the MEVs through REML. We have shown, by performing a comprehensive simulation study and by analysing a real-world case study, that by incorporating MEVs into a machine learning model can lead to increase in prediction accuracy. In particular, our results from the simulation study and the real-world

case study lead us to understand that the average size and the degree of the variability of the MEVs had a significant effect on the results of FRK, FPPR and FRF. That is, larger, and more heterogeneous MEVs lead to greater performance of these models. The reliability of the MEVs did not have such a great effect on model performance.

Supplementary materials

The supplementary materials, Web Appendices A - F, can be downloaded from the journal version of this chapter:
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**, p. e100572. doi: `https://doi.org/10.1016/j.spasta.2021.100572`.

Chapter 3

Multivariate random forest for digital soil mapping

This chapter is based on:

van der Westhuizen, S., Heuvelink, G.B.M., Hofmeyr, D.P. (2023). Multivariate random forest for digital soil mapping. *Geoderma*, **431**, p. e116365. doi: https://doi.org/10.1016/j.geoderma.2023.116365.

3.1 Introduction

Soil maps, and corresponding uncertainty maps, are important inputs that are used in various fields in environmental modelling, and can be effectively produced with digital soil mapping (DSM) (McBratney et al., 2003). It has become increasingly popular in DSM to produce these maps with machine learning models such as random forest (RF) (Breiman, 2001), and in the case of uncertainty maps, quantile regression forest (QRF) (Meinshausen, 2006). In such cases, however, soil properties are usually modelled in a univariate manner, that is, each soil property is modelled independently and as such, the covariance structure between soil properties is ignored (Wadoux et al., 2020a). Modelling soil properties in this way may lead to inconsistent predictions when maps of multiple soil properties are produced (Heuvelink et al., 2016), or to inconsistent stochastic simulations, as for example needed in Monte Carlo uncertainty propagation analyses (Heuvelink, 1998; Heuvelink et al., 2007; Van den Berg et al., 2012). In the case of modelling soil organic carbon (SOC) and total nitrogen (TN), Heuvelink et al. (2016) has found that when SOC and TN are mapped independently it may lead to unrealistic carbon-nitrogen (C:N) ratios. Maintaining a realistic C:N ratio is important because it influences the rate of residue decomposition and the nitrogen cycle in the soil (Weil & Brady, 2017).

In DSM, when multiple soil properties are mapped in a multivariate manner the underlying covariance or correlation structure between the soil properties is taken into consideration. Several examples of multivariate mapping in DSM exist, with co-kriging probably the most widely used technique (Goovaerts, 1999; Heuvelink et al., 2016; Vasat et al., 2010; Ge et al., 2015). Co-kriging, a geostatistical approach, explicitly models the underlying spatial autocorrelation of the soil properties as well as the correlation between the soil properties, but it requires certain strict conditions to be met (Goovaerts, 1999). Other examples of multivariate mapping in DSM include structural equation modelling, which can explicitly incorporate pedological knowledge allowing it to predict more than one soil property simultaneously (Angelini et al., 2017), multivariate adaptive regression splines (Söderström et al., 2016; Nawar et al., 2015), and partial least squares regression (Nawar et al., 2015). The use of machine learning models to simultaneously map more than one soil property is less common, which is somewhat surprising as there are several multivariate extensions of popular machine learning models used in DSM today. Examples of mapping multiple soil properties with machine learning include the use of convolutional neural networks (Wadoux, 2019; Padarian et al., 2019), and to use convolutional neural networks to map a soil property at multiple depths (Wadoux et al., 2019). However, other popular machine learning models in DSM, such as support vector regression and RF, both of which can be extended to handle multivariate responses (Xu et al., 2013; Segal & Xiao, 2011; Cevid et al., 2022; Ishwaran et al., 2008), have to the best of our knowledge not been used for multivariate soil mapping. In this paper we only focus on RF as it is the most preferred machine learning model in DSM today (Wadoux et al., 2020a). Multivariate prediction

with RF has been implemented in other fields such as economic forecasting (Pierdzioch & Risse, 2020), biochemistry (Swanson et al., 2012), drug sensitivity prediction (Wan & Pal, 2013) and ecology (Miller et al., 2014).

In this paper we illustrate the multivariate extension of RF, denoted as MRF, with simultaneous modelling of topsoil (0-20 cm) SOC and TN. We compare the multivariate approach to the alternative in which SOC and TN are modelled independently with two separate RF models. In the same way we also compare regression co-kriging (RCK) to two separate regression kriging (RK) models. We decided to use SOC and TN, because they are closely related and their ratio has considerable influence on soil quality (Reeves, 1997; Bonfante et al., 2019; Srivastava et al., 2017). However, other multivariate soil properties could also have been selected for this study, such as soil nutrients, heavy metals, soil hydrological properties or soil biological properties.

Multivariate models will be compared to their univariate counterparts by means of stochastic simulations conditional on the covariates and the calibration data. We will compare how well the models are able to reproduce the joint distribution of SOC and TN at a specific point in geographic space by analysing the correlation structure of SOC and TN in the simulations, and consequently also analyse how this influences the distribution of the C:N ratio. We anticipate that a model which incorporates the correlation structure between soil properties should be able to better maintain the correlation structure in its outputs as opposed to a model which deals with each soil property independently. Note that this comparison cannot be done with the predictions as the predictions correspond to the conditional expectations of SOC and TN. However, in addition to comparing the models on how the reproduce the joint distribution of SOC and TN, we also evaluated the models in terms of prediction accuracy by computing common univariate and multivariate cross-validation statistics for the predictions of SOC and TN.

3.2 Materials and Methods

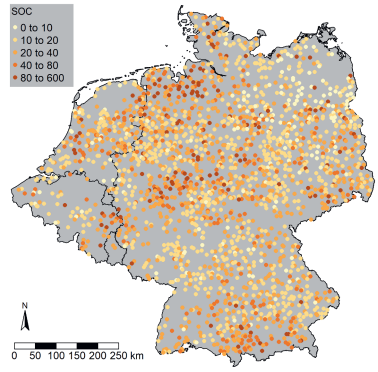
3.2.1 Soil profile data and environmental covariates

In this study we used topsoil (0-20 cm) SOC and TN observations from the 2018 Land Use and Coverage Area frame Survey[1] (LUCAS). LUCAS for soil has soil observations at about 22 000 locations across Europe, and also produces observations on other physicochemical properties of soil such as coarse fragments, particle-size distribution and phosphorus content (Orgiazzi et al., 2018). We only considered the region consisting of Belgium, the Netherlands, Luxembourg (Benelux Union) and Germany, an area of approximate size 433 703 km^2. This region has 2 216 locations (on average 1 sampling location per 195 km^2), shown in Figure 3.1, where both SOC and TN were measured. It should be noted that

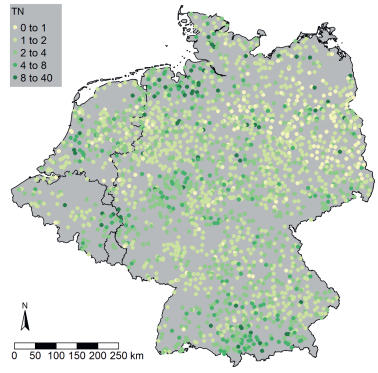
[1] `https://esdac.jrc.ec.europa.eu/projects/lucas`

some LUCAS locations might only have a SOC or only a TN observation, and in such a case, for the purpose of using a multivariate machine learning model, the observation is discarded. In Figure 3.1 we noted the sample locations are well dispersed throughout the region of interest, and that the observations have a moderate spatial correlation in the region of interest. With regards to the C:N ratio, shown in Figure 3.1c, we noted higher C:N ratios in the north-western and north-eastern regions of Germany, as well as in the eastern part of the Netherlands.

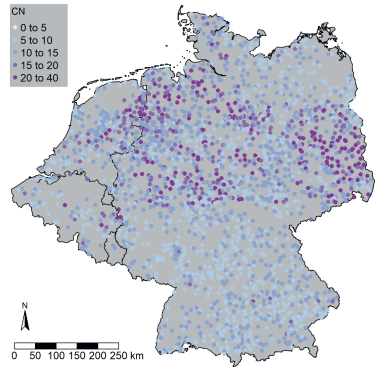
A collection of 173 covariates (at $1\,km \times 1\,km$ spatial resolution) was used to represent the soil-forming factors. These covariates include maps of different vegetation indices, elevation, surface temperatures, and precipitation, mostly the same as the ones used in SoilGrids (Poggio et al., 2021). Of this list, 48 covariates were removed due to near-zero variance. We also removed a set of 20 covariates that were highly correlated with another covariate (i.e., a correlation coefficient of more than 0.90). Therefore, the models in the subsequent sections were calibrated on 105 covariates.



(a)



(b)



(c)

Figure 3.1: LUCAS observation locations in Benelux and Germany. All 2 216 locations have both a (0-20 cm) SOC and a TN observation from which we also determined the C:N ratio. (a) SOC observations ($g.kg^{-1}$). (b) TN observations ($g.kg^{-1}$). (c) C:N ratios.

3.2.2 Regression kriging and regression co-kriging

In this section, and for the rest of the paper, vectors are written as lower case boldface letters, matrices are written as capital boldface letters, and random variables and vectors are indicated by non-boldface upper case letters. For a detailed outline of RK and RCK we refer the reader to Webster & Oliver (2007); Stein & Corsten (1991), but we also provide an overview in Web Appendix A. Here, for RK we note that, given a vector $\mathbf{z}$ of response values available at the n observation locations, $\mathbf{s}_1, \ldots, \mathbf{s}_n$, the best linear unbiased prediction at a prediction location, $\mathbf{s}_0$, is given by

$$\hat{z}(\mathbf{s}_0) = \mathbf{x}(\mathbf{s}_0)^T\hat{\boldsymbol{\beta}} + \mathbf{c}^T\mathbf{C}^{-1}\left(\mathbf{z} - \mathbf{X}\hat{\boldsymbol{\beta}}\right), \tag{3.1}$$

where $\mathbf{x}(\mathbf{s}_0)$ is a $p+1$ vector of covariate values at prediction location $\mathbf{s}_0$ with p the number of covariates, $\mathbf{C}$ is an $n \times n$ variance-covariance matrix of the residuals, η, at the observation locations, $\mathbf{c}$ is a vector of covariances between the residuals at the observation locations and the prediction location, and $\mathbf{X}$ is the $n \times (p+1)$ design matrix. It should be noted that η is assumed here to be a zero-mean, isotropic and second-order stationary stochastic residual with a spatial correlation characterised by a variogram. In addition, both $\mathbf{c}$ and $\mathbf{C}$ are determined from the variogram of η, and $\hat{\boldsymbol{\beta}}$ is the generalised least squares (GLS) estimate of the trend.

With RCK, we have q spatially and cross-correlated response variables, assembled in a vector $Z^c(\mathbf{s})$ for all $\mathbf{s} \in \mathcal{D}$, that must be jointly modelled (Pebesma, 2004). The multivariate best linear unbiased prediction of $\mathbf{Z}_0^c$ is given by

$$\hat{\mathbf{Z}}_0^c = (\mathbf{X}_0^c)^T\hat{\boldsymbol{\beta}}^c + (\mathbf{C}_0^c)^T(\mathbf{C}^c)^{-1}\left(\mathbf{Z}^c - \mathbf{X}^c\hat{\boldsymbol{\beta}}^c\right), \tag{3.2}$$

where $\hat{\boldsymbol{\beta}}^c$ is determined with the GLS estimator (Stein & Corsten, 1991). For the definitions of $\mathbf{X}_0^c$, $\hat{\boldsymbol{\beta}}^c$, $\mathbf{C}_0^c$, $\mathbf{C}^c$, $\mathbf{Z}^c$ and $\mathbf{X}^c$ we refer the reader to Web Appendix A.

Implementation of the kriging models

Before the kriging models were implemented we transformed the SOC and TN responses by using the (natural) *log* transformation. This is because, and we discuss this in Section 3.3.1, both SOC and TN have positively-skewed distributions. The kriging models, as well as the RF and MRF models discussed in the following subsection, were calibrated and assessed with a 10-fold nested cross-validation. In the outer-loop of the nested cross-validation the entire data set is 10 times repeatedly divided into a training and a test set. Then, for the inner-loop, the training data is 10 times repeatedly divided into a calibration and validation set. The calibration and validation sets are used to calibrate and select the optimal model, and then, the selected model is assessed with the outer-loop's test set. For the RK and RCK models, a model was calibrated by selecting the best subset

of covariates with stepwise regression with the Akaike's information criterion (Bozdogan, 1987), as well as by accounting for multicollinearity by removing covariates with variance inflation factors above 10.

Within a particular calibration, the regression coefficients and the variograms and cross-variograms of the residuals were obtained with an iterative estimation method that involves first estimating the variograms and cross-variograms from the ordinary least squares regression residuals, next estimating the GLS regression coefficients, recomputing the residuals and refitting the variograms on the GLS residuals, and so on. The procedure is repeated until convergence is achieved, which usually requires only a few iterations. This procedure is widely used in geostatistics (Webster & Oliver, 2007). To obtain the variance-covariance matrices in Eq. (3.1), we used a spherical variogram to model the spatial correlation of the residuals. In addition, to obtain the variance-covariance matrices in Eqs. (3.2), we used the linear model of coregionalisation. This procedure requires variograms and cross-variograms to be fitted with an appropriate variogram model, as described in Pebesma (2004); Wackernagel (2010), so that all variograms and cross-variograms were fitted with the spherical shape and have the same range.

We used the functionality in the `gstat` package (Pebesma, 2004) in `R` (R Core Team, 2020) to fit the RK and RCK models, as well as to simulate realisations conditional on the calibration data using sequential Gaussian simulation (Webster & Oliver, 2007; Heuvelink et al., 2016). To fit the variograms we used the `automap` package (Hiemstra et al., 2008) and to fit the cross-variogram we implemented the default settings with the `gstat` package.

3.2.3 Random forest

Overview of random forest and multivariate random forest

RF, a popular model for regression problems in DSM, uses a large collection of tree-based regression models. Tree-based models partition the covariate space, $\mathcal{R}^p$, into multiple non-overlapping (multidimensional) rectangles, say $R_1, \ldots, R_L$. Each rectangle is associated with a terminal node, or leaf, of a tree, and the trees are fit by recursive binary splitting based on minimising a measure of *impurity* (Hastie et al., 2008). The impurity of a leaf is determined in relation to the distribution of the response variable associated with the observations whose covariate vectors lie therein. In particular a leaf has a high degree of impurity if the associated responses have a high degree of dispersion, which in the case of regression is most frequently determined by the SS criterion. The impurity of a (leaf) node, say $l \in \{1, \ldots, L\}$, is thus defined by

$$SS_l = \sum_{i:\mathbf{x}(\mathbf{s}_i)\in R_l} (z(\mathbf{s}_i) - \bar{z}_l)^2, \tag{3.3}$$

where $\bar{z}_l$ is the average of the values $\{z(\mathbf{s}_i) : \mathbf{x}(\mathbf{s}_i) \in R_l\}$.

Trees are known to be universal approximators, in that they can approximate any given function to an arbitrary degree of precision, given enough leaf nodes. This makes them attractive options for estimating very flexible regression functions. However, it is also known that because of their discontinuity they have a high degree of estimation variance (Hastie et al., 2008). RF, and other models based on bootstrap-aggregating (bagging) mitigate the effects of high estimation variance by basing their predictions on an average from multiple tree-based models, each one fitted on a different random bootstrap sample from the observations. To better leverage the variance reducing effect of averaging, RFs add additional randomness between the different trees by randomly selecting covariates, with the so-called m_{try} parameter, on which a split in a tree can be based. This has the effect of reducing the correlation between the outputs of the trees and so decreasing the variance of their averages (Breiman, 2001).

An extremely useful property of RF models, and one leveraged by QRF models, is that the predictions are based on weighted averages of the responses from the observations. Furthermore these weights are determined based on the spatial distribution of the covariates, in that the weight associated with the observation at $\mathbf{s}_i$ in predicting the response value for a given vector of covariates $\mathbf{x}(\mathbf{s}_0) \in \mathcal{R}^p$, will be high if $\mathbf{x}(\mathbf{s}_i)$ and $\mathbf{x}(\mathbf{s}_0)$ fall in the same leaf node in a relatively large proportion of trees in the ensemble. Formally, if $T_1, ..., T_M$ denote the trees in the ensemble, and $R^t = \{R_1^t, ..., R_{L_t}^t\}$ is the set of rectangles associated with the leaf nodes in the t-th tree, T_t, then we can define the (unique) leaf node for each tree which contains the point $\mathbf{x}(\mathbf{s}_0)$ by $l(\mathbf{x}(\mathbf{s}_0), R^t), t = 1, ..., M$. The prediction made by the model for the response associated with $\mathbf{x}(\mathbf{s}_0)$ is then

$$\frac{1}{M}\sum_{t=1}^{M} \bar{z}_{l(\mathbf{x}(\mathbf{s}_0),R^t)} = \sum_{i=1}^{n} w_i(\mathbf{x}(\mathbf{s}_0))z(\mathbf{s}_i); \tag{3.4}$$

$$w_i(\mathbf{x}(\mathbf{s}_0)) = \frac{1}{M}\sum_{t=1}^{M} \frac{1_{\{l(\mathbf{x}(\mathbf{s}_0),R^t)=l(\mathbf{x}(\mathbf{s}_i),R^t)\}}}{\sum_{j=1}^{n} 1_{\{l(\mathbf{x}(\mathbf{s}_0),R^t)=l(\mathbf{x}(\mathbf{s}_j),R^t)\}}}, \tag{3.5}$$

where $1_{\{l(\mathbf{x}(\mathbf{s}_0),R^t)=l(\mathbf{x}(\mathbf{s}_i),R^t)\}}$ is equal to one if both $\mathbf{x}(\mathbf{s}_0)$ and $\mathbf{x}(\mathbf{s}_i)$ are allocated to the same leaf in tree T_t, and zero otherwise. In essence the points falling into the same leaf node as a point $\mathbf{x}(\mathbf{s}_0)$ are seen as having an associated response whose distribution approximates that of the conditional distribution of the response variable given covariate values equal to $\mathbf{x}(\mathbf{s}_0)$. Indeed points allocated to the same leaf node by a tree lie in the same rectangle and will tend to be similar, and so it is reasonable to assume the conditional distributions of the responses will be similar.

Meinshausen (2006) observed that using the observed responses as a weighted sample with weights given by Eq. (3.5) provides consistent estimation of the entire conditional distribution of the response, given covariates equal to $\mathbf{x}(\mathbf{s}_0)$, under relatively mild conditions. That is, let $F(z|\mathbf{x}(\mathbf{s}_0))$ be the conditional distribution of $Z(\mathbf{s}_0)$ given the covariates,

$\mathbf{x}(\mathbf{s}_0)$, then we have

$$\hat{F}(z|\mathbf{x}(\mathbf{s}_0)) = \sum_{i=1}^{n} w_i(\mathbf{x}(\mathbf{s}_0)) 1_{\{z(\mathbf{s}_i) \leq z\}}, \tag{3.6}$$

where the w_i are the weights defined in Eq. (3.5). This is known as QRF, and it allows us to sample approximately from the conditional distribution of the response for any vector of covariate values, as we discuss in the following subsection.

To model an outcome that is multivariate, Segal & Xiao (2011) proposed extending the tree-based models by modifying the impurity measure to account for the joint distribution of the multiple response variables. Where Segal & Xiao (2011) used a simple modified sum-of-squares criterion which only accounts for the covariance between the different response variables, Cevid et al. (2022) proposed a splitting criterion based on a fast Fourier-approximation of the maximum mean discrepancy (MMD) test (Zhao & Meng, 2015). The MMD test is a two-sample test which can detect arbitrary differences between two multivariate distributions which means that it is also able to account for non-linear dependence among the response variables. For details concerning the MMD splitting criterion we refer the reader to Cevid et al. (2022); Zhao & Meng (2015).

To construct a MRF model, the same methodology is followed as in the univariate case, except an ensemble of multivariate regression trees is fitted. The prediction of a MRF is similar to the prediction of Eq. (3.4), one just needs to replace the $\bar{z}_{l(\mathbf{x}(\mathbf{s}_0),R^t)}$ and $z(\mathbf{s}_i)$ with their multivariate counterparts. Therefore, as in the univariate case, the multivariate prediction is also estimated by computing a weighted average of the observed responses with weights equal to those in Eq. (3.5).

Conditional stochastic simulation with random forests

Sampling from the (univariate or multivariate) conditional distribution of the response variable(s), given the covariates, as estimated by a RF or MRF model, is relatively straightforward. As discussed above, a forest model estimates the conditional distribution of $Z(\mathbf{s}_0)$ given $\mathbf{x}(\mathbf{s}_0)$ as a weighted sample of the observed responses in which the weight associated with $z(\mathbf{s}_i)$ is itself a weighted average of the inverses of the sizes of the leaves in the forest which contain both $\mathbf{x}(\mathbf{s}_i)$ and $\mathbf{x}(\mathbf{s}_0)$. An alternative way to view this is that in order to estimate the conditional expectation of the response at a prediction location, $\mathbf{s}_0$, RF treats the response as coming from a mixture distribution in which the mixture components are given by the empirical distributions of the sets $\{z(\mathbf{s}_i) : \mathbf{x}(\mathbf{s}_i) \in l(\mathbf{x}(\mathbf{s}_0), R^t)\}_{t=1}^{M}$, and with uniform mixture proportions.

To draw a sample from this conditional distribution, $\hat{F}(z|x(\mathbf{s}_0))$, one can simply sample a tree from the forest uniformly at random, and then sample an observation uniformly at random from the leaf node to which $\mathbf{x}(\mathbf{s}_0)$ is allocated. Formally, let U be a uniform random variable on $\{1, \ldots, M\}$, and then let V be uniformly distributed on the set $\{i :$

$\mathbf{x}(\mathbf{s}_i) \in R^U_{l(\mathbf{x}(\mathbf{s}_0),R^U)}\}$. Then $z(\mathbf{s}_V)$ is a draw from the conditional distribution of $Z(\mathbf{s}_0)$ given covariates $\mathbf{x}(\mathbf{s}_0)$, as estimated by the forest model. Note that this can also be achieved for the multivariate case in that $\mathbf{z}(\mathbf{s}_V)$ would be a draw from the multivariate conditional distribution of $\mathbf{Z}(\mathbf{s}_0)$.

Implementation of random forest models

Similar to the kriging models, the RF and MRF models were calibrated and assessed with a 10-fold nested cross-validation. For calibration, we determined the optimal hyper-parameters, that is, the m_{try} parameter, and the minimum node size (Breiman, 2001; Hastie et al., 2008). The number of trees for MRF and the two RF models was set equal to 500. Note that for the calibration of the RF and MRF models, SOC and TN were scaled to have unit variance. For the RF-SOC and RF-TN models, the hyper-parameter selection was based on minimising the mean square error (MSE) for each model separately, while for the MRF model, the hyper-parameters were selected by minimising the trace of the MSE matrix (TMSE) (Pierdzioch & Risse, 2020). Suppose that matrix $\mathbf{A}$ represents an $(n_{val} \times q)$ matrix of prediction errors of the validation set, where n_{val} is the number of observations in the validation set, then we compute $TMSE = trace(\mathbf{A}^T\mathbf{A})$, which is the sum of the diagonal elements of $\mathbf{A}^T\mathbf{A}$.

We fitted the RF and MRF models as well as performed the stochastic simulation with the `drf` package (Michel & Ćevid, 2021) in the `R` programming language. For the MRF model we used the fast Fourier MMD splitting criterion and for the RF model we used Eq. (3.3). The main reason why we used the `drf` package is because of its functionality to readily obtain the weights in Eq. (3.6) which are essential for performing the stochastic simulations, and to the best of our knowledge, this functionality is not available in other popular packages which implements RF modelling.

3.2.4 Model evaluation

Evaluation by stochastic simulation

For each test set in the nested cross-validation, 1000 conditional stochastic simulations were drawn, the methodology of which was discussed in Sections 3.2.2 and 3.2.3. From these simulations we then determined the Pearson correlation coefficient between SOC and TN and obtained summary statistics for the C:N ratio. Finally, the results were averaged across the outer 10 folds.

Evaluation by prediction accuracy

The simulations themselves were also averaged to obtain predictions of SOC, TN and C:N for each of the test locations. For the RK and RCK models the simulations were first back-transformed with $e^{\widehat{logSOC}}$ and $e^{\widehat{logTN}}$ before the simulations were averaged. Since

we already had the simulations, it was more efficient to calculate the means of these than to make predictions. To evaluate prediction accuracy we used both univariate and multivariate validation statistics. In terms of assessing the models in a univariate manner, we used the mean error (ME), root MSE (RMSE), and the model efficiency coefficient (MEC) (Wadoux et al., 2018). For the multivariate models, the C:N predictions were determined from the SOC and TN simulations, that is, for each simulation pair, we determined the C:N ratio, and then we average these to obtain the C:N predictions. For the univariate models the C:N predictions were determined by dividing the predictions of SOC by TN.

For multivariate accuracy assessment we used a metric called the Mahalanobis distance (MAH) (McLachlan, 1999) in that

$$\text{MAH} = \sqrt{\sum_{j=1}^{n}(\hat{\mathbf{z}}_j - \mathbf{z}_j)^T \mathbf{V}^{-1}(\hat{\mathbf{z}}_j - \mathbf{z}_j)}, \tag{3.7}$$

where $\hat{\mathbf{z}}_j$ is a q vector containing the model predictions. The measure in Eq. (3.7) is a multivariate counterpart of the univariate RMSE.

Evaluation by comparison of prediction and uncertainty maps

We also evaluated the models on the basis of prediction maps and interquartile range (IQR) maps of the C:N ratio. The maps were produced by fitting the RF and MRF models on the full data set for the Benelux region and Germany. The m_{try} and minimum node size hyper-parameters were set equal to the ones that resulted in the smallest test MSE or TMSE in the cross-validation. We drew simulations, conditional on the covariates, according to the methodology discussed in Sections 3.2.2 and 3.2.3 for the entire region from which the predictions and IQRs were derived. The C:N map for RF was determined by taking the predicted map of SOC and dividing it by TN (similar to how the predictions were determined in the previous section).

3.3 Results

3.3.1 Exploratory data analysis of SOC and TN

Descriptive statistics for SOC, TN and the C:N ratio are shown in Table 3.3. The first and third quartiles define the middle 50% of a set of ordered observations. Therefore, half of the SOC observations were between $13.9\,g \cdot kg^{-1}$ and $33.6\,g \cdot kg^{-1}$, half of the TN observations between $1.3\,g \cdot kg^{-1}$ and $2.7\,g \cdot kg^{-1}$ and the middle 50% of the C:N ratio values were between 9.5 and 13.8. Both of the distributions of SOC and TN were heavily skewed to the right as the differences between, for example, the respective maxima and the third quartiles were much larger than the differences between the corresponding

minima and the first quartiles. This heavy skewness was also evident by observing that both respective means were larger than the corresponding medians and confirmed by the marginal distributions shown in Figure 3.2. Approximately 96% of SOC observations were under $100g \cdot kg^{-1}$, and about 98% of TN observations were under $10g \cdot kg^{-1}$.

It is also important to note that the relative dispersions of SOC and TN were quite large as the coefficients of variation for both SOC and TN were above 1. The relative dispersion of SOC was also larger than TN with the coefficient of variation (CV) for SOC being equal to 1.36 while for TN the CV was equal to 1.03. Due to the heavy skewness of SOC and TN we also provided the mean absolute deviation (MAD) to measure the variation in SOC and TN. For SOC the MAD was equal to $12.3\,g \cdot kg^{-1}$, for TN it was $0.9\,g \cdot kg^{-1}$, and for C:N it was equal to 2.2. The results for MAD suggest that the standard deviation (SD) was heavily influenced by the long right tails in both SOC and TN.

The Pearson correlation coefficient between SOC and TN was equal to 0.927, indicating that the relationship between SOC and TN was positive and strongly linear. This is confirmed by the scatter plot shown in Figure 3.2. However, the linear relationship between SOC and TN becomes weaker for values of SOC higher than $200\,g \cdot kg^{-1}$ and for values of TN higher than $10\,g \cdot kg^{-1}$.

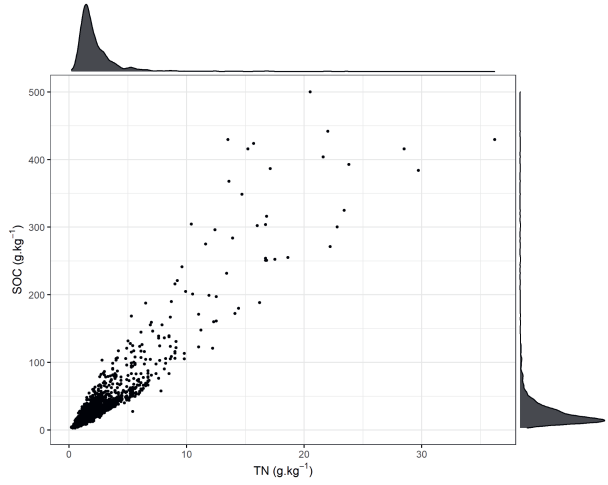


Figure 3.2: Scatter plot of LUCAS SOC and TN observations (shown in $(g.kg^{-1})$) in Benelux and Germany ($n = 2\,216$) with densities plotted on the margins.

3.3.2 Modelling results

The averaged cross-validation residual variogram parameters for the RK and RCK models are shown in Table 3.1. For both of the RK-SOC and RK-TN models we noted that the residuals had weak spatial correlations with an average nugget-to-sill ratio of 86.3% for

the SOC RK model, and 84.9% for the TN RK model. Similar conclusions were made for the variogram parameters for the RCK model. Note the large sill of 0.300 of the cross-variogram, which confirms the strong positive correlation between SOC and TN. The averaged cross-validation correlation between the RCK residuals of the SOC and TN models across the calibration sets was equal to 0.902 which indicated that the residuals between the models were highly correlated. The linear model of coregionalisation, fitted with a spherical shape, are shown in Figure 3.3. The results are mostly stable, except for one of the folds where we noted a higher sill.

Table 3.1: Averaged results of the inner folds of the nested cross-validation showing the parameters of the detrended variograms for RK and RCK. All variograms were fitted with a spherical model.

Model	Variable	Sill	Nugget-to-sill ratio (in %)	Range (in km)
RK	$log(SOC)$	0.376	86.3	63.3
	$log(TN)$	0.280	84.9	54.3
RCK	$log(SOC)$	0.383	83.7	63.3
	$log(TN)$	0.288	88.7	63.3
	$log(SOC)$ vs $log(TN)$	0.300	89.2	63.3

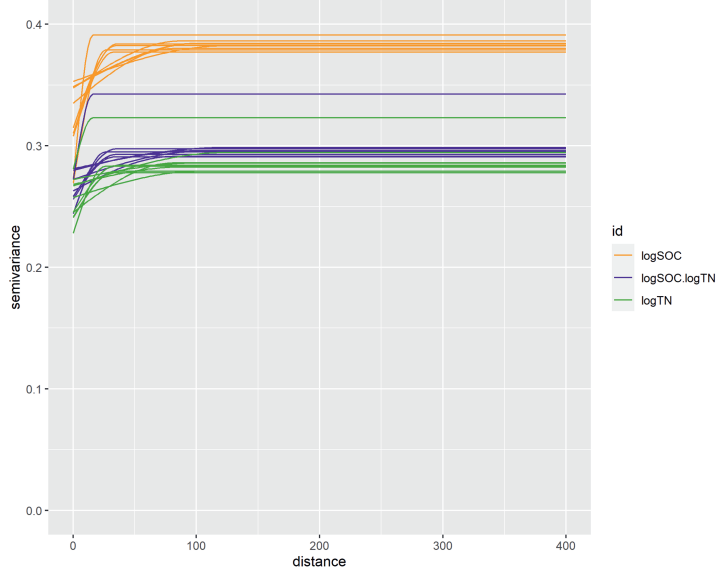


Figure 3.3: Fitted variograms of the residuals of the SOC regression models (orange), the TN regression models (green), and the cross-variograms (purple). Each case shows 10 variograms, as derived from 9 out of 10 folds used in cross-validation.

In Table 3.2 we present the stepwise regression selected covariates from the cross-validation

results, as well as the top 15 covariates according to the averaged cross-validation variable importance rankings for the RF-SOC, RF-TN and MRF models. The importance rankings are shown from 1, most important, to 15, least important. Table 3.2 also presents the soil forming factors (Jenny, 1941). For the stepwise regression procedures we show the number of times that a particular covariate was selected in the cross-validation. Note that for each stepwise regression procedure covariates were selected from the total list of 105 covariates.

There was more agreement between the SOC and TN stepwise regression procedures for covariates classified in "Organisms", than covariates classified in "Mixed". For example, the ESA land cover map for 2010, the long-term SD of the Landsat enhanced vegetation index (EVI) for July and August, the long-term yearly average EVI, and the net primary productivity were selected within all 10 calibration sets for both SOC and TN stepwise regressions, while the mean yearly MODIS MIR band 4, the long-term averaged mean surface reflectance for May and October were only selected by the SOC stepwise regression. Agreements were also noted between the stepwise regressions and the variable importance rankings of the random forest models. For instance, the ESA land cover map for 2010, which was selected 10 times by the stepwise regression procedures, was also ranked as the most important covariate by the RF-SOC, RF-TN and MRF models. There was also agreement between the stepwise regression for SOC and the importance rankings of the RF-SOC model, and between the stepwise regression for TN and the RF-TN model. With regards to the rankings of the MRF model, although relatively mixed, it coincided more with the rankings of the RF-TN model compared to the rankings of the RF-SOC model.

The cross-validation calibration results for the RF-SOC, RF-TN and MRF models consistently indicated that a minimum node size equal to 1 provided the best performance. The m_{try} hyper-parameter showed the best performance when calibrated to 25 and 27 for the RF-SOC and RF-TN models, respectively, while m_{try} for the MRF model showed the best results when calibrated to 27.

Table 3.2: Covariates selected within cross-validation by the stepwise regression (SR) procedures for the SOC and TN kriging models, as well as the variable importance rankings (i.e., from most important, 1, to least important, 15) for the RF and MRF models averaged over the cross-validation. For the stepwise regressions, the number of times each covariate was selected across the 10 folds are shown. The soil-forming factor (Jenny, 1941) is also presented. For details concerning the covariates refer to Poggio et al. (2021).

Soil-forming factor	Covariate	SR-SOC	SR-TN	RF-SOC	RF-TN	MRF
Climate	Bioclimatic zones	-	8	-	-	-
	Long-term averaged yearly mean cloud cover	4	5	15	-	-
	Long-term averaged mean surface daytime temperature for August	10	-	7	-	15
	Long-term SD monthly surface daytime temperature for May	-	-	-	13	-
	Long-term SD monthly surface daytime temperature for July	-	-	-	15	-
	Precipitation of warmest quarter	-	2	14	7	9
	SD yearly solar radiation	-	7	11	11	13
	Total precipitation for April	6	-	12	3	5
	Total precipitation for May	-	1	-	-	-
	Total precipitation for September	10	7	10	8	8
	Total precipitation for December	9	10	-	10	14
	Total yearly solar radiation	-	2	8	-	-
Organisms	ESA land cover map for 2010	10	10	6	4	3
	Global forest change	-	2	-	-	-
	Long-term average of MODIS EVI for July and August	10	10	1	1	1
	Long-term SD of MODIS EVI for May and June	1	-	-	-	-
	Long-term yearly average MODIS EVI	10	10	4	5	4
	Long-term yearly SD MODIS EVI	-	-	-	6	10
	Net primary productivity	10	10	9	12	12
	Normalised difference water index	5	-	-	-	-
Relief	Digital elevation model (DEM)	10	8	13	2	6
	Surface roughness	-	-	-	14	-
Mixed	Mean yearly MODIS MIR band 4	10	-	5	-	11
	Long-term averaged mean surface reflectance for May	10	-	2	9	2
	Long-term averaged mean surface reflectance for October	10	-	3	-	7

3.3.3 Model evaluation

Simulation results

The summary statistics calculated from the RK, RCK, RF and MRF simulations are presented in Table 3.3. The RF and MRF models performed well in reproducing the univariate distributions. For example, for both SOC and TN, the quantiles from the simulations matched the quantiles of the data almost exactly, with one exception of $P_{0.95}$ which was slightly overestimated by both models. The RF and MRF models also performed well to reproduce the central tendency and the dispersion of the distributions with the means, SDs and MADs that also almost exactly matched the statistics of the original data set.

The kriging models performed worse. For example, the medians overestimated the central tendency in SOC and TN while the means underestimated the central tendency in SOC and TN. We also noted that the kriging models underestimated the dispersion of the distributions in terms of the SDs. For example, we noted that RK and RCK produced SDs for SOC of $25.0\,g \cdot kg^{-1}$ and $25.6\,g \cdot kg^{-1}$, which were much smaller than the SD of the original data set (i.e., $44.4\,g \cdot kg^{-1}$). This is because the back-transformation of the

kriging simulations failed to reproduce the heavy positive tails of SOC and TN. This is evident when the MADs are considered, which matched that of MADs of the original data more closely compared to the SDs.

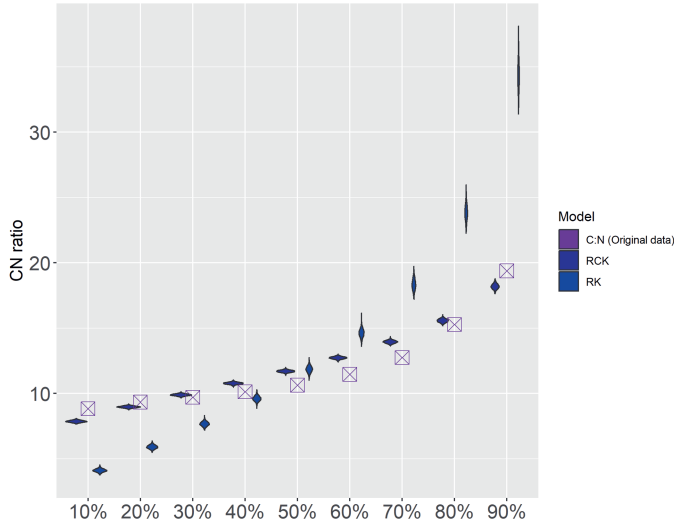
Table 3.3: Summary statistics for SOC (in $(g \cdot kg^{-1})$), TN (in $(g \cdot kg^{-1})$), and C:N calculated from the LUCAS data are presented, as well as summary statistics calculated from simulations for RK, RCK, RF and MRF.

Model	$P_{0.05}$	Q_1	Q_2	Q_3	$P_{0.95}$	Mean	SD	CV	MAD
SOC	9.1	13.9	20.7	33.6	84.2	32.6	44.4	1.36	12.3
RK	7.3	14.1	22.9	37.4	75.7	30.0	25.0	0.83	15.5
RCK	7.3	14.2	22.9	37.3	75.9	30.1	25.6	0.85	15.1
RF	9.1	14.0	20.8	33.7	86.6	32.4	42.9	1.31	12.3
MRF	9.1	14.0	20.8	33.7	86.8	32.5	43.1	1.32	12.4
TN	0.9	1.3	1.8	2.7	5.9	2.5	2.5	1.03	0.9
RK	0.7	1.3	1.9	2.9	5.4	2.4	1.7	0.70	1.1
RCK	0.8	1.3	1.9	2.9	5.3	2.3	1.6	0.73	1.1
RF	0.9	1.3	1.8	2.7	5.8	2.5	2.5	1.00	1.0
MRF	0.9	1.3	1.8	2.7	5.9	2.5	2.5	0.99	0.9
C:N	8.4	9.5	10.6	13.8	22.8	12.5	4.6	0.37	2.2
RK	3.1	6.8	11.9	20.8	46.3	16.8	16.7	0.99	9.1
RCK	7.6	9.6	11.6	14.3	19.8	12.4	3.9	0.31	3.5
RF	3.2	7.0	11.4	19.4	51.5	18.4	27.5	1.47	7.9
MRF	8.4	9.5	10.7	13.8	22.5	12.5	4.6	0.37	2.2

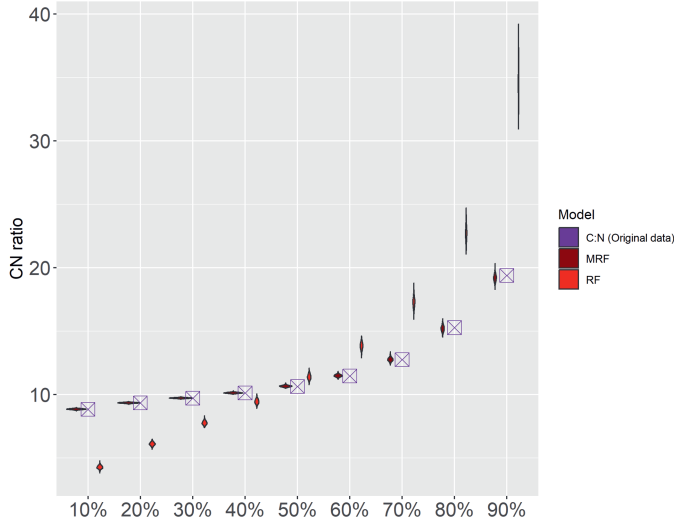
Table 3.3 clearly shows that the MRF and RCK models, in comparison to their univariate counterparts, were superior in terms of reproducing the distribution of the C:N ratio. For example, the mean was strongly overestimated by RK and RF. The superiority was especially true for the MRF model which was the only model to reproduce the distribution of the C:N ratio almost exactly. We present the C:N ratio results also visually in Figure 3.4. Figure 3.4a presents the simulation results for the RK and RCK models for the 10th up to the 90th percentile, and Figure 3.4b presents it for the RF and MRF models. In each panel of Figure 3.4 we present the same percentiles for the C:N ratio calculated from the LUCAS data. In both panels we noted that the multivariate models were superior in maintaining the structure of the C:N ratios compared to the corresponding univariate models. This was especially evident in Figure 3.4b where the simulation results of the MRF model followed the percentiles of the original data almost exactly. The RF model underestimated the C:N ratio for percentiles lower than 50%, and overestimated it for percentiles higher than 50%. The MRF model was also superior in maintaining the C:N ratio compared to the RCK model. This was evident in Figure 3.4a where we noted that the RCK C:N ratio results slightly overestimated the C:N ratios of the original data set from the 30th to the 80th percentiles. We also noted a small underestimation of the C:N ratio simulation results for the 10th percentile for the RCK model. Similar to the

RF model, RK also showed a severe underestimation for the lower percentiles, and an overestimation for the larger percentiles.

The rows of Figure 3.5 present the first four out of 1 000 simulations to depict the relationship between SOC and TN as simulated by the various models. The average of the correlation coefficients for all simulations were equal to 0.225, 0.898, 0.108, and 0.925, for the RK, RCK,RF, and MRF models, respectively. From Figure 3.5 it is clear that the multivariate models, i.e., MRF and RCK, were superior in terms of maintaining the correlation structure between SOC and TN. The MRF was also producing a better represented correlation structure than RCK, with the average correlation of the MRF model (i.e., 0.925) being closer to the correlation of the original data (i.e., 0.927) compared to the correlation of the RCK model (i.e., 0.898).



(a)



(b)

Figure 3.4: Violin plots showing the C:N ratio simulation results for (a) the RK and RCK models; as well as (b) the RF and MRF models. The 10th to 90th percentiles were determined for each model and each simulation. The corresponding percentiles calculated from the original data are also shown.

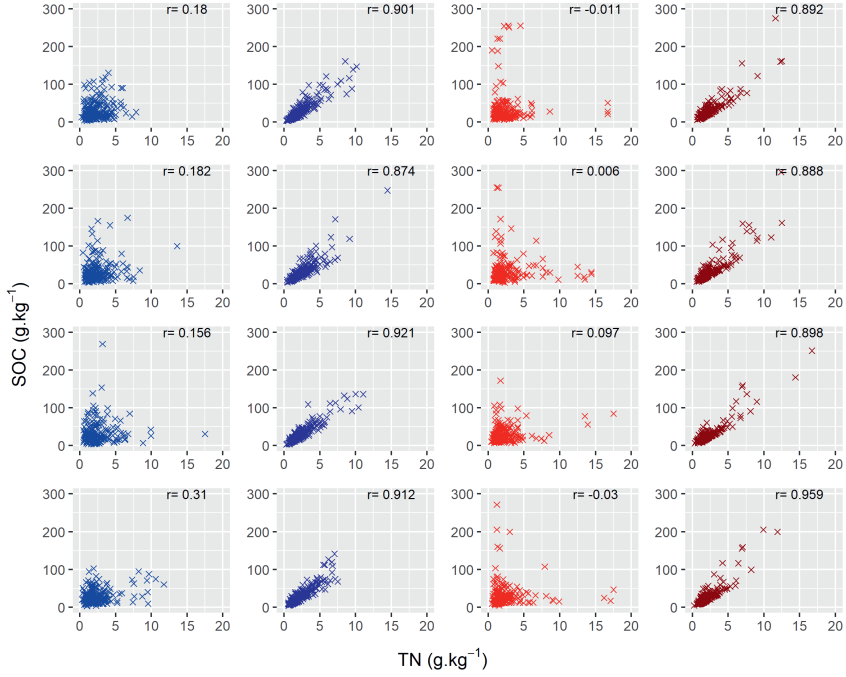


Figure 3.5: Simulation results for the first four simulations depicting the relationship between SOC and TN shown from left to right for RK, RCK, RF, and MRF. The average correlation coefficients calculated from all the simulations were equal to 0.225, 0.898, 0.108, and 0.925 for the RK, RCK, RF, and MRF models, respectively. The correlation between SOC and TN in the original data set is equal to 0.927. The individual correlation coefficients (r) are also shown in each plot.

Accuracy results

The results for the validation statistics, ME, RMSE, MEC, for each of SOC, TN and C:N, are presented in Table 3.4. The table also shows the results for MAH calculated from the SOC and TN predictions. The accuracy between the RK and RCK models, when predicting either SOC or TN, were comparable, which was observed by considering any of the validation statistics. For example, the RMSE for the RK-SOC model was equal to 41.23, while for the RCK-SOC model, the RMSE was equal to 41.71. Although it did seem that the RK models overall performed marginally better compared to the RCK model, for both SOC and TN, when considering the RMSE together with the MEC and MAH. On the other hand, RCK outperformed RK on all validation statistics when predicting the C:N ratio. The accuracy of the RF and MRF models were also comparable. This can be seen for example by considering the RMSE which, when predicting TN, were both equal to 2.26 for the RF and MRF models. The MAH validation statistic, which takes

the covariance structure between SOC and TN into account, did show improvement when MRF was compared to the predictions of the RF-SOC and RF-TN models. However, these improvements were marginal. With regards to comparing the kriging models to the random forest models, we noted that the RF and MRF models performed better overall when predicting SOC and TN, but when we consider the C:N ratio, RCK performed best.

Table 3.4: Prediction accuracy results for ME, RMSE, and MEC of the RK, RCK, RF and MRF models. The results for MAH, calculated for the predictions of SOC and TN, are also shown.

	ME			RMSE			MEC			MAH
	SOC	TN	C:N	SOC	TN	C:N	SOC	TN	C:N	
RK	-2.59	-0.11	0.32	41.23	2.37	3.63	0.138	0.126	0.362	21.21
RCK	-2.48	-0.07	0.04	41.71	2.40	3.43	0.117	0.097	0.436	21.49
RF	-0.14	-0.01	0.72	39.76	2.26	3.65	0.194	0.211	0.359	19.81
MRF	-0.03	-0.01	-0.01	40.04	2.26	3.58	0.193	0.212	0.389	19.46

Mapping results

The C:N ratio maps produced by the RF and MRF models are displayed in Figure 3.6. The separate SOC and TN maps are presented in Web Appendix B, while the maps for RK and RCK are presented in Web Appendix C. Although it falls outside the scope of this study to provide detailed interpretations of all the maps, we do outline here the most prominent features. We do not observe any prominent differences between Figures 3.6a and 3.6b in terms of spatial patterns, but we do note overall larger C:N ratios for the RF model compared to the MRF model. This is especially visible in the northern regions of Germany, and in the east of the Netherlands. The larger C:N ratios for the RF model is understandable as we noted in Table 3.3 that the RF model overestimated the mean C:N ratio. In Figures 3.6c and 3.6d we also note similar spatial patterns, and that the RF model produced larger values for the IQR compared to the MRF model. This again is not surprising as we noted in Table 3.3 and in Figure 3.4b that the RF model underestimated the C:N ratio below the median, and overestimated the C:N ratio above the median, which then leads to larger IQR values.

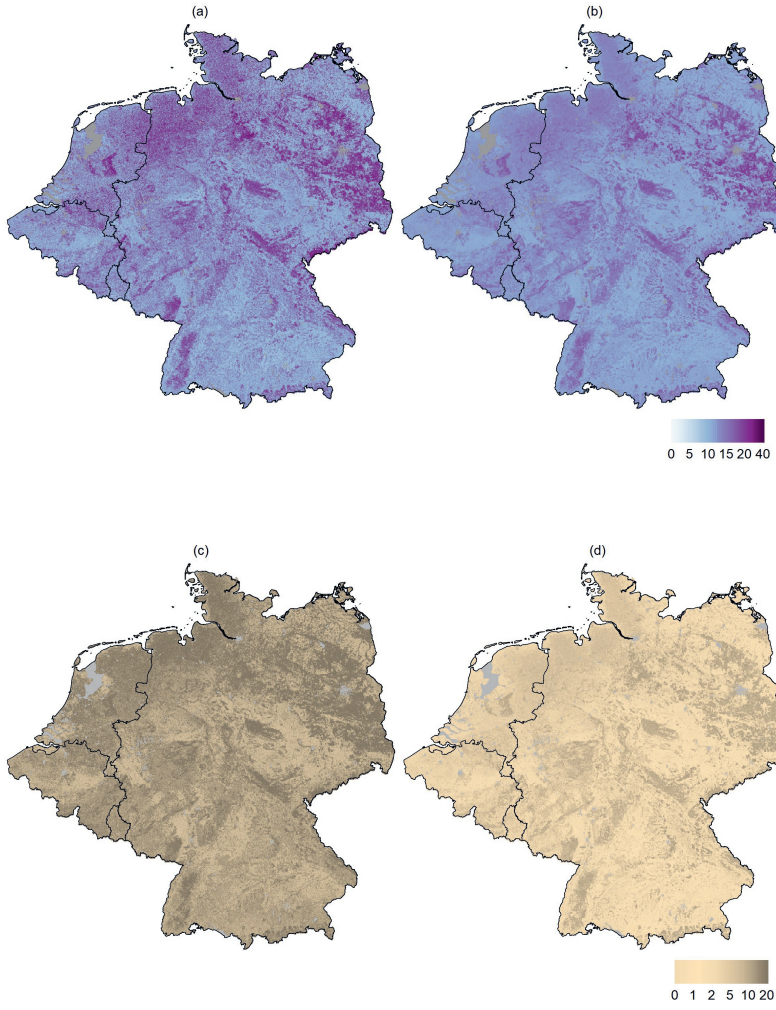


Figure 3.6: Predicted C:N maps for (a) RF; and (b) MRF; and corresponding IQR maps for (c) RF; and (d) for MRF.

In Figure 3.7 we present the C:N ratio density plots of 10 simulations (Sim) as well as the predictions (Pred) of both the RF and MRF models for Benelux and Germany. In this figure we also present the C:N ratio density of the LUCAS data. The densities of the RF simulations and predictions were much wider compared to the MRF densities, and therefore are also somewhat unrealistic. This results also confirmed the wider IQR

values produced by the RF model in Figure 3.6. The density of the MRF simulations also reproduced the density of the LUCAS C:N ratio almost exactly. We also note that the modes of the predictions of the RF and MRF models do not coincide with the mode of the LUCAS density. However, the means of the predictions of the RF and MRF models were more reflective of the LUCAS C:N ratio mean. The mean of the predictions for RF was equal to 13.2, while for the MRF model it was equal to 12.6 (in Table 3.3 we noted that the mean C:N for LUCAS was equal to 12.5).

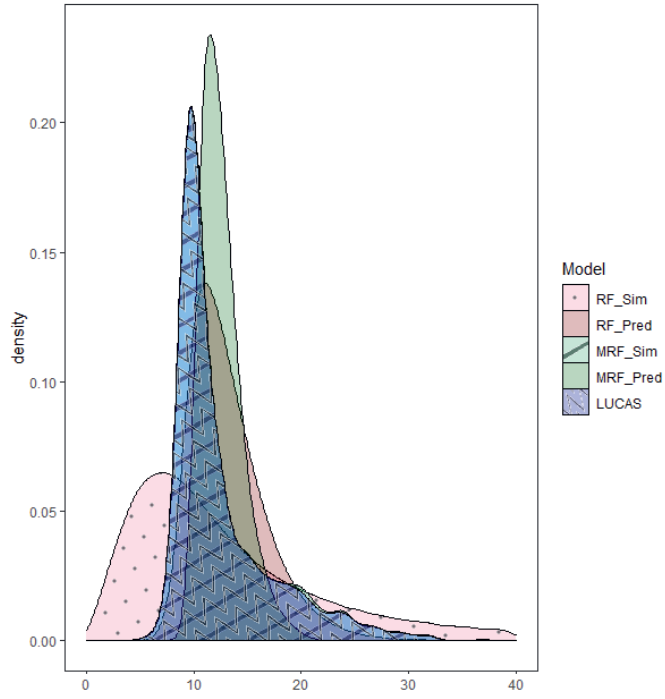


Figure 3.7: C:N ratio density plots of the conditional simulations (Sim) and predictions (Pred) of the RF and MRF models for Benelux and Germany. The density plot of the C:N ratio of the LUCAS data is also shown.

3.4 Discussion

This is the first study in DSM, to the best of our knowledge, where multiple soil properties are modelled simultaneously with a MRF model. We also compared the MRF model to a RCK model, as well as to the case where soil properties were modelled with separate RF models and separate RK models. In this study we only focused on SOC and TN, but our approach can easily be extended to other DSM case studies, such as, mapping soil nutrients (Hengl et al., 2017; Iticha & Takele, 2019), heavy metals (Taghizadeh-Mehrjardi et al.,

2021; Papritz & Dubois, 1999; De Sousa Mendes et al., 2022; Jia et al., 2019), mapping soil water content at different pressures (Wegehenkel, 2005; Huisman et al., 2002; Turek et al., 2022), mapping a soil property at different depths (Wadoux et al., 2019), or studies related to soil health, for example in Prout et al. (2021, 2022) where the ratio of SOC to clay is of importance. From the results of the stochastic simulations, as observed in Table 3.3 and in Figure 3.4, it was clear that the multivariate models, especially MRF, were superior in terms of reproducing the SOC and TN correlation in the data, and hence, also superior in producing more realistic C:N ratios. This is understandable as the multivariate models explicitly incorporate information from the correlation structure of SOC and TN. For the MRF model this happens when the trees are constructed with a multivariate splitting criterion, for example with the MMD test (Cevid et al., 2022), while for RCK, the correlation between SOC and TN is incorporated with the variance-covariance matrices estimated from the cross-variogram. The C:N ratio simulations produced by the univariate models were often unrealistic. For example, in Figure 3.7, we noted that a large proportion of the C:N simulations of Benelux and Germany produced by the RF model were less than 7, and some C:N ratios were even larger than 500, and this does not make sense for this region (Matschullat et al., 2018).

Reproducing the correlation structure between SOC and TN is important, because SOC and TN are vital soil components which have considerable influence on soil quality, and producing accurate realisations of SOC and TN, while maintaining a realistic C:N ratio, is essential (Reeves, 1997; Bonfante et al., 2019; Srivastava et al., 2017; Rawls et al., 2003). The added value of the MRF model is therefore when the joint distribution of the soil properties must be considered when multiple soil properties are simulated. This comes from the resultant ability to obtain samples from the estimated joint conditional distribution of SOC and TN at each location, arising from the approach discussed in Section 3.2.3. Note that estimates relating to the distribution of the C:N ratio cannot be obtained from predictions of SOC and TN alone, since these predictions correspond with estimates for the conditional expectations of SOC and TN at each location whose ratios do not tell us about the distribution of individual realisations of C:N ratio, nor even of the mean C:N ratio. Therefore, conditional simulation as performed in this paper can be important for subsequent DSM analyses such as Digital Soil Assessment (Kidd et al., 2015; Okonkwo et al., 2018; Rabot et al., 2022a) and Soil Function Assessment (Rabot et al., 2022b; Greiner et al., 2018, 2017) both of which rely on well defined (joint) distributions of the soil properties.

Although the main purpose of this study was to compare the various models in terms of stochastic simulations, we also compared the models on the basis of prediction accuracy. The accuracy of the MRF model, as indicated by the ME, RMSE and MEC validation statistics, was comparable to the accuracy of the RF-SOC and RF-TN models. We also noted a slight improvement in the MAH validation statistic when MRF was compared to the RF predictions, but these improvements were marginal. Similar results were also

found, for example in Pierdzioch & Risse (2020), where the authors noted improvement in MAH when MRF was compared to univariate RFs when forecasting metal returns. Also, in Wan & Pal (2013), the authors compared MRF to the RF predictions with the mean absolute error, and this improvement was also only marginal. For the kriging models it seemed that the RK-SOC and RK-TN models performed slightly better compared to the RCK model. This was not that surprising as it is well known that the advantage of co-kriging over regular kriging is when one variable, the so-called primary variable, is undersampled in relation to a secondary variable, and in our study this was not the case (Goovaerts, 1999; Webster & Oliver, 2007). We also noted that the accuracy of the RF and MRF models was better compared to that of the RK and RCK models.

For the DSM practitioner these results suggest that multiple soil properties can be simulated with a MRF model and the joint distribution will be well reproduced in the simulations. However, if the purpose of the DSM exercise is to only make predictions of individual soil properties, we do not recommend MRF. This is because we noted no significant improvement in the prediction accuracy of the MRF model compared to the RF-SOC and RF-TN models. With MRF there is also added complexity and inexperienced users might make mistakes that could lead to a decrease in model accuracy. In addition, in Wadoux et al. (2020a) the authors noted that the size of the calibrated MRF model will increase dramatically if the number of soil properties increases. In our experience, this was not a big issue as we calibrated the MRF model with the TMSE. We found that the amount of computation time needed to calibrate the MRF was similar to the amount of computation that was required to calibrate the individual univariate RF models. It should also be noted that with MRF variable importance cannot be individually interpreted for each soil property in the model. This was also noted by Wadoux et al. (2020a).

In this study we only considered two soil properties, and these were also strongly correlated. Further research would be required to investigate how MRF will perform in terms of prediction when an increased number of soil properties are included, as well as when the correlation structure between the soil properties are not that strong. In the case of RCK the model does not deal well when too many response variables are included or when the responses are not highly correlated (Goovaerts, 1999). Therefore, this could potentially be an important advantage of MRF over RCK. Further research will also be required to investigate how other machine learning models, for instance, multivariate support vector regression, might perform when mapping soil properties simultaneously.

3.5 Conclusion

In this paper we proposed the use of MRF in DSM to jointly simulate multiple soil properties, conditional on the environmental covariates. We have seen, when we compared MRF to RF models that modelled the soil properties separately, that the added value of MRF is that the joint distribution of the soil properties were much better reproduced

in the simulations. This is important for subsequent DSM analyses such as Digital Soil Assessment and Soil Function Assessment. In case of prediction, we concluded that modelling multiple soil properties with separate RF models may still be better due to the added model complexity of MRF and due to seeing no significant improvements in the prediction results of MRF compared to RF. Although we only considered a case study of SOC and TN observations, the proposed methodology to simulate soil properties in a multivariate manner can be extended to other DSM studies involving multiple correlated soil properties.

Supplementary materials

The supplementary materials, Web Appendices A - C, can be downloaded from the journal version of this chapter:
van der Westhuizen, S., Heuvelink, G.B.M., Hofmeyr, D.P. (2023). Multivariate random forest for digital soil mapping. *Geoderma*, **431**, p. e116365. doi: `https://doi.org/10.1016/j.geoderma.2023.116365`.

Chapter 4

Mapping soil thickness by accounting for right-censored data with machine learning

This chapter is based on:

van der Westhuizen, S., Heuvelink, G.B.M., Hofmeyr, D.P., Poggio, L., Nussbaum, M., Brungrad, C. (2024). Mapping soil thickness by accounting for right-censored data with machine learning.

Under review, *European Journal of Soil Science.*

4.1 Introduction

Over the last decade, the number of applications of machine learning in DSM has increased dramatically, and it is now common to use machine learning to map key soil properties, such as soil organic carbon, nitrogen and pH (Minasny & McBratney, 2016a). One reason for this is because "off-the-shelf" machine learning models, in particular random forest (RF) models, often produce soil maps with greater prediction accuracy compared to other models such as multiple linear regression or geostatistical models (Wadoux et al., 2020a). A soil property that is also important to map is the thickness of soil. Soil thickness is generally defined as the depth from the soil surface to the lithic or paralithic contact (Department of Agriculture, 2017), but the definition may be different depending on the region or country. It provides vital information for studies involving for example carbon storage (Greiner, 2018), crop suitability assessment (Fan et al., 2016) and land management (Greiner, 2018). However, soil thickness data are often right-censored which means that the true soil thickness is larger than the sampling depth (Chen et al., 2019; Malone & Searle, 2020b). Right-censored soil thickness data occur for two main reasons: (1) determination of soil depth may not be part of the study for which samples are obtained, in which case surveyors may be instructed to auger no deeper than a given depth as that is the part of the soil which is of interest; (2) soil sampling to depth could be expensive and needs specialised equipment, and even with specialised equipment there is a maximum feasible sampling depth (Malone & Searle, 2020b). It is important to account for the censored nature of soil thickness data, because if right-censored data are treated as if they were true measurements, then predictions may severely underestimate the true soil thickness (Shangguan et al., 2017; Vaysse & Lagacherie, 2015).

Mapping soil thickness has been a subject of considerable effort, and there are two predominant approaches found in the scientific literature - mechanistic modelling and statistical modelling. In this paper we will only focus on the latter approach, and for further information about mechanistic models we refer the reader to Minasny & McBratney (1999); Schoorl et al. (2002); Minasny & McBratney (2006); Bonfatti et al. (2018). Statistical methods that have been used to model soil thickness range from traditional methods such as principal component analysis, correlation analysis and linear regression (Chaplot et al., 2010; Odeh et al., 1991a; Moore et al., 1993; Zhang et al., 2018; Florinsky et al., 2002), to geostatistical models (Odeh et al., 1995; Kuriakose et al., 2009), and machine learning models, such as RF (Tesfa et al., 2009; Baltensweiler et al., 2021), quantile regression forest (Liu et al., 2022), support vector machines (Suleymanov et al., 2021), cubist and gradient boosting (Mulder et al., 2016). None of the above mentioned studies used statistical models which accounted for the fact that some of the soil thickness observations are right-censored (Chen et al., 2019; Malone & Searle, 2020b).

There are however studies that used statistical methods that incorporate strategies for dealing with right-censored data in mapping soil thickness. Kempen et al. (2015) predicted

peat thickness in the Netherlands and corrected for censored observations by adding simulated values from a beta distribution. Lacoste et al. (2016) accounted for right-censored data by adding a fixed value of 30 cm to the soil depths that were censored. Brungard et al. (2021) converted soil thickness data to soil depth classes, i.e., a categorical variable. Another way of dealing with right-censored soil thickness data was described in Chen et al. (2019), which used random survival forest (SRF) (Ishwaran et al., 2008) to map the probability that soil thickness exceeds certain depths (i.e., probabilistic prediction). While the study by Chen et al. (2019) did not generate maps of soil thickness itself, the authors did offer guidance on how to create these maps from the SRF model output. This included suggestions such as calculating the median from the predicted probability distribution or using the soil thickness values obtained at a predefined probability threshold.

In principle, knowing the exceedance probabilities (i.e., probability to exceed a certain depth) at all depths can be used to obtain point and interval predictions of soil thickness, because the exceedance probabilities define the cumulative probability distribution of soil thickness, but SRF and other survival models only produce exceedance probabilities for depths observed in the calibration data set. To the best of our knowledge, no study in DSM has produced soil thickness maps with SRF. Furthermore, it was mentioned in Malone & Searle (2020b) that SRF yielded unsatisfactory results, which led to the authors using a different modelling approach to map soil thickness. This was mainly due to the complex structure of soil thickness data in Australia, which included many rock outcrop and very deep soil observations (e.g., soils deeper than 10 m), as well as a large proportion of censored data. The authors decided to use an alternative non-survival approach where three separate RF models were used to map soil thickness. The first RF model classified the occurrence of rock outcrops, the second model predicted soil thickness within a 0-2 m range, while accounting for right-censored data with an approach similar to what was used in Kempen et al. (2015), and the third model classified the occurrence of deep soils. One drawback of this approach is that three separate models need to be calibrated which is more complex.

In general, standard approaches to survival modelling require that the cause of censoring is "non-informative" in order for their estimates to be unbiased (Leung et al., 1997; Clark et al., 2003; Willems et al., 2018). If the censoring is related to the response variable then the censoring mechanism is said to be informative. In the context of modelling soil thickness, this can occur for instance, when obstacles such as gravel prevent the soil surveyor from reaching the actual soil thickness. Although it can be challenging to ascertain if censoring is informative or non-informative in studies involving modelling of right-censored soil thickness data, we contend that it could be informative in some situations. Nevertheless, we wish to emphasise that even if the assumption of non-informative censoring is violated, survival models like SRF can still be used for prediction, as ignoring the censored nature of the data would be far worse (Leung et al., 1997).

Modelling results may also be less accurate if the proportion of censored is large. In Willems et al. (2018), the authors conducted a simulation study, and observed noteworthy bias in the estimated probability function of the response variable when the proportion of censored data was as large as 35% when using standard survival methods. Therefore, an additional survival method to SRF might also be needed to model right-censored soil thickness data. In Malone & Searle (2020b) the proportion of censored data was close to 40% which could also have contributed to the poor results of SRF.

In this paper we propose an inverse probability of censoring weighting (IPCW) method, similar to Vock et al. (2016), to model soil thickness with a regular RF model, but with the calibration data weighted according to the exceedance probabilities of the response variable. Using a regular RF model has the advantage of producing soil thickness maps without the intermediate steps of obtaining point predictions from the exceedance probabilities as required with SRF. The method we propose involves two main stages. First, it estimates a survival function of censored soil thickness data from which it then calculates the inverse probability of censoring (IPC) weights. Second, it incorporates these weights with the RF model. It should be noted that the IPCW method can also be used with other machine learning models.

The main objective of this paper is to illustrate the use of the IPC weighted RF, from now on denoted as WRF, and compare it to SRF and other strategies for dealing with right-censored data for mapping soil thickness. The models are compared by means of a comprehensive synthetic simulation study in which we investigate model performance, for example with different proportions of censored data. The models are also compared in two real-world case studies, one from Zurich (Switzerland) and one from Maine (USA).

4.2 Material and methods

4.2.1 Handling censored data

In the statistical literature there are four main approaches for dealing with right-censored data. These are: (1) Complete-data analysis; (2) Imputation; (3) Dichotomous data analysis; and (4) Survival analysis. Note that data can also be left-censored, interval-censored or truncated. In this paper we only focus on right-censored soil thickness data and how it is accounted for with survival analysis. For a detailed discussion on the different types of censoring and truncation and how to handle them we refer the reader to Turkson et al. (2021); Leung et al. (1997); Tobin (1958); McDonald & Moffitt (1980).

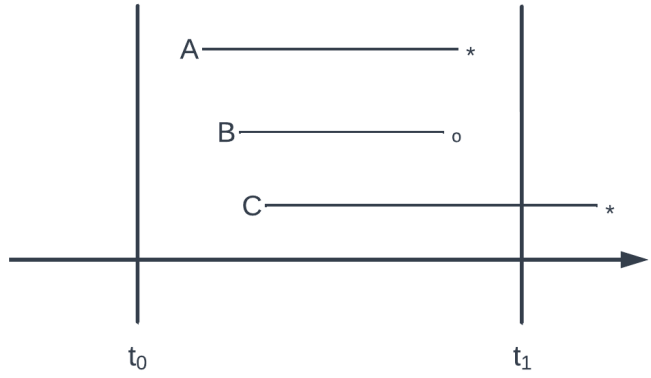
In the complete-data analysis approach, censored data are ignored, and only the observations in the data set that are non-censored are considered (i.e., hard observations). This is the most simple approach for dealing with censored data, but it could also be very inefficient because there is a loss of information when the sample size is lowered and censored

data are unused. Moreover, it can also lead to biased results if the censored observations were not censored at random, that is, if the censoring mechanism is possibly informative (Turkson et al., 2021).

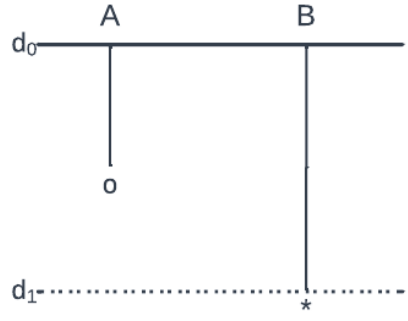
Imputation refers to the process of filling in data for individuals or observations that were censored. This method is popular with missing data, but may not be appropriate for censored data. For right-censored data, two extreme approaches to impute censored data are to either use left-point or right-point imputation. The former assumes that the true observation will be observed almost immediately after censoring, i.e. the surveyed depth is then used, while the latter assumes that the true observation is very unlikely to be observed and therefore a very large value for the censored observation is imputed. If this approach is considered, stringent assumptions about the censoring mechanism are therefore required (Leung et al., 1997). This added complexity makes this approach not an ideal choice for handling censored data.

With the dichotomous data analysis approach only the occurrence versus the nonoccurence of the true observation is analysed with a classification model (i.e., logistic regression). The problem with this approach is again a loss of information, because the observation values themselves are ignored, and we only model the probability of observing the true observation. Recall that a part of the modelling approach in Malone & Searle (2020b) analysed the occurrence of rock outcrops and deep soils as dichotomous variables.

Survival analysis refers to a collection of statistical methods with the goal of modelling a time-to-event response variable, also referred to as survival time (Kleinbaum et al., 2012). For example, in an epidemiological study involving cancer, the response variable could be the time until a patient goes out of remission. For some patients the recorded data might be right-censored. This could happen when patients were lost to follow up, or if they did not experience the event of interest while the study was conducted. In Figure 4.1a an example of an epidemiological study is illustrated. As indicated on the horizontal axis, the duration of the study is from time t_0 to t_1, where "1" here refers to the end of the study. Subject A ends with an asterisk (*) which indicates an event of interest and subject B ends with an open point (o) which indicates an event other than the event of interest. In the case of subject A, the time until the event of interest falls within the observation period, and therefore the time of occurrence of the event of interest is known exactly, that is, not censored. Subject B is right-censored, because an event other than the event of interest occurred within the observation period. Subject C is also right-censored, because the event of interest occurred after time t_1 which is after the observation period ended.



(a)



(b)

Figure 4.1: Graphical illustration of survival analysis and types of censored observations in (a) a typical epidemiological study, and in (b) a DSM soil thickness study.

In Figure 4.1b, an illustration of determining soil thickness, usually measured in distance (i.e., cm, or m), is presented. Soil thickness at d_0 represents the soil surface and the line at d_1 represents the true soil thickness, assuming it is constant in this simple illustration. The observation at location A is censored, because the true soil thickness was not reached. At location B we reached the true soil thickness, and therefore this observation is not censored.

4.2.2 Survival analysis of right-censored soil thickness data

In the context of modelling soil thickness the response variable is the depth at which the true soil thickness occurs, formally denoted as $D(\mathbf{s})$, with $\mathbf{s}$ a location in a region of interest, $\mathcal{A}$. For simplicity, we omit the location $\mathbf{s}$ from the functions explained in this section. The survival function, $S(d)$, conveys information of the probability that the true depth exceeds a depth of d,

$$S(d) = P(D > d). \tag{4.1}$$

It is important to note that theoretically the survival function is equal to one at the surface, i.e., $S(0) = 1$, and then decreases monotonically as d increases so that $\lim_{d\to\infty} S(d) = 0$. It should also be noted that if there are locations with the lower soil boundary at the surface (e.g., rock outcrops) then $S(0)$ could be smaller than 1. The cumulative distribution function, that is, the reciprocal of Eq. (4.1), is given by

$$\begin{aligned} F(d) &= P(D \leq d), \\ &= 1 - S(d). \end{aligned} \tag{4.2}$$

In survival analysis, two other functions that are also often used include the hazard function and the cumulative hazard function (CHF). However, these functions are not used in this study and hence not defined. We refer the reader to (Kleinbaum et al., 2012) for detailed discussions about these functions.

Right-censoring occurs when the true soil thickness remains unknown when we augered to depth, d. Let us denote by $C(\mathbf{s})$ the maximum depth that we intend to sample at location $\mathbf{s} \in \mathcal{A}$, even where $C(\mathbf{s}) > D(\mathbf{s})$. We then write an observation as $\{Y(\mathbf{s}), \delta(\mathbf{s}), \mathbf{x}(\mathbf{s})\}$, where $Y(\mathbf{s}) = \min(D(\mathbf{s}), C(\mathbf{s}))$, and $\delta(\mathbf{s}) = I(D(\mathbf{s}) \leq C(\mathbf{s}))$ and $\mathbf{x}(\mathbf{s})$ is a p-vector of covariates. That is, at each sampling location the depth reached, $Y(\mathbf{s})$, is recorded, as is an indicator $\delta(\mathbf{s})$ which tells us whether that depth was reached because the true soil thickness was obtained, that is, $Y(\mathbf{s}) = D(\mathbf{s})$, or not, in which case $Y(\mathbf{s}) = C(\mathbf{s})$. Note also that it is assumed that $\mathbf{x}(\mathbf{s})$ is known at all $\mathbf{s}$ in $\mathcal{A}$, including all sampling locations, regardless whether an observation is censored or not.

Our aim is to predict $\{D(\mathbf{s}), \mathbf{s} \in \mathcal{A}\}$ from the covariates $\mathbf{x}(\mathbf{s})$, using a model trained on calibration data $\{y(\mathbf{s}_i), \delta(\mathbf{s}_i), \mathbf{x}(\mathbf{s}_i)\}$ at sampling locations $\mathbf{s}_i$ for $i = 1, \ldots, n$. Here, $y(\mathbf{s}_i)$, $\delta(\mathbf{s}_i)$ and $\mathbf{x}(\mathbf{s}_i)$ are defined as before, and n is the sample size. Throughout this paper, it should be noted that we refer to a random variable with an uppercase letter, e.g., D, C or Y, and a realisation of that variable by a lowercase letter, e.g., d, c or y. Two types of prediction functions include point predictors and probabilistic predictors. If we denote $\mathbf{s}_0$ to represent a prediction location, a point predictor will specify a value for $d(\mathbf{s}_0)$, whereas a probabilistic predictor will provide an estimate of the conditional probability distribution for $D(\mathbf{s}_0)$.

4.2.3 Random survival forest for predicting soil thickness

For a detailed outline of SRF we refer the reader to Ishwaran et al. (2008), and for details on how SRF has been implemented for modelling soil thickness we refer to Chen et al. (2019). Here, we give a brief outline. SRF is an adaptation of the RF model and works on the same principles (Ishwaran et al., 2008). That is, trees are grown using bootstrap samples, a subset of covariates is randomly selected when tree nodes are split, and the final ensemble is determined by a statistic computed on the observations in the terminal nodes. Two key differences between RF and SRF for regression include: (1) instead of growing trees with the sum-of-squares splitting rule, SRF grows survival trees with the log-rank splitting rule (Ishwaran et al., 2008) that is based on the log-rank test which is a statistical test used to compare the survival functions of two groups (Segal, 1988; LeBlanc & Crowley, 1993); and (2) an ensemble statistic for SRF is the ensemble survival function. It should be noted that other splitting criteria are also available for SRF (Ishwaran et al., 2008).

Formally, consider a decision tree fit for an SRF model. Let h be a terminal node, and let $\{d_{1,h} < d_{2,h} < \cdots < d_{j,h} < \cdots < d_{m_h,h}\}$ be the m_h distinct (non-censored) depth observations in h. Furthermore, for $d_{j,h}$, $1 \leq j \leq m_h$, let $a_{j,h}$ be the number of (non-censored) observations with depth equal to $d_{j,h}$, and let $b_{j,h}$ be the number of observations whose depths (censored or not) exceed $d_{j,h}$ (note that $b_{j,h}$ is decreasing with j). Then, at a given node h, $1 - \frac{a_{j,h}}{b_{j,h}}$ may be interpreted as an estimate of the conditional probability of exceeding $d_{j,h}$, given previous depths have been reached. The survival function for h is then estimated from the Kaplan-Meier estimator (Kaplan & Meier, 1958; Ishwaran et al., 2008)

$$\hat{S}_h(d) = \prod_{j:d_{j,h} \leq d} \left(1 - \frac{a_{j,h}}{b_{j,h}}\right). \tag{4.3}$$

Given a covariate vector, $\mathbf{x}(\mathbf{s}_0)$, the survival function, $S(d|\mathbf{x}(\mathbf{s}_0))$, is estimated by dropping $\mathbf{x}(\mathbf{s}_0)$ down the tree, and then the terminal node statistics are determined with

$$\hat{S}(d|\mathbf{x}(\mathbf{s}_0)) = \hat{S}_h(d), \text{if } \mathbf{x}(\mathbf{s}_0) \in h.$$

Finally, the ensemble survival function is determined by averaging over all the trees. That is, if $\hat{S}_b(d|\mathbf{x}(\mathbf{s}_0)))$ is the estimate for the b-th tree, then the ensemble estimate is calculated with

$$\hat{S}_e(d|\mathbf{x}(\mathbf{s}_0)) = \frac{1}{B}\sum_{b=1}^{B} \hat{S}_b(d|\mathbf{x}(\mathbf{s}_0)) \tag{4.4}$$

where B is the total number of trees.

Note that although the SRF model does not provide a point prediction, it is worth noting that for a non-negative random variable, say Z, with survival function S_Z, we have $E[Z] =$

$\int_0^\infty P(Z > z)dz$. We can therefore obtain a prediction for the value of $d(\mathbf{s}_0)$ using

$$\hat{d}(\mathbf{s}_0) = \int_0^\infty \hat{S}_e(z|\mathbf{x}(\mathbf{s}_0))dz. \tag{4.5}$$

However, since $\hat{S}_e(d|\mathbf{x}(\mathbf{s}_0))$ is only estimated up to $d = \max\{d(\mathbf{s}_1), \ldots, d(\mathbf{s}_n)\}$, attention needs to be paid to how we can extend this to all larger d, that is, it is necessary to complete the function by extrapolating into the tail. For simplicity we assumed an exponential tail and estimated this by fitting a local-linear model to the pairs $\{(d(\mathbf{s}_i), \log(\hat{S}_e(d(\mathbf{s}_i)|\mathbf{x}(\mathbf{s}_0))))\}$, for $i = 1, \ldots, n$, using a kernel smoother as implemented in the `R` package `FKSUM` (Hofmeyr, 2022). Then, if ℓ is the resulting local-linear estimate we complete the tail of the estimated survival function by setting, for depth d^* larger than the greatest depth in the set of observations,

$$\hat{S}_e(d^*|\mathbf{x}(\mathbf{s}_0)) = \exp\left(\ell(d^*)\right). \tag{4.6}$$

4.2.4 Inverse probability of censoring weighting

In addition to SRF for point prediction in the presence of censoring, we propose an IPCW method which is well established in the statistical literature (Braekers & Veraverbeke, 2005; Huang & Wolfe, 2002; Zheng & Klein, 1994), and has been applied in other domains such as the health sciences (Bandyopadhyay et al., 2014). IPCW may reduce the bias in predictions caused by censoring by correcting for locations where soil thickness was censored, by giving extra weight to the locations where soil thickness was not censored. Specifically, each sampling location is weighted by the inverse of an estimate of the probability of having remained uncensored until depth, d. Locations where soil thickness was censored receive zero weight, except if the censoring occurred beyond a predefined depth, τ (explained below). One of the advantages of the IPCW method is that it can readily be applied with most "off-the-shelf" machine learning models as long as they can include observation weights. Below we provide an outline of the IPCW method, and for more details and mathematical proofs we refer the reader to (Vock et al., 2016).

We estimate the censored depth survival function, $S(c)$, with the Kaplan-Meier estimator (Kaplan & Meier, 1958) analogous to Eq. (4.3), but here it is estimated from the censored soil thickness data and from the entire calibration set (instead of just the node in a tree)

$$\hat{S}(c) = \prod_{j:c_j<c} \left(1 - \frac{a_j^*}{b_j}\right), \tag{4.7}$$

where a_j^* is the number of observations that were censored at depth c_j, and b_j is the number of observations whose depths exceed c_j. Note that the definition of b_j is the same as $b_{j,h}$ in Eq. 4.3, but a_j^* here refers to the number of observations that were censored, as opposed to $a_{j,h}$ in Eq. (4.3), which was the number of observations where the soil

thickness was reached (in node h). In addition, it should be noted that here, the Kaplan-Meier method provides a marginal estimate of $S(c)$, that is, irrespective of the covariates. Next, for each observation we define an IPC weight (Vock et al., 2016),

$$w(\mathbf{s}_i) = \begin{cases} \frac{1}{\hat{S}(\min(Y(\mathbf{s}_i),\tau))}, & \text{if } \delta(\mathbf{s}_i) = 1 \text{ or } C(\mathbf{s}_i) \geq \tau, \\ 0, & \text{otherwise.} \end{cases} \tag{4.8}$$

We select τ to be the largest value which does not result in extremely large weights given to some observations which could detrimentally affect the performance of the model. For example, the value for τ could be such that $\hat{S}(\tau) = 0.1$, which means that the resulting weights from Eq. (4.8) would not be larger than 10. The τ parameter can potentially be fine-tuned, e.g., with cross-validation, but this will require additional steps as the censored data need to be accounted for. For this reason the τ parameter was not fine-tuned in this study. The final step in the IPCW method is simply to incorporate the weights in Eq. (4.8) with a machine learning model. In this paper we used the weights with a regular RF model.

4.2.5 Synthetic simulation study

Synthetic data simulation experiments were performed in the `R` programming language (R Core Team, 2020) to compare the performance of SRF and WRF, and other modelling approaches, to account for right-censored soil thickness data under two main censoring scenarios. The first scenario is when censoring occurred at a fixed depth mimicking a situation where a survey follows a protocol to always reach a predefined depth, and the second is when censoring occurred at different depths for surveys without this constraint. In both of these scenarios we investigated model performance with various sample sizes, n, proportions of censored data, ρ, different values for the depth at which censoring occurred, λ, and two censoring mechanisms (non-informative and informative).

The first modelling approach to which SRF and WRF was compared was a random forest model with all data (ARF), that is, regardless if data were censored or not all data were treated as non-censored data. The second approach was a random forest model with only hard data (HRF), that is, censored data were excluded when the RF model was fitted. This approach was introduced in Section 4.2.1, i.e., the complete-data analysis approach. The reason we compared SRF and WRF to ARF and HRF was to show how the true soil thickness might be underestimated when survival data are not appropriately utilised (i.e., potential bias), and to show how the accuracy might be affected.

In each experiment we simulated a synthetic data set on a unit square discretised into a 100×100 grid comprising $N = 10\,000$ response values. First, we generated values, $z(\mathbf{s}_k)$, for $k = 1 \ldots N$, from a zero-mean Gaussian process with a covariance structure governed by an isotropic spherical variogram model with a nugget of 0.1, a partial sill of

0.9, and a range of 0.2. We also created three covariates which were functions of $z(\mathbf{s}_k)$. We then transformed the values, $z(\mathbf{s}_k)$, to follow a log-normal distribution which would then represent soil thickness, $d(\mathbf{s}_k)$. For more details concerning how the covariate values were calculated and how soil thickness was simulated from a log-normal distribution we refer the reader to Web Appendix A. A sample consisting of, $d(\mathbf{s}_i), \mathbf{x}(\mathbf{s}_i)$ for $i = 1, \ldots, n$, was obtained by randomly selecting either $n = 400$ or $n = 800$ of the N grid values. Note that for the i-th sample observation, $\mathbf{x}(\mathbf{s}_i)$ would consist of three values for the covariates. This is similar to how synthetic data were simulated in Chapter 2.

In the first scenario, for a certain censoring depth, λ, we randomly selected ρ proportion of the observations that were larger than λ, and replaced the observation with λ. To illustrate, suppose $\rho = 0.2$, and $\lambda = 60$, then 20% of the observations that were larger than 60 were randomly selected and replaced with a value of 60. In the second scenario, observations to be censored were selected in the same way as in the first scenario, that is, we randomly selected ρ proportion of the observations that were larger than λ, but then these were not set to a fixed censoring depth, but rather assigned various censoring depths. Specifically, a set of percentiles was determined from the actual depths before being censored. Each of the selected observations were then assigned one of the percentiles that was also larger than λ. Figure 4.2 provides an illustration of the difference between the two scenarios. For Scenario 1 a mixed discrete-continuous distribution is presented with the probability mass indicated by the light-yellow colour (censored data at a fixed depth), while the non-censored data is presented by the dark orange distribution. The two scenarios are shown for a censoring proportion of 0.6 and for different depths $(60, 90, 120)$. It is clear that as the censoring depth increases the distributions of the censored data move to the right, separating from that of the non-censored distributions.

In both scenarios, we also explored the impact of an informative censoring mechanism as opposed to a non-informative one. In the non-informative case, the selection of observations for censorship was done entirely at random. In the informative case, the probability of an observation being selected to be censored is inversely proportional to one of the covariates (refer to Web Appendix A). The simulations were conducted using the following parameter values: $n = \{400, 800\}, \rho = \{0, 0.1, 0.3, 0.6, 0.9\}, \lambda = \{60, 90, 120\}$. It is important to note that when $\rho = 0$ it means that there were no censored observations, and that the censoring depths, λ, are multiples of $\{1.0, 1.5, 2.0\}$ times the mean of the simulated soil thickness values on the grid. Finally, each combination was repeated 200 times, resulting in a total of $60 \times 200 = 12\,000$ simulations.

The response values in a selected sample are denoted by $\{y(\mathbf{s}_i), \delta(\mathbf{s}_i)\}$ for $i = 1, \ldots, n$, where $\delta(\mathbf{s}_i)$ has the same definition as in Section 4.2.2. The modelling approaches (ARF, HRF, SRF, WRF) were fitted on $\{y(\mathbf{s}_i), \delta(\mathbf{s}_i)\}$ along with $\mathbf{x}(\mathbf{s}_i)$ the generated covariate values. In the case of ARF and HRF, $\delta(\mathbf{s}_i)$ was ignored and the models were either fitted on all or only the non-censored observations. All the random forest models were

implemented with the `randomForestSRC` package (Ishwaran et al., 2008) in `R`. In addition, for the random forest models, the hyper-parameter, m_{try}, as well as the minimum node size and the number of trees were kept at the default values, i.e., 1 (when $p = 3$), 5 and 500, respectively. For WRF we used the `case.wt` functionality in `randomForestSRC` to incorporate the weights estimated with Eq. (4.8), and τ was set such that $\{\hat{S}(\tau) = 0.1\}$. Finally, predictions were generated for the entire grid and the models were evaluated with the mean error (ME) and root mean square error (RMSE)

$$\text{ME} = \frac{1}{N}\sum_{k=1}^{N}\left(d(\mathbf{s}_k) - \hat{d}(\mathbf{s}_k)\right), \tag{4.9}$$

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{k=1}^{N}\left(d(\mathbf{s}_k) - \hat{d}(\mathbf{s}_k)\right)^2}, \tag{4.10}$$

where the $\hat{d}(\mathbf{s}_k), k = 1 \ldots N$ are the predictions over the entire grid. Note that the $d(\mathbf{s}_k)$ are the known true soil thickness values, which are known in this simulation study, and therefore it was not required to account for censoring of the test data in the evaluation step as opposed to a real-world case study.

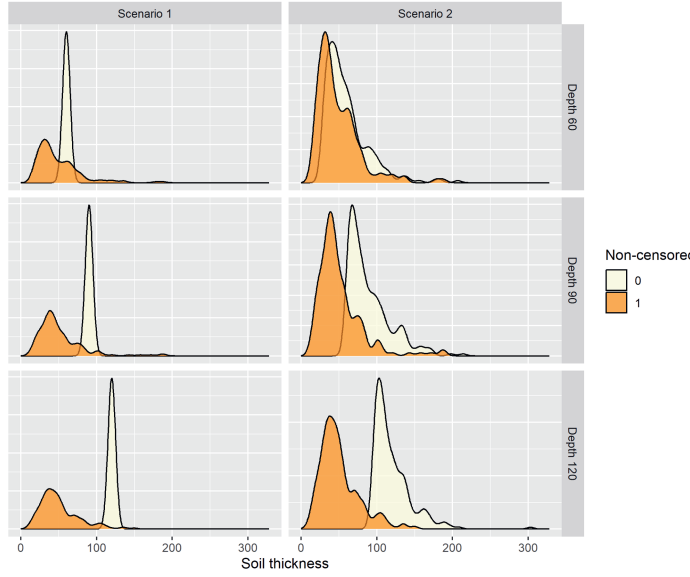


Figure 4.2: Scenarios with fixed censoring (left) and different censoring depths (right) are depicted with probability distributions from one of the simulations. Non-censored distributions are depicted with dark orange while the censored distributions are depicted by light-yellow. For Scenario 1, a mixed discrete-continuous distribution is presented with the mass indicated by the light-yellow colour (censored data). The two scenarios are shown for a censoring proportion of 0.6 and for different depths $(60, 90, 120)$.

4.2.6 Real-world applications

The first real-world application is from Maine, USA, which is located in the North-Eastern region of the country and has a surface area of approximately $91\,646\,km^2$. The case study consisted of 5 666 sampling locations (on average 1 sampling location per $16\,km^2$) as shown in Figure 4.3 that were purposely chosen by soil surveyors during initial mapping and traditional soil survey update efforts. The dot sizes in Figure 4.3 are proportional to the soil thickness and the dark orange dots represent locations where soil thickness was observed. There were 3 856 locations where soil thickness was censored which meant that 68.1% of the observations were right-censored. Locations where the true soil thickness was recorded occurred mostly in the north-central and central regions of the study region. Soil thickness was defined as the depth from the soil surface (including any organic horizons) to a lithic (i.e., bedrock) contact. It should be noted that censoring occurred at a fixed depth of $165\,cm$, and there were only seven observations that were non-censored that were deeper than $165\,cm$ (these ranged from $170\,cm$ to $213\,cm$).

The covariates for this case study were prepared at $5\,m$ pixel resolution consisting in

various derivatives of a digital elevation model as well as hydrologic properties (see Web Appendix B for more information). For computational purposes, the covariates were resampled to a resolution of 250 m. Additionally, six observations with missing values in the covariates were excluded, resulting in a total of 5 660 observations available for model training.

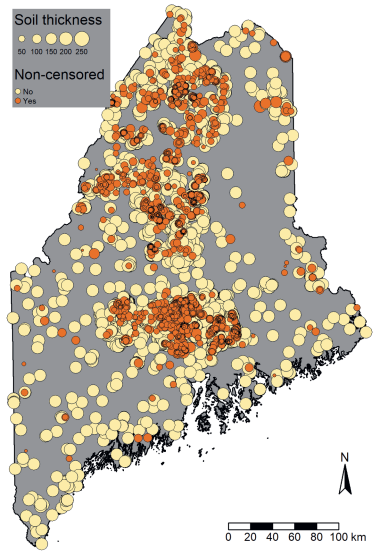


Figure 4.3: Soil thickness (in cm) observations in Maine, USA. The dot sizes are proportional to the soil thickness and the dark orange dots represent locations where soil thickness was observed.

For the second real-world application we used soil thickness data from the arable land of the Canton of Zurich located in the North-East of Switzerland. The surface area of this region is approximately 1 729 km^2. The majority (75 %) of the data set originated from fully described and analysed soil profiles recorded for a conventional detailed soil mapping campaign in 1988–1997. We complemented this data with older surveyed locations, but not older than from 1975 what resulted in 3 924 observations (Service center NABODAT, 2022), with thus on average 1 sampling location per 0.44 km^2 (Figure 4.4). Soil thickness was derived for this study from the in-situ recorded horizon qualifiers according to Swiss soil classification (Jäggli et al., 1998). Soil thickness was defined as the upper limit of the first occurring horizon considered unstructured parent material (i.e. horizons with only C or R qualifiers, excluding transition horizons) (Jäggli et al., 1998). When no single C or R qualifier was recorded, then an observation was considered censored with the surveyed depth available only. In Figure 4.4, non-censored observations are shown as dark orange

and censored observations as light-yellow. Soil profile locations were purposively chosen by pedological experts to support conventional map polygon delineation (Jäggli et al., 1998). Both sets of observations are well-dispersed throughout the arable land of the Canton of Zurich. The proportion of censored observations for this case study was equal to 61.1%. The covariates used for this case study are presented in Web Appendix C.

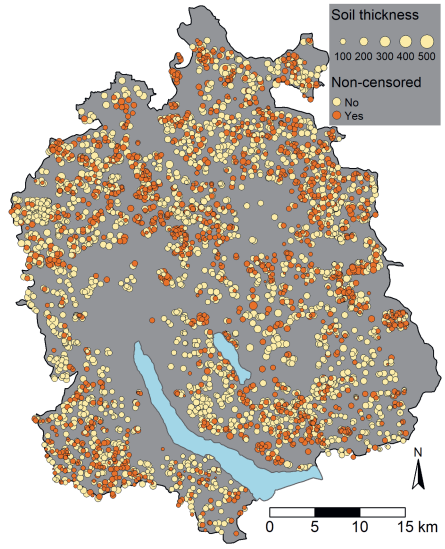


Figure 4.4: Location of soil thickness recordings on the arable land of the Canton of Zurich, Switzerland.

4.2.7 Model calibration and evaluation

As with the synthetic simulation study in Section 4.2.5, we employed the modelling approaches, ARF, HRF, SRF, and WRF in the two case studies introduced in Section 4.2.6. The models were calibrated and assessed with a k-fold nested cross-validation. In the outer-loop of the nested cross-validation, the data set is repeatedly divided (k times) into training and test sets. Then, for the inner-loop, the training data is repeatedly divided (k times) into a calibration and validation set. The calibration and validation sets are used for model calibration, that is, selecting the optimal hyper-parameters (we used the RMSE as the evaluation metric). Then, the selected model is assessed with the test set of the outer-loop. This is the same approached used in Chapter 3. It should be noted that ARF, SRF, and WRF were calibrated on the non-censored and the censored data, while HRF was calibrated only on the non-censored data. For each model we tuned the m_{try}, and the minimum node size. We used the default value for the number of trees as used in the `RandomForestSRC` package. In addition, for WRF we set τ, such that $\hat{S}(\tau) = 0.1$.

In this paper, we used $k = 10$ in the cross-validation.

To assess the models (with the test sets in the outer-loops) we used standard performance metrics, that is, the ME, RMSE and the concordance correlation coefficient (CCC) (Lawrence & Lin, 1989). These metrics were determined on three variations of the test sets. The first was to treat censored data as true observations, thereby ignoring the censored nature of the data. This is will not give a true reflection of model performance, and the ARF modelling approach will most likely perform best in this situation. However, to account for the fact that the test data might be censored, the second variation was to calculate the performance metrics only on non-censored data. Thirdly, we also calculated the metrics on test data that were first truncated to a certain threshold, d_{val}. The truncation is applied on non-censored observations in the test set, as well as on censored observations when the censored depth is at least as large as the threshold, d_{val}. For instance, suppose $d_{val} = 100\,cm$ then an observation (and the corresponding prediction) with soil thickness values larger than d_{val} will be truncated to $100\,cm$, otherwise the observation (and prediction) stay the same. In both case studies we performed testing with $d_{val} = \{100\,cm; 150\,cm\}$. We used this second approach as it might be of interest for a map user to be only interested in soil thickness down to a certain depth after which the knowledge of the exact soil thickness becomes less pertinent. For example, in wheat production, an agronomist might be interested to know what soil thickness is up to a value of $100\,cm$ (Fan et al., 2016).

For the case studies we used the same software to implement the models as in Section 4.2.5. It should be noted that in survival analysis, model evaluation is usually done with Harrell's concordance index (C-index) as it accounts for censoring (Harrell et al., 1982). However, for the calculation of the C-index for a model, an estimate of $S(d)$ (or the CHF) is required. Therefore, we did not evaluate the models in this paper with this approach, as SRF is the only model that can directly produce this output. For an example of the C-index in soil thickness modelling, we refer the reader to (Chen et al., 2019).

4.2.8 Evaluation by comparison of prediction maps

We also evaluated the models on the basis of prediction maps. The maps were produced by fitting the random forest models on all the training data (i.e., 5 660 observations for the Maine case study, and 3 924 for Zurich case study), except HRF which was fitted only on the non-censored data. Two sets of maps were produced by using different hyper-parameters. For the first set, the models were fitted with the hyper-parameters that resulted in the smallest test RMSE calculated on the evaluation approach that only considered the non-censored data. The second set of maps was produced with the models with hyper-parameters obtained with the smallest test RMSE based on the truncated approach with $d_{val} = 100\,cm$. For the latter, the map values were truncated to $d_{val} = 100\,cm$.

4.3 Results

4.3.1 Synthetic simulation study

The RMSE results for the censoring scenarios (constant versus non-constant) for the non-informative censoring mechanism are presented in Table 4.1. In this table we also present the results for a RF model that was fitted with the true soil thickness which can be used as a baseline to compare with the other models. In addition, for each unique combination of simulation parameters, the best modelling approach is highlighted in boldface. The ME results are presented in Web Appendix A.

The RMSE results of the RF model indicated that regardless of the value of ρ, and λ (and the censoring scenario), ARF, HRF, SRF, and WRF were compared against baselines of 24.0, and 22.3 when the sample size, n, was set to 400, and 800, respectively. When $\rho = 0.0$, the RMSE of ARF and HRF were comparable to that of the baseline. SRF performed worse than ARF and HRF when $\rho = 0.0$ for both censoring scenarios, while WRF produced similar results to that of ARF and HRF.

A close look at Table 4.1 revealed the following highlights concerning the effect of simulation parameters. In general, it was observed that, for a specific censoring depth, sample size, and censoring scenario, an increase in ρ led to a corresponding increase in the RMSE. As expected this confirms a decrease in model performance as the information content of the data decreases due to fewer known soil thickness observations available to the models. For example, in the case of the first censoring scenario, when $\lambda = 60$, $\rho = 0.3$ and $n = 400$, the RMSE is equal to 26.5, 25.0, 26.3 and 24.4 for ARF, HRF, SRF and WRF, respectively, and when ρ increased to 1 the RMSE increased to 39.0, 51.4, 38.9 and 39.0. For a larger sample size, at a given censoring proportion and depth, it was expected to note a decrease in RMSE. However, when $\rho = 1$ the RMSE values were similar between the two cases of $n = \{400, 800\}$. Note that when $\rho = 1$ it means that all observations larger than or equal to λ were censored. Therefore, we do not expect an increase in model performance for a larger sample size. Finally, smaller RMSE values were noted in the case of higher λ values, especially for larger values of ρ. This is attributed to the increase in the number of censored observations when the censoring depth, λ is smaller.

Table 4.1: RMSE results for the synthetic simulation study for both censoring scenarios (constant versus non-constant censoring depths) and for the non-informative censoring mechanism. Results are shown for censoring proportion $\rho = \{0, 0.3, 0.6, 0.9, 1\}$, censoring depth $\delta = \{60, 90, 120\}$ and sample size $n = \{400, 800\}$. Model results include RF fitted on the true soil thickness, and ARF, HRF, SRF, WRF, fitted on censored soil thickness. For each unique combination of simulation parameters, the best approach is presented in boldface.

Parameters			$n = 400$					$n = 800$				
Scenario	λ	ρ	RF	ARF	HRF	SRF	WRF	RF	ARF	HRF	SRF	WRF
1	60	0.0	24.0	24.0	24.0	26.4	**23.7**	22.3	22.4	22.4	24.1	**22.2**
		0.1	24.0	24.5	24.2	26.3	**23.7**	22.3	23.0	22.5	23.9	**22.3**
		0.3	24.0	26.5	25.0	26.3	**24.4**	22.3	25.2	23.3	24.1	**22.9**
		0.6	24.0	30.5	27.4	26.6	**26.3**	22.3	30.0	25.4	**24.7**	**24.7**
		0.9	24.0	36.4	36.3	**32.5**	35.4	22.3	36.0	34.1	**32.2**	33.3
		1.0	24.0	39.0	51.4	**38.9**	39.0	22.3	**38.4**	51.9	39.2	**38.4**
	90	0.0	24.0	24.0	23.9	26.3	**23.6**	22.3	22.3	22.3	23.8	**22.1**
		0.1	24.0	24.3	24.1	26.4	**23.7**	22.3	22.6	22.4	23.9	**22.2**
		0.3	24.0	25.2	24.8	26.2	**24.2**	22.3	23.6	22.9	23.9	**22.5**
		0.6	24.0	27.7	26.7	26.7	**25.4**	22.3	26.9	24.9	24.6	**24.1**
		0.9	24.0	30.4	32.6	**28.1**	31.7	22.3	30.2	30.5	**26.5**	30.0
		1.0	24.0	**31.5**	41.7	31.6	31.6	22.3	**31.4**	44.3	33.5	31.5
	120	0.0	24.0	24.0	24.0	26.4	**23.7**	22.3	22.3	22.3	23.9	**22.1**
		0.1	24.0	24.1	24.1	26.1	**23.6**	22.3	22.6	22.5	24.0	**22.2**
		0.3	24.0	24.5	24.5	26.0	**23.9**	22.3	23.3	23.1	24.2	**22.6**
		0.6	24.0	26.0	26.0	26.5	**25.1**	22.3	24.1	23.6	23.9	**23.0**
		0.9	24.0	27.4	29.1	**26.8**	28.0	22.3	26.5	27.3	**24.9**	26.4
		1.0	24.0	**28.2**	32.7	28.8	**28.2**	22.3	**27.2**	31.7	27.4	**27.2**
2	60	0.0	24.0	24.0	24.0	26.4	**23.7**	22.3	22.4	22.4	24.1	**22.2**
		0.1	24.0	24.1	24.2	26.0	**23.7**	22.3	22.5	22.5	23.7	**22.2**
		0.3	24.0	24.6	25.0	26.4	**24.2**	22.3	23.1	23.2	24.2	**22.7**
		0.6	24.0	**25.3**	27.1	28.2	25.8	22.3	**24.2**	25.2	26.5	24.4
		0.9	24.0	**26.9**	34.3	35.5	29.2	22.3	**25.8**	31.9	38.4	28.5
		1.0	24.0	**27.5**	41.1	31.0	31.0	22.3	**26.5**	39.7	31.0	30.5
	90	0.0	24.0	24.0	23.9	26.3	**23.6**	22.3	22.2	22.3	23.8	**22.1**
		0.1	24.0	24.0	24.2	26.0	**23.6**	22.3	22.4	22.4	23.7	**22.2**
		0.3	24.0	24.3	24.8	25.9	**24.1**	22.3	22.7	22.9	23.7	**22.6**
		0.6	24.0	**24.9**	26.6	26.9	25.4	22.3	**23.6**	24.6	25.3	23.8
		0.9	24.0	**25.4**	31.5	28.5	27.0	22.3	**24.4**	29.3	29.8	26.0
		1.0	24.0	**25.6**	35.9	27.6	28.4	22.3	**24.5**	34.5	26.6	27.3
	120	0.0	24.0	24.1	24.0	26.4	**23.7**	22.3	22.3	22.3	23.9	**22.1**
		0.1	24.0	24.0	24.2	25.9	**23.7**	22.3	22.4	22.5	23.9	**22.2**
		0.3	24.0	24.2	24.7	25.9	**24.1**	22.3	22.9	23.1	24.0	**22.7**
		0.6	24.0	**24.6**	25.8	26.2	25.0	22.3	**22.8**	23.6	23.8	23.1
		0.9	24.0	**24.8**	28.9	26.5	25.5	22.3	**23.5**	26.8	24.9	24.2
		1.0	24.0	**25.1**	30.8	26.8	26.2	22.3	**23.6**	29.1	24.7	24.6

In terms of model performance in the first censoring scenario, WRF consistently demonstrated superior results than SRF and the other modelling approaches (when $\rho \leq 0.6$, and regardless of censoring depth and sample size, with the only exception at $\rho = 0.6$ and $n = 800$ in which case SRF and WRF were comparable). Then, when $\rho = 0.9$ (regardless of censoring depth and sample size) SRF exhibited the best performance, and when $\rho = 1$ the results of ARF, SRF and WRF were comparable. Under the second censoring scenario, WRF was consistently superior in comparison to SRF and the other modelling approaches (when $\rho \leq 0.3$, and regardless of censoring depth and sample size). Then, when $\rho \geq 0.6$, ARF was superior, surpassing WRF. HRF and SRF performed poorly in the second censoring scenario, especially when λ was equal to 60. The only exception was when $\rho = 1$ wherein SRF produced comparable results to WRF.

The ME results, presented in Web Appendix A, indicated that for larger values of ρ ($0.6 \leq \rho \leq 1$) and smaller values for λ ($60 \leq \lambda \leq 90$), ARF, HRF and WRF increasingly underestimated the true soil thickness, while SRF increasingly overestimated it. The reason for the overestimation of SRF for larger ρ is because more observations were sooner "reached" in Eq. (4.3), because they were censored at λ. This then led to larger survival probabilities for data with true depths larger than λ. However, in case of the first censoring scenario, SRF underestimated the true soil thickness. This is because in Eq. (4.3) no distinct values larger than depth, λ were available.

The results pertaining to the informative censoring mechanism are presented in Web Appendix A. Notably, the outcomes of ARF, HRF, and WRF were relatively unaffected by the specific censoring mechanism employed. Conversely, SRF showcased inferior performance in the presence of the informative censoring mechanism in the first censoring scenario. Specifically, SRF yielded less accurate and more biased predictions. Nevertheless, in the second censoring scenario, the censoring mechanism had minimal impact on SRF.

4.3.2 Results for the real-world applications

Maine case study

The distribution of soil thickness is illustrated in Figure 4.5. The figure shows a histogram only for the non-censored observations with a dashed line representing the fixed censoring depth of 165 cm. The plots indicated that the distribution was skewed to the right. It is important to point out that almost 100% of the data with a depth of at least 165 cm were censored (only seven observations above this threshold were non-censored).

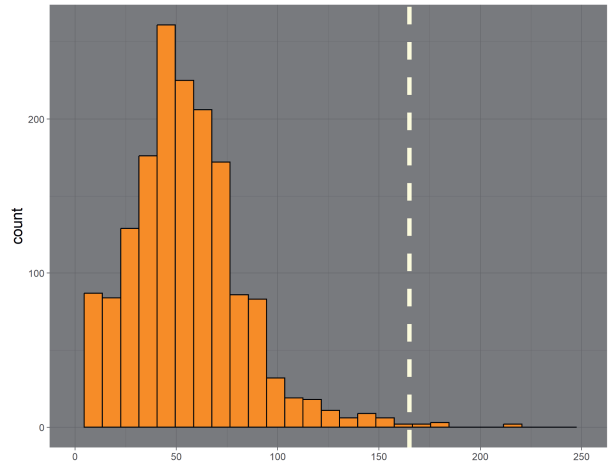


Figure 4.5: Distribution of soil thickness in Maine, USA. A histogram only for the non-censored data (dark orange) is shown, with the light-yellow dashed line representing the fixed censoring depth of 165 cm.

Table 4.2 displays the cross-validation (outer-loop) results for the Maine case study. The table showcases the ME, RMSE, and CCC values for the ARF, HRF, SRF, and WRF models, assessed as discussed in Section 4.2.7. When model testing was performed with the first approach (treating censored data as true observations), both ARF and WRF outperformed HRF and SRF. This is evident from the RMSE values, which were 44.7 and 44.4 for ARF and WRF, respectively, while HRF and SRF had RMSE values of 88.1 and 83.8, respectively. In case of testing on only the non-censored data, it is apparent that HRF and SRF were superior. Lastly, when testing with the truncated data, we once again observed that ARF and WRF exhibited superior performance in both cases of $d_{val} = \{100\,cm; 150\,cm\}$.

Table 4.2: Cross-validation results for the Maine case study. Evaluation metrics were calculated for the outer-loops of the nested cross-validation. The results are shown when using all the data as well as for two alternative strategies, that is, evaluating on non-censored data, and with truncated observations and predictions ($d_{val} = \{100\,cm; 150\,cm\}$).

Evaluation	Metric	ARF	HRF	SRF	WRF
All	ME	-1.3	-70.8	-65.2	-0.9
(Censored and non-censored)	RMSE	44.7	88.1	83.8	44.4
	CCC	0.527	0.079	0.087	0.534
Non-censored	ME	49.9	0.6	5.6	49.7
	RMSE	61.3	27.4	29.3	61.4
	CCC	0.162	0.354	0.287	0.163
Val. depth ($100\,cm$)	ME	11.6	-26.1	-20.9	11.6
	RMSE	27.2	36.9	33.5	27.2
	CCC	0.343	0.207	0.233	0.351
Val. depth ($150\,cm$)	ME	6.1	-60.5	-55.0	6.2
	RMSE	39.8	76.1	71.7	39.7
	CCC	0.533	0.092	0.106	0.538

The soil thickness maps for the Maine case study are displayed in Figure 4.6. For each of ARF, HRF, SRF, and WRF, two maps are shown. Maps (a), (b), (c), and (d) of Figure 4.6 were produced with ARF, HRF, SRF and WRF, respectively, fitted with the hyper-parameters which resulted in the smallest test RMSE calculated with the second evaluation approach discussed in Section 4.2.7 (non-censored data only). Maps (e), (f), (g), and (h) were produced with ARF, HRF, SRF, and WRF, fitted with the smallest test RMSE calculated with the third evaluation approach (i.e., truncated data). The latter four maps were then also truncated to $100\,cm$. Note that it falls outside the scope of this study to provide detailed interpretations of all the maps, but we do outline the most prominent features with regards to the aim of this study, which is to compare the results of WRF to SRF and the other modelling strategies (for both case studies).

The maps produced by ARF and WRF were very similar and showed no noteworthy differences. This is because with the WRF model, τ was set to $160\,cm$ which meant that all censored data of $165\,cm$ received a weight, and thereby were included in the calibration of the model. For ARF and WRF, the deepest soils were observed in the central, north-eastern, and south-western regions of Maine. HRF and SRF produced maps with much smaller soil thickness values, especially HRF. Most predictions produced by these two models were also less than $100\,cm$. The 90-th percentile of the predictions of HRF was $78.5\,cm$ while for the SRF model it was $93.7\,cm$. It is therefore clear that HRF and SRF severely underestimated (deeper) soil thickness in this study.

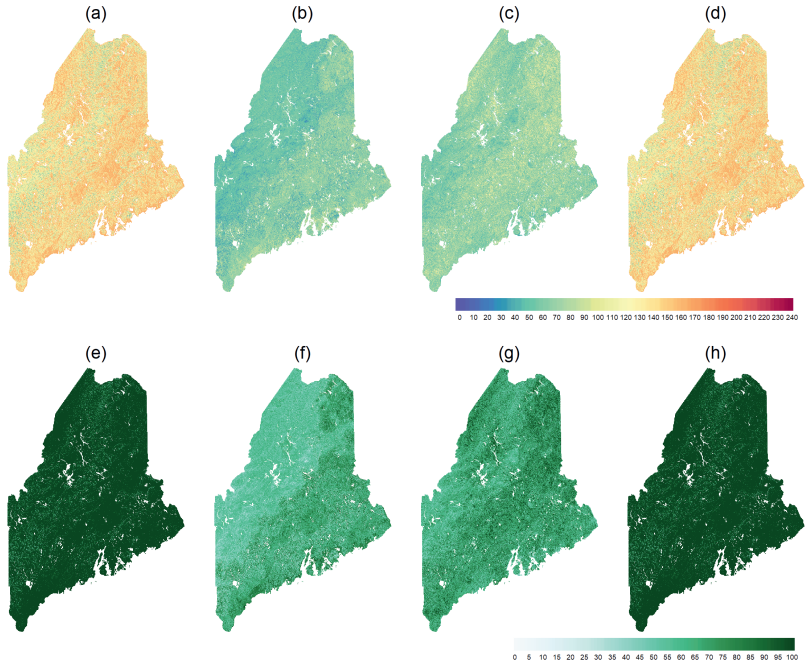


Figure 4.6: Predicted soil thickness maps for the Maine case study for (a) ARF; (b) HRF; (c) SRF; (d) WRF, fitted with the hyper-parameters which resulted in the smallest test RMSE calculated with the non-censored data, and (e) ARF; (f) HRF; (g) SRF; (h) WRF, fitted with the smallest test RMSE calculated with the truncated data. The latter four were also truncated to 100 cm.

Switzerland case study

Figure 4.7 presents the distribution of the soil thickness data. It can be seen that the shape of the distributions of the non-censored and censored data are similar (slightly right-skewed), but the distribution of the censored data (light-yellow) lies more to the right. It is important to point out that about 75 % of the data that were greater than the overall mean of 75 cm were censored.

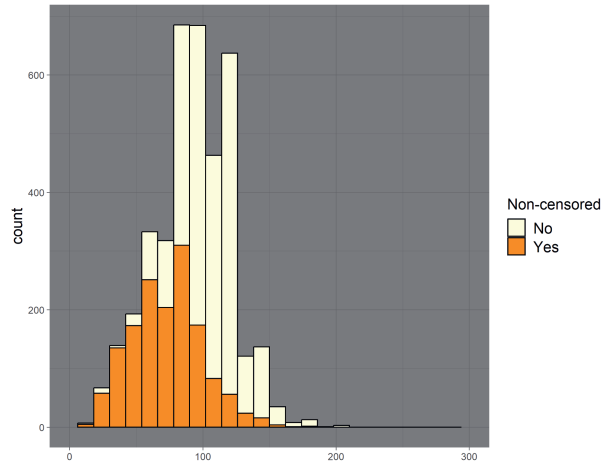


Figure 4.7: Distribution of censored and non-censored soil thickness in the Switzerland case study. Two histograms are presented, one for the censored data and one for the non-censored soil thickness data.

Table 4.3 presents the cross-validation results for the Switzerland case study. When the evaluation was done with the first evaluation approach (treating censored data as true observations), ARF demonstrated superior performance as expected. This is evident from the RMSE value which was lower than that of other models. In contrast to the Maine case study, we observed that HRF and SRF performed relatively well when testing with all the data, achieving RMSE values of 32.3 and 28.0, respectively. In terms of evaluation with only the non-censored data, HRF exhibited the best performance with an RMSE of 24.1, followed by WRF with an RMSE of 25.4. When considering evaluation with the truncated data, ARF, SRF and WRF were comparable in the 100 cm case. We also noted that SRF and WRF produced the least biased results in the 100 cm case. In the 150 cm case WRF performed best with a RMSE of 26.0 and a CCC of 0.427. However, in comparison to HRF, the difference between the RMSE values was negligible.

Table 4.3: Cross-validation results for the Switzerland case study. Evaluation metrics were calculated for the outer-loops of the nested cross-validation. The results are shown when using all the data as well as for two alternative testing strategies, that is, evaluation on non-censored data, and evaluation with truncated observations and predictions ($d_{val} = \{100\,cm; 150\,cm\}$).

Evaluation	Metric	ARF	HRF	SRF	WRF
All	ME	-0.1	-17.0	-5.9	-8.5
	RMSE	26.3	32.3	28.0	29.1
	CCC	0.333	0.174	0.294	0.326
Non-censored	ME	15.0	0.2	8.5	6.5
	RMSE	28.3	24.1	26.3	25.4
	CCC	0.269	0.288	0.267	0.365
Val. depth ($100\,cm$)	ME	6.9	-8.0	1.4	-1.6
	RMSE	20.0	21.4	19.4	19.8
	CCC	0.356	0.275	0.376	0.389
Val. depth ($150\,cm$)	ME	12.9	-2.0	6.4	4.8
	RMSE	29.0	26.4	27.6	26.0
	CCC	0.306	0.293	0.296	0.427

The maps produced by ARF, HRF, SRF, and WRF for the Switzerland case study are presented in Figure 4.8. As with with the Maine case study, HRF and SRF produced lower soil thickness values compared to that of ARF and WRF. The truncated maps of HRF and SRF in (f) and (g) are therefore very similar to (b) and (c), respectively. ARF and WRF produced larger values of soil thickness, especially WRF as seen in the Western and East-central parts of the Canton of Zurich.

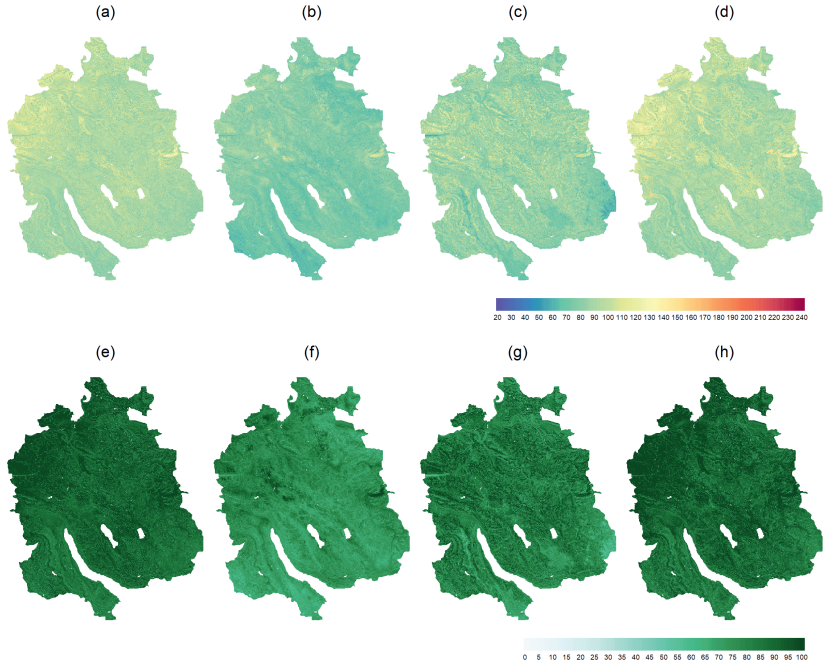


Figure 4.8: Predicted soil thickness maps for the Switzerland case study for (a) ARF; (b) HRF; (c) SRF; (d) WRF, fitted with the hyper-parameters that resulted in the smallest test RMSE calculated on the non-censored data, and (e) ARF; (f) HRF; (g) SRF; (h) WRF, fitted with the smallest test RMSE calculated on the truncated data. The latter four were also truncated to 100 cm.

4.4 Discussion

Handling right-censored data poses a great challenge as it relates to a situation with data that have reduced information. While there has been some exploration in the literature of DSM on how to deal with censored data, as evidenced by studies such as Lacoste et al. (2016); Kempen et al. (2015); Chen et al. (2019), users should be mindful that results will often fall short of optimal when compared to results that would have been obtained with data that were not censored (data for which the true observations are known), particularly when the proportion of censored data is large, and if censoring occurs at shallower depths. This is because less information of the true soil thickness is known, and if the censored nature of the data is then not accounted for, predictions will not reflect the true soil thickness process. The synthetic simulation study confirmed this. In Table 4.1, when the proportion of censored data was larger ($\rho \geq 0.3$) and when censoring occurred at shallower depths (i.e., $\lambda = 60$), all of the models produced much poorer results compared to the

baseline (when the true soil thickness was known and used). In a different simulation study that aimed to assess the estimation of several survival curves across different proportions of censored data (Willems et al., 2018), also revealed that results were notably worse when the proportion of censored data was larger than 0.35.

The novel application of using an IPCW random forest model to map soil thickness has to the best of our knowledge not been used in DSM. The WRF model performed well in the simulation study which investigated model performance for different censoring scenarios. Specifically, WRF consistently demonstrated superior performance when the proportion of censored data (that exceeded a specified threshold) was at most 0.6 for a fixed censoring depth, and no more than 0.3 for varying censoring depths. The superior performance of WRF is attributed to the model's ability to directly account for censored data by assigning larger weights to the non-censored data. This results in an overall improved fitted random forest model and, consequently, more accurate predictions. For larger proportions, in case of a Scenario 1, WRF produced comparable results to the other models, however, it is worth noting that SRF consistently produced the best results when $\rho = 0.9$. Recall that in Scenario 1, the selected observations to be censored (those with values exceeding a certain depth), all received the same censored value, while in Scenario 2, the selected observations were assigned different censored depth values.

In the simulation study, in the case where all observations that exceeded a certain depth were censored ($\rho = 1$), the results revealed that the best option is to either use ARF or WRF in case of Scenario 1, and to use ARF in case of Scenario 2. The reason for the comparable performance between ARF and WRF in the first scenario, is because the τ parameter allowed all censored observations to be included in the calibration step by giving equal weight to all. These findings suggest that due to the limited information available about the true soil thickness when $\rho = 1$, using a survival-related model like SRF may not necessarily confer an advantage. In such cases, a user might as well opt for an approach like ARF or a modelling approach in which the occurrence of the true (deep) soil thickness is modelled with an dichotomous data analysis approach as used in Malone & Searle (2020b).

The effect of the censoring mechanism was minimal and mainly affected SRF in the first censoring scenario. This is because under an informative censoring mechanism, censored depths provide prognostic information about the true soil thickness which can lead to the survival model producing biased results (Leung et al., 1997; Willems et al., 2018; Kleinbaum et al., 2012).

In the simulation study, model evaluation could be carried out using the true soil thickness data, but this will not be feasible in real-world applications. While Harrell's concordance index (C-index) is a common choice for model evaluation in survival analysis (Harrell et al., 1982), it requires an estimate of the survival function or the CHF. Within the context of DSM, this method was for example used in Chen et al. (2019). Another

method for evaluating models with censored data involves using an IPCW approach, as used in Graf et al. (1999). Although we explored this method in our study, its preference for WRF due to a similar formulation in the model, providing an unfair advantage, led to notably superior results compared to other models. Consequently, we opted to exclude it from our study. Instead, we adopted two alternative and meaningful evaluation strategies in the DSM domain. The first involved evaluating only with non-censored data, while the second involved evaluation with data truncated to specific depths - $100\,cm$ and $150\,cm$. This decision was influenced by the potential interest of a map user in soil thickness down to a certain depth, beyond which the exact thickness may be less important.

The Maine real-world case study, similar to the first scenario in the simulation study where $\rho = 1$ and $\lambda = 120$, indicated that ARF and WRF were comparable and outperformed HRF and SRF. This was evident from Table 4.2, specifically when the models were evaluated with all of the data in test set (using non-censored data and treating censored data as true), as well as when the models were evaluated with the truncated data. As noted before, this is not unexpected as limited information available about the true soil thickness is known. Therefore, a modelling approach like the one used in Malone & Searle (2020b) could also be more appropriate, but such an approach could possibly be further improved by modelling the smaller depths with WRF instead of using a regular random forest with imputations from a beta distribution (Kempen et al., 2015).

In the Zurich case study, both WRF and SRF performed relatively well in terms of the RMSE metric, especially when the evaluation was conducted using truncated data. However, both survival models did overestimate the true soil thickness (more so in the case of SRF). A similar conclusion was made in Malone & Searle (2020b) concerning SRF. It is noteworthy that conclusions for the Zurich case study are not as straightforward because several aspects are occurring simultaneously. For instance, it is likely that censoring is informative and exhibits a spatial correlation structure with the observation depths. Moreover, censoring is not at a fixed depth as for the Maine case study and occurs at low soil thickness values already.

Further research is needed to also derive uncertainty maps for the models employed in the two case studies. For the random forest models, this can be accomplished using methods outlined in Hengl et al. (2018). Regarding SRF, one straightforward approach involves extracting soil thickness from the survival function, such as obtaining the 5% and 95% percentiles in case of a 90% prediction interval.

4.5 Conclusion

Soil thickness data are often right-censored, indicating that the sampling depth is smaller than the true soil thickness. In this paper, we proposed an IPC weighted machine learning model to address this issue, assigning extra weight to non-censored data and zero weight

to censored data, unless censoring occurred beyond a predefined depth. This model involves two stages. First, it estimates a survival function of censored soil thickness data from which it then calculates the weights. Second, it incorporates these weights with a machine learning model. In this paper, we used these weights with a random forest model. We compared the proposed model with a SRF and other strategies for dealing with right-censored data. The models were evaluated in a synthetic simulation study under various censored scenarios. The results of the simulation study showed that WRF demonstrated superior performance when the proportion of censored data (that exceeded a certain depth) was equal to or less than 0.6. The models were also assessed in two case studies using metrics (ME, RMSE, CCC) that were computed on test data sets that were adjusted for censoring. With these studies we also demonstrated that WRF is a viable option for modelling right-censored soil thickness data.

Supplementary materials

The supplementary materials, Web Appendices A - C, can be downloaded from the following link: https://github.com/CSVDW/Soil-thickness-modelling.

Chapter 5

Biplots for understanding machine learning predictions in digital soil mapping

This chapter is based on:

van der Westhuizen, S., Heuvelink, G.B.M., Gardner-Lubbe, S., Clarke, C.E. (2024). Biplots for understanding machine learning predictions in digital soil mapping.

Under review, *Ecological Informatics.*

5.1 Introduction

Soil maps play a crucial role in various fields by providing valuable information about the spatial distribution of soil properties. A widely used tool for generating these maps is digital soil mapping (DSM) (McBratney et al., 2003), which often makes use of machine learning models like the random forest (RF) model (Minasny & McBratney, 2016a). The reason for its wide use is that machine learning models can effectively capture complex nonlinear relationships between soil properties and environmental covariates, leading to more accurate soil maps compared to traditional statistical models such as multiple linear regression and geostatistical models (Wadoux et al., 2020a). However, unlike traditional models, a notable drawback of machine learning models is that they are often considered as "black-box" methods due to their limited ability to provide comprehensive interpretations for their predictions. In this paper, we adopted the definition provided by Belle & Papantonis (2021) for "black-box" machine learning models, which refers to models that are not simulateable by a human, lack decomposability, and are algorithmically nontransparent. This definition includes models like RF, support vector machines, and multi-layered neural networks.

Explainable machine learning (XML) is a rapidly growing field in machine learning literature. It focuses on methods designed to understand the predictions made by machine learning models (Biecek & Burzykowski, 2021). The field of XML is extensive, and for a comprehensive overview we refer to Biecek & Burzykowski (2021) and Belle & Papantonis (2021). XML primarily consists of model-specific and model-agnostic methods which can then be used to make local and/or global interpretations. A local interpretation refers to when the goal is to assess how predictions are made for individual cases, i.e., spatial locations, but perhaps also for different smaller sub regions in the region of interest. On the other hand, a global interpretation offers insights into the overall behaviour of the model, i.e., over the entire region of interest, such as by considering the overall importance of the environmental covariates (variable importance statistics) (Biecek & Burzykowski, 2021).

Model-specific XML approaches utilise the intrinsic properties of the machine learning model itself to provide explanations. For instance, in tree ensembles like RF, approaches such as feature relevance are used to explain the model's output. Feature relevance, indicated by variable importance statistics, is widely utilised to determine the most relevant and influential covariates when making predictions. Feature relevance is also commonly used in DSM, and studies like Subramanian & Rajendra (2022) and Zizala et al. (2022) have employed variable importance statistics, such as mean decrease in node impurity, to gain insights into which covariates were deemed important by a tree ensemble. However, a notable drawback of model-specific methods is their limited applicability, as they can only be used with a specific machine learning model. For instance, in the case of a RF model, one can easily obtain feature relevance through conventional variable importance

metrics, whereas for models like support vector machines, these methods are not readily accessible (Belle & Papantonis, 2021).

The second type of XML is model-agnostic methods, i.e., techniques not reliant on the specific architecture of a machine learning model. According to Belle & Papantonis (2021), model-agnostic approaches can be categorised into three main types: (1) explanation by simplification, (2) feature relevance, and (3) visual explanations. In the first category one widely used method for interpreting model predictions is known as Local Interpretable Model-Agnostic Explanations (LIME) (Ribeiro et al., 2016) which employs a local approximation to understand the behavior of a machine learning model locally. It achieves this by creating a linear model or a decision tree around the predictions and utilising this surrogate model to explain the underlying machine learning model. One drawback of model simplification methods is that the approximation cannot always be quantitatively assessed (Belle & Papantonis, 2021), making empirical demonstrations necessary to evaluate the accuracy of the approximation.

In the second category, one of the most prominent contributions to XML are Shapley values (Shapley, 1953), which originate from coalitional Game Theory. In essence, a Shapley value represents the average marginal contribution of a covariate to a prediction. Wadoux & Molnar (2022) and Padarian et al. (2020); Wadoux et al. (2023) demonstrated the use of Shapley values in the context of DSM and demonstrated the applicability of Shapley values in both local and global model interpretation (Wadoux & Molnar, 2022). However, their applicability can be limited when used for contructing partial dependence plots, as this then assumes covariates to be uncorrelated (Molnar, 2020).

In the third category, two popular visualisation methods are Individual Conditional Expectation (ICE) and Partial Dependence (PD) plots (Goldstein et al., 2015; Friedman, 2001). ICE plots visualise how a prediction changes for different values of a specific covariate using line plots. Each line in an ICE plot corresponds to a spatial location and is generated by holding the values of all other covariates constant except for the one of interest. Therefore, ICE plots are used for local interpretations. In contrast, a PD plot demonstrates how predictions behave on average as a function of one or more covariates. PD plots are used for global interpretations. Later on in the paper we point out that Shapley values can also be used as a visualisation approach. In the context of DSM, Wadoux & Molnar (2022) demonstrated the use of ICE and PD plots when modeling soil organic carbon (SOC). A limitation of these visualisation methods is that they also assume covariates to be uncorrelated. For cases where covariates are correlated, an alternative visualisation approach for global interpretations is the Accumulated Local Effect (ALE) plot (Apley & Zhu, 2020). But the ALE plot is limited to depicting one, two or three covariates at a time.

In this paper we propose the use of principle component analysis (PCA) biplots as a model-agnostic visualisation approach for understanding the predictions made by a machine

learning model. A biplot is a powerful visualisation tool that is often used to seek patterns in multivariate data (Gower et al., 2011). Our proposed biplot methodology allows a user to investigate machine learning model predictions both at sub regional scale and globally, and does not require restrictions such as covariates to be uncorrelated. Theoretically, there is also no limit to the number of covariates that can be viewed at a time, but adding too many covariates may clutter the biplot and hinder optimal visualisation. Furthermore, an analytically derived goodness-of-fit is provided which allows the user to evaluate the accuracy of the approximation. Previous work by Rowan (2019); Rodwell et al. (2021); Wei et al. (2023) used biplots to understand the predictions of a machine learning model in the case of classification, but to the best of our knowledge, no work has been done to use biplots to elucidate machine learning predictions in regression. Biplots have also been used in several fields such as soil science (Odeh et al., 1991b; McBratney et al., 2018), finance (Van der Merwe, 2020; Rodwell et al., 2021), engineering (Aldrich et al., 2004) and archaeology (Wurz et al., 2005), but not yet in DSM. The aim of this paper is therefore to introduce the methodology of using biplots for understanding machine learning predictions in DSM, and to compare biplots to other XML methods like ICE and PD plots, ALE plots and Shapley values.

The article is structured in the following way: in Section 5.2 we provide a brief introduction to PCA biplots, introduce the proposed methodology of using PCA biplots to explore the predictions of a machine learning model, and also introduce a real-world case study in which we map SOC in South Africa. In Section 5.3 we present our results of modelling SOC in South Africa with a RF model, as well as interpreting the predictions with biplots and other XML methods. In Section 5.4 we provide a general discussion, and in Section 5.5 we present a summary of our conclusions.

5.2 Material and methods

5.2.1 Introduction to biplots

Biplots is an extension of multi-dimensional scaling which provides an optimal two-dimensional representation of a multivariate data set (Gower et al., 2011). A biplot consists of points and lines. The points represent the observations and act as a measure of similarity between the observations in the data set. That is, the closer two observations are in the original data set, the more similar the corresponding points will be in the biplot. The lines reflect the variables and act as axes on the two-dimensional display, which allows the data points to be read by projecting them perpendicularly onto the axes.

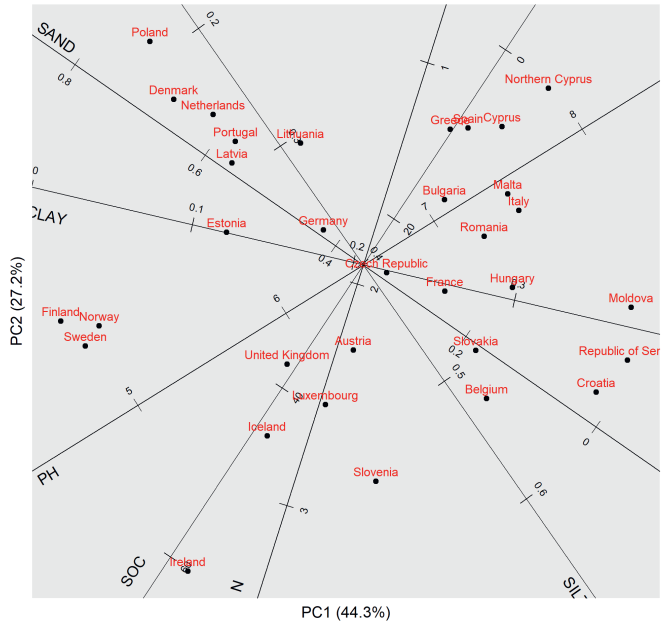


Figure 5.1: A biplot showing an optimal two-dimensional display of the LUCAS soil data aggregated by country. For each country the median of the particular soil property was obtained.

An example of a biplot is provided in Figure 5.1. The biplot is constructed from the 2018 Land Use and Coverage Area frame Survey[1] (LUCAS) soil data set which has been aggregated by country. That is, for each country we determined the median SOC, total nitrogen, pH and particle-size fractions. The points in Figure 5.1 are labelled by the names of the countries, and the variables are displayed as black axes with tick marks. The vertical and horizontal sides of the graph are called scaffolding axes and are irrelevant to the interpretation of the points and axes in the biplot. Note that a biplot can also represent the variables as arrows that originate at the centre, $(0; 0)$. This is commonly referred to as a Gabriel biplot (Gabriel, 1971), and we present an example of such a biplot in Web Appendix A. A biplot, such as in Figure 5.1, is referred to as a Gower biplot (Gower & Hand, 1996), upon which the methodology proposed in this paper is based on.

In Figure 5.1, the angle between the lines (more specifically, the cosine of the angle between the lines) approximates the correlation between the variables. An angle close to 90° or 270° represents a weak or zero correlation, and an angle close to 0° or 180° represents a strong or perfect correlation of 1 or -1, respectively. In Figure 5.1 we note a strong

[1] `https://esdac.jrc.ec.europa.eu/projects/lucas`

positive correlation between SOC and nitrogen, and a weak correlation between pH and silt. It should be noted that clay and sand have a strong negative correlation. This is indicated by the tick marks on the corresponding axes which increase in the opposite directions (i.e., an angle close to 180°). Each point can be read off from a variable axis by projecting it perpendicularly to the line. For example, if Poland, Denmark, and the Netherlands are projected onto the Sand axis we note high values for the median sand content for these countries while in comparison we note low values for Croatia, Serbia and Moldova. The distance between two points approximates the Euclidean distance between those two points in the multivariate space. Therefore, points that are far away from each other have a large Euclidean distance, and vice versa. This property of a biplot allows the user to detect clusters. For example, we note that Finland, Norway and Sweden are close to each other, but far away from, say Croatia. Note that the variables in the data set that was used to construct this biplot were standardised to have zero-means and variances of one.

5.2.2 Mathematical background for PCA biplots

Suppose that $\mathbf{X}$, an $(n \times p)$ standardised data matrix (i.e., columns are zero-mean and have unit variances), with n the number of observations or locations, p the number of covariates and $n \geq p$, is to be represented in an r-dimensional display, with $r < p$. PCA can be used to approximate $\mathbf{X}$ in r dimensions, denoted by $\hat{\mathbf{X}}$, such that the least squares error between $\mathbf{X}$ and $\hat{\mathbf{X}}$ is a minimum. PCA makes use of the singular value decomposition (Everitt et al., 2001) of $\mathbf{X}$

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T, \tag{5.1}$$

where $\mathbf{U}$ is a $(n \times k)$ orthonormal matrix with $k = \min(n, p)$, and $\mathbf{V}$ is a $(p \times p)$ orthonormal matrix with the eigenvectors of $\mathbf{X}$. $\mathbf{D}$ is a $(n \times k)$ matrix with the singular values of $\mathbf{X}$, λ_k, as the (k, k) entries for $k = 1, 2, \ldots, \min(n, p)$. Note that when $k = p$, $\mathbf{D}$ reduces to a diagonal $(p \times p)$ matrix. Note also that the singular values of $\mathbf{X}$ are the square roots of the eigenvalues of $\mathbf{X}^T\mathbf{X}$ (Everitt et al., 2001). The approximated data matrix is determined with

$$\hat{\mathbf{X}} = \mathbf{U}\mathbf{D}\mathbf{J}\mathbf{V}^T, \tag{5.2}$$

where $\mathbf{J}$ is a $(p \times p)$ matrix written as

$$\mathbf{J} = \begin{bmatrix} \mathbf{I}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}.$$

and $\mathbf{I}_r$ is a $(r \times r)$ identity matrix. Finally, to obtain a biplot, the principal component scores, (i.e., the points in the biplot) $\mathbf{XVJ} = \mathbf{UDJ}$, are plotted. To add the lines, also called the principal component loadings, the columns of $\mathbf{VJ}$ are plotted. To obtain

a two-dimensional display one would set $r = 2$ so that two principal components are plotted. Note that the first two are usually selected, but the first and third or other binary combinations of principal components can also be visualised in a two-dimensional display.

To assess the goodness-of-fit of the biplot one can obtain the "quality" of the biplot. This is calculated as the proportion of the variances of the columns in $\mathbf{X}$ explained by $\hat{\mathbf{X}}$(Gower et al., 2011)

$$\frac{\text{trace}(\mathbf{\Sigma J})}{\text{trace}(\mathbf{\Sigma})}, \tag{5.3}$$

where $\mathbf{\Sigma} = \mathbf{D}^2$ is the variance-covariance matrix of $\mathbf{XV}$. Eq. (5.3) yields a value between zero and one with one indicating a perfect fit. The quality of the biplot in Figure 5.1 was 0.715 which is the sum of the variances explained by the first two principal components, $0.443 + 0.272 = 0.715$.

Predictivity of a variable axis or a point in a biplot refers to how well the axis or point is approximated in the two dimensional display. The predictivity of the axes and the points of a biplot are given by (Gower et al., 2011)

$$\text{diag}(\hat{\mathbf{X}}^T\hat{\mathbf{X}})(\text{diag}(\mathbf{X}^T\mathbf{X}))^{-1}, \tag{5.4}$$

and

$$\text{diag}(\hat{\mathbf{X}}\hat{\mathbf{X}}^T)(\text{diag}(\mathbf{X}\mathbf{X}^T))^{-1}, \tag{5.5}$$

respectively. Eqns. (5.4) and (5.5) yield values between zero and one for each axis and point, respectively. A value close to one means that the corresponding axis or point is very accurately approximated in the biplot. The predictivity of the axes for the biplot in Figure 5.1 are shown in Table 5.1 while the predictivity of the points are shown in Table 5.2. For example, these results illustrate that the axis for sand is very accurately represented in the biplot while the axis for N is not so accurate. In addition, the point for Finland is very accurate and hence is reliable, but the point for the Czech Republic is very inaccurately represented and care should be taken to interpret or project this point onto an axis.

Table 5.1: Axes predictivity of the LUCAS biplot.

Variable	Predictivity
Sand	0.970
pH	0.758
SOC	0.747
Silt	0.654
Clay	0.646
N	0.517

Table 5.2: Point predictivity of the LUCAS biplot.

Finland	Denmark	Latvia	Sweden	Italy	Poland
0.992	0.980	0.978	0.966	0.964	0.951
Spain	Netherlands	Slovenia	Greece	Lithuania	Hungary
0.929	0.928	0.900	0.899	0.898	0.886
N. Cyprus	Estonia	Portugal	Norway	Cyprus	Austria
0.878	0.855	0.851	0.829	0.828	0.803
Serbia	Slovakia	Croatia	France	Luxembourg	Malta
0.803	0.792	0.790	0.738	0.687	0.686
Ireland	Bulgaria	Germany	Belgium	UK	Moldova
0.666	0.624	0.605	0.600	0.537	0.400
Romania	Iceland	Czech Republic			
0.269	0.225	0.082			

One disadvantage of utilising a biplot arises when the data matrix is large as this can result in a crowded representation, thus impeding effective visualisation. In Figure 5.2(a), a biplot with all LUCAS observations is shown (non-aggregated, there are 21 977 observations). Examining for patterns, especially in the lower section of the graph, is challenging. In Figure 5.2(b), an enhancement has been made to the biplot by replacing the points with a density plot, providing a clearer representation. For example, in the lower section of the graph it is now clearer that the majority of the points lie more to the right, and that particle-size fractions are mostly responsible for this trend from left to right. In Figure 5.2(c), the colour of the points in the biplot corresponds to the levels of the coarse fragments (CF) variable. The CF variable was initially not included in the analysis, and is now used to depict the results of Figure 5.2(a) with different colours according to the values of CF. The same depiction is shown in Figure 5.2(d), but the levels of the CF variable are depicted with alpha-bags. An alpha-bag is a contour that encloses the innermost α proportion of the observations in a biplot (Rousseeuw et al., 1999; Gardner et al., 2005). The purpose of alpha-bags is to quantify class separation. This is determined by identifying the largest α value, say α^*, which defines the $\alpha^*\%$ innermost bivariate data points within one class that overlaps with only the $(100 - \alpha)\%$ most extreme bivariate data points contained in the other classes. For further details on the construction of alpha-bags we refer the reader to Aldrich et al. (2004) and Gower et al. (2011). From Figures 5.2(c) and (d) it is noted that particle-size fractions are mostly responsible for the CF configurations from left to right.

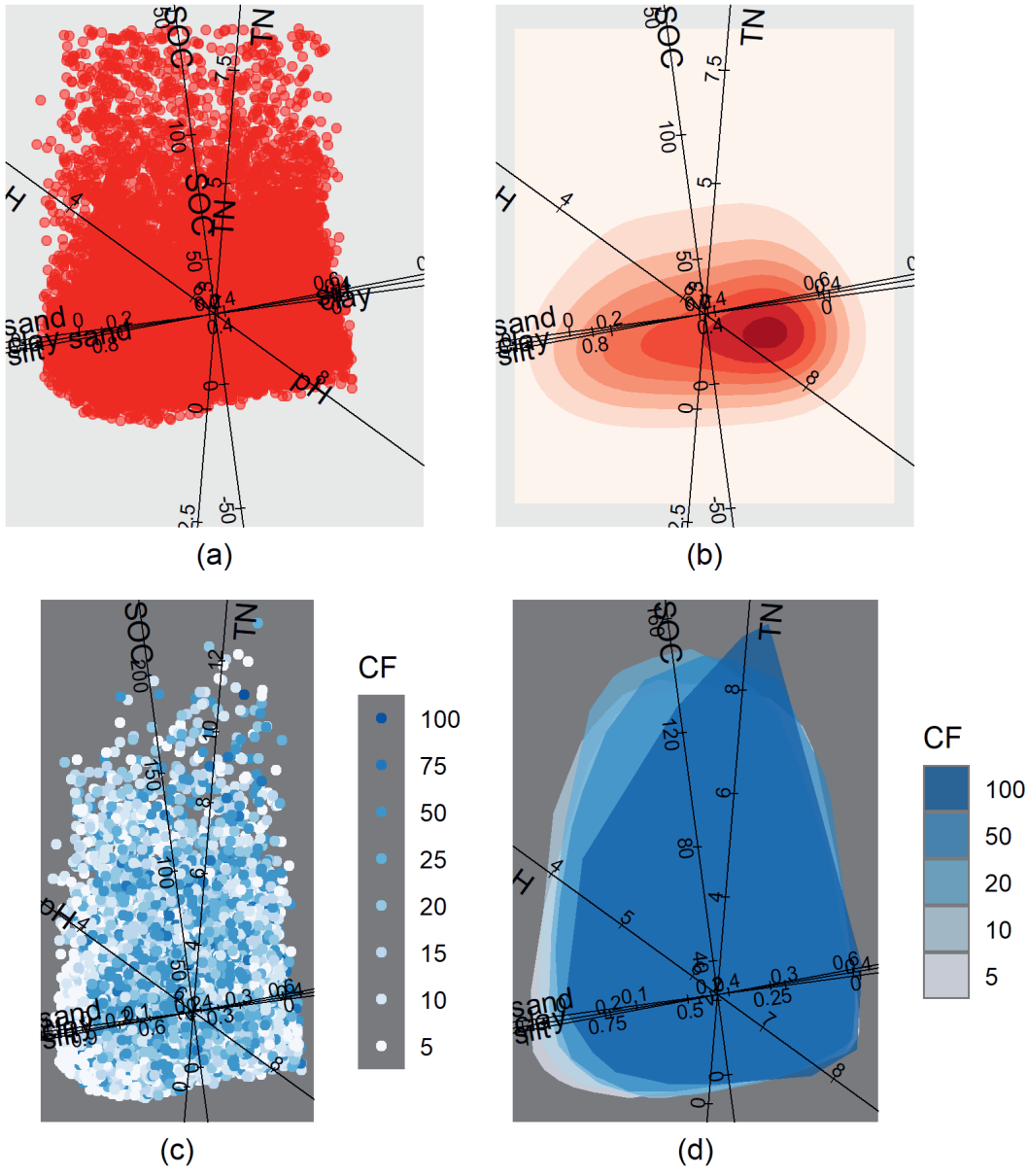


Figure 5.2: (a) A cluttered biplot showing a two-dimensional display of over 22 000 observations in the LUCAS data set. (b) A biplot of the LUCAS data and points shown with a bivariate density. (c) The biplot in (a) with points configured by the levels of CF (in %). (d) The biplot in (d) is shown with 70% alpha-bags.

5.2.3 Biplots for understanding machine learning predictions

Let the soil property of interest at any location $\mathbf{s}$ in a geographical region $\mathcal{D}$ be denoted by $Y(\mathbf{s})$, for $\mathbf{s} \in \mathcal{D}$. We model the soil property with a machine learning model, denoted by $f(\cdot)$, such that

$$Y(\mathbf{s}) = f\left(\mathbf{x}(\mathbf{s})\right) + \epsilon(\mathbf{s}), \tag{5.6}$$

where $\mathbf{x}(\mathbf{s})$ is a p-dimensional vector of covariates at $\mathbf{s}$, and $\epsilon(\mathbf{s})$ is assumed a zero-mean normal random variable (Wadoux et al., 2020a). In DSM, our main aim is then to obtain a prediction function, $\hat{f}(\cdot)$, by means of minimising some loss function, which can then be used to predict Y at all locations in $\mathcal{D}$.

With our methodology we propose to construct a PCA biplot with the environmental covariates (i.e., covariate maps) that were used to produce the final prediction map, $\{\hat{y}(\mathbf{s}) \; \forall \; \mathbf{s} \in \mathcal{D}\}$. Therefore, $\mathbf{X}$ in Eq. 5.1 would refer to an $(N \times p)$ matrix, with N the total number of grid cells in the region of interest. Note that the covariate data should first be normalised to have zero means and unit variances. In the biplot, the covariates then act as the axes and the points are shown with various colours according to the predictions, $\hat{y}(\mathbf{s})$. Covariate maps usually consist of millions or even more grid cells which could make the biplot difficult to interpret (e.g., Figure 5.2(c)). Therefore, to improve the visualisation of the biplot, we propose to use alpha-bags to represent the predictions. Note that our methodology can also be used with validation or test data in which case visualisation might not be a problem, and therefore predictions (or the residuals) may be shown with or without alpha-bags.

In the situation where the number of covariates, p, is very large, it can result in an overcrowded biplot. In such cases, it is possible to choose a subset of q covariates, with $q < p$, before conducting the PCA. We explore this further in Section 5.3.

Biplots used to analyse machine learning predictions are model-agnostic as one simply needs to supply the covariates along with a vector consisting of the corresponding machine learning predictions. Hence, this method is not dependent on the architecture of a specific model that was used to produce the predictions.

5.2.4 Other model-agnostic visualisation methods

Individual conditional expectation and partial dependence plots

An ICE curve indicates how a prediction at a certain location changes for changes in a covariate(s) of interest (Goldstein et al., 2015). Suppose that $\mathbf{X}_S$ is a subset of $\mathbf{X}$ that consists of q covariates of interest, and $\mathbf{X}_{-S}$ is the compliment of $\mathbf{X}_S$ so that $f(\mathbf{X}) = f(\mathbf{X}_S, \mathbf{X}_{-S})$ (usually $q = \{1, 2\}$). Then, let $\mathbf{x}_S$ be the values of $\mathbf{X}_S$ at a location of interest, and let $\mathbf{x}_{-S}$ be all other covariate values at that location. The ICE curve visualises how a prediction changes when only the values of the covariate of interest, $\mathbf{x}_S$, changes. That

is, we visualise $\hat{f}(\mathbf{x}_S^*, \mathbf{x}_{-S})$ as a function of $\mathbf{x}_S^*$, where $\mathbf{x}_S^*$ takes on a grid of values in the range of $\mathbf{X}_S$, and $\mathbf{x}_{-S}$ are kept fixed. Note that the ICE curve is a local interpretation method as it shows how a prediction changes at a specific location. It is also common to centre ICE curves to improve visualisation

$$\hat{f}_{cICE}(\mathbf{x}_S^*, \mathbf{x}_{-S}) = \hat{f}(\mathbf{x}_S^*, \mathbf{x}_{-S}) - \hat{f}(x_0, \mathbf{x}_{-S}),$$

where x_0 refers to a baseline value which is usually set to the minimum in $\mathbf{x}_S$ (Goldstein et al., 2015).

A PD function is formally defined as the expectation of $\hat{f}(\mathbf{x}_S, \mathbf{X}_{-S})$ over the marginal distribution of $\mathbf{X}_{-S}$ (Friedman, 2001). That is,

$$f_{PD}(\mathbf{x}_S) = E_{\mathbf{x}_{-S}}\left(\hat{f}(\mathbf{x}_S, \mathbf{X}_{-S})\right).$$

The PD curve is estimated by averaging ICE curves over all $\mathbf{x}_{-S}(\mathbf{s}_i)$ in that

$$\hat{f}_{PD}(\mathbf{x}_S^*) = \frac{1}{N}\sum_{i=1}^{N} \hat{f}(\mathbf{x}_S^*, \mathbf{x}_{-S}(\mathbf{s}_i)). \tag{5.7}$$

The PD curve is used as a global interpretation method as it shows the average marginal effect of a covariate on the predictions over all locations. As with ICE curves, PD curves can also be centred

$$\hat{f}_{PD}(\mathbf{x}_S^*) = \hat{f}_{PD}(\mathbf{x}_S^*) - \hat{f}_{PD}(x_0, \mathbf{x}_{-S}(\mathbf{s}_i)), \tag{5.8}$$

Finally, recall that for ICE and PD plots we assume that the $\mathbf{x}_S$ are not correlated to $\mathbf{x}_{-S}$ (Hastie et al., 2008).

Accumulated local effect

PD plots may give unrealistic results of the effect of a covariate on the model predictions if that covariate is correlated to another. This is because the PD plot averages over the marginal distribution of the covariates in set $-S$. ALE plots have been proposed when covariates are correlated, and they are based on the idea of integrating over averaged local effects in small neighbourhoods in $\mathbf{x}_S$ (Apley & Zhu, 2020). Formally, let z_0, be some starting point then for up to a point x_S in the covariate of interest, the accumulated local effect is defined as

$$f_{ALE}(x_S) = \int_{z_0}^{x_S} E_{\mathbf{X}_{-S}|\mathbf{x}_S}\left(\frac{\partial \hat{f}(\mathbf{x}_S, \mathbf{X}_{-S})}{\partial \mathbf{x}_S}|\mathbf{x}_S = z\right) dz, \tag{5.9}$$

where z defines a point in the covariate of interest at which the local effect is first determined with the partial derivatives, then averaged (the expectation over $\mathbf{x}_{-S}$ conditional on $\mathbf{x}_S$), and then integrated up to x_S.

The partial derivatives in Eq. 5.9 are approximated by finite differences of predictions within K intervals in $\mathbf{x}_S$. The lower and the upper bounds of the k-th interval are given by $k-1$ and k, for $k = 1, \ldots, K$. If a value within the covariate of interest is located in the k-th interval, then that value is replaced by the bounds of the interval while the other covariates are kept constant. The finite differences are then given by $\hat{f}(z_k, \mathbf{x}_{-S}(\mathbf{s}_i)) - \hat{f}(z_{k-1}, \mathbf{x}_{-S}(\mathbf{s}_i))$. At point x_S, the ALE is then given by (Apley & Zhu, 2020)

$$\hat{f}_{\text{ALE}}(x_S) = \sum_{k=1}^{k(x_S)} \frac{1}{N_S(k)} \sum_{i:x_S(\mathbf{s}_i)\in[z_{k-1};z_k]} \left(\hat{f}(z_k, \mathbf{x}_{-S}(\mathbf{s}_i)) - \hat{f}(z_{k-1}, \mathbf{x}_{-S}(\mathbf{s}_i)) \right), \tag{5.10}$$

where $k(x_S)$ denotes the index of the interval in which the value x_S falls, and $N_S(k)$ denotes the number of observations (grid cells) inside the k-th interval. To improve visualisation, the ALE curve can be centred with

$$\hat{f}_{\text{cALE}}(x_S) = \hat{f}_{\text{ALE}}(x_S) - \frac{1}{N} \sum_{i=1}^{N} \hat{f}_{\text{ALE}}(\mathbf{x}_S(\mathbf{s}_i)). \tag{5.11}$$

For more details concerning the approximation concerning the ALE curve we refer the reader to (Apley & Zhu, 2020; Molnar, 2020).

Shapley values

Recall that Shapley values originate from coalitional game theory (Shapley, 1953). Within the context of machine learning, suppose that in a game where the prediction is the "payout", Shapley values distributes the payout among the covariates which are considered to be the players in the game. Recall that S denotes the subset of covariates of interest. Suppose that $S \subseteq \{1, \ldots, p\} \setminus \{ j\}$ is a subset that does not include the j-th covariate, then, for observation $\mathbf{x}$, a Shapley value for covariate j is given by

$$\phi_j(\mathbf{x}) = \sum_{S\subseteq\{1,\ldots,p\}\setminus\{j\}} \frac{|S|!(p-|S|-1)!}{p!} \left(\hat{f}(\mathbf{x}_{S\cup\{j\}}) - \hat{f}(\mathbf{x}_S) \right), \tag{5.12}$$

where $|S|$ is the size of the subset that excludes the j-th covariate, $S \cup \{j\}$ is the subset with the j-th covariate added, and $\hat{f}(\mathbf{x}_S)$ is the prediction function where covariates not in S are marginalised (similarly for $\hat{f}(\mathbf{x}_{S\cup\{j\}})$) (Molnar, 2020; Wadoux & Molnar, 2022). The solution to Eq. 5.12 is computationally very demanding as it requires the sum of the marginal contribution of more than $2^p - 1$ combinations. Work done by for example Strumbelj & Kononenko (2014) provides ways to approximate Eq. 5.12 with a Monte Carlo method.

Shapley values are used for local explanations and are interpreted as the average contribution of a covariate to the difference between the prediction and the average prediction. However, they can also be combined to perform global interpretations. For instance, the average of the absolute values of Shapley values can be used as a feature importance method, or the average of Shapley values plotted against values of a covariate of interest may be interpreted as the partial dependence (Molnar, 2020). Another use of Shapley values involves mapping Shapley values at all locations in a region of interest so that a spatial pattern is obtained (Wadoux et al., 2023). However, such an approach would be very computationally demanding.

5.2.5 Real-world case study and practical implementation

In this study we mapped topsoil SOC for South Africa with a RF model. Studies that also mapped SOC for South Africa with machine learning include Odebiri et al. (2023) and Venter et al. (2021). Understanding the spatial variation of SOC in South Africa is important for applications in ecology, hydrology and agronomy (Department of Environmental Affairs, 2015; Schulze & Schütte, 2020; Du Preez & van Huyssteen, 2011). It is also a key parameter in climate change studies, as soil carbon forms the largest part of the terrestrial carbon pool in South Africa (Department of Environmental Affairs, 2015). The topsoil (A Horizon) SOC data (shown in %) used in this study were from 7 714 profiles which were extracted from the South African Profile Database of the Agricultural Research Council - Institute of Soil Climate and Water. A full description of this database is described in Paterson et al. (2015). The sampling locations, shown in Figure 5.3, are well dispersed throughout the country with on average 1 sampling location per $152\,km^2$ except in the Northern Cape, i.e., North-Central region, which is mostly characterised by vast expanses of arid and semi-arid landscapes, with on average 1 location per $912\,km^2$. We had 109 covariates (at $250m \times 250m$) which included maps of various vegetation indices, an elevation model, precipitation and surface temperatures, etc. The covariates used are mainly the same as those used in SoilGrids 2.0 (Poggio et al., 2021).

The RF model was implemented with the `randomForestSRC` package in `R`. We decided to use RF as it is one of the most widely used machine learning models in DSM (Wadoux et al., 2020a). The model was calibrated and assessed with a 10-fold nested cross-validation, and with the calibration, we determined the optimal m_{try} and minimum node size hyper-parameters by minimising the mean square error (MSE). The number of trees was set equal to 500. The model was evaluated with the outer loops of the nested cross-validation with commonly used validation statistics in DSM, that is, with the mean error (ME), root MSE (RMSE), model efficiency coefficient (MEC), and the concordance correlation coefficient (CCC) (Wadoux et al., 2020a). The final SOC map was produced by fitting the RF model on the entire data set using the identified optimal hyper-parameters that minimised the validation MSE.

PCA biplots with alpha-bags were constructed to explain the predictions made by the RF model at national and regional scale. Due to the large number of covariates and to enhance the visualisation and interpretability of the biplot, we chose a subset of covariates to serve as axes in the biplot. However, there is a total of $2^p - 2$ possible subsets available to visualise (excluding the full- and null sets). We therefore selected the top 8 performing covariates based on their variable importance statistics as suggested by the RF model (Breiman, 2001), considering it a good starting point. We also selected a subset of covariates that represented the `clorpt` soil formation model of Jenny (1941). It is important to note that the aim of this paper is to compare biplots to other XML methods for a given subset of covariates. Therefore, any subset of covariates could have been chosen. One important aspect to understand about biplots is that the configuration will change when covariates are removed or added to the biplot. This characteristic of the biplot is further explored in Section 5.4. Biplots were constructed with the `ordr` (Brunson, 2023) and `ggplot2` packages (Wickham, 2016) in `R`. All biplots were constructed with 90% alpha-bags. Examples of `R` code to construct biplots are presented in Web Appendix F.

We compared the biplots to ICE and PD plots as well as to ALE plots and Shapley values. For these methods we used the `iml` (Molnar et al., 2018) and `fastshap` packages (Greenwell, 2023) in `R`. For calculation of Shapley values we used 100 Monte Carlo simulations which is sufficient as indicated in Wadoux et al. (2023). Shapley values needed to be approximated at about 20 million grid cells (the number of grid cells in the entire country) which is computationally infeasible. Therefore, to speed up computations, we conducted a systematic grid sampling similar to what was performed in Wadoux et al. (2023). This resulted in 400 000 grid cells at which the Shapley values were estimated.

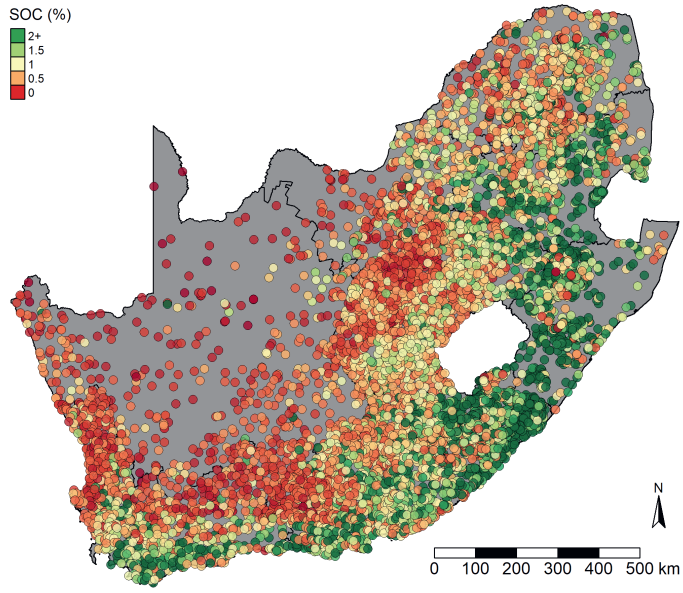


Figure 5.3: Sampling locations in South Africa (on average 1 sampling location per 152 km^2). Points were configured by the distribution of SOC% with green indicating larger SOC percentages.

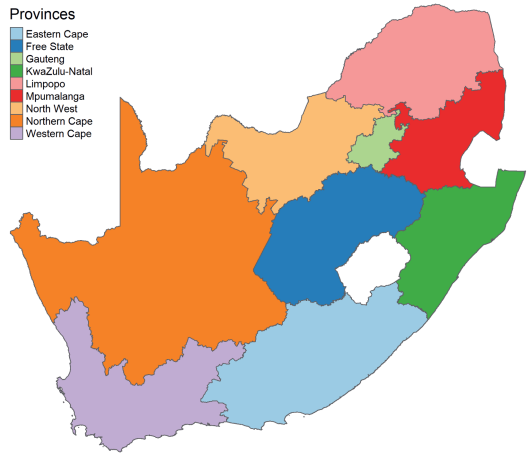


Figure 5.4: Provinces of South Africa.

5.3 Results

5.3.1 Modelling results

Figure 5.5(a) presents a density plot illustrating the distribution of topsoil SOC in South Africa. Additionally, Figure 5.5(b) displays separate density plots for the provinces of the country. The distribution of SOC across the entire nation was skewed to the right, with roughly 90% of observations falling below 2.5%. KwaZulu-Natal exhibited the highest concentration of SOC, followed by Mpumalanga, while the Northern Cape displayed the lowest concentration with approximately 90% of observations falling below 0.8%.

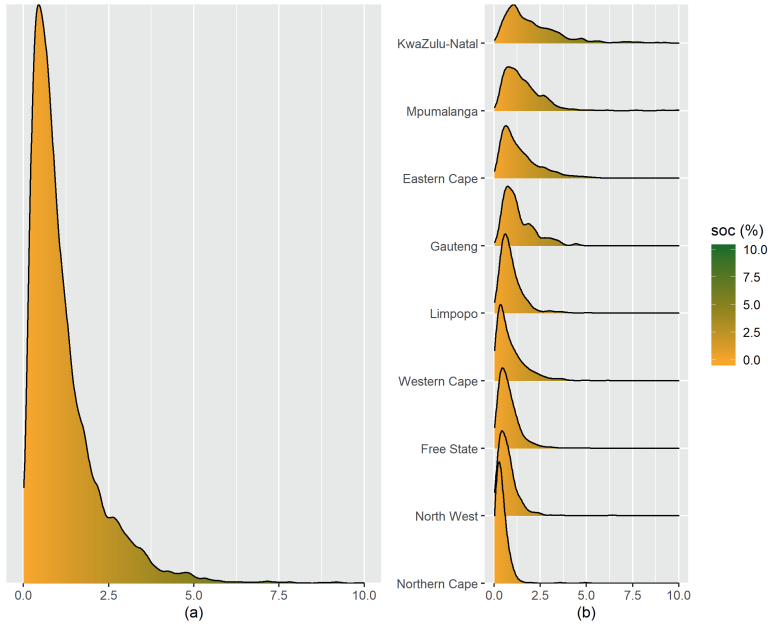


Figure 5.5: (a) Density plot of topsoil SOC for South Africa. (b) SOC densities shown for each province separately.

The results of the outer-folds of the 10-fold nested cross-validation are presented in Web Appendix B. The RF model produced a ME of 0.027%, RMSE of 0.715%, a MEC of 0.552 and a CCC of 0.705. The results also indicated that the validation results were relatively consistent between the outer-folds. The cross-validation calibration consistently indicated that a minimum node size equal to 5 provided the best performance, while the m_{try} hyper-parameter produced the best performance when calibrated to 51. These parameter values were used with the RF model when it was fitted on the entire data set. This produced the final predicted SOC map shown in Figure 5.6. The model predicted larger concentrations

of SOC along the southern and eastern coast of South Africa as well as in the province of KwaZulu-Natal compared to the rest of the country. This was not unexpected as the point map in Figure 5.3 indicated general higher quantities of topsoil SOC along the southern and westerns coasts as well in the provinces of KwaZulu-Natal and Mpumalanga to the east.

We highlighted two regions of interest in Figure 5.6. The first was in the region of the Kahalari desert to the north-west, and the second was in the province of KwaZulu-Natal to the east. The reason why the first region was of interest was because of unexpected predictions made by the RF model. It is generally accepted that this arid region has low quantities of SOC (Department of Environmental Affairs, 2015), and therefore it should be further investigated why the model predicted higher than expected values of SOC in this region. The second region in KwaZulu-Natal was selected since it might be of interest to understand how the covariates relate to the predictions in this region, which is generally regarded as an area of interest concerning SOC (Department of Environmental Affairs, 2015; Schulze & Schütte, 2020).

We selected the top 8 covariates as determined by the variable importance statistics of the RF model which were used to analyse the predictions. Sorted by decreasing importance, these covariates were: (1) index of mean annual cloud cover (CCYRAVG), (2) land surface elevation (DEM) in meters, (3) valley depth (VDP) in decameters, (4) index of average monthly cloud cover for December (CC12AVG), (5) index of average monthly cloud cover for September (CC09AVG), (6) downslope curvature (CRV) in kilometers, (7) long-term standard deviation of the monthly daytime surface temperatures for September (LSTD09STD) in Kelvin, and (8) long-term standard deviation of the monthly nighttime surface temperature for December (LSTN12STD) in Kelvin. For a detailed description of these covariates we refer the reader to Poggio et al. (2021). The list of the top 40 variables is presented in Web Appendix B.

To aid in the interpretations of the biplots and other XML methods discussed in this paper, a correlation matrix, determined from the calibration data, is provided in Web Appendix B. This matrix shows the pairwise correlations between SOC, the top 8 covariates and other climate covariates. The reason for adding additional covariates was because some of the top 8 covariates like CCYRAVG and LSTD09STD, were hard to interpret with respect to soil formation. For this reason additional climate covariates such as average precipitation and ground temperature were also included. Finally, to simplify graph labels as well as interpretations of graphs, we omitted the units in which the covariates were measured for the remainder of the paper. To aid in the discussion, a matrix showing the correlations between SOC and only the top 8 covariates are provided in Table 5.3.

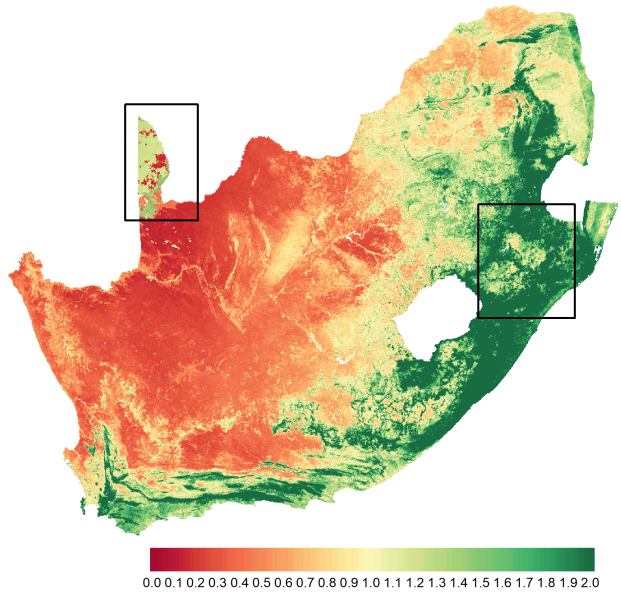


Figure 5.6: Predicted topsoil SOC map of South Africa with the RF model. To improve visualisation all predictions above 2% received the same colour. The two regions highlighted by the rectangular boxes are analysed in Section 5.3.3.

From Table 5.3 we noted that CCYRAVG, CC12AVG and CC09AVG shared moderately strong positive correlations with SOC. The correlation between SOC and LSTD09STD was negative. There were also high correlations between the covariates. For example, between CC12AVG and LSTD09STD we noted a moderately strong correlation of -0.628. Such high correlations could indicate dependence between covariates which will be problematic for XML methods that require covariates to be mostly uncorrelated. From the larger correlation matrix it was noted that cloud cover covariates were strongly positively correlated to precipitation (between 0.65 and 0.85), while long-term standard deviation of the monthly daytime surface temperatures in September and December were negatively correlated to precipitation (between -0.5 and -0.6). Long-term standard deviation of the monthly nighttime surface temperatures were mostly uncorrelated to precipitation but moderately correlated to the EVI index (0.35).

Table 5.3: Correlation matrix of SOC and the top 8 RF predictors, calculated from the calibration data.

	CCYRAVG	DEM	VDP	CC12AVG	CC09AVG	CRV	LSTD09STD	LSTN12STD
SOC	0.539	0.082	-0.024	0.423	0.430	-0.296	-0.171	0.273
CCYRAVG		0.057	-0.033	0.829	0.604	-0.356	-0.450	0.462
DEM			-0.386	0.331	-0.478	0.089	0.056	-0.181
VDP				-0.271	0.337	-0.208	0.134	0.084
CC12AVG					0.130	-0.156	-0.628	0.319
CC09AVG						-0.408	0.034	0.469
CRV							0.074	-0.235
LSTD09STD								-0.077

5.3.2 Unravelling the random forest predictions at national scale

Interpretations with bipots

The reader should be aware that the primary objective of this research was not to offer extensive explanations for how covariates interact with SOC or to make causal inferences. While biplots and other XML methods can reveal specific correlations, they should not be misconstrued as causal relationships, as pointed out by (Wadoux et al., 2020b). Instead, the paper is focused on investigating correlations through biplots and shedding light on essential differences between the interpretations of biplots and other XML methods.

We first analysed the predictions of the RF model at national scale, i.e., for the entire country. The PCA biplot for understanding the predictions of the RF model is presented in Figure 5.7 in which we noted that the alpha-bags vary mostly horizontally from left to right, indicating high to low predictions for SOC. It appeared that CCYRAVG was the covariate that mainly led to these variations. Specifically, the predictions of SOC was mostly less than 0.8% when CCYRAVG was less than 2 500, and the majority of higher SOC predictions (larger than 1.1%) occurred when CCYRAVG was roughly larger than 4 250. CC12AVG and LSTD09STD also somewhat contributed to these variations, with their axes connecting from the top-left to the bottom-right corner of the biplot. LSTD09STD was negatively correlated to CC12AVG and to the SOC predictions, as the direction of the LSTD09STD axis moved in the opposite direction for larger SOC values. This is confirmed by the correlations in Table 5.3. CC12AVG and LSTD09STD also seemed to be mostly uncorrelated to VDP and DEM, while a strong correlation is apparent between VDP and DEM. In addition, the biplot also illustrated that relief factors such as VDP and DEM did not explain much of the national SOC variation.

The overall quality of the biplot in Figure 5.7 was 0.604. The respective predictivity measures of the axes of the covariates are shown in the caption of Figure 5.7. The most predictive axis was that of CC12AVG with a predictivity measure of 0.932, while the least predictive axis was that of CRV with a predictivity measure of 0.326. In case of the latter, the reader should exercise caution when interpreting the results, as the reliability

of projecting points onto the CRV axis is relatively low.

The biplot with the covariates based on the `clorpt` model is presented in Web Appendix C. One interesting highlight from this biplot was that it confirmed that relief factors such as DEM and VDP did not contribute much to the national variation in SOC predictions.

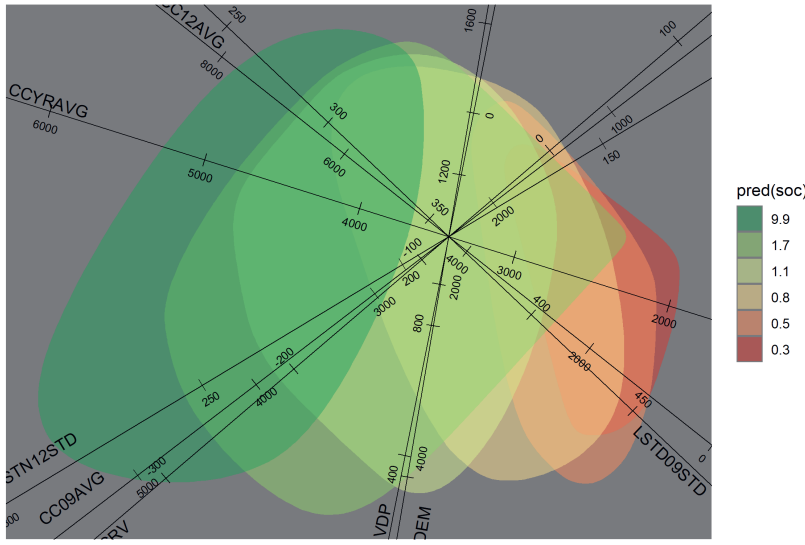


Figure 5.7: PCA biplot showing the relationship between the top 8 covariates and the predictions made by the RF model with 90% alpha-bags. Predictions for SOC are presented in %. The quality of the biplot is 0.604. The respective predictivity measures of the covariates were (sorted from highest to lowest): 0.932 for CC12AVG, 0.919 for CCYRAVG, 0.754 for CC09AVG, 0.526 for VDP, 0.503 for DEM, 0.493 for LSTD09STD, 0.386 for LSTN12STD, and 0.326 for CRV.

The proposed biplot methodology can also be used to understand how covariates relate to the residuals. In Figure 5.8 we present biplots, produced from the calibration data, that depict the residuals of the RF model. For these biplots we used 70% alpha-bags. The biplot in Figure 5.8(a) shows the residuals and the biplot in Figure 5.8(b) shows the absolute values of the residuals. These biplots can provide insight into which covariates led to larger residuals as well as to which covariates led to under- or overestimation of SOC. For example, CCYRAVG, CC12AVG and LSTD09STD led much to the left-to-right variation

of the residuals (Figure 5.8a) and absolute value of the residuals (Figure 5.8b).

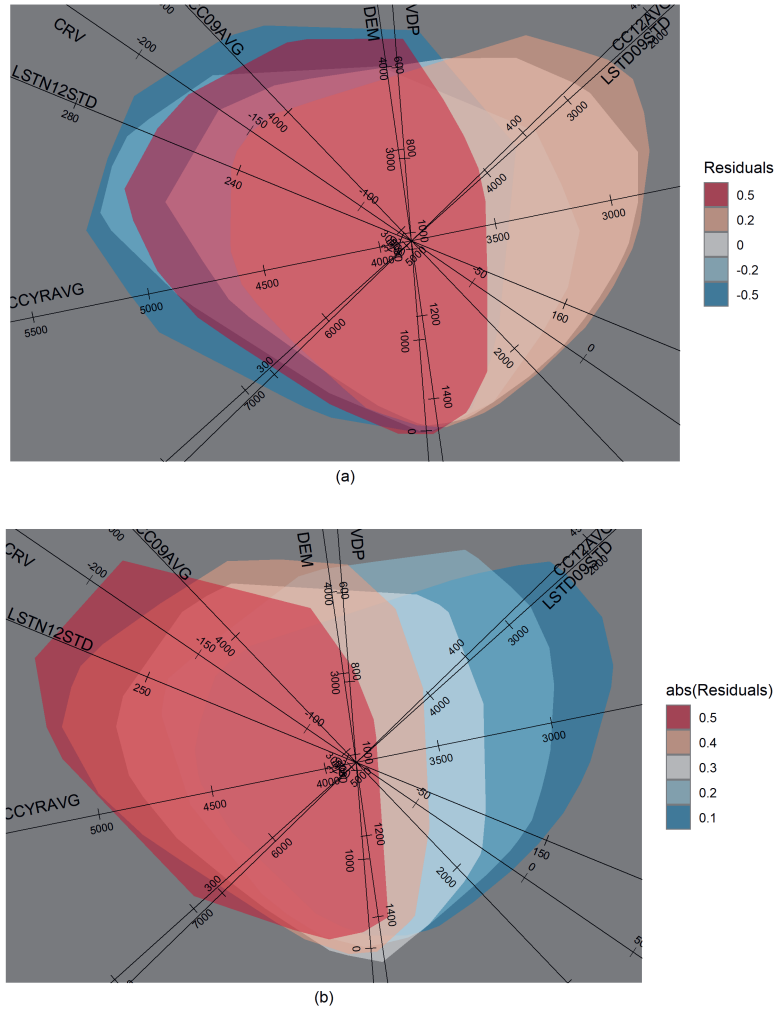


Figure 5.8: PCA biplots with 70% alpha-bags showing the relationship between the top 8 covariates and (a) the residuals; (b) the absolute values of the residuals of the RF model.

Interpretations with other XML methods

Centred ICE curves with the corresponding centred PD curves for the top 8 covariates according to the RF model are presented in Figure 5.9. For CCYRAVG we noted that the PD curve was mostly constant up to approximately 4 500 at which point there was a small

increase in the predictions on average. In terms of the ICE curves (mainly used for local interpretations) some locations experienced a larger increase in the predictions of SOC. This result mostly coincided with that of the biplot in Figure 5.7, that is, around a value between approximately 4 000 and 4 500 we noted a large jump in the predictions of SOC. However, in the biplot we also noted a gradual increase in predictions when CCYRAVG increased from 2 000 to 4 000, which was not that prominent in Figure 5.9.

Some of the results of the PD curves were not that reflective of the biplot in Figure 5.7. For instance, the PD curves for DEM and VDP experienced an increase and decrease in predictions at approximately 1 500 and $-2\,500$, respectively, while certain ICE curves showed a sharp increase after 1 500 for DEM, and a sharp decrease at $-5\,000m$ for VDP. In the biplot it was clear that VDP and DEM could not be used at national scale to explain SOC predictions. Nonetheless, for CRV it was interesting to note a drop in the PD curve at approximately zero which was also visible in the biplot. However, in the biplot it was clearer that CRV was also somewhat negatively correlated to the SOC predictions (also seen in Table 5.3).

It is also important to note from Figure 5.9 that for larger values of LSTD09STD there was a general increase in SOC predictions, especially at 450, where a sharper increase in predictions was observed. In the biplot in Figure 5.7, we noted that when LSTD09STD was about 450, SOC predictions was smaller than 0.5%, and when LSTD09STD was between 250 and 300, predictions were mainly bigger than 1.1%. Table 5.3 shows that the relationship between SOC and LSTD09STD was negative, and this was also confirmed by the biplot (the axis of LSTD09STD moved in the opposite direction as SOC predictions increased). Therefore, conclusions from the ICE and PD curves were possibly misleading. One explanation for this could be due to the fact that LSTD09STD was correlated to CC12AVG, as shown in Table 5.3 and Figure 5.7 (an angle of almost 360°).

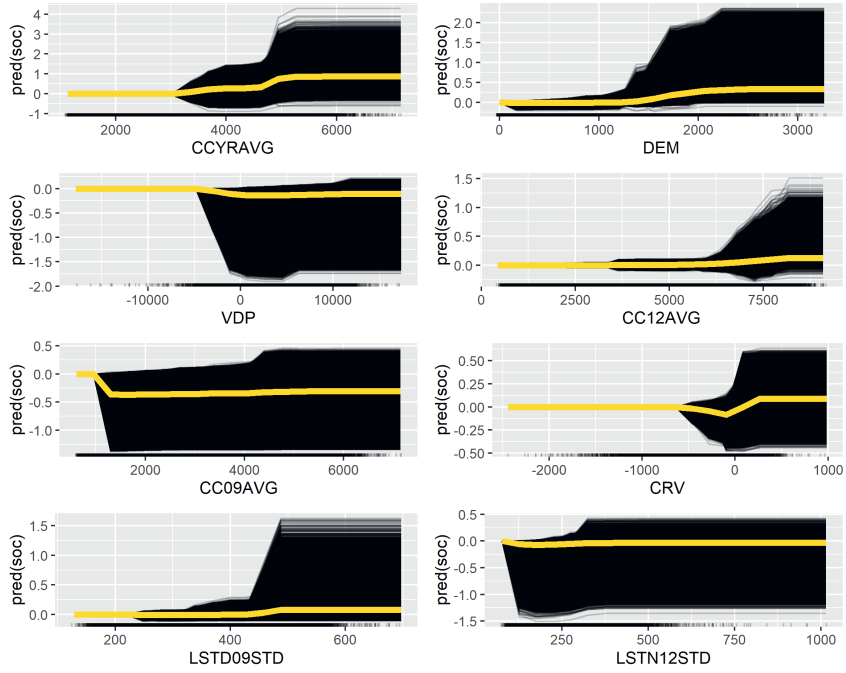


Figure 5.9: Centred independent conditional expectation (ICE) curves shown in black with the corresponding partial dependence (PD) curves in yellow for the top 8 covariates according to the RF model. Predictions for SOC are presented in %.

The ALE plots for the top 8 covariates are presented in Figure 5.10. An ALE curve is interpreted as the effect of a covariate on the predictions for observations within a certain interval or neighbourhood. For example, the ALE estimate of CCYRAVG showed that for large values (i.e., 5 000 to 6 000), predictions of topsoil SOC were about 0.5% higher compared to the average prediction. In contrast to the biplot in Figure 5.7, the ALE curve for DEM clearly indicated that predictions of SOC were increasingly above average as DEM increased from 1 500. In addition, the ALE curve for VDP showed a decreasing trend as VDP increased to 0 which was not that apparent in the biplot in Figure 5.7. Similar to the PD and ICE curves in Figure 5.9, the ALE plot for LSTD09STD also showed above average predictions when the covariate was larger than 450, and less than average when it was less than 400. Recall that the biplot pointed out that the relationship between LSTD09STD and the SOC predictions was negative.

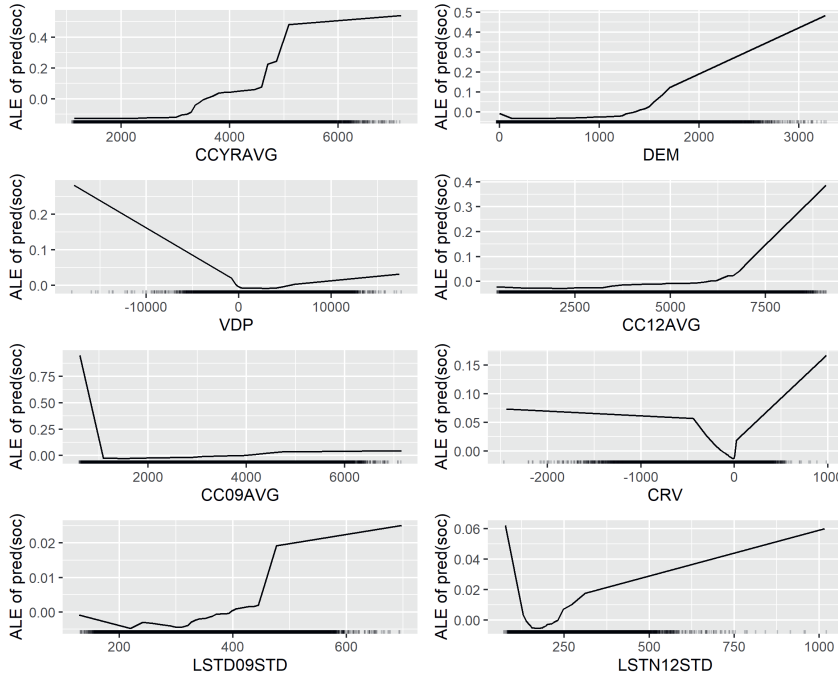


Figure 5.10: Accumulated local expectation (ALE) plots the top 8 covariates in the RF model.

Shapley values are used to analyse the contribution of each covariate to the prediction compared to the average prediction at a certain location. As discussed in Section 5.2.4, Shapley values can also be combined to show the partial dependence on a particular covariate. In Figure 5.11, the partial dependence curves of the top 8 covariates are presented. For example, a general increase in SOC predictions for larger values of CCYRAVG (i.e., more than 4 500) was noted. Then, as noted with the ICE, PD and ALE curves, it was again observed that there was an increase in the predictions as DEM increased (at 1 500), and as LSTD09STD increased (at 450). Shapley values can also be combined to reveal important covariates (Biecek & Burzykowski, 2021). In Web Appendix D we present a variable importance plot derived from Shapley values.

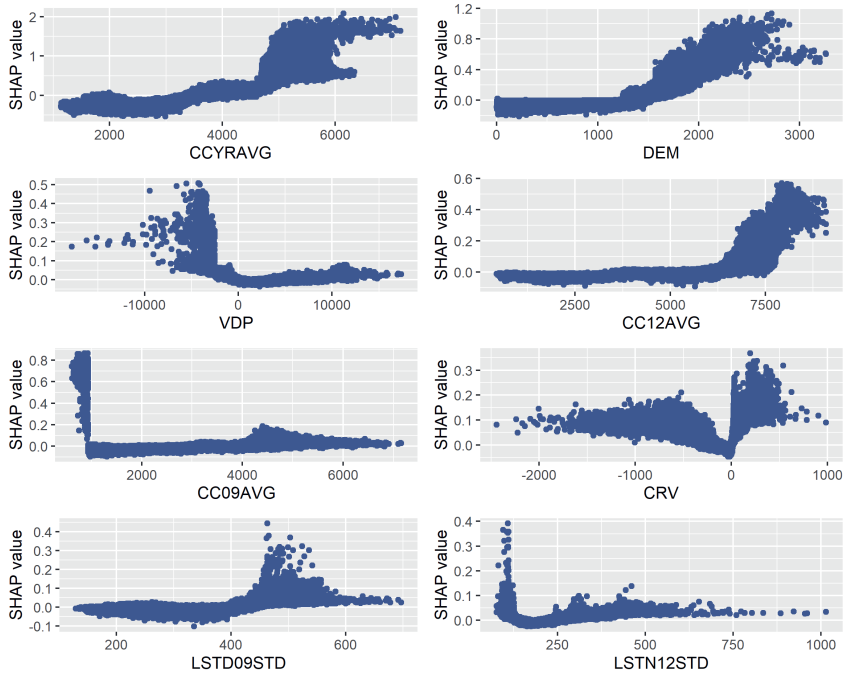


Figure 5.11: Partial dependence plots with Shapley values for the top 8 covariates in the RF model.

5.3.3 Unravelling the random forest predictions at regional scale

In Section 5.3.1 we noted larger than expected predictions in the Kahalari region. In addition, the selected region in KwaZulu-Natal is of interest for SOC as it is one the regions with the highest levels of SOC in South Africa. It is therefore natural to attempt to explain how the covariates in the machine learning model relate to the predictions for these two regions. To understand the predictions at regional scale, we constructed biplots with the top 8 covariates as indicated by the RF model. In Web Appendix C, we present biplots for these two regions based on a subset of covariates that represented Jenny's `clorpt` model.

The biplot for the Kahalari region is shown in Figure 5.12, while the biplot for the highlighted region in KwaZulu-Natal is shown in Figure 5.13. Due to the very low predictivity of CRV in both biplots this covariate was omitted when both biplots were constructed. The quality of the first biplot was 0.617, that of the second was 0.646. The predictivity of the axes are shown in the captions of the two figures. The most predictive axis of the Kalahari biplot and the KwaZulu-Natal biplot was that of CCYRAVG with a predictivity measure of 0.909 and 0.964, respectively. The least predictive axes of both biplots were

that of LSTD09STD. In case of the Kalahari biplot this covariate had a predictivity of 0.0.059 and for the KwaZulu-Natal biplot it was 0.148. Therefore, it is recommended not to interpret the axis of this covariate.

Figure 5.12, based on the Kalahari region, showed mostly larger SOC predictions in the upper section of the biplot. The axis of VDP was mainly responsible for the top-to-bottom variation of the SOC predictions, indicating that valley depth greatly influenced SOC predictions. Specifically, when VDP was larger than 2 000 most predictions were smaller than 0.4%. Other covariates that also somewhat contributed to the top-to-bottom variation were CC09AVG and LSTN12STD. For the region in KwaZulu-Natal, Figure 5.13 illustrates that CC12AVG, CCYRAVG and VDP were mainly responsible for the high-to-low variation (right-to-left) of SOC predictions. For example, when CC12AVG and CCYRAVG were larger than 7 000 and 5 250, respectively, SOC predictions were mainly larger than 2.8%. In addition, when VDP was smaller than 0, SOC predictions were also mainly larger than 2.8%. It was interesting to note that relief factors such as VDP contributed more notably to the explanation of SOC predictions at the regional scale as opposed to the national scale. This was also confirmed in the regional biplots in Web Appendix C, which also illustrated the significance of VDP at the regional scale.

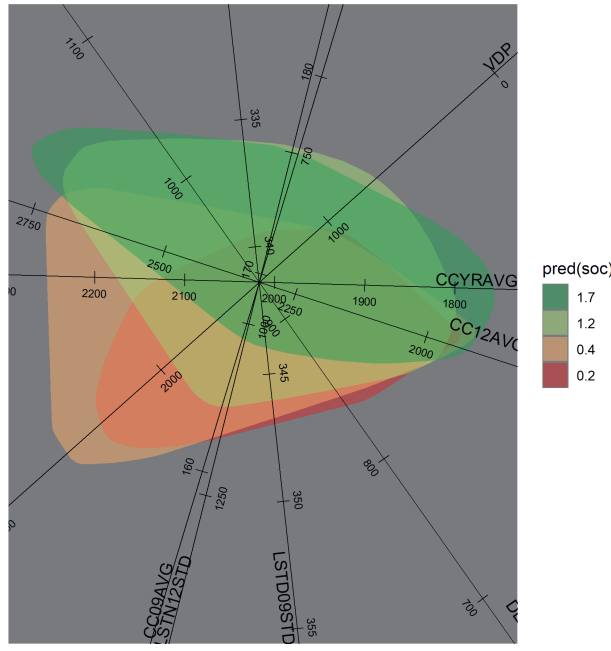


Figure 5.12: PCA biplot showing the relationship between covariates and the predictions made by the RF model for the Kalahari region. Predictions for SOC are presented in %. The quality of the biplot is 0.617. The respective predicticity measures of the covariates were (sorted from highest to lowest): 0.909 for CCYRAVG, 0.842 for DEM, 0.825 for CC09AVG, 0.794 for CC12AVG, 0.789 for VDP, 0.103 for LSTN12STD, and 0.059 for LSTD09STD.

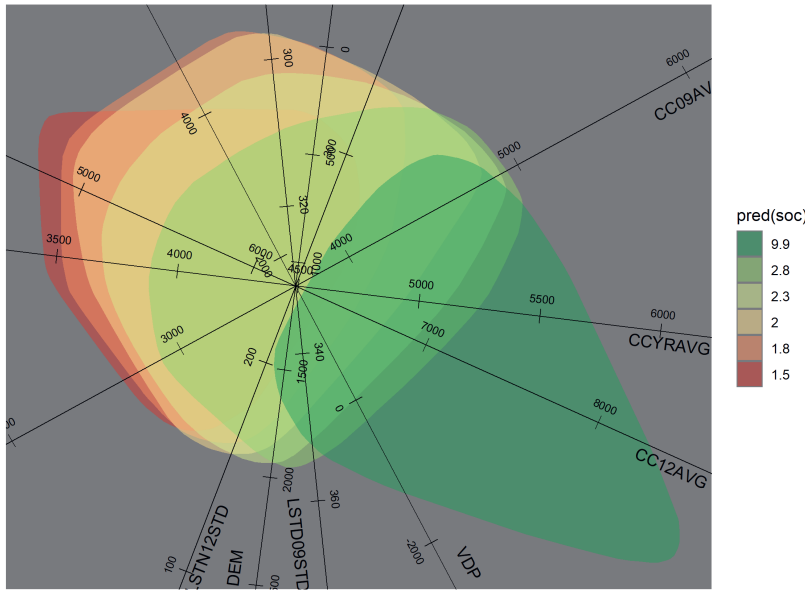


Figure 5.13: PCA biplot showing the relationship between covariates and the predictions made by the RF model for the highlighted region in KwaZulu-Natal. Predictions for SOC are presented in %. The quality of the biplot is 0.646. The respective predictivity measures of the covariates were (sorted from highest to lowest): 0.964 for CCYRAVG, 0.889 for CC09AVG, 0.862 for CC12AVG, 0.818 for DEM, 0.456 for VDP, 0.384 for LSTN12STD, and 0.148 for LSTD09STD.

The biplot in Figure 5.12 failed to explain which covariate(s) might have led to the poor predictions in the Kalahari region. In Figure 5.14, the same biplot in Figure 5.7 is presented but with the grid cells of the two highlighted regions added as points. The grid cells of the Kalahari region were clustered to the right along the axis of CCYRAVG while the KwaZulu-Natal region was clustered to the left. This biplot showed how CCYRAVG, CC12AVG and LSTD09STD might have led to the larger than expected predictions. Along the axes of these covariates many of the locations in the Kalahari region received a prediction of up to 1.1% which was originally noted in Figure 5.6.

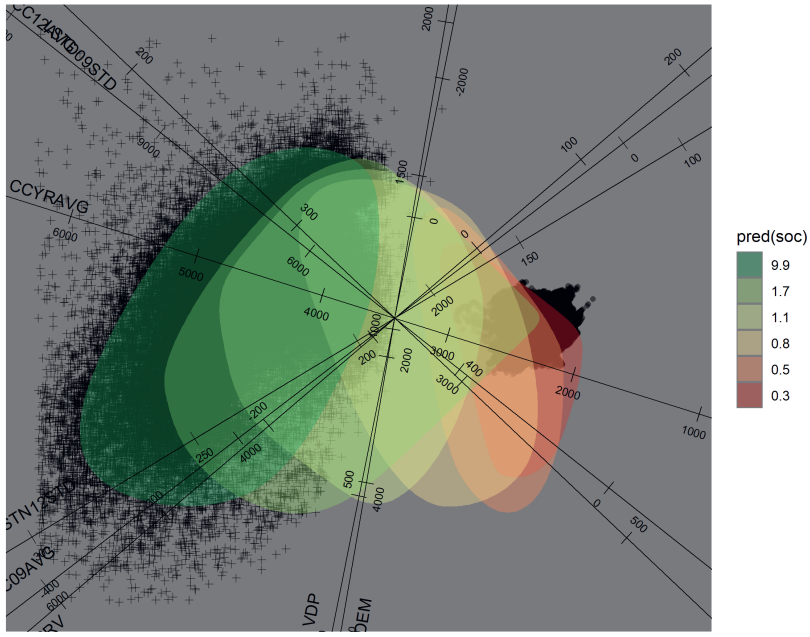


Figure 5.14: The biplot in Figure 5.7 depicted with the grid cells of the Kalahari region (dots) and the KwaZulu-Natal region (crosses).

We also produced ICE, PD and ALE curves and plots based on Shapley values for the Kahalari region and the region in KwaZulu-Natal. These results are in Web Appendix E.

5.4 Discussion

In this study, PCA biplots were introduced as a model-agnostic visualisation method for understanding "black-box" (Belle & Papantonis, 2021) regression machine learning predictions. The proposed methodology was also compared to various other XML methods, including ICE and PD curves, ALE curves, and Shapley values. It is essential to emphasise that none of these methods are suitable for identifying causal relationships; they can only reveal correlations. While important covariates may provide valuable insights into soil formation, their causal implications should be evaluated through experiments, as highlighted in Wadoux et al. (2020b), or with methods based on observational data as described in Hernan & Robins (2020) and Huber (2023). In case of performing experiments we refer readers to the extensive body of literature available on experimental design, with one excellent example being the work by Montgomery (2012).

All methods discussed in this paper were evaluated in a case study in which topsoil SOC was mapped in South Africa with a RF model. The produced SOC map, very similar to the one in Venter et al. (2021) which was also produced with a RF model, showed high concentrations of SOC along the southern- and eastern coastlines, majority of KwaZulu-Natal and the eastern regions of Mpumalanga. The map was produced from 109 covariates (Poggio et al., 2021). Therefore, subsets of covariates were selected which were used with the biplots and other XML methods. In theory, there is no limit to the number of axes that can be incorporated into a biplot, that is, if p covariates are available then all of the p covariates can be visualised. However, the judicious addition of axes is crucial to maintain the interpretability of the biplot and to prevent clutter. Additionally, it is important to recognise that the configuration of the biplot may change if a covariate is added or removed from the PCA. That is, the relation between a certain covariate and the configuration of the predictions is dependent on which other covariates are included in the biplot. This is illustrated with a web application created for this paper (Shiny app[2]). The application was built on the calibration data set, and also allows one to analyse the effect of including different covariates on the biplot configuration for global interpretations. Furthermore, the application also demonstrates how varying proportions and the number of alpha-bags can impact the biplot configuration.

With regards to the national interpretation of the predictions, the biplot in Figure 5.7 revealed that cloud cover covariates (CCYRAVG and CC12AVG) were important. Cloud cover showed a strong correlation with climate variables like precipitation (Web Appendix B). This observation further validates the existing body of literature emphasising the significance of climate factors in the prediction of SOC (Lamichhane et al., 2019). Similar trends were observed by the ICE, PD, and ALE curves, and the partial dependence curves derived from Shapley values.

With the effects of the climate covariates taken into account, the biplots indicated relief factors to be less important at national scale and more pertinent at regional scale. This is not unexpected as DSM literature has shown that climate covariates are the most important driving factors of SOC at national to continental scale (Minasny et al., 2013), and that relief and land management covariates appear to increase in importance when predicting SOC at regional scale (Lamichhane et al., 2019; Minasny et al., 2013). The other XML methods failed to indicate this which might give the impression to inexperienced users that relief is important when SOC predictions in South Africa are analysed at national scale. For example, Figures 5.9, 5.10 and 5.11 all indicated that SOC predictions were marginally affected by relief factors VDP, DEM and CRV at national scale.

An explanation for the different findings between the biplots and the other XML methods could be linked to the fact that covariates are highly correlated (Hastie et al., 2008). Methods such as ICE, PD curves, and Shapley values assume that covariates are uncor-

[2] `https://stephanvdw.shinyapps.io/shinyapp/`

related. When this condition is not met, for example, some covariates show interactions or are correlated, then averaging over the marginal distribution of the covariates in set $-S$ might give misleading results with regards to the effect of the covariate of interest on the model predictions (Molnar, 2020). Note that while this "independence" is not strictly required for the ALE curve (Apley & Zhu, 2020), it yielded a comparable yet less pronounced conclusion concerning the relief factors. It should therefore be emphasised that a key advantage of biplots over ICE and PD curves and Shapley values is that the condition of uncorrelated covariates is not required. In biplots, the dependence structure between covariates is explicitly accounted for (Gower et al., 2011), enabling biplots to uncover essential patterns that other XML methods might miss or misrepresent entirely. Another example of this was the interpretation of the effect of the long-term standard deviation of the monthly daytime surface temperatures for September (LSTD09STD) on SOC predictions. The calibration data suggested a negative relationship between LSTD09STD and SOC, and while this relationship was correctly depicted in the biplot in Figure 5.7, the other XML methods produced misleading results.

Biplots can also be used to relate the residuals of a machine learning model to the covariates. For example, in Figure 5.8, it was observed that the annual average cloud cover played a noteworthy role in accounting for the variability in the residuals of the RF model. This is an important advantage of biplots over other XML methods which can only also show the effect of covariates on the predictions of a machine learning model.

Another important advantage of a biplot over other XML methods is the availability of goodness-of-fit measures (Gower et al., 2011), i.e., the quality of a biplot and the predictivity of the axes and points. For example, the quality of the biplot in Figure 5.7 was 0.604 which meant that 60.4% of the variances in $\mathbf{X}$ was explained by the 2-dimensional approximation. It is difficult to establish a general rule of thumb regarding what level of quality is considered sufficient, as the quality will ultimately depend on the variation in $\mathbf{X}$. However, the quality can be used as a selection measure between different biplots to aid in finding the best biplot for various sets of covariates. Nonetheless, additional research would be required to investigate appropriate measures of quality for biplots when used for analysing machine learning predictions in DSM.

A shortcoming of the biplots presented in this paper was the fact that axes are only interpreted as linear with respect to the points. That is, for a given axis in a biplot, the variation of the SOC predictions can only be understood in a linear manner. The other XML methods were able to show possible nonlinear relationships between SOC predictions and changes in the covariates. In addition, in Wadoux et al. (2023) the authors presented partial dependence plots with Shapley values which illustrated nonlinear trends between SOC stock predictions and covariates such as elevation, temperature and precipitation (Figure 3). It is possible to extend biplots to nonlinear biplots (Gower & Harding, 1988; Vines, 2015), but this extension was outside the scope of this paper.

In contrast to Shapley values, the biplot methodology (and ICE, PD and ALE curves) do not produce variable importance measures. Therefore, if the number of covariates in $\mathbf{X}$ is large, a subset of covariates would be required, and it might be unclear which covariates to analyse first. With that said, it should be mentioned that the computational demand of Shapley values is considerably large. Estimating Shapley values for all covariates over all grid cells to establish variable importance would be infeasible in many DSM projects.

The computational demands of producing a biplot is minimal, even on large data sets. In this paper we were able to produce biplots on the covariate maps, which consisted of about 20 million grid cells, in several minutes. Whereas Shapley values required a grid sampling of the covariate maps in order to become computationally feasible. Even then, computation time of Shapley values was large. For example, the Shapley values used to produce the results in Figure 5.11 took approximately 67 hours on a 8 core desktop computer with 256GB of RAM. In Wadoux et al. (2023) it took about 6 days to compute Shapley values for 800 000 grid cells with 500 Monte Carlo simulations.

Biplots can be used to reveal how covariates affect a prediction at a specific location. For example, a point in a biplot can be projected onto the axes to reveal the corresponding values in the covariates (Gower et al., 2011). However, for the biplots presented in this paper this would be impractical as the point configurations in the biplots were represented with alpha-bags. This is similar to studies such as in Wurz et al. (2005) which represented archaeological data with various alpha-bags. Furthermore, it is not clear how to visually link the point in the biplot (principal component score) to an actual location point in a region of interest. In Wadoux et al. (2023) Shapley values were mapped directly so that covariate importance could also be studied geographically. Further research would therefore be required to visually link the points in the biplot to the locations in soil maps.

5.5 Conclusion

A biplot is a powerful tool for multivariate data exploration, and in this study it was used as an effective method for understanding machine learning predictions in DSM. Note that our proposed methodology can be used to understand machine learning predictions, regardless of domain of application. Biplots can be used with several covariates and may provide insight into predictions both locally and globally. It was also illustrated that the biplot revealed important patterns in machine learning predictions of topsoil SOC which other XML methods such as ICE, PD and ALE plots and Shapley values failed to detect. This is because biplots do not require covariates to be uncorrelated. Biplots also provide measures on goodness-of-fit which allows a user to evaluate the quality of the biplot as well as the predictivity of the axes and points. It was also shown how biplots can be used to analyse the residuals of a machine learning model.

Supplementary materials

The supplementary materials, Web Appendices A - F, can be downloaded from the following link: https://github.com/CSVDW/PCA-Biplots-for-machine-learning-predictions.

Chapter 6

Synthesis

"*That, Dr.Calvin, is the fundamental difference between robots and human beings. Humans are inherently irrational.*"

Isaac Asimov - I, Robot

6.1 Introduction

In Chapter 1, I argued that machine learning is increasingly used in fields such as digital soil mapping (DSM), mainly due to its ability to produce maps with greater accuracy compared to traditional statistical models. However, this increased use of machine learning could also bring about the potential for misuse, especially when faced with certain challenges. In this thesis, I identified four such challenges to show how machine learning can be enhanced: modelling uncertain soil data, modelling more than one soil property simultaneously, modelling right-censored soil data, and model interpretability. Furthermore, in Chapter 1, I pointed out common pitfalls related to the use of machine learning. Although the aim of this thesis is not to give an overview of these pitfalls and a comprehensive list of solutions to avoid it, but in light of the identified challenges, I do address certain important pitfalls including model selection, model optimisation, model evaluation and data leakage, and model interpretability.

The structure of this concluding chapter is as follows: in Section 6.2 I give an overview of the results of the thesis and review the objectives, with the corresponding research questions, that were presented in Chapter 1. In this section I also discuss a few aspects on how the methodology implemented with each of the four objectives can be improved. In Section 6.3 I place the thesis within the wider context of DSM and the general use of machine learning in scientific domains. I give my personal reflections and present my final conclusion.

6.2 Overview of findings and potential improvements

6.2.1 Machine learning for measurement error-contaminated soil data

The first research question presented in Chapter 1 raised the issue whether measurement errors can be effectively accounted for with a machine learning model. In Chapter 2, a novel maximum likelihood framework was presented which allowed measurement errors to be accounted for. Specifically, the framework uses weights that are well defined to account for measurement errors with a maximum likelihood framework (Buonaccorsi, 2010) that will lead to efficient estimates of model parameters. It is important to note that in Chapter 2, I argued that studies such as Wadoux et al. (2019) and Hengl et al. (2018) also incorporated weights to account for measurement errors, but opposed to these studies, the weights used in our proposed framework will minimise the measurement error-adapted loss function (Eq. 2.14), thereby proving that measurement errors can be effectively accounted for.

The second research question focused on whether accounting for measurement errors pro-

duces higher prediction accuracy compared to models that do not take measurement errors into account. In Chapter 2, the advantage of filtering out measurement errors with machine learning with regards to prediction accuracy was not overwhelming. The synthetic simulation study revealed that the average size as well as the relative variation of the measurement error variances had significant effects on the results of the error-filtered models. The error-filtered models produced better results than the models that did not account for measurement error when the size and the relative variability of the measurement error variances were larger. However, in the Namibian case study the differences were small, and therefore there was no real improvement in the prediction accuracy of the models when measurement uncertainty was accounted for. It was also noted that the geostatistical model was able to take more advantage of the provided measurement error variances compared to the machine learning models. More recently in Takoutsing & Heuvelink (2022) it was noted that a random forest model implemented with the proposed framework did not improve prediction accuracy when it was compared to another random forest model that ignored the measurement errors. The authors hypothesise that this was due to the measurement error variances being too small. It is therefore worthwhile to investigate if it ever will be the case that the uncertainty in soil data will be such that the error-filtering (machine learning) models will be able to take significant advantage of it. Further research will be required to investigate the typical characteristics of the measurement error variances of various soil properties obtained from various sources (Van Leeuwen et al., 2021).

One big limitation of this chapter was not to conduct an analysis on the impact of accounting for measurement errors on the uncertainty of the models. Often in DSM too little emphasise is placed on the prediction variance of a model. This is expected with machine learning, because the prediction variance cannot be calculated. However, with random forests, several methods have been proposed to estimate the prediction error variance, with one being the use of a quantile regression forest (Meinshausen, 2006; Vaysse & Lagacherie, 2017). In Somarathna et al. (2018) the authors noted a remarkable decrease in prediction uncertainty when they accounted for measurement errors with a linear mixed model. Therefore, further research is required to investigate the impact of using the proposed framework on prediction uncertainty. Another drawback of the proposed framework is the assumption that measurement errors are independent. This is unrealistic and the spatial correlation of these errors should be accounted for in future research. I suspect that if there is a significant spatial correlation, the impact of accounting for measurement errors on prediction accuracy with proposed framework will be more pronounced.

To use the proposed framework, it is required to know or least have some estimate of the variances of the measurement errors. However, it is often the case that such information is poor or even unavailable to the DSM practitioner (Van Leeuwen et al., 2021). This is especially problematic since soil data are more frequently obtained with methods that produce measurements with larger uncertainties (i.e., infrared spectroscopy and citizen

science). Moreover, it is expected that other sources of error in the DSM error budget (Nelson et al., 2011) is in decline, such as covariate error. This is mostly due to the advancements in remote sensing technology. As a result, measurement errors will play an increasingly significant role in DSM, and it is therefore becoming ever more imperative for soil (e.g., field and laboratory) data producers to report the uncertainty of soil measurements.

6.2.2 Multivariate random forest as a viable option for multivariate mapping of more than one soil property

In Chapter 3, the investigation of the multivariate random forest entailed an in-depth explanation of the underlying theory of the model, and how it can be used to perform conditional stochastic simulations. This is important because I wanted to investigate if the multivariate model is better able to maintain the correlation structure in its output compared to when each soil property is modelled separately with different random forest models (Research Question 1). In the case of modelling soil organic carbon (SOC) and total nitrogen, the multivariate model was superior in maintaining not just the correlation between carbon and nitrogen, but it also produced more realistic carbon-nitrogen ratios. I performed the same comparison with geostatistical models in which regression co-kriging was compared to two separate regression kriging models, and found similar results, in that the multivariate model was better. It is worth noting here that the multivariate random forest model was also superior in this respect to the regression co-kriging model. Ensuring that the correlation structure among soil properties is maintained holds significance for subsequent DSM analyses, including Digital Soil Assessment (Kidd et al., 2015; Okonkwo et al., 2018; Rabot et al., 2022a) and Soil Function Assessment (Rabot et al., 2022b; Greiner et al., 2018, 2017). This is because such analyses depend on well-defined joint distributions of soil properties.

Furthermore, in Chapter 3, the multivariate models were also compared to the univariate models with regards to prediction accuracy (Research Question 2). When focusing solely on mapping each soil property independently, the conclusion was drawn that the multivariate models did not exhibit any advantage over the univariate models concerning prediction accuracy. However, when extending the task to include mapping derivatives, such as the carbon-nitrogen ratio explored in Chapter 3, the multivariate models outperformed the univariate models. This finding holds significance as it could potentially extend to other combinations of soil properties, such as soil organic carbon (SOC) and clay, where the ratio of SOC to clay holds importance in studies related to soil health (Prout et al., 2021, 2022). Such insights could prove beneficial for DSM projects like SoilGrids 2.0[1] by ISRIC, which globally maps multiple soil properties. Generating maps that can be used to further map derivatives, such as the carbon-nitrogen or SOC-clay ratio, could greatly benefit its user base.

[1] `https://soilgrids.org/`

Chapter 3 presented compelling evidence that the multivariate random forest model was able to maintain the correlation between SOC and nitrogen in its output. In this thesis, I only investigated these two (strongly correlated) soil properties. It will therefore be interesting to investigate the performance of the multivariate random forest model when more than two soil properties are modelled, and perhaps when the soil properties are not that strongly correlated. My suspicion is that if sufficient data are available, the multivariate model will generally perform well. This is because a random forest model is based on trees which are known to closely approximate any given regression function. The increase in the number of soil properties might however be computationally too costly (Wadoux et al., 2020a), especially if a large dataset is needed to efficiently estimate all model parameters.

6.2.3 Accounting for right-censored soil thickness data with machine learning

In Chapter 4, an overview of the random survival forest model was presented which included a proposed method for its utilisation in predicting soil thickness. The viability of the random survival model depended on the type of censoring and the proportion of censored data. For example, in the synthetic simulation study, the model performed poorly when the proportion of censored data was too large when censoring occurred at multiple depths, or when the censoring mechanism was informative (Kleinbaum et al., 2012). Moreover, in Chapter 4, an alternative modelling approach was also proposed in which soil thickness data were modelled with a (regular) random forest model that weighted the non-censored data according to the probability of exceeding certain depths. This model also incorporated an additional parameter which allowed censored data that exceeded a predefined depth to be included in the calibration of the model. The proposed model performed relatively well in the synthetic simulation study, especially when the proportion of censored data was not too large. It also performed well in the case studies from Switzerland and the USA, especially when the models were validated with the truncated approach.

For modelling and mapping soil properties with right-censored data further research is required for validating models. In this thesis I only used two simple methods to validate models. The first was to validate with the non-censored data, and the second was to validate only down to a certain depth. The latter would be of interest to map users who are only interested in soil thickness to a depth of say, $120\,cm$. However, these methods are sub-optimal as censoring is directly not taken into account, and therefore additional methods should be investigated. One approach is to use an inverse probability of censoring weighting method which gives more weight to non-censored data in the validation process (Graf et al., 1999). As in Chapter 2, one big limitation of Chapter 4 was also not to investigate prediction uncertainty between the different random forest models. It would have been interesting to compare prediction uncertainty under various simulation param-

eters, such as proportion of censored data. Since only random forest models were used in this chapter, obtaining estimates of uncertainty would have been possible (Meinshausen, 2006).

6.2.4 Visually explain machine learning predictions with respect to multiple covariates

Chapter 5 presented a biplot methodology for understanding predictions made by machine learning models. To illustrate its use, a case study of mapping topsoil SOC in South Africa with a random forest model was considered. The biplot was able to explain predictions of the machine learning model by visualising the predictions with various environmental covariates simultaneously. Although it should be noted that the case study included 109 covariates, and so a subset of the covariates had to be selected before the biplot could be constructed, the biplot methodology can handle many more covariates compared to alternative explainable machine learning (XML) methods such as partial dependence plots (Friedman, 2001), accumulated local effect plots (Apley & Zhu, 2020), and partial dependence plots produced with Shapley values (Shapley, 1953; Molnar, 2020). The biplot was used to explain predictions at global scale (i.e., over the entire study area) as well as at local scale (i.e., over different sub-regions of the country). I also showed how the biplot can be used to relate the residuals of a machine learning model to the covariates. In addition, the biplot methodology also provided goodness-of-fit statistics (Gower et al., 2011), which are an important advantage over other XML methods, such as partial dependence plots and Shapley values. For these reasons, biplots can be regarded as an exceptional tool for analysing machine learning predictions. Since the biplot also explicitly takes the covariance structure between the covariates into account it also revealed trends in the predictions that other XML methods missed. This is because most other XML visualisation methods require covariates to be uncorrelated. For example, in the case study on mapping SOC in South-Africa the biplot indicated that relief factors were of lesser importance on a national scale but gained greater relevance at the regional level.

Biplots proved to be a promising tool for understanding model predictions. Nonetheless, one aspect that requires further investigation is to improve the biplot methodology for understanding predictions in the spatial domain. For example, it is possible to map Shapley values, which provides meaningful information to interpret predictions spatially (Wadoux et al., 2023). In this thesis, spatial information was not incorporated into the biplot, other than by showing biplot results for sub-regions. Further research is required to visually link the points in the biplot to locations in a soil map. One way to address this is to include the (x, y) coordinates as axes, but this may be difficult as the directions will most probably not reflect a standard compass. Another promising line of research is to investigate nonlinear biplots for analysing predictions. In this thesis, the standard biplot with linear axes was used, but it is possible to extend biplots to have nonlinear axes (Vines, 2015), which could possibly give more meaningful interpretations.

6.3 Reflection

This thesis primarily focused on tackling several key challenges inherent to machine learning in DSM. Nevertheless, its relevance extends beyond DSM and finds applications in various areas in which machine learning is applied, as well as in the broader domains of machine learning research itself. In this section, I delve into the contributions of this thesis, also highlighting common pitfalls related to the use of machine learning introduced in Chapter 1 and demonstrating through examples how this thesis offers some guidance. Finally, I contextualise the thesis within a wider scope by exploring potential avenues for promising future research, before presenting my final conclusion.

6.3.1 Misuse of machine learning

In many scientific disciplines, the adoption of machine learning has dramatically expanded over the past two decades. In DSM, this is also true as the increase of machine learning is primarily driven by the need for more accurate soil maps (Minasny & McBratney, 2016b). However, the environment is complex, and even with the use of machine learning, it is often difficult to obtain satisfactory accuracy levels when using it to make predictions. In Chapter 2, I provided an overview of several sources of error that could propagate through the model, thereby leading to lower prediction accuracy. One such source is model error which in part refers to environmental covariates (i.e., model inputs) that do not fully explain the variation in the response variable (i.e., the model output). Although I argued that for instance due to technological advancements in remote sensing, model error is in decline, but it has yet to reach a point at which covariates explain a sufficient amount of the variation in the response. This is important because often machine learning is misused in that not much accuracy is gained by using a more complicated model (i.e., the covariates only explain a limited portion of the variation in the response). Consequently, this leads to a questionable trade-off between the increase in prediction accuracy, and model interpretability.

The increase in the usage of machine learning, along with its misuse, is also attributed to the proliferation of user-friendly software packages like those accessible on `CRAN` for the `R` programming language (R Core Team, 2020). These tools often negate the necessity for a deeper understanding of mathematical and statistical concepts. For this reason, statisticians and other data experts are less relied upon for data analysis and modelling. The likelihood of misuse becomes more pronounced, particularly when dealing with data presenting additional challenges, such as contamination with measurement errors or right-censored data. In DSM, reviews by Wadoux et al. (2020a) and Wadoux et al. (2021) highlighted the need for further research with machine learning in challenges related to sampling design, accounting for spatial information, multivariate mapping, uncertainty analysis, and interpreting predictions. While this thesis has addressed several of these challenges (see Section 6.2 for an overview), it does not claim to provide an exhaustive

list of solutions. Instead, the thesis aims to offer guidance on addressing some challenges to mitigate the misuse of machine learning. Additionally, it signals the necessity for further model development capable of handling such complexities.

Model selection

The misuse of machine learning can result in common pitfalls, with one notable concern being related to model selection. As discussed in Chapter 1, some of the issues associated with model selection involve adapting a model to address specific challenges. For instance, when dealing with uncertain data, expertise is crucial to recognise the need for appropriately addressing uncertainty to prevent biased outcomes. Effective model selection also involves considering a variety of model types since machine learning models are data-driven, and what proves effective in one study may not be optimal for another. In DSM, there is often a tendency to choose inadequately among models for a given task, a situation exacerbated when selecting a model solely based on its popularity in the literature (examples in (Odebiri et al., 2023; Venter et al., 2021; Liu et al., 2022)). Additionally, it is advisable to compare machine learning models with at least one traditional statistical model, such as regression kriging (Zhu et al., 2023), to prevent unnecessary loss of model interpretability. A traditional geostatistical model like regression kriging has the added advantage of explicitly accounting for spatial variation which is often ignored in machine learning. In Chapters 1 and 2, I argued that spatial variation in covariates has the potential to offer significant insights into the spatial variation in the response, provided a suitably complex model is used. However, with regards to proper model selection, there remains room for enhancement, particularly in the realm of machine learning that can effectively account for spatial variation (refer to the Section 6.3.2 for a further discussion).

Model optimisation

Model optimisation (i.e., hyper-parameter tuning) is an important step when a machine learning model is used, because the choice of the values for the hyper-parameters can greatly affect the model's accuracy and reliability. In DSM, model optimisation is usually done with a grid-search (Wadoux et al., 2020a). In this thesis, I also mostly used a grid-search to find optimal hyper-parameter values for several different machine learning models. Of the various pitfalls related to the use of machine learning, model optimisation is one of the least occurring ones in DSM.

Typical grid-searches could be computationally very demanding, and so a "random search" could be beneficial to improve efficiency and to broaden the hyper-parameter value space (Zhu et al., 2023). Other methods that allow for more "intelligent" model optimisation include methods such as genetic algorithm (Wu et al., 2016), and Bayesian optimisation (Wadoux et al., 2019). It is important to note that in face of additional data challenges,

one should be mindful on how to perform efficient model optimisation. For example, in Chapter 3, when I modelled multiple soil properties, it was more efficient to use a "multivariate loss function" (i.e., the trace of the mean square error matrix) that has to the best of my knowledge not been used in DSM.

Model evaluation and data leakage

With model evaluation it is important to use several evaluation metrics to show model performance. In DSM, it is common to use at least one metric, and the most commonly used ones include the mean error, (root) mean square error, and the model efficiency coefficient (Wadoux et al., 2020a). Evaluation metrics are often obtained with k-fold cross-validation, but it is important to note that metrics obtained with this method could lead to over-optimistic results. In the data-splitting phase, it is important to have three different model developing sets, namely, the training and validation sets (used in the calibration or training phase), and the test set (used in the testing phase). With a standard cross-validation, performed over the entire data set, there are only two sets and if a test set is not introduced then the evaluation metrics will be over-optimistic. In Chapters 3 and 4, I used a nested cross-validation which ensures that the outer-loops (i.e., test sets) are completely unseen by the models. Cross-validation results can also produce over-optimistic results when the data are spatially clustered. In such a case, it is recommended to use blocked spatial cross-validation if the data exhibits strong clustering (De Bruin et al., 2022). As with model optimisation, it is also important to be mindful of additional data challenges. For example, if data are right-censored, then standard evaluation metrics cannot be used, as these may then favour models that ignore the censored nature of the data (Chapter 4).

Another pitfall is data leakage which refers to when a model gets information from unseen data during the calibration phase which can then lead to biased results. This happens for example when variable scaling or variable selection is performed on the entire data set before data-splitting. In DSM, this is a common pitfall and examples of scaling variables before data-splitting occurred in Liu et al. (2022); Tesfa et al. (2009), and variable selection leading to data leakage occurred in Tesfa et al. (2009); Suleymanov et al. (2021); Odebiri et al. (2023); Venter et al. (2021). I need to point out here that in Chapter 3 of this thesis, variable selection was performed before data-splitting which is sub-optimal, but the aim of this chapter was to showcase the functionality of the multivariate random forest model, and to compare it to various other models, and not necessarily to obtain the best predicted soil map.

Model interpretability

With the exception of feature importance, unravelling the black box is relatively new in DSM, with only several implementations of some popular XML methods (e.g., par-

tial dependence plots, Shapley, accumulated local effects) to understand machine learning predictions (for examples, refer to (Wadoux et al., 2023; Wadoux & Molnar, 2022; Odebiri et al., 2023)). It is worth noting that these methods are also excellent tools to convey information about the predictions to non-machine learning experts. However, caution should be taken to not confuse correlation for causal relationships. As noted in Chapter 5, causal relationships should be evaluated through experiments (Wadoux et al., 2020b), or with methods based on causal inference (see discussion in the next section). To increase model interpretability one could also simply use a simpler model which allows for the parameter estimates of the model to be interpreted, for instance, regression kriging. As discussed earlier, model interpretability is also closely related to model selection, underlining the importance of comparing machine learning models with traditional statistical models to mitigate the potential loss of interpretability.

6.3.2 Future research

The initial excitement surrounding the innovative application of a machine learning model in a specific domain has subsided. Roughly two decades ago, scientists enthusiastically embraced the machine learning bandwagon, and for many, it proved worthwhile, as they were among the first to experiment with a model that yielded promising results. This significant phase in scientific history marked a rapid absorption of machine learning to enhance predictions and identifying trends. However, this adoption may have come at the cost of neglecting traditional (spatial) statistics (Heuvelink & Webster, 2022). For instance, contrary to traditional statistics, machine learning does not directly account for spatial correlation, and obtaining uncertainty estimates is not as straightforward.

Fast forward to the present, nearly halfway through the 2020s, and there seems to be a shift in the narrative, or perhaps it has already occurred. While machine learning continues to be employed to improve prediction accuracy, the current focus leans more towards gaining process knowledge from these models. This shift is evident in recent years with the rapid rise and application of XML. An intriguing development is also the resurgence of causal inference (Huber, 2023), originally popularised at the beginning of the 21st century, gaining prominence in computer sciences in recent years (Rudin et al., 2021). There is also an increase of research endeavors that integrate machine learning with traditional statistical methodologies to better understand the variable of interest. For instance, using machine learning to also account for spatial correlation (Hengl et al., 2018; Sekulic et al., 2020; Saha & Datta, 2023).

In the context of DSM, two noteworthy research opportunities for machine learning are starting to emerge. Firstly, exploring causal inference within the realm of machine learning promises to enrich our understanding of the underlying soil formation processes. Secondly, addressing spatial correlation with machine learning methods to better quantify the relationship a soil property has with its surroundings. Although some literature on

causal inference (Yuan et al., 2022) and addressing spatial correlation (Hengl et al., 2018; Wadoux et al., 2019; Sekulic et al., 2020; Saha & Datta, 2023) already exist in DSM, there is certainly more research required. For example, for addressing spatial correlation, research is required to improve prediction accuracy as well as estimating spatial correlation. In Chapter 3, the multivariate random forest produced very promising results to retain the correlation structure of the soil properties. It would be interesting to use the multivariate random forest, to model in addition to a soil property, also output variables related to the spatial distribution of the soil property.

6.4 Final conclusion

Mapping soil is crucial in ensuring its preservation, and DSM is an exceptional tool for this purpose. It has also played a significant role in reducing uncertainty in soil maps. While machine learning undeniably plays a significant role in DSM, it also introduces new challenges, and in this thesis, several of these were addressed. This included accounting for uncertainty in soil data, modelling more than one soil property simultaneously, addressing the issue of right-censored soil thickness data, and interpreting machine learning predictions. Given the significance of these challenges, this thesis has made a substantial contribution to improving the application of machine learning in DSM. As the utilisation of machine learning in DSM continues to grow, so too will the associated challenges. Therefore, just as the "Three Laws of Robotics" guide the interaction between machines and humans (Asimov, 1950), an ongoing commitment to research is essential to ensure the proper advancement of machine learning practices necessary in DSM and pedometrics.

References

Aldrich, C., Gardner, S., & Le Roux, N. J. (2004). Monitoring of metallurgical process plants by using biplots. *AIChE Journal*, *50*, 2167–2186.

Anderson, J. A. (1995). *An Introduction to Neural Networks*. Cambridge: MIT press.

Angelini, M. E., Heuvelink, G. B. M., & Kempen, B. (2017). Multivariate mapping of soil with structural equation modelling. *European Journal of Soil Science*, *68*, 575–591. doi: 10.1111/ejss.12446.

Apley, D. W., & Zhu, J. (2020). Visualizing the Effects of Predictor Variables in Black Box Supervised Learning Models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *82*, 1059–1086. doi: 10.1111/rssb.12377.

Asimov, I. (1950). *I, Robot*. Garden City, New York: Doubleday.

Baltensweiler, A., Walthert, L., Hanewinkel, M., Zimmermann, S., & Nussbaum, M. (2021). Machine learning based soil maps for a wide range of soil properties for the forested area of switzerland. *Geoderma Regional*, *27*, e00437. doi: 10.1016/j.geodrs.2021.e00437.

Bandyopadhyay, S., Wolfson, J., Vock, D., Vazquez-Benitez, G., Adomavicius, G., Elidrisi, M., Johnson, P., & O'Connor, P. (2014). Data mining for censored time-to-event data: A bayesian network model for predicting cardiovascular risk from electronic health record data. *Data Mining and Knowledge Discovery*, *29*. doi: 10.1007/s10618-014-0386-6.

Banwart, S. et al. (2014). Benefits of soil carbon: report on the outcomes of an international scientific committee on problems of the environment rapid assessment workshop. *Carbon Management*, *5*, 185–192. doi: 10.1080/17583004.2014.913380.

Barnes, M. (2015). Protect biodiversity, not just area. *Nature*, *526*. doi: 10.1038/526195e.

Belle, V., & Papantonis, I. (2021). Principles and practice of explainable machine learning. *Frontiers in Big Data*, *4*. doi: 10.3389/fdata.2021.688969.

Biecek, P., & Burzykowski, T. (2021). *Explanatory model analysis: explore, explain, and examine predictive models*. CRC Press.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. New York: Springer.

Bonfante, A., Terribile, F., & Bouma, J. (2019). Refining physical aspects of soil quality and soil health when exploring the effects of soil degradation and climate change on biomass production: an italian case study. *Soil*, *5*, 1–14.

Bonfatti, B. R., Hartemink, A. E., Vanwalleghem, T., Minasny, B., & Giasson, E. (2018). A mechanistic model to predict soil thickness in a valley area of rio grande do sul, brazil. *Geoderma*, *309*, 17–31. doi: 10.1016/j.geoderma.2017.08.036.

Bozdogan, H. (1987). Model selection and akaike's information criterion (aic): The general theory and its analytical extensions. *Psychometrika*, *52*, 345–370. doi: 10.1007/BF02294361.

Braekers, R., & Veraverbeke, N. (2005). Cox's regression model under partially informative censoring. *Communications in Statistics - Theory and Methods*, *34*, 1793–1811. doi: 10.1081/STA-200066346.

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

Brungard, C., Nauman, T., Duniway, M., Veblen, K., Nehring, K., White, D., Salley, S., & Anchang, J. (2021). Regional ensemble modeling reduces uncertainty for digital soil mapping. *Geoderma*, *397*, 114998. doi: 10.1016/j.geoderma.2021.114998.

Brunson, J. C. (2023). *ordr: A 'tidyverse' Extension for Ordinations and Biplots*. R package version 0.1.1.0001.

Buonaccorsi, J. P. (2010). *Measurement error : models, methods, and applications*. Chapman & Hall/CRC interdisciplinary statistics series. Boca Raton: CRC Press.

Cevid, D., Michel, L., Näf, J., Bühlmann, P., & Meinshausen, N. (2022). Distributional random forests: Heterogeneity adjustment and multivariate distributional regression. *Journal of Machine Learning Research*, *23*, 1–79. doi: https://doi.org/10.48550/arXiv.2005.14458.

Chaplot, V., Lorentz, S., Podwojewski, P., & Jewitt, G. (2010). Digital mapping of a-horizon thickness using the correlation between various soil properties and soil apparent electrical resistivity. *Geoderma*, *157*, 154–164. doi: 10.1016/j.geoderma.2010.04.006.

Chen, S., Mulder, V. L., Martin, M. P., Walter, C., Lacoste, M., Richer-de Forges, A. C., Saby, N. A. P., Loiseau, T., Hu, B., & Arrouays, D. (2019). Probability mapping of soil thickness by random survival forest at a national scale. *Geoderma*, *344*, 184–194. doi: 10.1016/j.geoderma.2019.03.016.

Christensen, W. F. (2011). Filtered kriging for spatial data with heterogeneous measurement error variances. *Biometrics*, *67*, 947–957.

Clark, T. G., Bradburn, M. J., Love, S. B., & Altman, D. G. (2003). Survival analysis part i: basic concepts and first analyses. *British journal of cancer*, *89*, 232–238. doi: 10.1038/sj.bjc.6601118.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*, 273–297. doi: 10.1007/BF00994018.

Cressie, N. A. C. (1993). *Statistics for Spatial Data*. New York: John Wiley & Sons.

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. doi: 10.1016/j.ecoinf.2022.101665.

De Marsily, G. (1986). *Quantitative Hydrogeology*. San Diego, California: Academic Press.

De Sousa Mendes, W., Demattê, J. A., De Resende, M. E. B., Chimelo Ruiz, L. F., César De Mello, D., Fim Rosas, J. T., Quiñonez Silvero, N. E., Ferracciú Alleoni, L. R., Colzato, M., Rosin, N. A., & Campos, L. R. (2022). A remote sensing framework to map potential toxic elements in agricultural soils in the humid tropics. *Environmental Pollution*, *292*, 118397. doi: 10.1016/j.envpol.2021.118397.

Delhomme, J. P. (1978). Kriging in the hydrosciences. *Advances in Water Resources*, *1*, 251–266.

Department of Agriculture, U. S. (2017). *National soil survey handbook, title 430-VI*.

Department of Environmental Affairs, S. A. (2015). *National Terrestrial Carbon Sink Assessment*. Pretoria, South Africa.

Du Preez, C. C., & van Huyssteen, M.-P., C. (2011). Land use and soil organic matter in south africa 1: A review on spatial variability and the influence of rangeland stock production. *South African Journal of Science*, *107*, 27–34. doi: 10.4102/sajs.v107i5/6.354.

Everitt, B., Dunn, G. et al. (2001). *Applied multivariate data analysis* volume 2. Wiley Online Library.

Fan, J., McConkey, B., Wang, H., & Janzen, H. (2016). Root distribution by depth for temperate agricultural crops. *Field Crops Research*, *189*, 68–74. doi: 10.1016/j.fcr.2016.02.013.

FAO (2020). State of knowledge of soil biodiversity–status, challenges and potentialities, summary for policy makers., . doi: 10.4060/cb1928en.

Florinsky, I. V., Eilers, R. G., Manning, G. R., & Fuller, L. G. (2002). Prediction of soil properties by digital terrain modelling. *Environmental Modelling & Software*, *17*, 295–311. doi: 10.1016/S1364-8152(01)00067-6.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, *29*, 1189 – 1232. doi: 10.1214/aos/1013203451.

Friedman, J. H., & Stuetzle, W. (1981). Projection pursuit regression. *Journal of the American Statistical Association*, *76*, 817–823.

Gabriel, K. R. (1971). The biplot graphic display of matrices with application to principal component analysis. *Biometrika*, *58*, 453–467.

Gardner, S., Le Roux, N. J., & Aldrich, C. (2005). Process data visualisation with biplots. *Minerals Engineering*, *18*, 955–968. doi: 10.1016/j.mineng.2004.12.010. Solid-Liquid Separation '04.

Ge, Y., Wang, J. H., Heuvelink, G. B. M., Jin, R., Li, X., & Wang, J. F. (2015). Sampling design optimization of a wireless sensor network for monitoring ecohydrological processes in the babao river basin, china. *International Journal of Geographical Information Science*, *29*, 92–110. doi: 10.1080/13658816.2014.948446.

Goldstein, A., Kapelner, A., Bleich, J., & Pitkin, E. (2015). Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational and Graphical Statistics*, *24*, 44–65. doi: 10.1080/10618600.2014.907095.

Goovaerts, P. (1999). Geostatistics in soil science: state-of-the-art and perspectives. *Geoderma*, *89*, 1–45.

Gower, J. C., Gardner Lubbe, S., & Le Roux, N. J. (2011). *Understanding biplots*. John Wiley & Sons.

Gower, J. C., & Hand, D. J. (1996). *Biplots*. Chichester: Chapman and Hall.

Gower, J. C., & Harding, S. A. (1988). Nonlinear biplots. *Biometrika*, *75*, 445–455.

Graf, E., Schmoor, C., Sauerbrei, W., & Schumacher, M. (1999). Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine*, *18*, 2529–2545. doi: 10.1002/(SICI)1097-0258(19990915/30)18:17/18<2529::AID-SIM274>3.0.CO;2-5.

Greenwell, B. (2023). *fastshap: Fast Approximate Shapley Values*. R package version 0.1.0.

Greiner, L. (2018). *Soil function assessment for Switzerland*. Ph.D. thesis ETH Zurich.

Greiner, L., Keller, A., Grêt-Regamey, A., & Papritz, A. (2017). Soil function assessment: review of methods for quantifying the contributions of soils to ecosystem services. *Land Use Policy*, *69*, 224–237. doi: 10.1016/j.landusepol.2017.06.025.

Greiner, L., Nussbaum, M., Papritz, A., Zimmermann, S., Gubler, A., Grêt-Regamey, A., & Keller, A. (2018). Uncertainty indication in soil function maps – transparent and easy-to-use information to support sustainable use of soil resources. *SOIL*, *4*, 123–139. doi: 10.5194/soil-4-123-2018.

Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K. L., & Rosati, R. A. (1982). Evaluating the yield of medical tests. *JAMA*, *247*, 2543–2546.

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation

and to related problems. *Journal of the American Statistical Association*, *72*, 320–338.

Hastie, T., Tibshirani, R., & Friedman, J. (2008). *The Elements of Statistical Learning*. Stanford, California: Springer.

Hengl, T. (2021). *landmap: Automated Spatial Prediction using Ensemble Machine Learning*. R package version 0.0.7.

Hengl, T., Leenaars, J. G. B., Shepherd, K. D., Walsh, M. G., Heuvelink, G. B. M., Mamo, T., Tilahun, H., Berkhout, E., Cooper, M., Fegraus, E., Wheeler, I., & Kwabena, N. A. (2017). Soil nutrient maps of sub-saharan africa: assessment of soil nutrient content at 250 m spatial resolution using machine learning. *Nutrient Cycling in Agroecosystems*, *109*, 77–102.

Hengl, T., Nussbaum, M., Wright, M. N., Heuvelink, G. B. M., & Gräler, B. (2018). Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. *PeerJ 6:e5518*, . doi: 10.7717/peerj.5518.

Hernan, M. A., & Robins, J. M. (2020). *Causal Inference: What If*. Chapman and Hall/CRC, Boca Raton.

Herrick, J. E. et al. (2016). The land-potential knowledge system (landpks): mobile apps and collaboration for optimizing climate change investments. *Ecosystem Health and Sustainability*, *2*, e01209. doi: 10.1002/ehs2.1209.

Heuvelink, G. B. M. (1998). *Error Propagation in Environmental Modelling with GIS*. CRC press.

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. doi: 10.1080/13658810601063951.

Heuvelink, G. B. M., Kros, J., Reinds, G. J., & De Vries, W. (2016). Geostatistical prediction and simulation of european soil property maps. *Geoderma Regional*, *7*, 201–215. doi: https://doi.org/10.1016/j.geodrs.2016.04.002.

Heuvelink, G. B. M., & Webster, R. (2022). Spatial statistics and soil mapping: A blossoming partnership under pressure. *Spatial Statistics*, *50*, 100639. doi: 10.1016/j.spasta.2022.100639.

Hiemstra, P. H., Pebesma, E. J., Twenhofel, C. J. W., & Heuvelink, G. B. M. (2008). Real-time automatic interpolation of ambient gamma dose rates from the dutch radioactivity monitoring network. *Computers & Geosciences*, . doi: 10.1016/j.cageo.2008.10.011.

Hofmeyr, D. P. (2022). Fast kernel smoothing in R with applications to projection pursuit. *Journal of Statistical Software*, *101*, 1–33. doi: 10.18637/jss.v101.i03.

Huang, X., & Wolfe, R. A. (2002). A frailty model for informative censoring. *Biometrics*, *58*, 510–250. doi: 10.1111/j.0006-341x.2002.00510.x.

Huber, M. (2023). *Causal analysis: Impact evaluation and Causal Machine Learning with applications in R*. MIT Press.

Huisman, J. A., Snepvangers, J. J. J. C., Bouten, W., & Heuvelink, G. B. M. (2002). Mapping spatial variation in surface soil water content: comparison of ground-penetrating radar and time domain reflectometry. *Journal of Hydrology*, *269*, 194–207. doi: 10.1016/S0022-1694(02)00239-1.

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., & Lauer, M. S. (2008). Random survival forests. *Ann. Appl. Statist.*, *2*, 841–860.

Iticha, B., & Takele, C. (2019). Digital soil mapping for site-specific management of soils. *Geoderma*, *351*, 85–91. doi: 10.1016/j.geoderma.2019.05.026.

Jabro, J. D. (1992). Estimation of saturated hydraulic conductivity of soils from particle size distributions and bulk density data. *American Society of Agricultural and Biological Engineers*, *35*, 557–560. doi: 10.13031/2013.28633.

Jäggli, F., Peyer, K., Pazeller, A., & Schwab, P. (1998). *Grundlagenbericht Zur Bodenkartierung Des Kantons Zürich*. Technical Report, .

Jenny, H. (1941). *Factors of Soil Formation: A System of Quantitative Pedology*. New York: McGraw-Hill.

Jia, X., Hu, B., Marchant, B. P., Zhou, L., Shi, Z., & Zhu, Y. (2019). A methodological framework for identifying potential sources of soil heavy metal pollution based on machine learning: A case study in the yangtze delta, china. *Environmental Pollution*, *250*, 601–609. doi: 10.1016/j.envpol.2019.04.047.

Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, *53*, 457–481.

Kempen, B., Brus, D. J., & De Vries, F. (2015). Operationalizing digital soil mapping for nationwide updating of the 1:50,000 soil map of the netherlands. *Geoderma*, *241-242*, 313–329. doi: 10.1016/j.geoderma.2014.11.030.

Kidd, D., Webb, M., Malone, B., Minasny, B., & McBratney, A. B. (2015). Digital soil assessment of agricultural suitability, versatility and capital in tasmania, australia. *Geoderma Regional*, *6*, 7–21. doi: 10.1016/j.geodrs.2015.08.005.

Kleinbaum, D. G., Klein, M. et al. (2012). *Survival analysis: a self-learning text* volume 3. Springer.

Kuriakose, S. L., Devkota, S., Rossiter, D. G., & Jetten, V. G. (2009). Prediction of soil depth using environmental variables in an anthropogenic landscape, a case study in the western ghats of kerala, india. *Catena*, *79*, 27–38.

Lacoste, M., Mulder, V. L., Richer-De-Forges, A. C., Martin, M. P., & Arrouays, D. (2016). Evaluating large-extent spatial modeling approaches: A case study for soil depth for france. *Geoderma Regional*, *7*, 137–152.

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. doi: 10.1016/j.geoderma.2019.05.031.

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. doi: 10.1111/j.1365-2389.2005.00768.x.

Laslett, 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. *Journal of Soil Science*, *41*, 451–471. doi: 10.1111/j.1365-2389.1990.tb00079.x.

Lawrence, I., & Lin, K. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, *45*, 255–268.

LeBlanc, M., & Crowley, J. (1993). Survival trees by goodness of split. *Journal of the American Statistical Association*, *88*, 457–467.

Leenaars, J. G. B., Kempen, B., van Oostrum, A. J. M., & Batjes, N. H. (2014). African soil profiles database: a compilation of georeferenced and standardised legacy soil profile data for sub-saharan africa. In D. Arrouays, N. McKenzie, J. Hempel, A. R. de Forge, & A. B. McBratney (Eds.), *GlobalSoilMap - Basis of the global spatial information system* (pp. 51–57). Orléans: Taylor & Francis Group.

Leung, K. M., Elashoff, R. M., & Afifi, A. A. (1997). Censoring issues in survival analysis. *Annual review of public health*, *18*, 83–104.

Lingwall, J. W., & Christensen, W. F. (2007). Pollution source apportionment using a priori information and positive matrix factorization. *Chemometrics and Intelligent Laboratory Systems*, *87*, 281–294. doi: 10.1016/j.chemolab.2007.03.007.

Liu, F., Yang, F., Zhao, Y., Zhang, G., & Li, D. (2022). Predicting soil depth in a large and complex area using machine learning and environmental correlations. *Journal of Integrative Agriculture*, *21*, 2422–2434. doi: 10.1016/S2095-3119(21)63692-4.

Malone, B., & Searle, R. (2020a). Improvements to the australian national soil thickness map using an integrated data mining approach. *Geoderma*, *377*, 114579. doi: 10.1016/j.geoderma.2020.114579.

Malone, B., & Searle, R. (2020b). Improvements to the australian national soil thickness map using an integrated data mining approach. *Geoderma*, *377*, 114579. doi: 10.1016/j.geoderma.2020.114579.

Matschullat, J. et al. (2018). Gemas: Cns concentrations and c/n ratios in european agricultural soil. *Science of The Total Environment*, *627*, 975–984. doi: 10.1016/j.scitotenv.2018.01.214.

Mazzetti, C., & Todini, E. (2009). Combining weather radar and raingauge data for

hydrologic applications. In P. Samuels, S. Huntington, W. Allsop, & J. Harrop (Eds.), *Flood Risk Management: Research and Practice*. London: Taylor and Francis Group.

McBratney, A. B., Fajardo, M., Malone, B. P., Bishop, T. F. A., Stockmann, U., & Odeh, I. O. A. (2018). Effective multivariate description of soil and its environment. In *Pedometrics* (pp. 87–112). Cham: Springer International Publishing. doi: 10.1007/978-3-319-63439-5_4.

McBratney, A. B., Mendonça Santos, M. L., & Minasny, B. (2003). On digital soil mapping. *Geoderma*, *117*, 3–52. doi: 10.1016/S0016-7061(03)00223-4.

McDonald, J. F., & Moffitt, R. A. (1980). The uses of tobit analysis. *The review of economics and statistics*, (pp. 318–321).

McLachlan, G. J. (1999). Mahalanobis distance. *Reson*, *4*, 20–26. doi: 10.1007/BF02834632.

Meinshausen, N. (2006). Quantile regression forests. *Journal of Machine Learning Research*, *7*, 983–999.

Michel, L., & Ćevid, D. (2021). *drf: Distributional Random Forests*. R package version 1.1.0.

Miller, K., Huettmann, F., Norcross, B., & Lorenz, M. (2014). Multivariate random forest models of estuarine-associated fish and invertebrate communities. *Marine Ecology Progress Series*, *500*, 159–174. doi: {10.3354/meps10659}.

Minasny, B., & McBratney, A. B. (1999). A rudimentary mechanistic model for soil production and landscape development. *Geoderma*, *90*, 3–21. doi: 10.1016/S0016-7061(98)00115-3.

Minasny, B., & McBratney, A. B. (2006). Mechanistic soil–landscape modelling as an approach to developing pedogenetic classifications. *Geoderma*, *133*, 138–149. doi: 10.1016/j.geoderma.2006.03.042. Advances in landscape-scale soil research.

Minasny, B., & McBratney, A. B. (2016a). Digital soil mapping: A brief history and some lessons. *Geoderma*, *264*, 301–311. doi: 10.1016/j.geoderma.2015.07.017. Soil mapping, classification, and modelling: history and future directions.

Minasny, B., & McBratney, A. B. (2016b). Digital soil mapping: A brief history and some lessons. *Geoderma*, *264*, 301–311.

Minasny, B., McBratney, A. B., Malone, B. P., & Wheeler, I. (2013). Chapter one - digital mapping of soil carbon. In D. L. Sparks (Ed.), *Advances in Agronomy* (pp. 1–47). Academic Press volume 118 of *Advances in Agronomy*. doi: 10.1016/B978-0-12-405942-9.00001-3.

Molnar, C. (2020). *Interpretable Machine Learning: A Guide for Making Black Box Models Interpretable*. Raleigh, USA: Lulu Press.

Molnar, C., Bischl, B., & Casalicchio, G. (2018). iml: An r package for interpretable machine learning. *JOSS*, *3*, 786. doi: 10.21105/joss.00786.

Montgomery, D. C. (2012). *Design and Analysis of Experiments*. John Wiley & Sons, New York.

Moore, I. D., Gessler, P. E., Nielsen, G. A. E., & Peterson, G. A. (1993). Soil attribute prediction using terrain analysis. *Soil science society of America journal*, *57*, 443–452.

Mulder, V. L., Lacoste, M., Richer-de Forges, A. C., & Arrouays, D. (2016). Globalsoilmap france: High-resolution spatial modelling the soils of france up to two meter depth. *Science of The Total Environment*, *573*, 1352–1369. doi: 10.1016/j.scitotenv.2016.07.066.

Nawar, S., Buddenbaum, H., & Hill, J. (2015). Digital mapping of soil properties using multivariate statistical analysis and aster data in an arid region. *Remote Sensing*, *7*, 1181–1205. doi: 10.3390/rs70201181.

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.

Nocita, M. et al. (2015). Chapter four - soil spectroscopy: An alternative to wet chemistry for soil monitoring. (pp. 139–159). Academic Press volume 132 of *Advances in Agronomy*. doi: 10.1016/bs.agron.2015.02.002.

Odebiri, O., Mutanga, O., Odindi, J., & Naicker, R. (2023). Mapping soil organic carbon distribution across south africa's major biomes using remote sensing-topo-climatic covariates and concrete autoencoder-deep neural networks. *Science of The Total Environment*, *865*, 161150. doi: 10.1016/j.scitotenv.2022.161150.

Odeh, I. O. A., Chittleborough, D. J., & McBratney, A. B. (1991a). Elucidation of soil-landform interrelationships by canonical ordination analysis. *Geoderma*, *49*, 1–32.

Odeh, I. O. A., Chittleborough, D. J., & McBratney, A. B. (1991b). Elucidation of soil-landform interrelationships by canonical ordination analysis. *Geoderma*, *49*, 1–32.

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.

Okonkwo, E. I., Corstanje, R., & Kirk, G. J. D. (2018). Digital soil assessment for quantifying soil constraints to crop production: a case study for rice in punjab, india. *Soil Use and Management*, *34*, 533–541. doi: 10.1111/sum.12446.

Orgiazzi, A., Ballabio, C., Panagos, P., Jones, A., & Fernández-Ugalde, O. (2018). Lucas soil, the largest expandable soil dataset for europe: a review. *European Journal of Soil Science*, *69*, 140–153. doi: 10.1111/ejss.12499.

Padarian, J., McBratney, A. B., & Minasny, B. (2020). Game theory interpretation of

digital soil mapping convolutional neural networks. *SOIL*, *6*, 389–397. doi: 10.5194/soil-6-389-2020.

Padarian, J., Minasny, B., & McBratney, A. B. (2019). Using deep learning for digital soil mapping. *SOIL*, *5*, 79–89. doi: 10.5194/soil-5-79-2019.

Papritz, A., & Dubois, J. R. (1999). Mapping heavy metals in soil by (non-) linear kriging: an empirical validation. In *geoENV II—geostatistics for environmental applications* (pp. 429–440). Springer.

Paterson, G., Turner, D., Wiese, L., van Zijl, G., Clarke, C., & Van Tol, J. (2015). Spatial soil information in south africa: Situational analysis, limitations and challenges. *South African Journal of Science*, *111*, 1–7. doi: 10.17159/sajs.2015/20140178.

Pebesma, E. J. (2004). Multivariable geostatistics in s: the gstat package. *Computers & Geosciences*, *30*, 683–691. doi: 10.1016/j.cageo.2004.03.012.

Pierdzioch, C., & Risse, M. (2020). Forecasting precious metal returns with multivariate random forests. *Empirical Economics*, *58*, 1167–1184. doi: 10.1007/s00181-018-1558-9.

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. doi: 10.5194/soil-7-217-2021.

Prasad, R., & Power, J. F. (1997). *Soil fertility management for sustainable agriculture*. Boca Raton, Florida: CRC Press.

Prout, J. M., Shepherd, K. D., McGrath, S. P., Kirk, G. J. D., & Haefele, S. M. (2021). What is a good level of soil organic matter? an index based on organic carbon to clay ratio. *European Journal of Soil Science*, *72*, 2493–2503. doi: 10.1111/ejss.13012.

Prout, J. M., Shepherd, K. D., McGrath, S. P., Kirk, G. J. D., Hassall, K. L., & Haefele, S. M. (2022). Changes in organic carbon to clay ratios in different soils and land uses in england and wales over time. *Scientific Reports*, *12*. doi: 10.1038/s41598-022-09101-3.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing Vienna, Austria.

Rabot, E., Guiresse, M., Pittatore, Y., Angelini, M., Keller, C., & Lagacherie, P. (2022a). Development and spatialization of a soil potential multifunctionality index for agriculture (agri-spmi) at the regional scale. case study in the occitanie region (france). *Soil Security*, *6*, 100034. doi: 10.1016/j.soisec.2022.100034.

Rabot, E., Guiresse, M., Pittatore, Y., Angelini, M., Keller, C., & Lagacherie, P. (2022b). Development and spatialization of a soil potential multifunctionality index for agriculture (agri-spmi) at the regional scale. case study in the occitanie region (france). *Soil Security*, *6*, 100034. doi: 10.1016/j.soisec.2022.100034.

Rawls, W. J., Pachepsky, Y. A., Ritchie, J. C., Sobecki, T. M., & Bloodworth, H. (2003). Effect of soil organic carbon on soil water retention. *Geoderma*, *116*, 61–76. doi: 10.1

016/S0016-7061(03)00094-6.

Reeves, D. W. (1997). The role of soil organic matter in maintaining soil quality in continuous cropping systems. *Soil and Tillage Research*, *43*, 131–167. doi: https://doi.org/10.1016/S0167-1987(97)00038-X.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* KDD '16 (pp. 1135—-1144). New York, NY, USA: Association for Computing Machinery. doi: 10.1145/2939672.2939778.

Rodwell, D. T., Van der Merwe, C. J., & Gardner-Lubbe, S. (2021). Categorical cva biplots. *Computational Statistics & Data Analysis*, *163*, 107299. doi: 10.1016/j.csda.2021.107299.

Rossiter, D. G., Liu, J., Carlisle, S., & Zhu, A. (2015). Can citizen science assist digital soil mapping? *Geoderma*, *259-260*, 71–80. doi: 10.1016/j.geoderma.2015.05.006.

Rousseeuw, P. J., Ruts, I., & Tukey, J. W. (1999). The bagplot: A bivariate boxplot. *The American Statistician*, *53*, 382–387. doi: 10.1080/00031305.1999.10474494.

Rowan, A. (2019). Unravelling black box machine learning methods using biplots.

Rudin, C., Chen, C., Chen, Z., Huang, H., Semenova, L., & Zhong, C. (2021). Interpretable machine learning: Fundamental principles and 10 grand challenges. `arXiv:2103.11251`.

Saha, B. S., A., & Datta, A. (2023). Random forests for spatially dependent data. *Journal of the American Statistical Association*, *118*, 665–683. doi: 10.1080/01621459.2021.1950003.

Schennach, S. M. (2016). Recent advances in the measurement error literature. *Annual Review of Economics*, *8*, 341–377. doi: 10.1146/annurev-economics-080315-015058.

Schoorl, J. M., Veldkamp, A., & Bouma, J. (2002). Modeling water and soil redistribution in a dynamic landscape context. *Soil Science Society of America Journal*, *66*, 1610–1619. doi: 10.2136/sssaj2002.1610.

Schulze, R. E., & Schütte, S. (2020). Mapping soil organic carbon at a terrain unit resolution across south africa. *Geoderma*, *373*, 114447. doi: 10.1016/j.geoderma.2020.114447.

Segal, M., & Xiao, Y. (2011). Multivariate random forests. *WIREs Data Mining and Knowledge Discovery*, *1*, 80–87. doi: 10.1002/widm.12.

Segal, M. R. (1988). Regression trees for censored data. *Biometrics*, *44*, 35–47.

Sekulic, A., Kilibarda, M., Heuvelink, G. B., Nikolić, M., & Bajat, B. (2020). Random forest spatial interpolation. *Remote Sensing*, *12*. doi: 10.3390/rs12101687.

Service center NABODAT (2022). *Swiss Soil Dataset – Documentation Version 6 (April*

2022). Technical Report, . Https://nabodat.ch/index.php/de/service/bodendatensatz.

Shangguan, W., Hengl, T., Mendes De Jesus, J., Yuan, H., & Dai, Y. (2017). Mapping the global depth to bedrock for land surface modeling. *Journal of Advances in Modeling Earth Systems*, *9*, 65—88. doi: 10.1002/2016MS000686.

Shapley, L. S. (1953). 17. a value for n-person games. In H. W. Kuhn, & A. W. Tucker (Eds.), *Contributions to the Theory of Games (AM-28), Volume II* (pp. 307–318). Princeton: Princeton University Press. doi: doi:10.1515/9781400881970-018.

Söderström, M., Sohlenius, G., Rodhe, L., & Piikki, K. (2016). Adaptation of regional digital soil mapping for precision agriculture. *Precision Agriculture*, *17*, 588–607.

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.

Srivastava, P., Singh, R., Tripathi, S., Singh, P., Singh, S., Singh, H., Raghubanshi, A. S., & Mishra, P. K. (2017). Soil carbon dynamics under changing climate — a research transition from absolute to relative roles of inorganic nitrogen pools and associated microbial processes: A review. *Pedosphere*, *27*, 792–806. doi: 10.1016/S1002-0160(17)60488-0.

Stein, A., & Corsten, L. C. A. (1991). Universal kriging and cokriging as a regression procedure. *Biometrics*, *47*, 575–587.

Stein, M. L. (1999). *Interpolation of spatial data: Some theory on kriging*. New York: Springer.

Strumbelj, E., & Kononenko, I. (2014). Explaining prediction models and individual predictions with feature contributions. *Knowl Inf Syst*, *41*, 647–665. doi: 10.1007/s1 0115-013-0679-x.

Subramanian, D., & Rajendra, H. (2022). Digital mapping of soil texture classes using random forest classification algorithm. *Soil Use and Management*, *38*, 135–149. doi: 10.1111/sum.12668.

Suleymanov, A., Abakumov, E., Suleymanov, R., Gabbasova, I., & Komissarov, M. (2021). The soil nutrient digital mapping for precision agriculture cases in the trans-ural steppe zone of russia using topographic attributes. *ISPRS International Journal of Geo-Information*, *10*. doi: 10.3390/ijgi10040243.

Swanson, R. K., Xu, R., Nettleton, D., & Glatz, C. E. (2012). Proteomics-based, multivariate random forest method for prediction of protein separation behavior during cation-exchange chromatography. *Journal of Chromatography*, *1249*, 103–114. doi: 10.1016/j.chroma.2012.06.009.

Taghizadeh-Mehrjardi, R., Fathizad, H., Ali Hakimzadeh Ardakani, M., Sodaiezadeh, H.,

Kerry, R., Heung, B., & Scholten, T. (2021). Spatio-temporal analysis of heavy metals in arid soils at the catchment scale using digital soil assessment and a random forest model. *Remote Sensing*, *13*. doi: 10.3390/rs13091698.

Takoutsing, B., & Heuvelink, G. (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. doi: 10.1016/j.geoderma.2022.116192.

Tesfa, T. K., Tarboton, D. G., Chandler, D. G., & McNamara, J. P. (2009). Modeling soil depth from topographic and land cover attributes. *Water Resources Research*, *45*.

Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica: journal of the Econometric Society*, (pp. 24–36).

Turek, M. E., Poggio, L., Batjes, N. H., Armindo, R. A., De Jong van Lier, Q., De Sousa, L., & Heuvelink, G. (2022). Global mapping of volumetric water retention at 100, 330 and 15000 cm suction using the wosis database. *International Soil and Water Conservation Research*, . doi: 10.1016/j.iswcr.2022.08.001.

Turkson, A. J., Ayiah-Mensah, F., & Nimoh, V. (2021). Handling censoring and censored data in survival analysis: A standalone systematic literature review. *International Journal of Mathematics and Mathematical Sciences*, *2021*.

Van den Berg, F., Tiktak, A., Heuvelink, G. B. M., Burgers, S. L. G. E., Brus, D. J., de Vries, F., Stolte, J., & Kroes, J. G. (2012). Propagation of uncertainties in soil and pesticide properties to pesticide leaching. *Journal of Environmental Quality*, *41*, 253–261. doi: 10.2134/jeq2011.0167.

Van der Merwe, C. J. (2020). *Classifying yield spread movements in sparse data through triplots*. Ph.D. thesis Stellenbosch University.

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*, *n/a*. doi: 10.1111/ejss.13137.

Vasat, R., Heuvelink, G. B. M., & Boruvka, L. (2010). Sampling design optimization for multivariate soil mapping. *Geoderma*, *155*, 147–153. doi: 10.1016/j.geoderma.2009.07.005.

Vaysse, K., & Lagacherie, P. (2015). Evaluating digital soil mapping approaches for mapping globalsoilmap soil properties from legacy data in languedoc-roussillon (france). *Geoderma Regional*, *4*, 20–30. doi: 10.1016/j.geodrs.2014.11.003.

Vaysse, K., & Lagacherie, P. (2017). Using quantile regression forest to estimate uncertainty of digital soil mapping products. *Geoderma*, *291*, 55–64. doi: 10.1016/j.geoderma.2016.12.017.

Venter, Z. S., Hawkins, H., Cramer, M. D., & Mills, A. J. (2021). Mapping soil organic

carbon stocks and trends with satellite-driven high resolution maps over south africa. *Science of The Total Environment*, *771*, 145384. doi: 10.1016/j.scitotenv.2021.145384.

Vines, S. K. (2015). Predictive nonlinear biplots: Maps and trajectories. *Journal of Multivariate Analysis*, *140*, 47–59. doi: 10.1016/j.jmva.2015.04.010.

Vock, D. M., Wolfson, J., Bandyopadhyay, S., Adomavicius, G., Johnson, P. E., Vazquez-Benitez, G., & O'Connor, P. J. (2016). Adapting machine learning techniques to censored time-to-event health record data: A general-purpose approach using inverse probability of censoring weighting. *Journal of Biomedical Informatics*, *61*, 119–131. doi: 10.1016/j.jbi.2016.03.009.

Wackernagel, H. (2010). *Multivariate Geostatistics : An Introduction with Applications*. (3rd ed.). Berlin: Springer.

Wadoux, A. M. J.-C. (2019). Using deep learning for multivariate mapping of soil with quantified uncertainty. *Geoderma*, *351*, 59–70. doi: 10.1016/j.geoderma.2019.05.012.

Wadoux, A. M. J.-C., Brus, D. J., & Heuvelink, G. B. M. (2018). Accounting for non-stationary variance in geostatistical mapping of soil properties. *Geoderma*, *324*, 138–147. doi: 10.1016/j.geoderma.2018.03.010.

Wadoux, A. M. J.-C., Heuvelink, G. B. M., Lark, R. M., Lagacherie, P., Bouma, J., Mulder, V. L., Libohova, Z., Yang, L., & McBratney, A. B. (2021). Ten challenges for the future of pedometrics. *Geoderma*, *401*, 115155. doi: 10.1016/j.geoderma.2021.115155.

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. doi: 10.1016/j.earscirev.2020.103359.

Wadoux, A. M. J.-C., & Molnar, C. (2022). Beyond prediction: methods for interpreting complex models of soil variation. *Geoderma*, *422*, 115953. doi: 10.1016/j.geoderma.2022.115953.

Wadoux, A. M. J.-C., Padarian, J., & Minasny, B. (2019). Multi-source data integration for soil mapping using deep learning. *SOIL*, *5*, 107–119. doi: 10.5194/soil-5-107-2019.

Wadoux, A. M. J.-C., Saby, N. P. A., & Martin, M. P. (2023). Shapley values reveal the drivers of soil organic carbon stock prediction. *SOIL*, *9*, 21–38. doi: 10.5194/soil-9-21-2023.

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. doi: 10.1111/ejss.12909.

Wan, Q., & Pal, R. (2013). A multivariate random forest based framework for drug sensitivity prediction. In *2013 IEEE International Workshop on Genomic Signal Processing and Statistics* (pp. 53–53). doi: 10.1109/GENSIPS.2013.6735929.

Webster, R., & Oliver, M. A. (2007). *Geostatistics for Environmental Scientists*. (2nd ed.). Chichester, West Sussex: John Wiley & Sons.

Wegehenkel, M. (2005). Validation of a soil water balance model using soil water content and pressure head data. *Hydrological Processes: An International Journal*, *19*, 1139–1164.

Wei, Z., Li, X., Sun, M., Guo, R., Liu, G., Xu, Z., & Cheng, Y. (2023). Discriminating chert origins using machine-learning approaches. *Geological Journal*, *58*, 2403–2417. doi: 10.1002/gj.4753.

Weil, R., & Brady, N. (2017). *The Nature and Properties of Soils*. (15th ed.).

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Willems, S., Schat, A., van Noorden, M., & Fiocco, M. (2018). Correcting for dependent censoring in routine outcome monitoring data by applying the inverse probability censoring weighted estimator. *Statistical Methods in Medical Research*, *27*, 323–335. doi: 10.1177/0962280216628900.

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*, 1–17. doi: 10.18637/jss.v077.i01.

Wu, J., Teng, Y., Chen, H., & Li, J. (2016). Machine-learning models for on-site estimation of background concentrations of arsenic in soils using soil formation factors. *Journal of Soils and Sediments*, *16*, 1787–1797. doi: 10.1007/s11368-016-1374-9.

Wurz, S., Van Peer, P., Le Roux, N. J., Gardner, S., & Deacon, H. (2005). Continental patterns in stone tools: A technological and biplot-based comparison of early late pleistocene assemblages from northern and southern africa. *African Archaeological Review*, *22*, 1–24. doi: 10.1007/s10437-005-3157-3.

Xu, S., An, X., Qiao, X., Zhu, L., & Li, L. (2013). Multi-output least-squares support vector regression machines. *Pattern Recognition Letters*, *34*, 1078–1084. doi: 10.1016/j.patrec.2013.01.015.

Yuan, K. et al. (2022). Causality guided machine learning model on wetland ch4 emissions across global wetlands. *Agricultural and Forest Meteorology*, *324*, 109115. doi: https://doi.org/10.1016/j.agrformet.2022.109115.

Zhang, W., Hu, G., Sheng, J., Weindorf, D. C., Wu, H., Xuan, J., Yan, A., & Gu, Z. (2018). Estimating effective soil depth at regional scales: Legacy maps versus environmental covariates. *Journal of Plant Nutrition and Soil Science*, *181*, 167–176. doi: 10.1002/jpln.201700081.

Zhao, J., & Meng, D. (2015). Fastmmd: Ensemble of circular discrepancy for efficient two-sample test. *Neural Computation*, *27*, 1345–1372. doi: 10.1162/NECO_a_00732.

Zheng, M., & Klein, J. P. (1994). A self-consistent estimator of marginal survival functions based on dependent competing risk data and an assumed copula. *Communications in Statistics - Theory and Methods*, *23*, 2299–2311. doi: 10.1080/03610929408831387.

Zhu, J., Yang, M., & Ren, Z. J. (2023). Machine learning in environmental research: Common pitfalls and best practices. *Environmental Science & Technology*, *57*, 17671–17689. doi: 10.1021/acs.est.3c00026.

Zizala, D., Minařík, R., Skála, J., Beitlerová, H., Juřicová, A., Reyes Rojas, J., Penížek, V., & Zádorová, T. (2022). High-resolution agriculture soil property maps from digital soil mapping methods, czech republic. *CATENA*, *212*, 106024. doi: 10.1016/j.catena.2022.106024.

Summary

Machine learning, characterised by data-driven models that identify patterns in data to execute specific tasks without explicit instructions or assumptions, is increasingly applied in environmental research, including in digital soil mapping (DSM). These models excel in quantifying complex relationships between inputs and outputs, and thereby often surpass the predictive accuracy of traditional statistical models like linear regression. However, the increased use of machine learning in this field brings forth challenges for DSM practitioners, and it is not clear how to address some of these issues with machine learning. In addition, users of machine learning also face common pitfalls such as data leakage, model selection and optimisation, and model interpretability. This thesis focuses on four opportunities within DSM to demonstrate how the use of machine learning can be enhanced. These opportunities include modelling uncertain soil data (**Chapter 2**), employing machine learning for multivariate mapping of soil (**Chapter 3**), addressing the modelling of right-censored soil thickness data (**Chapter 4**), and enhancing the interpretability of machine learning models (**Chapter 5**). In each instance, the objective is to tackle the challenges in an optimal statistical manner, predominantly addressing issues related to model selection and evaluation, model optimisation, and interpretability.

Soil data are frequently contaminated with measurement errors, which is mainly attributed to samples being increasingly obtained with low-cost techniques, such as infrared spectroscopy. When dealing with soil observations contaminated with measurement errors, it is important to account for these errors to ensure that model predictions reflect the underlying error-free process of interest. In **Chapter 2**, a two-stage maximum likelihood framework is introduced to address measurement errors in the response variable using machine learning. In this proposed framework, an estimate of the measurement error variance is required and is incorporated as a weight in the log-likelihood function. This weighting ensures that measurements with a higher measurement error variance contribute less to the calibration of the model. Furthermore, knowledge of the residual variance of the measurement error-free model is required. The framework therefore works in two stages, with the first stage involving obtaining an estimate of the model parameters, which is then used to estimate the residual variance. The second stage entails using the residual variance to derive new estimates of the model parameters. This process can be iterated between the two stages until satisfactory estimates are achieved. The proposed

framework is illustrated with a random forest model and a projection pursuit regression model in a synthetic simulation study as well as in a real-world case study of mapping clay content in Namibia. The machine learning models were also benchmarked against that of an error-filtering regression kriging model. Specifically, each model was compared to itself with and without accounting for measurement errors. The simulation study revealed that the average size and the relative variation of the measurement error variance greatly influenced the accuracy of the error-filtered machine learning models, and even more so for the error-filtered kriging model. In the case study, the error-filtered models were comparable to the ordinary models, which was mainly attributed to the measurement error variances being small and relatively constant.

It is often that soil maps are produced in a univariate manner, ignoring the underlying dependence structure between the respective soil properties. This way of mapping soil could lead to inconsistent model predictions. For instance, independently mapping and simulating soil organic carbon and total nitrogen may yield unrealistic carbon-nitrogen ratios. Despite the capability of several machine learning models to handle multiple responses, there remains limited research on the application of multivariate machine learning in DSM. One such model is a random forest. In **Chapter 3**, a case study involving European data, where soil organic carbon and total nitrogen are mapped, compares the performance of a multivariate random forest model to the case where each soil property is modeled separately using ordinary random forest models. This comparison employs stochastic simulations determined by sampling from the conditional distributions of the soil properties, given the covariates, as estimated by a quantile regression forest. The models are also assessed based on prediction accuracy using common metrics such as the root mean square error. Similar comparisons were made between regression co-kriging and regression kriging models. The simulations revealed that the multivariate random forest model (and the regression co-kriging model) was superior in maintaining the dependence structure between the soil properties compared to that of the univariate random forest models (and regression kriging models). In addition, by using metrics like the root mean square error, the accuracy of the multivariate and univariate models were comparable.

Chapter 4 deals with mapping soil thickness, which is often defined as the distance from the soil surface to bedrock. Soil thickness is crucial for various studies, including those related to land capability, carbon storage, crop suitability assessment, and infrastructure planning. However, soil thickness data are often right-censored, indicating that the true thickness exceeds the recorded auger depth. This may occur when determining soil depth is not a focal point of the study for which samples are collected, or when sampling to a particular depth is costly and requires specialised equipment. When mapping soil thickness, it is imperative to consider the right-censoring of data; otherwise, predictions may severely underestimate the actual depth. Survival models, commonly used in epidemiological studies to model time-to-event response variables, can effectively account for

censoring. In this chapter, the focus is on investigating the random survival forest and a proposed inverse probability of censoring weighting (IPCW) random forest model. The IPCW model operates on the principle of assigning weights to calibration data, where the weights are proportional to the probability of exceeding a certain depth. These models, along with ordinary random forest models fitted either on all of the data (censored and non-censored) or only the censored data, were compared in a synthetic simulation study and applied to two real-world case studies — one in Switzerland and another in Maine, USA. The simulation study revealed that the proportion of censored data had a significant effect on the results of the models. The ICPW random forest model also performed relatively well in the simulation study and in both case studies.

Machine learning models are considered to be "black-box" in nature as it is difficult to analyse how they produce their predictions. In DSM, it is important to provide explanations for predictions. For example, it could be vital to understand how a certain environmental covariate such as temperature contributed to the final map predictions of soil organic carbon. Explainable machine learning (XML) is a rapidly growing field with the purpose of unravelling the black box by providing insight in the model's process of producing predictions. Some of the most popular XML tools are model-agnostic and provide intuitive visualisations that depict the relationship between the predictions and the covariates. However, with many of these tools, only one (or two) covariate can be visualised at a time, and often covariates are required to be uncorrelated. This is problematic as it is commonly known that environmental covariates share complex relationships and may even interact, and if these relationships are not accounted for the resulting visualisation may give misleading results. The most popular XML tools also do not come with goodness-of-fit statistics, which is problematic because the quality of the representation cannot be known. In **Chapter 5**, the use of biplots is explored to address some of the issues with most XML methods. A biplot is a powerful multivariate tool which can visualise multiple covariates at a time, and does not require covariates to be uncorrelated. In addition, a biplot comes with goodness-of-fit statistics which allows a user to analyse the quality of the display. The use of biplots for understanding machine learning predictions is illustrated in a case study of mapping soil organic carbon in South Africa. The biplot methodology is also compared to other XML methods, such as partial dependence plots, accumulated local effect plots, and visualisations produced with Shapley values. The results presented in this chapter demonstrate that a biplot serves as a valuable tool for understanding model predictions, providing insights into the relationships between multiple covariates and the model's predictions.

Chapter 6 gives an overview of the main findings of this thesis, provides recommendations for future research, and presents a reflection on the use of machine learning in DSM and in the wider context of environmental research. This thesis illustrates the effective use of machine learning in a number of fundamental challenges in DSM in relation to model selection, evaluation, optimisation and model interpretability. The main conclusion of the

thesis is that machine learning will continue to play a vital role in DSM, necessitating ongoing efforts to address the associated challenges. Continued research of the enhancement of the use of machine learning is therefore essential to ensure the robust development of statistical models in DSM and pedometrics.

Opsomming

Masjienleer, gekenmerk deur datagedrewe modelle wat patrone in data identifiseer om spesifieke take uit te voer sonder uitdruklike instruksies of aannames, word toenemend toegepas in omgewingsnavorsing, insluitend in digitale grondkartering (DGK). Hierdie modelle blink uit in die kwantifisering van komplekse verwantskappe tussen model insette en uitsette, en oortref dikwels die voorspellingsakkuraatheid van tradisionele statistiese modelle soos lineêre regressie. Die toenemende gebruik van masjienleer in hierdie veld bring egter uitdagings vir DGK-praktisyns na vore, en dit is nie duidelik hoe om sommige van hierdie kwessies met masjienleer aan te spreek nie. Daarbenewens staar gebruikers van masjienleer ook algemene struikelblokke in die gesig soos datalekkasie, modelseleksie en -optimalisering, en modelverklaarbaarheid. Hierdie tesis fokus op vier geleenthede binne DGK om te demonstreer hoe die gebruik van masjienleer verbeter kan word. Hierdie geleenthede sluit in die modellering van twyfelagtige gronddata (**Hoofstuk 2**), die gebruik van masjienleer vir meerveranderlike kartering van grond (**Hoofstuk 3**), die aanpak van die modellering van regsgesensoreerde gronddiktedata (**Hoofstuk 4**), en die verbetering van die verklaarbaarheid van masjienleermodelle (**Hoofstuk 5**). In elke geval is die doel om die uitdagings op 'n optimale statistiese manier aan te spreek, hoofsaaklik deur kwessies rakende modelseleksie en -evaluering, modeloptimalisering, en verklaarbaarheid te hanteer.

Gronddata is dikwels onsuiwer as gevolg van meetfoute, wat hoofsaaklik toegeskryf word aan monsters wat toenemend met goedkooptegnieke, soos infrarooi spektroskopie, verkry word. Wanneer 'n mens te doen het met grondwaarnemings wat onsuiwer is, is dit belangrik om hierdie foute in ag te neem om te verseker dat modelvoorspellings die onderliggende foutvrye proses van belang weerspieël. In **Hoofstuk 2** word 'n tweefase maksimumaanneemlikheidsraamwerk bekendgestel om meetfoute in die afhanklike veranderlike deur middel van masjienleer aan te spreek. In hierdie voorgestelde raamwerk moet 'n skatting van die meetfoutvariansie beskikbaar wees en word dit dan as 'n gewig in die logaanneemlikheidsfunksie ingesluit. Hierdie gewigte verseker dat metings met 'n hoër meetfoutvariansie minder tot die kalibrering van die model bydrae. Wete van die residuele variansie van die meetfoutvrye model is ook nodig. Die raamwerk werk dus in twee fases. Die eerste fase behels om 'n skatting van die modelparameters te kry wat dan gebruik word om die residuele variansie te skat. Die tweede fase behels dan die gebruik van

die residuele variansie om nuwe skattings van die modelparameters af te lei. Hierdie proses kan tussen die twee fases herhaal word totdat bevredigende skattings bereik word. Die voorgestelde raamwerk word met 'n "random forest" en 'n "projection pursuit" regressie model uitgebeeld. Hierdie uitbeelding vind plaas in 'n sintetiese simulasie studie asook in 'n werklikheidsgevallestudie wat die kartering van klei-inhoud in Namibië behels. Die masjienleermodelle word ook met regressie krigeïng modelle wat foute aanspreek gemeet. Elke model was spesifiek met homself vergelyk deur meetfoute in ag te neem en deur meetfoute te ignoreer. Die simulasie studie het getoon dat die gemiddelde grootte en die relatiewe variasie van die meetfoutvariansie die akkuraatheid van die foutgefiltreerde masjienleermodelle grootliks beïnvloed het, en selfs meer vir die gefiltreerde krigeïng mo delle. In die werklikheidsgevallestudie was die foutgefiltreerde modelle vergelykbaar met die gewone modelle, wat hoofsaaklik toegeskryf is aan die meetfoutvariansies wat klein en relatief konstant was.

Dit gebeur dikwels dat grondkaarte op 'n eenveranderlike manier geproduseer word, sonder om die onderliggende afhanklikheidstruktuur tussen die onderskeie grondeienskappe in ag te neem. Hierdie manier van grondkartering kan tot strydige modelvoorspellings lei. Byvoorbeeld, die onafhanklike kartering en simulering van organiese koolstof en stikstof kan onrealistiese koolstof-stikstof verhoudings oplewer. Ten spyte van die vermoë van verskeie masjienleermodelle om meervoudige veranderlikes te kan hanteer, bly daar beperkte navorsing oor die toepassing van meerveranderlike masjienleer in DGK. Een so 'n model is 'n "random forest". In **Hoofstuk 3**, in 'n Europese gevallestudie waarin organiese koolstof en stikstof gekarteer word, word hierdie meerveranderlike model met gewone modelle vergelyk. Hierdie vergelyking gebruik stogastiese simulasies wat geskep word deur uit die voorwaardelike verdelings van die grondeienskappe, gegee die onafhanklike veranderlikes en soos beraam deur 'n "quantile regression forest", steekproewe te neem. Die modelle word ook op grond van voorspellingsakkuraatheid beoordeel deur algemene maatstawwe soos die wortel van gemiddeldekwadraatfout te gebruik. Soortgelyke vergelykings word gemaak tussen regressie ko-krigeïng en regressie krigeïng modelle. Die simulasies toon dat die meerveranderlike "random forest" model (en die regressie ko-krigeïng model) beter vaar as die modelle wat die grondeienskappe apart hanteer. Daarbenewens, deur maatstawwe soos die wortel van gemiddeldekwadraatfout te gebruik, was die akkuraatheid van die meerveranderlike en eenveranderlike modelle vergelykbaar.

Hoofstuk 4 handel oor die kartering van gronddikte, gedefinieer as die afstand van die grondoppervlak tot die rotsbodem. Gronddikte is van kardinale belang vir verskeie studies, insluitend dié wat verband hou met grondvermoë, koolstofberging, gewasgeskiktheidsbeoordeling, en infrastruktuurbeplanning. Data oor gronddikte is egter dikwels regsgesensoreerd, wat beteken dat die ware diepte die opgetekende boordiepte oorskry. Dit gebeur wanneer die bepaling van gronddikte nie 'n fokuspunt van die studie is waarvoor monsters ingesamel word nie, of wanneer monsterneming tot 'n spesifieke diepte te duur is en gespesialiseerde toerusting vereis word. Wanneer gronddikte gekarteer word,

is dit noodsaaklik om die regsgesensorering van data in ag te neem; andersins kan voorspellings die werklike diepte heelwat onderskat. Oorlewingsmodelle, wat algemeen gebruik word in epidemiologiese studies om tyd-tot-gebeurtenis afhanklike veranderlikes te modelleer, kan doeltreffend vir sensorering voorsiening maak. In hierdie hoofstuk is die fokus op die ondersoek van die "random survival forest" model en 'n voorgestelde model wat van omgekeerde waarskynlikheid sensoreringsgewigte (OESW) gebruik maak. Die OESW model werk op die beginsel van die toewysing van gewigte aan kalibreringsdata, waar die gewigte eweredig is aan die waarskynlikheid om 'n sekere diepte te oorskry. Hierdie modelle word met gewone "random forest" modelle wat óf op al die data (gesensoreerd en nie-gesensoreerd) óf net op die gesensoreerde data gepas is, vergelyk. Hierdie vergelyking vind plaas in 'n sintetiese simulasie studie asook in twee werklikheidsgevallestudies — een in Switserland en 'n ander in Maine, VSA. Die simulasie studie het gewys dat die proporsie van gesensoreerde data 'n beduidende effek op die resultate van die modelle gehad het. Die OESW model het ook relatief goed in die simulasie studie en in albei gevallestudies gevaar.

Masjienleermodelle word dikwels beskou as "swart boks" modelle omdat dit moeilik is om te analiseer hoe hulle hul voorspellings produseer. In DGK is dit belangrik om verklarings vir voorspellings te verskaf. Byvoorbeeld, dit kan noodsaaklik wees om te verstaan hoe 'n sekere omgewingsveranderlike soos temperatuur bydra tot die finale kaartvoorspellings van organiese koolstof. Verklaarbare masjienleer (VML) is 'n vinnig groeiende veld met die doel om die swart boks oop te maak deur insig in die model se proses van voorspellingproduksie te bied. Sommige van die gewildste VML-instrumente is modelagnosties en bied visualiserings wat die verhouding tussen die voorspellings en die onafhanklike veranderlikes uitbeeld. Nietemin, met baie van hierdie instrumente kan slegs een (of twee) veranderlikes op 'n slag uitgebeeld word, en die instrumente vereis ook dat veranderlikes ongekorreleerd is. Dit is problematies aangesien dit algemeen bekend is dat omgewingsveranderlikes komplekse verhoudings deel en selfs kan wisselwerk. Hierdie verhoudings moet in ag geneem word anders kan die gevolglike visualisering misleidende resultate lewer. Die gewildste VML-instrumente kom ook nie met passingstoetse nie. Dit is problematies omdat die kwaliteit van die voorstelling nie bepaal kan word nie. In **Hoofstuk 5** word die gebruik van bi-stippings ondersoek om sommige van die kwessies met die meeste VML-metodes aan te spreek. 'n Bi-stipping is 'n kragtige meerveranderlike instrument wat meervoudige veranderlikes gelyktydig kan visualiseer, en vereis nie dat veranderlikes gekorreleerd is nie. Daarbenewens kom 'n bi-stipping met passingstoetse wat 'n gebruiker in staat stel om die kwaliteit van die visualisering te analiseer. Die gebruik van bi-stippings om masjienleervoorspellings te verstaan, word geïllustreer in 'n werklikheidsgevallestudie van die kartering van organiese koolstof in Suid-Afrika. Die bi-stipping metodiek word ook met ander VML-metodes vergelyk, soos byvoorbeeld die parsiële afhanklikheidsgrafieke, geakkumuleerde lokale effekgrafieke, en visualiserings vervaardig met Shapley-waardes. Die resultate van hierdie hoofstuk toon dat 'n bi-stipping

as 'n waardevolle instrument vir die verstaan van modelvoorspellings dien. Dit bied insigte in die verhoudinge tussen meervoudige onafhanklike veranderlikes en die model se voorspellings.

Hoofstuk 6 gee 'n oorsig van die hoofbevindinge van hierdie tesis, bied aanbevelings vir toekomstige navorsing, en gee nadenke oor die gebruik van masjienleer in DGK en in die wyer konteks van omgewingsnavorsing. Hierdie tesis illustreer die effektiewe gebruik van masjienleer in sommige fundamentele uitdagings in DGK. Hierdie hou in verband met modelseleksie, evaluering, optimalisering en modelverklaarbaarheid. Die hoofgevolgtrekking van die tesis is dat masjienleer sal voortgaan om 'n noodsaaklike rol in DGK te speel, en dit vereis aanhoudende pogings om die gepaardgaande uitdagings aan te spreek. Voortgesette navorsing om die gebruik van masjienleer te verbeter is dus noodsaaklik om die ontwikkeling van statistiese modelle in DGK en pedometrie te verseker.

Acknowledgements

On October 26, 2018, Andrei Rozanov responded to an email from Gerard Heuvelink, during which he recommended me as a candidate for a PhD position. If it were not for Andrei's kind recommendation, my studies at Wageningen would not have been possible, and I would have missed the opportunity to work under a remarkable supervisor. To Andrei, my sincerest thanks.

Words cannot express my gratitude towards Gerard. He was not only an exceptional supervisor but also became a good friend. Thank you, Gerard, for accepting me as your student, and for your invaluable input and expert guidance. It was a pleasure to host you in Stellenbosch, where I had the opportunity to share a bit of my beautiful South Africa with you.

To my co-supervisors, David Hofmeyr and Laura Poggio, thank you for your patience and exceptional guidance throughout this project. David, it was great to be colleagues in the same department for those three months before you left for Lancaster.

I would also like to express my gratitude to the following people. My colleagues and friends in the Department of Genetics, especially Willem Botes, the department chair, who gave me the freedom and encouragement to pursue my PhD. A special thank you also to Thanja Allison, Clint Rhode, Nathan McGregor, and Samantha Joao for your support throughout the years. To my new colleagues in the Department of Statistics and Actuarial Science, specifically Justin Harvey – my new departmental chair, thank you for giving me the freedom to finalise my PhD. Line Shrug, for becoming my friend in the cold dutch country. Jessica and Reinhardt, fellow South Africans who supported me while I was in the Netherlands. To my PhD peers at WUR and colleagues at ISRIC, thank you for making me feel at home when I visited Wageningen over the years. A special thank you to Cathy Clarke for allowing me to be part of her exciting career, and for supplying ideas and data so that I can continue to work on DSM projects. To my community at InVia with whom I ponder life's questions. To YourHeadsGone and Shabbi and many others, thank you for joining me in the World of Warcraft for the times I needed to escape. Thank you to Carmen Platt for designing the thesis cover.

My siblings, Franzel, and her husband Johan, Ian, and his wife Jeanette, Joalisha and

Hanya. You are the best brothers and sisters, and I count myself very lucky to have you in my life. To my parents, Stephan, Max, Liza, and Colin, and (soon to be) parents-in-law Pierre and Naomi, thank you for your support and guidance through it all. I love you very much.

To Jacques, the love of my life, and my best friend. Thank you for being there with me every step of the way.

To God, who through it all, everything is possible.

Life does not stop for you to work on your PhD thesis. You work long days, you work long nights, you teach unappreciative undergrads, you supervise wannabe postgrads, you attend too many meetings, you mark countless papers, you go for performance evaluations, your papers get rejected, you hope for a raise, you buy a house, you buy a second house, you renovate the first house, your dog unexpectedly gets puppies, your neighbours throw late night parties, you lose dear friends, you make new friends, you become an uncle, you start a new job, someone breaks into your house and steals your laptop, you get engaged, you fall off your bike (on the corner of Churchillweg and Hoevestein), proving that foreigners are not be trusted on bicycles, you drink and eat too much, you gain weight, you go on holidays, you play video games, you never feel rested, and you never stop thinking about the end. So, finally, to my PhD, thank you for making my life a bit more adventurous.

About the author

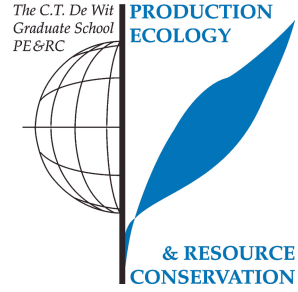
Stephan van der Westhuizen was born on 29 June 1989 in Ermelo in the province of Mpumalanga in South Africa. His family soon moved to Bredasdorp near the most southern tip of Africa where he grew up among farmers, fighter pilots and beach rats. Stephan went to Stellenbosch University where he studied Statistics and Economics and obtained his Bachelor of Commerce in 2010. He also obtained his master's degree in Statistics at Stellenbosch. After several years of working in the credit industry, Stephan was appointed in 2016 as a biometry lecturer at Stellenbosch University in the Faculty of Agrisciences. This was his first exposure to agricultural and environmental research. In 2019, he started focusing on applications in soil science which then led him to enroll for a PhD at Wageningen University where he specialised in machine learning in digital soil mapping. While working on his PhD, Stephan was also a full-time lecturer, presenting undergraduate and postgraduate courses in biometry, and consulting on agricultural research projects. In 2023, Stephan accepted a position in the Department of Statistics and Actuarial Science at Stellenbosch University where he presents courses on applied statistics and data science. After obtaining his PhD in 2024, Stephan continues to work, amongst others, on digital soil mapping research projects.

Peer-reviewed journal publications from this thesis

- 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**, p. e100572. doi: `https://doi.org/10.1016/j.spasta.2021.100572`.
- van der Westhuizen, S., Heuvelink, G.B.M., Hofmeyr, D.P. (2023). Multivariate random forest for digital soil mapping. *Geoderma*, **431**, p. e116365. doi: `https://doi.org/10.1016/j.geoderma.2023.116365`.

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 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 (4.5 ECTS)

- Enhancement of the use of machine learning in digital soil mapping

Post-graduate courses (3 ECTS)

- Introduction to GIS; Centre for Geographical Analysis, Stellenbosch University (2019)
- Uncertainty propagation in spatial and environmental modelling; PE&RC (2022)

Deficiency, refresh, brush-up courses (0.3 ECTS)

- Introduction to Latex; PE&RC (2020)

Invited review of journal manuscript (4 ECTS)

- Journal of Geophysical Research (2022)
- European Journal of Soil Science(2023)
- Geoderma (2023)
- Science of the Total Environment (2023)

Competence, skills and career-oriented activities (3.3 ECTS)

- Scientific communication; African Doctoral Academy, Stellenbosch University (2020)

- Postgraduate supervision; Division for Research Development, Stellenbosch University (2021)
- Reviewing a scientific manuscript; PE&RC (2022)

Scientific integrity/ethics in science activities (0.3 ECTS)

- Good research practice: research ethics and beyond; African Doctoral Academy, Stellenbosch University (2020)

PE&RC Annual meetings, seminars and PE&RC weekend/retreat (1.2 ECTS)

- PE&RC Weekend for first years (2019)
- PE&RC Day Nijmegen (2022)

Discussion groups / local seminars / or scientific meetings (18 ECTS)

- South African Statistical Association (2019-2023)
- Spatial statistics interest group in South Africa (2019-2023)
- Multivariate data analysis group (2019-2023)

International symposia, workshops and conferences (7.5 ECTS)

- Spatial statistics; Sitges, Spain (2019)
- International federation for classification societies; Porto, Portugal (2022)
- World congress of soil science; Glasgow, UK (2022)

Societally relevant exposure (0.9 ECTS)

- Present one day workshop in R to environmental scientists (2022, 2023)
- Co-hosted geostatistics workshop (2023)

Lecturing / supervision of practicals / tutorials (3.6 ECTS)

- Environmental data collection and analysis (2022)

BSc/MSc thesis supervision (3 ECTS)

- Digital soil mapping with uncertain (2020)

This research was supported by

1. The National Research Foundation of South Africa (Grant number: 129856).
2. Mitigate+: Research for Low Emissions Food Systems and the Global Research Alliance on Agricultural Greenhouse Gases (GRA) through their CLIFF-GRADS programme. Funding for Mitigate+ comes from the CGIAR Trust Fund.

Cover design by Carmen Platt, Unearthing Words

Printed by Proefschriftmaken.nl