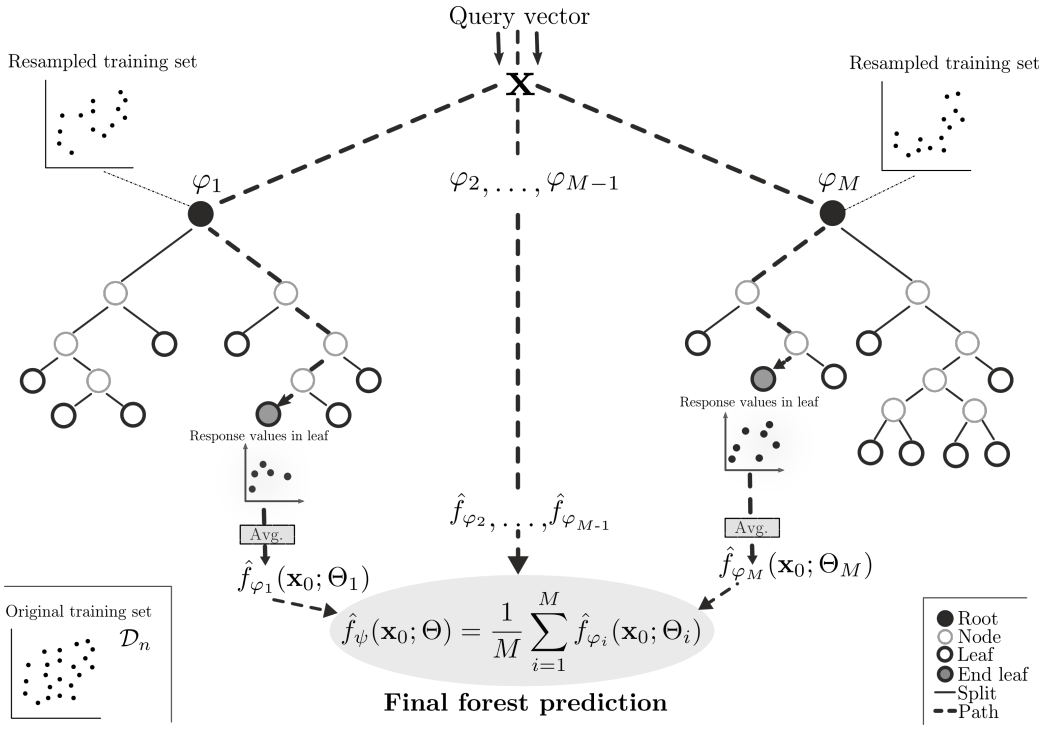
Geo-information Science and Remote Sensing

Thesis Report GIRS-2018-14

QUANTIFYING UNCERTAINTY OF RANDOM FOREST PREDICTIONS

A Digital Soil Mapping Case Study

Kees Baake



April, 2018

WAGENINGEN
UNIVERSITY & RESEARCH

Quantifying Uncertainty of Random Forest Predictions

A Digital Soil Mapping Case Study

Kees Baake

Registration number 89 03 08 022 110

Supervisors:

Gerard Heuvelink
Sytze de Bruin

A thesis submitted in partial fulfillment of the degree of Master of Science
at Wageningen University and Research Centre,
The Netherlands.

April, 2018
Wageningen, The Netherlands

Thesis code number: GRS-80436
Thesis Report: GIRS-2018-14
Wageningen University and Research Centre
Laboratory of Geo-Information Science and Remote Sensing

ABSTRACT

Random Forest is a type of machine learning algorithms that are known for making predictions with low errors. Random Forest (RF) has successfully been applied in a soil modelling context. Due to their black-box nature Random Forest models are difficult to interpret and the inherent modeling and input uncertainties are difficult to quantify. Within the last ten years statisticians discovered desirable properties of Random Forest that make the models more transparent, especially with regards to the quantification of prediction uncertainties. A literature review was done on the mathematical foundations of four uncertainty quantification techniques for Random forest predictions after which they underwent a qualitative assessment on the main criteria: scalability, usability and statistical rigor. Two techniques, Quantile Regression Forest (QRF) and Regression Kriging (RK) were chosen as most viable candidates mainly because they quantified the complete uncertainty, meaning they can be used for creating prediction intervals (PI). The other major reason being that both are widely available and easily implementable.

QRF and RK were both evaluated as (1) an overall assessment in the form of accuracy plots with derived summary statistics; (2) as a local assessment on spatial dispersal of outliers that consistently fall outside the PI; and (3) in terms of computation time scalability. This was done by averaging over 100 runs of 10-fold cross validation. A case study in eastern Australia (Edgeroi), characterized by a sampling design mix of systematic and clustered sampling, was selected for evaluation of the Random Forest prediction interval estimation by QRF and RK. After preprocessing steps pH and soil organic carbon content (SOC) were modeled with both a 14 covariate model (RF14) and 4 covariate model (RF4) with covariates falling within the soil forming factor categories of location, relief, vegetation, climate and parent material. For the overall uncertainty assessment multiple PIs were validated and for the local assessment, only the 0.9 probability level was investigated chosen because it is the predicament of the GobalSoilMap consortium.

In the overall uncertainty assessment both RK and QRF performed well on both with 4- and 14 covariate models with low absolute deviations ($<5\%$) from the accuracy plot $1:1$ (observed vs expected proportion in PI). QRF was often too optimistic: most of its observed proportion was below the $1:1$ line (>0.90). RK was too pessimistic and was mostly above the $1:1$ line (>0.90). No major differences in uncertainty quantification performance were observed between the modeling of pH and SOC although the predictive R^2 of the underlying Random Forest model varied largely between the two soil response variables (e.g. 0.41 vs 0.08 for RF4). However, the local uncertainty assessment did note substantial differences between pH and SOC for QRF and RK: pH seemed to be more clustered in regions of spatial outliers (RK) instead of being more dispersed (QRF). SOC did not find any major differences in spatial outlier dispersal between RK and QRF. In terms of scalability QRF doubled in computation time when the number of points to predict increased 10 fold. In general, the width maps of the 0.9-PI showed more detail and clear boundaries for QRF. Indicating that conditioned geographical data has a large effect on the magnitude of uncertainty. Other literature on QRF in soil science context also showed promising results under a more sparse sampling design. Thus, there are strong clues that QRF can be used as a new, flexible tool in the field of uncertainty modeling in spatial context.

ACKNOWLEDGEMENTS

On forehand I knew this thesis was going to be quite a challenge as my knowledge on predictive soil modeling and machine learning was very limited. The completion of this thesis has therefore been a huge conquest that I was not able to complete without the help of several decisive people. First, Sytze de Bruin who helped me break through some stalemate moments during the thesis and helped me differentiate between what is important and what is not. Second, Gerard Heuvelink who has been very helpful in making practical choices and helped to explain complicated topics with such ease that made almost everything crystal clear. Third, Tom Hengl who helped me choose a dataset and for giving me technical advice, especially during the proposal.

Much gratitude also goes out to my family, my dad for helping with designing some of the diagrams and my mother for the love and support. My wife, Josien Boetje has been a tremendous support throughout the whole thesis and helped me to follow through emotionally and also analytically where she could. Thank you all very much!

Kees Baake
Monday 16th April, 2018

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

GLOSSARY

ccdf Conditional cumulative distribution function.

DEM Digital Elevation Model.

DSM Digital Soil Mapping.

GSM GlobalSoilMap consortium.

PI Probability Interval.

PSM Predictive Soil Mapping.

QRF Quantile Regression Forest.

REML Restricted Maximum Likelihood.

RF Random Forest.

RK Regression Kriging.

SFM State Factor Model.

1 INTRODUCTION

The mapping of soils has historically relied on soil surveys, which consist of collecting soil samples at several locations to draw a soil map in discrete soil mapping units aided by visual photo-interpretation or terrain maps (Rowell, 2014). Much of the available soil polygon maps today are digitized from these legacy soil maps (Malone et al., 2016). Traditional soil surveys are centred around the State Factor Model (SFM) by Jenny (1941) which postulates that soil formation is dependent on parent material, climate, organisms, relief, and time (Hudson, 1992). The surveyors use their implicit knowledge on these soil forming factors to draw borders on a terrain map (Moore et al., 1993). Not only are traditional soil surveys time-consuming and expensive to conduct (Hartemink et al., 2010), they also suffer from three major scientific drawbacks. Firstly, conventional soil surveys cannot capture all relevant soil information as many of the soil forming processes are still not fully understood (Scull et al., 2003). Secondly, as the variability in soil properties over the landscape can be high, it becomes difficult to construct a complete representation through a traditional soil survey because there are only a limited number of total soil samples and therefore certain soil characteristics might be missed (Campbell & Edmonds, 1984; R. Wright & Wilson, 1979). Thirdly, a soil survey is difficult to reproduce since the expert's assumptions are implicit and tend to focus mainly on qualitative assessments of soil properties (Beckett & Burrough, 1971; Dijkerman, 1974). The soil science community therefore set clear goals to communicate their uncertainties to soil map users, but according to sources such as Wilder (1985) the community failed to adhere to these goals in practice.

Geostatistics is a branch of statistics brought into the field of soil science by Burgess and Webster (1980) specifically to tackle these issues by using a more objective linear interpolation method called Kriging using the auto-correlation of a soil property over distance. The technique comes with a major advantage: a measure of uncertainty is present for each newly interpolated location. Throughout the years, more techniques and information got included in geostatistical models to increase their performance. Especially the regression kriging variant offered a highly flexible approach to modeling by estimating a soil property from a combination of one or more soil covariates and kriging the regression residuals (Zaouche et al., 2017). Including such covariates into a statistical model implies using more information to explain soil forming factors. Therefore, instead of a mental soil forming model, a quantitative model was proposed by McBratney et al. (2003) to capture and summarize the different categories of soil covariates: Scorpan, standing for **S**oil property observations, **c**limate, **o**rganisms, **r**elief; **p**arent material, **a**ge and locatio**n**, respectively. Note that the scorpan model does not rely solely on regression kriging; kriging is just the location component of the scorpan model. Hence, a more general approach of soil mapping through covariates developed in what is named: Predictive Soil Mapping (PSM), or Digital Soil Mapping (DSM).

Scull et al. (2003) define PSM as the "development of a numerical or statistical model of the relationship among environmental variables and soil properties applied to a geographic data base to create a predictive map". Currently, spatial exhaustive soil forming data can be derived from the broad availability of optical, radar and lidar sensed data, that provide relatively cheap and accurate spatial information on

the value of many different soil covariates (Minasny & McBratney, 2016). For example, parent material (e.g. mineral composition) can be assessed through optical sensors (Solomon & Rock, 1985); Synthetic Aperture Radar (SAR) can determine soil moisture content, salinity or surface roughness (Dubois et al., 1995; Wagner et al., 2007). A Digital Elevation Model (DEM) can be computed from Lidar, Radar and legacy land surveys (Mulder et al., 2011) and serves the basis for many DEM derivatives (e.g. curvature and slope). Thus, there is no (direct) need to interpolate point location data as a full spatial grid can be constructed straight from these remotely sensed products.

Simple statistical techniques, however, do not recognize all patterns of information present within these spatially exhaustive covariates. Parts of the SFM are potentially non-linear and soil formation can be highly sensitive to small variations of soil factors (Addiscott & Tuck, 2001; Heuvelink & Webster, 2001; Webster, 2000). Modern machine learning techniques can improve the modeling of the non-linear relationships as they enable computers to recognize patterns in data without the need for a scientist to explicate relationships (Henderson et al., 2005; Minasny et al., 2008). Moreover, soil scientists have already demonstrated that maps produced by machine learning are, in general, more accurate than conventional soil maps (Lorenzetti et al., 2015; Bazaglia et al., 2013). An illustration on the potential of the implementation of these techniques is the recent application that uses PSM with machine learning techniques is ISRIC's SoilGrids (250m) platform that maps different soil characteristics of the whole world by using many different soil forming covariates as inputs (Hengl, Mendes de Jesus, et al., 2017).

Although PSM received much attention over the past years, there remain a couple of important conditions that must be satisfied to guarantee practical applicability. Uncertainty quantification is one of these important conditions (Minasny & McBratney, 2016) and very interesting one for research applications as it exposes information of the underlying soil forming mechanisms that can be used to measure the effect of sampling or modeling improvements. Quantifying these uncertainties does not only provide useful information for users but especially for scientists, engineers and policy makers to reduce risks associated with climate change, natural and man-made hazard prevention, food quantity, health and security that can use this to analyze and use this for risk assessment for decision makers (Hartemink et al., 2010). Hence, an essential questions is on how information on the inherent uncertainty of Machine Learning predictions can be distilled from the technique itself instead of relying on an additional spatial model as with regression kriging; whether this information is reliable in practice and if it can provide additional information that is currently unavailable to traditional techniques.

1.1 Problem statement

GlobalSoilMap (GSM), a global consortium to map the most important functional soil properties on a global scale through PSM specifically demands for the quantification of the uncertainty to support optimal decision making (Arrouays et al., 2014). The GSM requirement states that the prediction interval of a point should encompass the true value 9 out of 10 times. Machine learning algorithms often show high prediction accuracy, but current applications frequently omit to address the uncertainty of these predictions as the focus is mainly put on overall performance (e.g. Hengl, Mendes de Jesus, et al. (2017); Nussbaum et al. (2017); Were et al. (2015)). Soil modeling with machine learning should aim to include the uncertainty quantification component of their predictions by as this can lead to better decisions making. Furthermore, the quantified uncertainty could possibly lead to improvements in sampling designs, better choice of covariates; detection of uncertainty propagation or support the development of new DSM approaches as it can quantify its effect directly.

Frequently used machine learning algorithms to predict soil properties are Classification and Regression

Trees (CARTs) and their derived ensemble algorithms such as Random Forests (RF) (Malone et al., 2016). Regression trees can use a large number of predictors to train a model/tree with good results (James et al., 2013) and this fits the requirement in DSM of a multitude of soil forming factors well. Regression trees have multiple advantages (Kuhn & Johnson, 2013): (1) Easily implementable; (2) Handle many different predictor distributions; (3) No need for explicit relationship descriptions between the predictors and the response and (4) Implicit feature selection. Although regression trees are very dependent on their training dataset and are associated with a high variance in their predictions (effects highly differ when built on a different training set), ensemble methods such as Random Forest subsample the data, train multiple trees and aggregate all these individual tree results, which largely reduces the prediction error on a test set compared to individual tree estimates (Kuhn & Johnson, 2013). Therefore, the prediction accuracy as an overall measure becomes much higher than for an individual tree. Recent research established that Random Forest performs very well for soil property predictions in comparison with other techniques (see Lorenzetti et al., 2015; Nussbaum et al., 2017).

Although ensemble RF algorithms show very good predictive power, often with very low calibration and validation errors, it is not straightforward to quantify how uncertain a prediction at an unmeasured location is. Validation statistics do quantify model uncertainty as a whole, but they are merely a summary statistic of the overall model performance and have no notion of spatial explicitness. What is needed is an ability to predict new points with an estimate of the error margins at a requested probability level such the GSM stipulates. There are several approaches available that aim to quantify these prediction uncertainties per point, yet implementation of these methods with spatially explicit data for predicting soil properties is currently limited; the Vaysse and Lagacherie study (2017) is currently the only published study that tested uncertainty quantification of Random Forest under a sparse sampling scheme. Furthermore, a practical comparison of computation times and performance of Random Forest uncertainty quantification methods under different restrictions can clear up concerns and confusion on their applicability so researchers can make a better choice of research methods.

1.2 Research objective

The main objective of this research is to apply and evaluate methods for spatially explicit uncertainty quantification of Random Forest predictions on continuous soil properties.
This objective will be reached through exploring four research questions:

I. Which methods are available for quantifying uncertainty of random forest (RF) predictions and what is their mathematical foundation?

II. What are the most viable methods (a priori) for quantifying uncertainty based on the criteria scalability, usability and rigor?

III. When applied to a digitally soil mapped case study, are the results of the uncertainty quantification methods consistent with those obtained through validation test sets of set-aside samples?

IV. What is the scalability of the uncertainty quantification methods in terms of number of covariates and the number of prediction points.

The structure of this thesis is as follows: Chapter 2 describes the basis of the Random Forest algorithm in detail. Chapter 3 provides an introduction to what is exactly meant by uncertainty quantification and discusses the mathematical foundations of the techniques. This chapter will end with a qualitative assessment on which uncertainty quantification techniques are suitable in a spatial context. The next chapter (Chapter 4) outlines the materials, methods and strategy on how to evaluate the performance the

selected RF uncertainty quantification techniques in a spatial context as an overall measure and also on a local level, including a benchmark on computation times. Chapter 5 describes the chosen case study site in detail, the required preprocessing steps and ends with a presentation on the evaluation results. The thesis then concludes with a general discussion of all the results, a conclusion and, ultimately, some recommendations (Chapters 6, 7 & 8).

2 RANDOM FOREST

The main goal of this chapter is to provide a background on the Random Forest algorithm as it is needed to acquire a better understanding of why RF works and build the foundations of the RF uncertainty quantification methods in Chapter 3. Furthermore, this background helps to understand why the uncertainty quantification is not straightforward. This chapter first provides the framework of regression wherein the bias-variance trade-off is discussed needed for the understanding of the rationale behind Random Forest in general. Section 2.1 gives a formal definition of non-parametric regression and introduces the used symbols. Next, the basic principles of regression trees are explained in Section 2.2 with a focus on the splitting algorithm. The next Section 2.3 then introduces a simpler version of the RF algorithm called bagging and the chapter finishes with the complete RF algorithm.

2.1 Regression

Statistical regression is a set of techniques that allow the determination of how a dependent variable, Y, is affected by one or more independent variables, denoted as $\mathbf{X}$ (Fox, 1997). Often observations on $\mathbf{X}$ are easier to obtain than observations of Y and therefore the main idea is to use $\mathbf{X}$ to predict Y through a statistical model. This is why $\mathbf{X}$ are often called the predictors and Y the matched response or target variable. There will always be some inherent discrepancies (denoted as ϵ) between the dependent and independent variables so a statistical model needs to incorporate an error term (Fox, 1997). In other words, the goal is to construct a model f that maps $\mathbf{X} \to Y$ and leaves room for a random error, or equivalently: $Y = f(\mathbf{X}) + \epsilon$.

In practice, an estimate of the true regression function needs to be estimated as data on Y is much more limited than data on $\mathbf{X}$. This estimate function is denoted as $\hat{f}$. A prediction can be made in the form $\hat{Y} = \hat{f}(\mathbf{X})$, where $\hat{Y}$ represents the prediction at $\mathbf{X}$. To estimate the best possible model in regression, the expected squared error term is minimized − note that squared error instead of the absolute error is chosen because it has more convenient mathematical properties. Following Friedman et al. (2001), Equation 1 below gives the squared prediction error that underlies the regression minimization problem:

$$E\left[(Y - \hat{Y})^2\right] = E\left[f(\mathbf{X}) + \epsilon - \hat{f}(\mathbf{X})\right]^2 \tag{1}$$

After simplification this equation can be rewritten as:

$$E\left[(Y - \hat{Y})^2\right] = \underbrace{\left(f(\mathbf{X}) - \hat{f}(\mathbf{X})\right)^2}_{\text{Reducible error}} + \underbrace{\text{Var}[\epsilon]}_{\text{Irreducible error}} \tag{2}$$

After this decomposition it becomes visible that regression can only focus on reducing the left part of the decomposed error (the reducible error) as the other part was the inherent noise introduced at the beginning of the model and hence cannot be minimized (i.e. irreducible). Friedman et al. (2001) further decompose the reducible error into an additional bias and variance term:

$$E\left[(Y-\hat{Y})^2\right] = \underbrace{\text{Bias}[\hat{f}(\mathbf{X})]^2 + \text{Var}[\hat{f}(\mathbf{X})]}_{\text{Reducible error}} + \underbrace{\text{Var}[\epsilon]}_{\text{Irreducible error}} \tag{3}$$

Friedman et al. (2001) describe the bias of the estimated regression function $\hat{f}$ as the expression of error that is introduced when approximating a complicated real problem by a much simpler model. The variance of the estimated regression function is the amount by which $\hat{f}$ would change if it had a different data set available on which it was modeled. In the rest of this document it helps to keep these terms in mind as there is often a trade-off between the two. Furthermore, the understanding on how this error is decomposed becomes essential in the chapter on uncertainty quantificaiton later on (Chapter 3).

Parametric regression makes some assumptions on the function form of f and the regression problem therefore gets simplified to the optimal estimation of these parameters. Herein also lies the disadvantage of parametric regression, the assumptions might not always hold and the functional form used can be very different from the true f (Friedman et al., 2001). Hence, most parametric regression methods often have a high bias as the complexity of the models is often low. In the case that the methods are more complex this often leads to a high variance as the relationships are overfitted on a single training set. In contrast, non-parametric methods do not make these assumptions on the function form of f and to compensate they often need a larger data set to correctly model f from the available data (Friedman et al., 2001). Despite this need of a larger training set, non-parametric methods can achieve much more flexibility and allow for complex patterns to be modeled without necessarily leading to severe overfitting as is often the case with the parametric methods (Kuhn & Johnson, 2013). Regression trees and Random Forests are examples of non-parametric regression (Biau & Scornet, 2016).

Now, the general symbolic framework can be provided for the regression trees and its derived algorithms. Let c be the total number of predictors and let the complete predictor space be represented by $\mathcal{X} \subset \mathbb{R}^c$ such that every dimension $1, 2, \ldots c$ represents a distinct predictor. Now suppose that for $i \in \{1, 2, \ldots, n\}$, $\mathbf{X}_i \in \mathcal{X}$ represents an input predictor random vector that matches a random response $Y_i \in \mathbb{R}$. Then, with a total of n such pairs, a training sample $\mathcal{D}_n$ of independent random variables can be formed as $\mathcal{D}_n = ((\mathbf{X}_1, Y_1),\ \ldots (\mathbf{X}_n, Y_n))$. The goal is to use $\mathcal{D}_n$ to estimate the regression function f that maps $\mathcal{X} \to \mathbb{R}$ in such a way that as the number of observation response pairs in $\mathcal{D}_n$ approaches infinity, the squared error between the estimated regression function $\hat{f}$ and the observed response values approaches 0.

2.2 Regression trees

In essence, a regression tree is a computational model that is constructed by binary recursive partitioning (Louppe, 2014). Binary recursive partitioning is a method that repeatedly splits training data into two partitions at each step until a stopping criterion is met. At first, the complete predictor part of the training set $\mathcal{X}$ is grouped into a single partition, called the root node. The algorithm then evaluates binary partitions of this root using every possible partition on $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ in $\mathcal{X}$ such that no training data point can be in two partitions at the same time. The binary split that displays the minimal value of a special splitting metric (see more on splitting below) is selected (Louppe, 2014). The newly created nodes then undergo the same procedure (with the training points still available within that node) until each node hits a user-defined minimum node size. The final nodes are also called leaf nodes, a singular leaf will be denoted by ℓ, and define the final partitioning of the training set. The average of the response values present in each leaf node determine the final prediction. A more generalized prediction formula can be made by considering the whole training set by using indexed weights to calculate this average and will now be defined.

Suppose that an unknown query point $\mathbf{x}_0$ is dropped down the trained tree and falls into a leaf with index j, such that $\mathcal{X}_\ell^j$ represents the partition of all predictor observations in this leaf. Then, a weight can be calculated for *all* training set response observations. This weight is defined as 1 divided by the number of training points in this leaf and 0 if the response does not fall into this leaf. Now, a final prediction with the tree estimated regression function $\hat{f}$ can be given by the following equation by going over all response values in the training set (Eq 4):

$$\hat{f}(\mathbf{x}_0) = \frac{1}{n}\sum_{i=1}^{n} w_i(\mathbf{x}_0) \cdot y_i \tag{4}$$

Where the weights are defined as:

$$w_i(\mathbf{x}_0) = \frac{\mathbb{1}_{\{\mathbf{x}_i \in \mathcal{X}_\ell\}}}{\#\{k : \mathbf{x}_k \in \mathcal{X}_\ell\}} \tag{5}$$

In this equation for the weights (Eq 5), $\mathbb{1}$ denotes the indicator function that returns a 1 when $\mathbf{x}_i$ is in the pool of input predictor vectors at the leaf and 0 if not present as displayed in the subscript. The denominator gives the count of all input training vectors present in the leaf.

2.2.1 Splitting nodes

Growing a regression tree mainly depends on how well the nodes are split. Finding the global optimal split itself is a very computation heavy task (Friedman et al., 2001), therefore a greedy recursive algorithm is used to find local optimal binary splits. This greediness is defined as only looking at the current split, not consecutive splits. But what exactly defines such a binary split?

Definition *A binary split s of node t is a set of two non-empty subsets of $\mathcal{X}_t \subset \mathcal{X}$ at t such that every element $\mathbf{x}_j \in \mathcal{X}_t$ cannot be in both of these two subsets $(\mathcal{X}_{t_L}, \mathcal{X}_{t_R})$ simultaneously. For regression this translates to a threshold value ($H \in \mathbb{R}$) at a specific dimension (denoted by d) of $\{1, 2, \dots c\}$ at $\mathbf{x}_j$ such that $t_L = \{\mathbf{x}_l : \mathbf{x}_l^{\mathrm{d}} \leq H\}$ and $t_R = \{\mathbf{x}_r : \mathbf{x}_r^{(\mathrm{d})} > H\}$.*

The objective is to find the local optimal binary split. For regression the split to be made is almost always based on an "impurity" decrease criteria. There are other, similar criteria, but in regression context these are not often used, therefore they will not be described here. For a more detailed overview of splitting criteria see Shih (1999) for example. For regression the impurity function i in Equation 7 is the local estimate of the squared error loss for all training pairs still present at node t. This corresponds to the within node variance and therefore an optimal split is said to minimize the variance in the child nodes (Louppe, 2014). Let t represent a potential node to be evaluated for purity, then the impurity function $i(t)$ is given as:

$$i(t) = \frac{1}{n_t}\sum_{j=1}^{n_t}(y_j - \bar{y}_t)^2 \mathbb{1}_{\{\mathbf{x}_j \in \mathcal{X}_t\}} \tag{6}$$

In the equation above $\mathcal{X}_t$ is the subset of input predictor vectors at node t, and n_t is defined as the total number of such vectors. $\bar{y}_t$ is the average of all y in node t together. The indicator function $\mathbb{1}$ determines whether the response value will be considered or not.

The next step is the definition of the impurity decrease, which is computed by comparing the impurity of the parent node t to the child nodes t_L (left) and t_R (right). The impurity decrease is calculated by substracting the proportion of training samples multiplied by the impurity in the splitted child nodes t_L and t_R from the original impurity in node t (see Equation 6). Let n_{t_L}/n_t and n_{t_R}/n_t be the proportions of the number of training points in the left and right child nodes compared to the original n_t training points at node t, then the impurity decrease is written as:

$$\Delta i(s,t) = i(t) - \frac{n_{t_L}}{n_t} i(t_L) - \frac{n_{t_R}}{n_t} i(t_R) \tag{7}$$

The complete algorithm for the splitting procedure is given in Algorithm 1:

Algorithm 1. *Find the best split*

Input : Node to be evaluated: t, Predictor space of c_t dimensions and n_t number of points: $\mathcal{X}_t$
Function *FindBestSplit*$(t, \mathcal{X}_t)$
  Set the initial impurity decrease $\Delta = -10^{99}$;
  for $d = 1, \dots c_t$ **do**
    for $j = 1, \dots, n_t$ **do**
      Set the splitting threshold H equal to the value $\mathbf{x}_j^{(\mathrm{d})}$;
      Split t into child node $t_L = \{\mathbf{x}_l : \mathbf{x}_l^{(\mathrm{d})} \leq H\}$ and $t_R = \{\mathbf{x}_r : \mathbf{x}_r^{(\mathrm{d})} > H\}$;
      Compute the impurity decrease at $s_j^{(\mathrm{d})}$ using current partition:
       $\Delta i(s_j^{(\mathrm{d})}, t) = i(t) - \frac{n_{t_L}}{n_t} i(t_L) - \frac{n_{t_R}}{n_t} i(t_R)$
       where the impurity function i is defined according to equation 6;
      if $\Delta i(s_j^{(\mathrm{d})}, t) > \Delta$ **then**
        $\Delta = \Delta i(s_j^{(\mathrm{d})}, t)$;
        $s^* = s_j^{(\mathrm{d})}$
      end
    end
  end
return s^*

2.3 Bagging

Regression trees are very dependent on their training dataset and small changes in the training dataset can result in large deviations when comparing the new predictions with the original predictions (Kuhn & Johnson, 2013). Hence, regression trees are said to show high variance of the estimated regression function. Now suppose that to construct new training datasets, training sets are simulated by subsampling the original training set. Then, the average calculated over all newly constructed trees could decrease the error related to function estimation variance as the prediction becomes less dependent on the original training set. This is because the variance of the initially estimated regression function can be reduced as multiple trees have been grown on "different" training datasets. This technique is what Breiman (1996) introduced as bagging.

Bagging, short for bootstrap aggregating, starts by drawing a set of $a_b \leq n$ points randomly from the original training data, with replacement. This is repeated a total of B consecutive times. Bagging continues to train trees on each of these $1, \dots, B$ subsamples and proceeds to aggregate all these individual tree results into one final prediction. Let B be the total number of trees grown, $\{\varphi_b,\ b = 1, \dots, B\}$ represent all individual trees and *bag* be the collection of all these trees. Following notation of James et al. (2013) the bagging prediction at a query vector $\mathbf{x}_0$ is then computed by averaging over all B tree predictions:

$$\hat{f}_{bag}(\mathbf{x}_0) = \frac{1}{B} \sum_{b=1}^{B} \hat{f}_{\varphi_b}(\mathbf{x}_0) \tag{8}$$

As mentioned, Bagging largely reduces the error related to prediction variation compared to individual tree estimates (Kuhn & Johnson, 2013). Therefore, the prediction accuracy on a new query point becomes much higher as ensemble than for an individual tree. The drawback is that each bagged tree draws from an identical multivariate distribution, hence the expected value of a prediction at point $\mathbf{x}_0$ of

the aggregate of B such trees is equal to the expected value of an individual tree (Friedman et al., 2001). Furthermore, the fact that all predictors are considered in the splitting means that some predictors might dominate the splitting criterion and might oversimplify the amount of partitions of the training set where much more partitions could have been made.

2.4 Random Forest

Random Forest resembles the bagged tree procedure closely. The rationale is that by introducing an extra random perturbation during the splitting of a tree, predictions of individual trees can be decorrelated even more from each other than by the mere subsampling of the training set. Thus, Random Forest (Algorithm 2) not only draws a total of say M subsamples of a supposed size a_φ from the orginal data before training a new tree, RF also randomizes the partitioning procedure by only considering a dimension reduced subset, often called $\mathcal{M}_{try}$ in literature, of the original predictor space $\mathcal{X}$ per split (Breiman, 2001). Let the cardinality of $\mathcal{M}_{try}$ be represented by `mtry` ($\texttt{mtry} = |\mathcal{M}_{try}|$), then the dimensions of all the training input vectors decrease from c to `mtry`. The random variable Θ_i determines for tree i (φ_i) both the value of a_φ and which predictors get included in $\mathcal{M}_{try}$. The growing of the trees can be done in parallel as they are independent, making Random Forest in itself a scalable solution.

An overview of the complete Random Forest prediction workflow for a new prediction (at query point $\mathbf{x}_0$) is summarized in Figure 2.1: the query vector $\mathbf{x}_0$ is dropped down all trees that were trained on a resampled subset and ends up in the final leaf. The response values matching the input training vectors in this leaf are averaged, giving the individual tree prediction. Once all individual tree predictions are calculated, they are averaged to give the final random forest prediction.

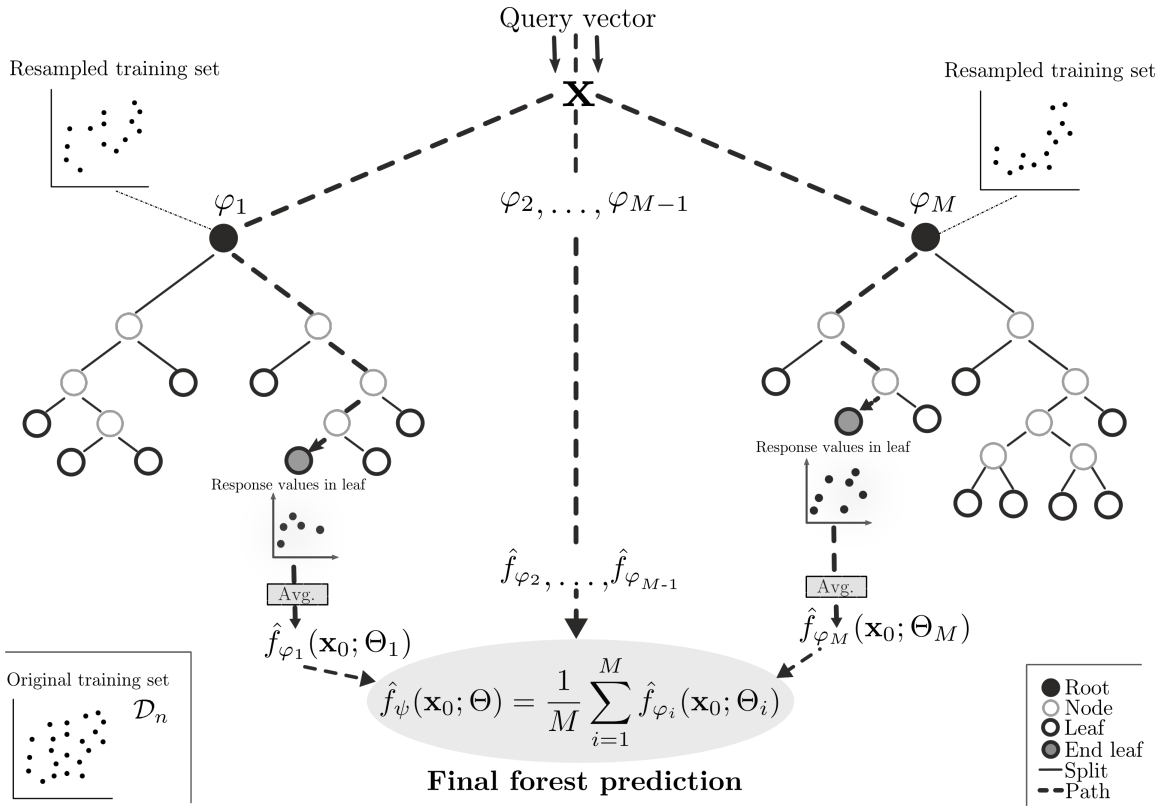


Figure 2.1. Diagram of a single Random Forest prediction.

2.4.1 Parameters

RF is controlled by only three parameters which make it easily implementable (Scornet, 2015). The first parameter is the minimum size of a node, called `nodesize`, that is used to determine what the terminal nodes (leaves) are. Second is the parameter that determines the total number of trees grown, defined as M. This forest size parameter is often set to a default of 500; a larger number will lead to a higher accuracy that asymptotically decreases after 1000 trees (Biau & Scornet, 2016). Furthermore, an increase in the number of trees in RF also leads to a linear increase computational cost (Biau & Scornet, 2016). The third parameter is the number of predictors to randomly consider per split, the parameter is named `mtry` and it is simply the cardinality of $\mathcal{M}_{try}$, equivalently: `mtry` $= |\mathcal{M}_{try}|$. In practice the size of the `mtry` is often either the number of predictors or the square root of the total number of predictors/covariates (M. Wright & Ziegler, 2015).

All steps needed for the complete algorithm for training a Random Forest and making a prediction at query vector $\mathbf{x}_0$ can now be described as follows (Algorithm 2).

Algorithm 2. *Random forest prediction (adapted from Biau & Scornet, 2016)*

Input : Training set: $\mathcal{D}_n$, total number of trees: M, number of predictors to choose after splitting: `mtry`, threshold below which cell is not split: `nodesize`, an independent random vector collection for each tree: $\Theta : \{\Theta_i : i \in 1, \ldots, M\}$, query vector $\mathbf{x}_0$

Output : Prediction of $y = f_\psi(\mathbf{x}_0)$

for *tree* φ_i, $i = 1, \ldots, M$ **do**
  Select a_φ points with (or without) replacement uniformly in $\mathcal{D}_n$ by consulting Θ_i and only use these points in the growing process of the current tree; Set a new ordered list $\mathcal{L}_{init}$ equal to the ordered root of the tree ($\mathcal{X}$) ;
  Set a new ordered list $\mathcal{L}_{final} = \emptyset$;
  while $\mathcal{L}_{init} \neq \emptyset$ **do**
    Let t be the first element in line from $\mathcal{L}_{init}$;
    if *Number of points in* t *are less than* `nodesize` *or if all remaining training points in* t *are equal* **then**
      Remove t from $\mathcal{L}_{init}$;
      Insert t to $\mathcal{L}_{final}$;
    else
      Select `mtry` times uniformly, without replacement, a predictor dimension ($\{1, \ldots, c\}$) by consulting Θ_i to construct $\mathcal{M}_{try} \subset \mathbb{R}^{mtry}$;
      Split t according to the FindBestSplit function (algorithm 1) with arguments $(t, \mathcal{M}_{try})$ and let t_L and t_R be the resulting cells.;
      Remove t from $\mathcal{L}_{init}$;
      Insert t_L and t_R into $\mathcal{L}_{init}$;
    end
  end
  Compute the predicted value $\hat{y} = f_\varphi(\mathbf{x}_0; \Theta_i)$ of the individual tree given $\mathbf{x}_0$ by setting $\hat{y} = y_\ell$ where ℓ corresponds to the leaf that $\mathbf{x}_0$ falls in as delineated by $\mathcal{L}_{final}$.
end
Compute the random forest estimate $f_\psi(\mathbf{x}_0)$ by aggregating the result of the individual trees.

The RF prediction is then given by a similar function as Equation 8 for a new prediction at query vector $\mathbf{x}_0$:

$$f_\psi(\mathbf{x}_0; \Theta) = \frac{1}{M} \sum_{i=1}^{M} f_{\varphi_i}(\mathbf{x}_0; \Theta_i) \tag{9}$$

3 UNCERTAINTY QUANTIFICATION

No prediction is free from errors, as every model is a simplified representation of reality. The prediction error can be tracked down to uncertainty introduced in a model either as a result of input uncertainty or during incomplete construction of a model. Thus, the modelling process is very dependent on training data, not only because of its uncertainties but also because the data needs to be a representative sample of the underlying populations (James et al., 2013). Representative in this case means that it samples from the complete distribution of the population and that the sample is large enough. Suppose that an experiment is replicated with no access to previous training data then training a new model on this training dataset will yield a different model than the model trained with the original training dataset; this was the error related to variance of the estimated function form. Furthermore, wrong assumptions on the relations within the training data or on the distributions of the covariate or response populations can also lead to an increase in prediction errors; the error related to bias.

Uncertainty quantification can be an ambiguous term as it does not specify what part of the uncertainty is quantified. Remember that the beginning of the previous chapter 2 shortly described the relationship between regression and the prediction error. Then, the error was broken down into two major pieces:

$$E\left[(Y-\hat{Y})^2\right] = \underbrace{\text{Bias}[\hat{f}(\mathbf{X})]^2 + \text{Var}[\hat{f}(\mathbf{X})]}_{\text{Reducible error}} + \underbrace{\text{Var}[\epsilon]}_{\text{Irreducible error}} \tag{10}$$

Instead of minimizing the prediction error, the objective in uncertainty quantification is to quantify how large these expression of errors could be for a newly predicted unobserved point, i.e. what is their associated uncertainty. There are two possibilities on what the term uncertainty quantification therefore means. The first is that it can aim to quantify the reducible error for new predictions which is called a confidence interval. The second is that uncertainty quantification can mean that all parts of the error are quantified, which is called a prediction interval.

Machine learning techniques are often highly effective in keeping the reducible error at a minimum as they require no assumptions to be made on the function form making the the predictions and can focus on fitting a training set specific model. Therefore, statistical inference of population parameters is difficult. This is in sharp contrast to traditional parametric regression techniques (e.g. linear regression) that assume a predetermined function form that requires the residuals to be normallly distributed. Once a function form is clear, classical statistical theory can be used to infer population parameters. So if nothing is known about the function, how is the uncertainty quantified? This leads to the question of which techniques are currently available for quantifying uncertainty in Random Forest predictions and what is their practical viability?

To answer these questions a literature review was conducted with the aim to describe the mathematical background of the uncertainty quantification techniques. This chapter starts by first describing the used methodology (3.1), then proceeds to describe all techniques. Ultimately, all techniques were assessed (Section 3.10) based on mostly practical criteria and the part of the error they quantify and accuracy to

pick two techniques to apply in a soil modeling case study.

3.1 Technique assessment

Identification of the Random Forest uncertainty quantification methods started with a literature search and review. For a total period of 10 hours the keywords "*uncertainty quantification*", "*random forest*", "*probability interval*", "*confidence interval*", "*prediction interval*", "*quantiles*" with "*machine learning*" and "*Random Forest*" were queried on the three different scientific literature search engines: Scopus, Web of Science and Google Scholar. All articles published within the last 10 years that are cited at least 5 times were selected with relevant abstracts were sub-selected for further inspection. Then, papers that contained a mathematical theory for some kind of uncertainty quantification were selected. This resulted in a total of four different methods for uncertainty quantification of random forest predictions, which were reviewed in greater detail:

- ◇ **Quantile Regression Forest**s (Meinshausen, 2006)
  Quantile Regression Forests (QRF) saves the spread of the response variable in the node to compute weights for constructing an empirical ccdf where prediction intervals are derived from.
- ◇ **Jackknife and Infinitesimal Jackknife** (Wager et al., 2014)
  Conceptualizes RF predictions as statistic so that the standard error of Random Forest predictions can be assessed by evaluating the average variability between RF predictions built on the whole training set and the RF predictions with the jackknifed training sets (that exclude one observation pair iteratively).
- ◇ **U-Statistic-based random forest** (Mentch & Hooker, 2016)
  By training multitude of trees on strict subsample combinations of the training set and averaging their results, RF can be seen as a U-statistic which are proven to be asymptotically normal. This asymptotic normality enables the quantification of U-statistic variance parameters that are used to estimate the variance of a RF prediction.
- ◇ **Kriging on the regression residuals**
  Building a geostatistical model on the RF regression residuals by combining the regression component with the interpolated residual component as the complete underlying statistical model to calculate the uncertainty of the RF prediction.

Special emphasis was put on what uncertainty the techniques quantify. After the methods for uncertainty quantification were identified and reviewed, they have undergone an assessment based on the main criteria Scalability, Usability & Rigor. These were divided using sub-criteria for a more detailed assessment. An overview of these criteria with motivation and explanation is given in Table 3.1 below. The main interest during the initial practical assessment was on the implementation and usability. A score form, guided by a rubric that determines the qualities per score level, grades each sub-criterion to make the assessment quantifiable. Additional qualitative observations were also supplied to highlight some practical issues that might occur when implementing the techniques.

Table 3.1. Overview of the assessment criteria

	Subcriteria	Description	Motivation
Scalability	Computation time	Does the algorithm require computationally heavy tasks that increase processing time?	The lower the computation time, the more chances it can be applied on large scale projects.
		Is the relation between computation time and number of co-variates, samples or cells linear, quadratic, etc.?	
		Can the algorithm be parallelized?	
	Flexibility	Can the algorithm be used in combination with other methods and can it be adjusted for different research goals?	The combination with other algorithms or possibilities for tuning parameters can increase the potential as every research case can have specific conditions that need to be addressed.
Usability	Availability	Is the algorithm available in a programming distribution (especially R?)	The availability in a software distribution will highly reduce development time that can be used for analysis instead of programming.
		Is the software released under open-source licensing?	
	Extensive	Does the package come with options for parameters and validation and how well are these documented?	Time spent on getting acquainted with the package will be largely reduced if the usability scores high.
Rigor	Completeness	Does the technique quantify the complete prediction or does it provide information on just one uncertainty component?	If the complete distribution can be inferred than determining prediction interval boundaries becomes fast and easy.
	Accuracy	Is the technique mathematically consistent?	Accuracy needs to be in practical margins to be useful for further applications.
		How fast does it converge to consistent predictions?	

3.2 Quantile Regression Forest

Quantile regression forest estimates the CCDF by using an empirical CCDF. Therefore, it quantifies the complete error given a certain input vector as it includes a conditional variance estimate for Y by using the information within the leaves. Hence, Meinshausen (2016) technique can be used for making prediction intervals and not for confidence intervals because the empircal ccdf provides no information on the uncertainty of the fit of the Random Forest model itself.

3.3 Underlying mathematics

Random forest approximates the conditional mean $E(Y|X = \mathbf{x}_0)$ by averaging the observations of the response variable Y in the terminal nodes (i.e. leaves). In contrast to this procedure, Quantile Regression Forest (QRF) does not average out the response variable Y, but keeps the complete distribution of all observed response values of every leaf of each tree in the forest (Meinshausen, 2016). The reason is that QRF aims to estimate the conditional probability function F by using the distribution of the response in the leaves of the tree.

This section will now summarize Meinshausen (2006) on how the quantiles are constructed. Meinshausen starts with the standard definition of the conditional probability function (ccdf) F:

$$F(y|\mathbf{X} = \mathbf{x}_0) = Prob(Y \leq y|\mathbf{X} = \mathbf{x}_0) \tag{11}$$

Knowledge about this function F enables the construction of a formula for computing quantiles. Let the α quantile be defined as $Q_\alpha(\mathbf{x}_0)$ such that the probability of Y less than or equal to $Q_\alpha(\mathbf{x}_0)$ is equal to α at query point $\mathbf{x}_0$, then $Q_\alpha(\mathbf{x}_0)$ can be expressed as the set of the lowest value y for which F is smaller than α:

$$Q_\alpha(\mathbf{x}_0) = \arg\min_y \{y : F(y|\mathbf{X} = \mathbf{x}_0) \leq \alpha\} \tag{12}$$

Constructing a prediction interval, say p, then simply entails the determination of the quantile boundaries of this interval. The lower boundary results in an $\alpha_l = \frac{1-p}{2}$ and the upper boundary $\alpha_u = \frac{1+p}{2}$. Let p-PI be the prediction interval of width p at a query vector $\mathbf{x}_0$, then the interval is computed using:

$$p\text{-PI}(\mathbf{x}_0) = [Q_{\alpha_l}(\mathbf{x}_0), \quad Q_{\alpha_u}(\mathbf{x}_0)] \tag{13}$$

In practice the true ccdf F cannot be determined. Therefore, an estimate of the ccdf $\hat{F}$ needs to be constructed, which is done empirically. Meinshausen (2006) takes two steps to construct $\hat{F}$. First, the Quantile Regression Forest algorithm iterates through each terminal node and counts the number of times the response observation appears in the terminal node/leaf. This number is then divided by the total number of observations that occur within the same leaf, resulting in a proportion. This proportion can be regarded as the weight of a single tree. Second, the derived proportion/weight for every response value of the original training set is aggregated over all trees in the forest, resulting in a weight that can be used to construct the final empirical conditional cumulative distribution function. This complete procedure is illustrated in Figure 3.1. Meinshausen (2006) summarizes the empirical conditional probability function $\hat{F}$ from the distribution of the training response pairs within every leaf by:

$$\hat{F}(y|\mathbf{X} = \mathbf{x}_0) = \sum_{j=1}^{n} w_j(\mathbf{x}_0) \mathbb{1}_{\{y_j \leq y\}} \tag{14}$$

Here, the indicator function $\mathbb{1}$ determines whether the weight will be counted or not, depending on the constraint $y_j \leq y$. This is done for all n training pairs of the training dataset $\mathcal{D}_n$. Each weight w_j is an average that is constructed by taking the sum over all weights per tree for which the query vector $\mathbf{x}_0$ falls into the leaf represented by ℓ:

$$w_j(\mathbf{x}_0) = \frac{1}{M} \sum_{i=1}^{M} w_j^{\varphi_i}(\mathbf{x}_0) \tag{15}$$

Note that this weight function (Eq 15) is indexed with respect to and calculated for *all observations in training data pairs* instead of the subsample on which every individual tree in the random forest is constructed. If it does not occur in a specific tree, then it gets a weight value of 0 assigned for such a tree. The following function calculates what the weight is for one specific tree:

$$w_j^{\varphi_i}(\mathbf{x}_0) = \frac{\mathbb{1}_{\{\mathbf{x}_j^{\varphi_i} \in \mathcal{X}_\ell^{\varphi_i}\}}}{\#\{k : {\mathbf{x}_0}_k^{\varphi_i} \in \mathcal{X}_\ell^{\varphi_i}\}} \tag{16}$$

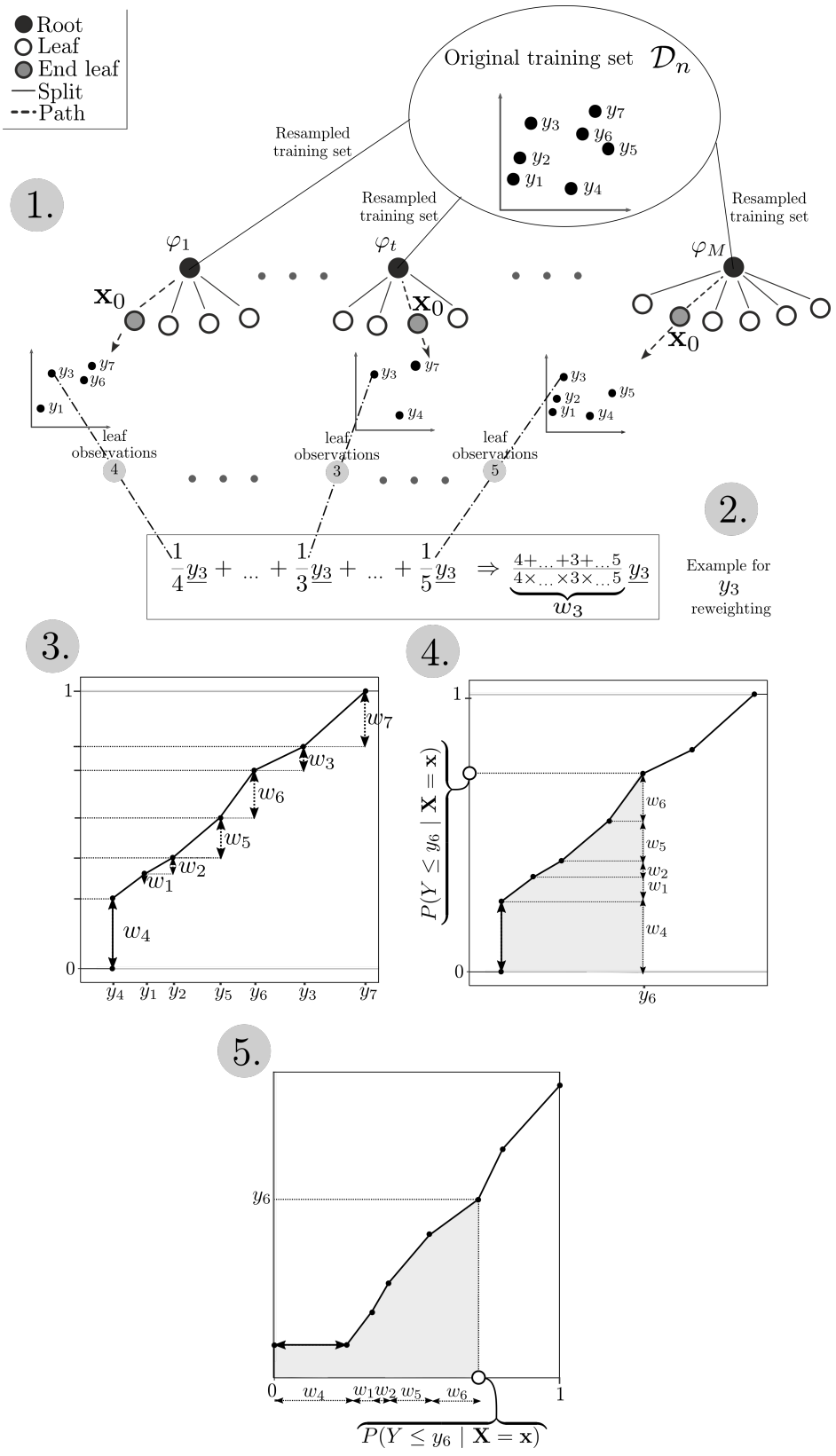


Figure 3.1. Example of the Quantile Regression Forest process.
(1) Drop an unknown query vector $\mathbf{x}_0$ down all trees in the forest; (2) Calculate weights for all response values of the end nodes (i.e. leaves) that $\mathbf{x}_0$ falls in (example given for the construction of w_3 for y_r); (3) Construct the empirical cumulative distribution on the condition $\mathbf{x}_0$; (4) Acquire the probability that the response is smaller than a threshold (say y_6); (5) Use the inverse of 4. to find an arbitrary quantile.

3.4 Jackknife and Infinitesimal Jackknife after Random Forest

Wager's jackknifing approach for uncertainty quantification of Random Forest predictions only considers the expected mean of the predictions from the individual trees that make up the forest prediction. In other words, its aim is to quantify the variance of the *expected* prediction. The central idea of Wager et al. (2014) is that estimating the mean of a statistic is, by the central limit theorem, safer to assume normal than to assume that the distribution of the target value conditioned on the input vectors is normal. For example, the variance of the target value within a certain partition can be unequal to other partitions leading to unreliable quantification of the uncertainty. The uncertainty of the expected prediction of the aggregate of trees in the forest is quantified rather than the uncertainty of the random forest as a whole. Several Random Forest predictions are simulated and thus the standard error can be estimated over the Random Forest predictions. The technique cannot be used for constructing prediction intervals, it can only construct a confidence interval as it aim is quantification of the reducible error.

3.5 Underlying mathematics

When dealing with small amounts of data, a common strategy is to simulate more data by the resampling of the original sample. Given that the underlying assumptions on which the resampling is implemented hold, inferring population parameters through resampling offers a simple, generalized model to estimate the distribution at each point. There are three of these subsampling procedures that can be combined with each other in order to quantify the uncertainty of random forest predictions.

Bootstrapping

Bootstrapping is a well known resampling technique that in essence, draws a prefixed number of times with replacement from the original training sample to construct a new set of samples that approximates the original sample (Hillis & Bull, 1993). Bootstrapping uses the original sample as a proxy to estimate the distribution of the actual population. Hence, bootstrapping is said to model inference of a population from sample data. All this is done under the assumption that the original sample is a close approximation of the actual population (Hillis & Bull, 1993).

Jackknifing

Jackknifing in itself is a resampling method specifically developed for estimating the bias and variance of an estimator at a specific query point (Efron, 1992a). Unlike bootstrapping, the jackknife does not draw with replacement but leaves only one observation out of the original sample to construct a total of $n-1$ new samples. Now, estimating the variance of an estimator is done by iterating over each input training point and averaging the predictions of the jackknifed samples that contain this input training point.

As an illustration the equation below (Eq 17) gives an estimate of the population variance using jackknifing. Let $\mathbf{s}$ be a random sample from a population and let θ be a statistic on $\mathbf{x}_0$ such as the variance, then $\hat{\theta}(-j)$ denotes the estimated outcome of the statistic without the j^{th} point and $\hat{\theta}$ the outcome with all points:

$$V_J(\mathbf{x}_0) = \frac{n-1}{n} \sum_{i=j}^{n} (\hat{\theta}_{(-j)}(\mathbf{x}_0) - \hat{\theta}(\mathbf{x}_0))^2 \tag{17}$$

3.6 Jackknife-after-bootstrap

The jackknife-after-bootstrap is slight alteration of the normal jackknife. Instead of leaving one element out of the original sample systematically, the bootstrapped samples now determine which element is left out (Efron, 1992b). This means that whereas in the regular jackknifing procedure every left out observation is absent in one and only one subsample, in the jackknife-after-bootstrap it could be left out in more than one of the bootstrapped subsamples. The estimation of the population parameters is done in a similar methodology as regular jackknifing. However, instead of one unique jackknifed subsample there now might be several subsamples that have to be aggregated first before calculating the final mean of the estimator statistic (Efron, 1992b). The paragraph below expounds on the procedure for calculating the jackknife-after-bootstrap variance estimates for random forest predictions and will follow the line of reasoning represented in Wager et al. (2014).

Let ψ stand for all trees in the random forest and each individual tree by an element of the set $\{\varphi_i : i \in 1, \ldots, M\}$ such that $f_{\varphi_i}(\mathbf{x}_0)$ gives an individual tree prediction at query point $\mathbf{x}_0$. Now the prediction of the random forest can be seen as a statistic of all M trees, represented by $\hat{\theta}_\psi$. The forest without the j^{th} observation pair is then denoted as $\hat{\theta}_{\psi_{(-j)}}$. Finally, let $\mathcal{X}_{\varphi_b}$ be the root of the b^{th} tree; now regarded as the b^{th} bootstrap sample. Then, the equation for the jackknife-after-bootstrap sample variance estimation of the random forest prediction at $\mathbf{x}_0$ can be written as Equation 18:

$$\hat{V}_\psi^{\text{J}}\left[\hat{\theta}_\psi(\mathbf{x}_0)\right] = \frac{n-1}{n}\sum_{j=1}^{n}\left(\hat{\theta}_{\psi_{(-j)}}(\mathbf{x}_0) - \hat{\theta}_\psi\right)^2 \tag{18}$$

Where $\theta_{\psi_{(-j)}}(\mathbf{x}_0)$ is defined as:

$$\theta_{\psi_{(-j)}}(\mathbf{x}_0) = \frac{\sum_{\{b:\mathbf{x}_j \notin \mathcal{X}_{\varphi_b}\}} f_{\varphi_b}(\mathbf{x}_0)}{\#\{b : \mathbf{x}_j \notin \mathcal{X}_{\varphi_b}\}} \tag{19}$$

Here, the denominator simply counts the number of occurrences that the j^{th} observation $\mathbf{x}_j$ is absent in the roots of all individual trees. The sum of the prediction of the trees where $\mathbf{x}_j$ is absent are then divided by this number to get the aggregate.

Equation 18 leads to a high bias, especially when the total number of observations in a bootstrapped subsample at the root of a tree is small according to Wager et al. (2014), which is a consequence of the Monte Carlo noise of the random parameters in RF. Therefore, an additional bias correction is given (Eq 20). The derivation of this bias correction is further explained in Wager et al. (2014):

$$\hat{V}_\psi^{\text{J-U}}\left[\hat{\theta}_\psi(\mathbf{x}_0)\right] = \underbrace{\hat{V}_\psi^{\text{J}}\left[\hat{\theta}_\psi(\mathbf{x}_0)\right]}_{\text{Biased term}} - \underbrace{(e-1)\cdot\frac{n}{M^2}\sum_{i=1}^{M}\left(f_{\varphi_i}(\mathbf{x}_0) - \hat{\theta}_\psi(\mathbf{x}_0)\right)^2}_{\text{Correction term}} \tag{20}$$

The standard error can then be computed by taking the square root of the variance estimate given above.

Infinitesimal jackknife

The non-parametric delta method, better known as the infinitesimal jackknife, is a slightly different variation of the jackknife (Efron, 1981). In contrast to the original jackknife, the infinitesimal jackknife does not leave one observation out, but reduces the weight of the observation by an infinitesimal amount for every observation repeatedly (Efron, 1981). Then it aims to calculate the estimator at a specific query point by averaging over all individual prediction metrics of the estimator. Summarized in Equation 21:

$$\hat{V}_{\psi}^{\mathrm{IJ}}\left[\hat{\theta}_{\psi}(\mathbf{x}_0)\right] = \sum_{j=1}^{n}\left(\frac{1}{M}\sum_{i=1}^{M}(\#\{j \in \mathcal{X}_{\varphi_i}\} - 1)\cdot(f_{\varphi_i}(\mathbf{x}_0) - \hat{\theta}_{\psi}(\mathbf{x}_0))\right)^2 \tag{21}$$

Also a bias correction for this equation exists. The derivation of this bias correction is further explained in Wager et al. (2014):

$$\hat{V}_{\psi}^{\mathrm{IJ-U}}\left[\hat{\theta}_{\psi}(\mathbf{x}_0)\right] = \underbrace{\hat{V}_{\psi}^{\mathrm{IJ}}\left[\hat{\theta}_{\psi}(\mathbf{x}_0)\right]}_{\text{Biased term}} - \underbrace{\frac{n}{M^2}\sum_{i=1}^{M}\left(f_{\varphi_i}(\mathbf{x}_0) - \hat{\theta}_{\psi}(\mathbf{x}_0)\right)^2}_{\text{Correction term}} \tag{22}$$

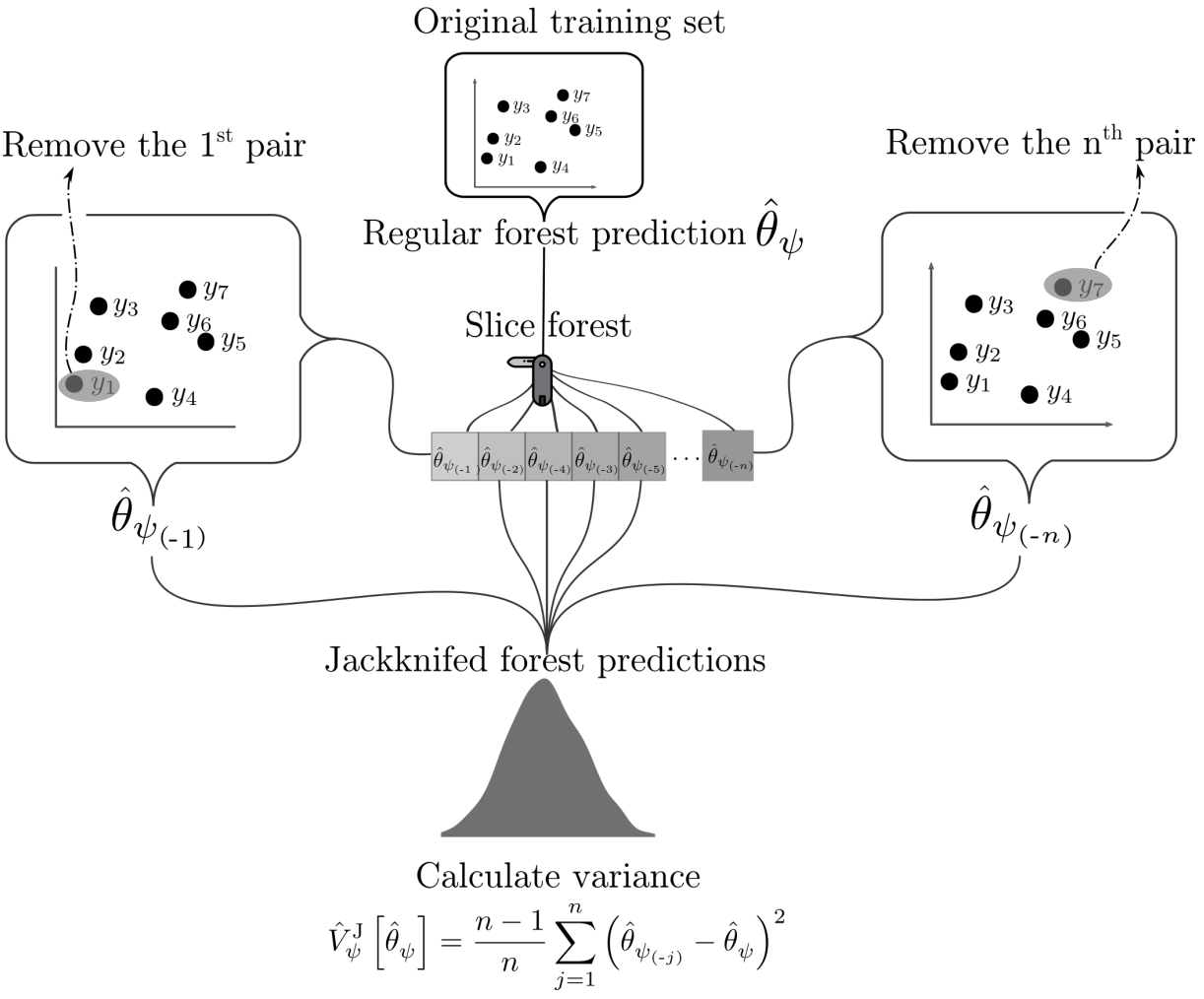


Figure 3.2. Overview of the jackknife-after-bootstrap without bias correction

Standard error

The underlying idea key to the approach of (Wager et al., 2014) is to estimate the standard error by taking the square root of the difference between a prediction at a query point $\mathbf{x}_0$ with a multitude of Random Forests and the expected value a prediction at $\mathbf{x}_0$. The expected value of its prediction is found by taking the mean over Random Forests predictions built on different samples. In practice, there are no new samples and hence, new samples are simulated by bootstrapping the training set and growing a new

Random Forest for each of these new bootstrapped training samples. This was the original idea of Sexton and Laake (2009) to estimate the variance of Random Forest prediction by first bootstrapping the training set B times and then train B Random Forests from these bootstrapped training sets. While theoretically founded, in practice, this training of multiple Random Forests is very computationally demanding.

The technique that Wager et al. (2014) propose makes use of a trick to circumvent the growing of new Random Forests through the emulation of Random Forest from the already trained, original Random Forest. This is done by selecting subsets of trees from the original Random Forest based on whether a training point is within such a subset (jackknife-after-bootstrap) and comparing this to the mean of all such Random Forest subsets. Wager et al. (2014) then compensate for the noise introduced through this emulation of new Random Forests through the addition of an extra bias-correction. After the bias correction the jackknife-after-bootstrap is still biased upward and the infinitesimal jackknife is biased downward. Therefore, the authors suggest to take the arithmetic mean between the jackknife and infinitesimal jackknife to calculate an unbiased estimate of the variance statistic of the predicted mean of several Random Forest predictions.

3.7 Random Forests as U-statistics

Mentch and Hooker (2016) showed that under a strict subsampling scheme, supervised ensembles such as random forests predictions resemble conditions for U-statistics enough that with the addition of certain lemmas they fall (indirectly) under Hoeffding's (1948) developed theory of U-statistics, which are proven to be asymptotically normal. This normal distribution can then be used to quantify the uncertainty related to the reducible error of the random forest prediction. Therefore, confidence intervals can be constructed through this method. The construction of prediction intervals is not possible as the U-statistic quantifies the expected aggregate of the tree predictions instead of quantifying the uncertainty of the Random Forest itself.

The mathematical foundations of the U-statistic based Random Forests are crudely summarized as the technique falls under advanced graduate statistics. Mentch and Hooker (2016) do only outline certain choices in their appendix, especially regarding unbiased expected RF prediction variance estimates. Hence, the motivation of choices was omitted here as well. For more information on U-statistics the work of Lee (1990) is recommended.

3.7.1 U-statistics

U-statistics is a special class of statistics that typically emerge from the theory of minimum-variance unbiased estimators; the "U" in U-statistics stands for unbiased. The main idea behind U-statistics is to draw a predetermined number of times throughout all combinatorial selections from the sample of size n. Subsequently, by averaging over the possible results of these subsamples an unbiased estimator of a statistic can be derived. Due to cumbersome notation later on let the conventional training set $\mathcal{D}_n$ now be replaced by S, which has observation pairs $S_1, \ldots, S_n$. Let θ be a statistic or population parameter of interest. Now suppose that a function h exists with $r \leq n$ arguments selected from S such that its expected value equals θ. Or, equivalently:

$$\theta = E[h(S_1, S_2, \ldots, S_r)] \tag{23}$$

Hoeffding (1948) then postulates that this expected value is unbiasedly approximated by considering all combinatorial subsamples of size r that can be drawn from the joint random original sample S (that had

size n). This means that a total of $\binom{n}{r}$ new samples can be selected from S. Now their average should estimate the minimum variance unbiased the statistic θ. This is summarized in Equation 24 that is named as the U-statistic with kernel h and rank r:

$$U_n = \binom{n}{r}^{-1} \sum_{i}^{\binom{n}{r}} h(S_{i_1}, S_{i_2}, \dots S_{i_r}) \tag{24}$$

Where $\{i_1, \dots, i_r\}$ are the indices that represent subsets of r different integers. In other words, $- \ i \in \{1, 2, \dots \binom{n}{r}\}$ denoting the i^{th} combination.

3.7.2 Bagged tree predictions as U-statistic

The weaker form of Random Forest, bagged trees, already closely resembles a U-statistic. This can be seen by replacing the function h by the function of an individual tree. Albeit a U-statistic with a very small number of combinations. With a slight alteration of the function definition of the original bagged prediction function f_{bag} the U-statistic kernel function can be applied on it. Let the specific individual tree (tree represented as φ_i) prediction functions for query vector $\mathbf{x}_0$, denoted as $f^{\varphi_i}_{\mathbf{x}_0}$. Note that instead of a function of $\mathbf{x}_0$, the individual tree prediction function $f^{\varphi_i}_{\mathbf{x}_0}$ is seen as a function with a subsample as input: $S_{i_1}, S_{i_2}, \dots S_{i_r}$. Then, the U-statistic kernel function estimator for the tree bagging procedure at query point $\mathbf{x}_0$, denoted as $U_n^{bag(\mathbf{x}_0)}$, maps $(\mathcal{X} \times \mathbb{R})^r \to \mathbb{R}$. Or, written in a single equation (Eq. 25):

$$U_n^{bag(\mathbf{x}_0)} = \binom{n}{r}^{-1} \sum_{i}^{\binom{n}{r}} f^{\varphi_i}_{\mathbf{x}_0}(S_{i_1}, S_{i_2}, \dots S_{i_{r_n}}) \tag{25}$$

Hence, bagged predictions are considered to be asymptotically normal as they can be written as a U-statistic and the individual predictions are independent of the order of the training data.

In practice, it is not possible to calculate the U-statistic when the number of training examples becomes large. This is a trivial consequence of the fact that the number of combinations rises rapidly for each increase in n. Moreover, the classical notion of U-statistic requires much more subsamples to be chosen than is the case for bagging. Mentch and Hooker (2016) tackle these problems by proposing an analogy to *incomplete* U-statistics. The incomplete U-statistic was proven to remain asymptotically normal under a set of conditions, even when the number of combinations drastically decreases (Janson, 1984). This incomplete U-statistic is constructed by drawing say m_n times (uniformly) from the original $\binom{n}{r}$ combinations. Mentch and Hooker (2016) note that the performance of this incomplete U-statistic is very dependent on the size of r and therefore, they allow r to scale together with the number of samples n which they call r_n. Equation 25 after rewriting now becomes the largely reduced:

$$U^{bag(\mathbf{x}_0)}_{n,r_n,m_n} = \frac{1}{m_n} \sum_{i}^{m_n} f^{\varphi_i}_{\mathbf{x}_0,r_n}(S_{i_1}, S_{i_2}, \dots S_{i_{r_n}}) \tag{26}$$

3.7.3 Random Forest predictions as U-statistics

Mentch and Hooker (2016) then discuss that the random perturbation component in Random Forest tree building limits the applicability of U-statistics. Bagging can fall under incomplete U-statistics by drawing from all possible subsample combinations, but Random Forest has an additional randomness. Therefore, they prove (no textual motivation) that if the expected value is taken with respect to the random perturbation parameters $\{\omega_i : i \in 1, \dots, m_n\}$ − and these are independently selected of the original sample − they conform to incomplete U-statistics. The reason is that the kernel function given in Equation 24 gets fixed as the expected value of a random variable is singular. Hence, the mean prediction

becomes asymptotically normally distributed. The U-statistic kernel for Random Forest predictions then takes the form:

$$U^{\psi(\mathbf{x}_0)}_{\omega;n,r_n,m_n} = E_\omega \left[\frac{1}{m_n} \sum_{i}^{m_n} f^{\varphi_i;\omega_i}_{\mathbf{x}_0,r_n}(S_{i_1}, S_{i_2}, \dots S_{i_r}) \right] \tag{27}$$

3.7.4 Variance estimation

Due to the asymptotic normality of U-statistics, the variance of the expected Random Forest prediction can be estimated. This is a special procedure that is quite complex, so only a crude summary is given. The main idea is that variance of the expected statistic can be estimated by looking at the joint variance of all predictions of RF models that have 1 or all r_n examples overlapping in their underlying subsamples but no clear motivation is given and will thus be omitted. Mentch and Hooker (2016) give the final variance estimate of the expected RF prediction as the average of the variance of RF models with 1 example overlap and all examples in overlap. Lee (1990) introduced a metric η_{d,r_n} that gives the variance of the expected value of the samples when they have d elements in common with each other, or equivalently d chosen points are fixed:

$$\eta_{d,r_n} = \text{Var}\left[E[h_{r_n}(S_1, \dots, S_{r_n}) \mid S_1 = s_1, \dots, S_d = s_d]\right] \tag{28}$$

For Random Forest it is necessary to estimate η by only using a certain number of Monte Carlo simulations for the drawing procedure, say m_n, and averaging due to computational difficulty. Using this, the estimation of Lee (1990) variance metric of common examples among subsamples 28 for RF prediction variance becomes:

$$\hat{\eta}_{d,r_n} = \text{Var}\left[\frac{1}{m_n} \sum_{i=1}^{m_n} f^{\varphi_i}_{\mathbf{x}_0,r_n}(\mathcal{S}_{\bar{\mathcal{Z}}^{(1)},i}), \dots, \frac{1}{m_n} \sum_{i=1}^{m_n} f^{\varphi_i}_{\mathbf{x}_0,r_n}(\mathcal{S}_{\bar{\mathcal{Z}}^{(n\bar{z})},i}) \right] \tag{29}$$

Here $\mathcal{S}_{\bar{\mathcal{Z}}^{(j)},i}$ denotes the i^{th} subsample that includes the j^{th} set of fixed points (represented by $\bar{Z}^{(j)}$).

Note that for the final variance estimate, the case $d = 1$ and $d = r_n$ need to be calculated. In the case that $d = r_n$, m_n simply equals 1 as all cases are identical. Now, the final variance estimate can be given by adding $\hat{\eta}_{r_n,r_n}$ and $\hat{\eta}_{1,r_n}$ after applying its correction:

$$Var(U^{\psi(\mathbf{x}_0)}_{\omega;n,r_n,m_n}) = \frac{r_n^2}{(\frac{n}{m_n})} \hat{\eta}_{1,r_n} + \hat{\eta}_{r_n,r_n} \tag{30}$$

Now that the variance is known, a normal distribution can be constructed by taking the square root of the variance as standard deviation and using the prediction as the mean.

3.8 Regression Kriging

Regression kriging (RK), as the name implies, is a combination of training a regression model and then performing an algorithm of "best linear unbiased prediction" (BLUP) on the regression residuals. Kriging entails a different approach from the modern machine learning predictions; it is based on spatial prediction instead of direct prediction. Kriging only considers the spatial correlation within the area between target variable observations for its prediction (note that target variable also implies regression residuals).

In overview, regression kriging first quantifies the explanatory variation and then builds the rest of the model spatially by solely looking at the unexplained variation, i.e. residuals (Hengl et al., 2018). Because information from regression is now regarded as the deterministic part of the model, the variance of the

kriging prediction error is assumed to now be independent of the underlying regression. Furthermore, kriging assumes that the mean stays the same over the search neighborhood, called the stationarity assumption (N. A. Cressie, 1993). Thus, as both parameters of mean and standard deviation (square root kriging prediction error variance) are estimated and the predicted target value is assumed to have a normal distribution, kriging can provide a complete measure of the total uncertainty at a certain location. Quantification of the complete uncertainty means that prediction intervals can be constructed. For constructing confidence intervals with this method, an additional step needs to be taken such as bootstrapping the sample to quantify the uncertainty of the fit of the spatial model (Paoli et al., 2003).

This section has a different structure then the sections on the other techniques as RK entails a different perspective from the direct uncertainty quantification approaches. In general the ordinary kriging approach is explained. This section is therefore larger in content as RK consists of a multitude of steps which will all be described in detail. These steps help to explain why regression kriging minimizes and simultaneously quantifies the complete uncertainty. First, a clear description of what a spatial model is will be given in Section 3.8.1. Section 3.9 will then discuss how a prediction is made. Section 3.9.1 details the background of how the minimization of the prediction error is achieved and should be seen more as a supplementary information. The final section explains how prediction intervals can be constructed with kriging.

3.8.1 Modeling spatial correlation.

Let a study area have a total of n locations $u_i,\ i\ = 1, \ldots, n$. Now suppose that with these locations a new location is to be predicted: u_0. Note that this notation is merely the location instead of the input query vector $\mathbf{x}_0$. Now suppose that the random variable Z, representing the target variable, depends solely on location or distance. Then, kriging assumes that the predicted target value $\hat{z}$ at location u_0 can be estimated from the values of the surrounding n sample observations.
For kriging to be accurate as a model, certain conditions need to be satisfied as assumptions are made on the distribution of the target variable throughout space. First, kriging assumes a constant mean over the whole search neighborhood (often the whole study site) (Van Beers & Kleijnen, 2003). Second, the target variable should be normally distributed. Third, the semivariance should only depend on the distance measure. Here, semivariance is defined as "variance of the difference between field values at two locations across realizations of the field (N. A. Cressie, 1993)". If these conditions are met, kriging should in theory produce a reliable model.

The degree to which a new location relates to the other n sample locations depends on the function γ that estimates the semivariance of the target value per distance increase (denoted by h). Typically this entails computing the squared differences between the known target values at the n sample locations that fall within different sets of distance lags. In practice, often the function $\gamma(h)$ is constructed by fitting a curve through the pairwise comparison of the semivariance between observations that are within the same distance lag (:

$$\gamma(h) = \frac{1}{2 \cdot |N(H_k)|} \sum_{(i,j) \in N(H_k)} \left(z(u_i) - z(u_j)\right)^2 \tag{31}$$

In Equation 31 above, $N(h) = \{(u_i, u_j) : ||u_i - u_j|| \in H_k\}$. The cardinality $|N(H_k)|$ returns the number of distinct elements in this set. That is, the set of all point pairs within a certain distance interval. Following Wackernagel (2003), lags are grouped into K disjoint lag distance intervals H_k such that the union $\cup_{k=1}^{K} H_k$ retrieves all distances in the target value set Z. In practice often only half of the diagonal of the study area extent is used. Figure 3.3 shows an example of a semivariogram with all its components.

Arguably, constructing the right variogram models will differ from person to person as other bin sizes and modeling curves can be chosen which will lead to difference in the weighted least squares errors (WLS). Furthermore, variogram fitting with WLS will lead to the fact that most bins in the center of the variogram model will have a stronger weight than bins towards the extremes (N. Cressie, 1985). Using a restricted maximum likelihood (REML) estimator for variogram eliminates some of these issues by looking at the empirical variogram cloud instead, therefore using the complete information available and not only the binned sample variogram. REML is centered on the assumption that the semivariance follows a multivariate Gaussian distribution. The parameters are chosen in such a way that minimizes the negative log-likelihood function (Kerry & Oliver, 2007). REML fitting for estimating curve parameters is more advanced represented here and therefore the work of Kerry and Oliver (2007) is adviced for further information.

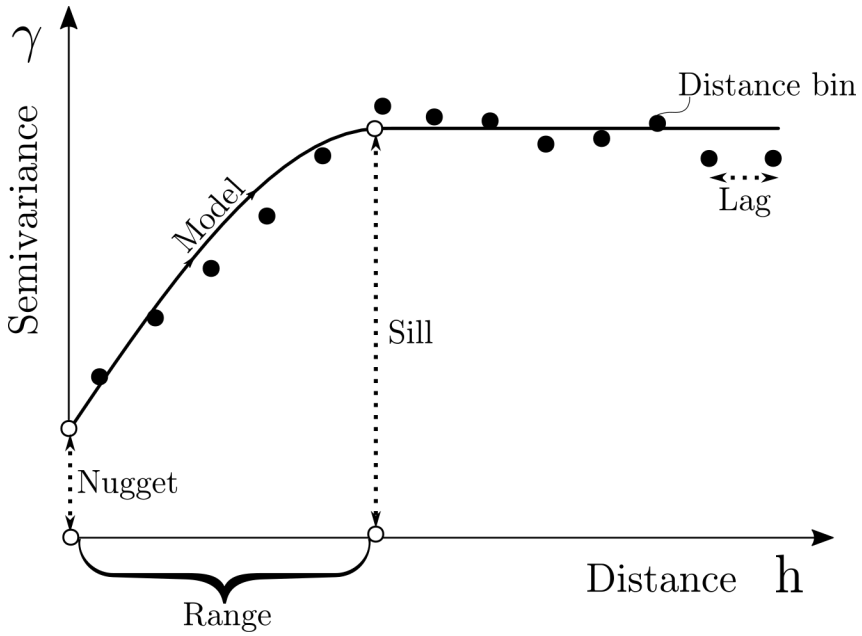


Figure 3.3. Variogram and its components.

3.9 Regression kriging predictions

In practice, auxilliary variables are often more abundant than response variables. Furthermore, in regression kriging they are required to be spatially exhaustive as it needs to coincide with the sample observations. Therefore, the goal is to use these auxiliary variables (also termed covariates) to model the mean of the response first. The uncertainty in such a prediction is then modeled by kriging the residuals (now considered the to be kriged target value) after substracting the predicted mean of the response value with the target value at the known locations (Hengl et al., 2007).

RK first uses statistical regression model to predict the target value from the auxiliary variables ($\mathbf{x}_u$) observed at at an unknown location u_0 as $\hat{f}(\mathbf{x}_u)$. In the case of Random Forest $\hat{f} = \hat{f}_\psi$ which was discussed in Section 2.4. Next, the spatial model of the regression residuals can be added. These residuals are are modeled by the variogram described in Section 3.8.1. The variogram is crucial in what the values as weights are used for predicting the increase in semivariance. Let these weights be defined as: $\{\lambda_i : i \in 1, \ldots, n\}$. Then kriging optimizes the weights in such a way that the prediction error (also called kriging variance) is minimized. The kriging variance σ^2_{OK} is simply the expected value of the difference between estimated target value at location u_0 and its true value:

$$\sigma^2_{OK}(u_0) = E\left[(\hat{Z}(u_0) - Z(u_0))^2\right] \tag{32}$$

Keeping this minimization in mind the following equation predicts the target value at location u_0:

$$\hat{z}(u_0) = \hat{f}(\mathbf{x}_u) + \hat{\epsilon}(u_0) \quad \Rightarrow \hat{z}(u_0) = \hat{f}(\mathbf{x}_u) + \sum_{i=1}^{n} \lambda_i \cdot \epsilon(u_i) \tag{33}$$

Let $\varepsilon(u_0)$ be a random variable that is normally distributed with unknown, but constant μ and standard deviation $\sqrt{\sigma^2_{RK}(u_0)} = \sigma_{RK}(u_0)$ that is dependent on the location say u_0. Now suppose that as regression function Random Forest is chosen ($\hat{f}_\psi$). Then, the final regression kriging model to estimate the randomly distributed target variable Z at location u_0 takes the form:

$$Z(u_0) = \hat{f}_\psi(\mathbf{x}_u) + \sum_{i=1}^{n} \lambda_i \cdot \epsilon(u_i) + \varepsilon(u_0) \tag{34}$$

3.9.1 Weight estimation

The only measure that still needs to be described is how the weights are estimated in Equation 34. The minimization of the kriging variance σ^2_{OK} must be done such that the prediction error is unbiased. For this reason, kriging is also called the best linear unbiased predictor. If the bias of the prediction error is equal to 0, then the kriging weights sum to 1 ($\sum_{i=1}^{n} \lambda_i = 1$). By substituting the estimate of $\hat{Z}(u_0)$ with a weighted prediction $\sum_{i=1}^{n} \lambda_i \cdot Z(u_i)$, kriging variance or expected squared prediction error (Eq 32) can be rewritten as:

$$\sigma^2_{OK}(u_0) = E\left[\left(\sum_{i=1}^{n} \lambda_i \cdot Z(u_i) - Z(u_0)\right)^2\right] \tag{35}$$

Factoring out the prediction error variance above (Eq 35) results in:

$$\sigma^2_{OK}(u_0) = \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i\lambda_j E\left[Z(u_i) \cdot Z(u_j)\right] - 2\sum_{i=1}^{n} \lambda_i E\left[Z(u_i) \cdot Z(u_0)\right] + E\left[(Z(u_0))^2\right] \tag{36}$$

Following Lichtenstern (2013) Equation 36) can be simplified as equation 37 below if and only if the weights sum to 1:

$$\sigma^2_{OK}(u_0) = -\sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i\lambda_j \underbrace{\frac{E\left[Z(u_i) - Z(u_j)\right]}{2}}_{\gamma(u_i - u_j)} + 2\sum_{i=1}^{n} \lambda_i \underbrace{\frac{E\left[Z(u_i) - Z(u_0)\right]}{2}}_{\gamma(u_i - u_0)} \tag{37}$$

Replacing the expected values in Equation 37 with the semivariance expressions below them (denoted by underbraces) results in:

$$\sigma^2_{OK}(u_0) = -\sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i\lambda_j\gamma(u_i - u_j) + 2\sum_{i=1}^{n} \lambda_i\gamma(u_i - u_0) \tag{38}$$

As the mean is unknown and kriging should provide an unbiased estimator minimization of the above equation 38 is not straightforward. The unbiasedness condition means that all weights sum to 1 ($\sum_{i=1}^{n} \lambda_i = 1$). By using this as a constraint on finding a minimum, the problem becomes easier. Inclusion of this constraint is done through the construction of a new augmented function called the Lagrangian $\mathfrak{L}_{\sigma^2_{\text{OK}}}$ that has an additional term with a new variable, called the Lagrange parameter (ϕ). This Lagrangian function is the original function (in this case σ^2_{OK}) augmented with an additional term

of the Lagrange parameter ϕ multiplied by the constraint subjected to 0. In this case the constraint will be subjected to 0 when $\sum_{i=1}^{n} \lambda_i - 1$ is calculated. Written in one equation:

$$\mathfrak{L}_{\sigma^2_{\mathrm{OK}}}(\lambda_1, \ldots, \lambda_n, \phi) \;=\; \sigma^2_{OK}(u_0) - \phi \cdot \left(\sum_{i=1}^{n} \lambda_i - 1 \right) \tag{39}$$

Minimization of the Lagrangian function is done by setting partial derivatives with respect to each individual weight equal to zero and an additional partial derivative with respect to the Lagrange parameter ϕ to 0. Working these steps out mathematically leads to what Matheron (1971) describes as the *kriging equations for stationary random function with unknown expectation.* The final kriging system of equations to compute both the Lagrange parameter and the optimal choice of weights:

$$\begin{aligned} &\sum_{i=1}^{n} \lambda_i \gamma(u_i - u_j) + \phi = \gamma(u_i - u_0) \quad \text{for } i = 1, 2, \ldots, n \\ &\sum_{i=1}^{n} \lambda_i = 1 \end{aligned} \tag{40}$$

3.9.2 Prediction interval estimation

The uncertainty estimate can simply be derived from the kriging variance. The Kriging system already chose the optimal weights to keep this kriging variance as low as possible as it built upon this condition. Therefore, by substituting the weights and the Lagrange parameter that arise by solving the kriging system (Equations 40) in the Equation 38, the kriging variance at a new query location u_0 becomes:

$$\sigma^2_K(u_0) = \sum_{i=1}^{n} \lambda_i \cdot \gamma(h_i) + \phi \tag{41}$$

The kriging prediction errors are assumed to be normally distributed. Therefore, a prediction interval can be constructed as the the kriging prediction itself is considered the mean (μ): $\hat{z}(u_0)$. The standard deviation is the square root of the kriging variance $\sqrt{\sigma^2_{OK}} = \sigma_{OK}$. Looking up the standard score z (see Table 3.2) for a specific p and filling this value in the following interval will give the final prediction interval boundaries defined as:

$$p\text{-PI} = [\mu - z\sigma, \mu + z\sigma] \tag{42}$$

Table 3.2. Prediction interval versus standard score.

p-value	Standard score
75%	1.15
90%	1.64
95%	1.96
99%	2.58

3.10 Method viability assessment results

In this section all methods are compared with each other according to the main criteria Scalability and efficiency; Practicality; and Robustness and Accuracy. The section consists out of two main parts. First, a table is given that summarizes the qualitative assessment based on the literature review and auxiliary

information on the uncertainty quantification methods such as experiments and software packages (Table 3.3). The choice is then also made for which RF uncertainty quantification technique is suitable. Second, a discussion is presented highlighting the strong and weak points.
The table below (table 3.3) summarizes the a priori qualitative literature assessment results.

Table 3.3. Viability assessment of RF uncertainty quantification methods.

	Subcriteria	**Score RK**	**Score QRF**	**Score JKIJ**	**Score USI**
Scalability	Computation time	+	−	+	+
	Flexibility	+	0	0	+
Usability	Availability	+	+	+	−
	Extensive	−	+	+	−
Rigor	Completeness	+	+	0	+
	Accuracy	0	+	0	0

For the case study the complete uncertainty needs to be quantified. This means that QRF and RK are the only candidates possible for further investigation.

3.10.1 Scalability

In general, the main computation issues can come from the training of a regression model such as Random Forest itself. QRF does not take a longer time to grow than normal RF as it does not require the averaging of all leafs in the forest (Mentch & Hooker, 2016). However, QRF needs to aggregate all weights for every unknown point and this can lead to substantial increases in computational time. For regression kriging computation time is in general small (Declercq, 1996), the kriging system needs to be solved which involves an inversion of a large matrix but once this is done computation time decreases considerably. Regarding flexibility, there are several approaches for kriging, such as reducing the neighborhood size (Declercq, 1996). However, the computations consist mostly of matrix multiplications and this might be tricky to paralllelize/multi-thread as it is a dependent calculation. Therefore, regression kriging is quite flexible for both small and large datasets. Quantile regression Forest is parallelizable for the tree growing process (M. Wright & Ziegler, 2015), but the reweighting for the calculation of the quantiles is not. The jackknifing method is fast as it does not built new random forests but does require that forests need to have a large number of trees to reduce Monte Carlo noise that can lead to bias (Wager et al., 2014); using large number of trees might not always feasible for large data sets. The U-statistics approach grows a large number of forests, but it does not require a lot of trees and thus is more feasibile for large datasets. There is currently no published information on how fast the technique is at time of writing this thesis and rules-of-thumb for parameter settings are also lacking.

3.10.2 Usability

There is an standalone package for QRF available and an implementation is also available in the computationally fast Ranger package that is a R wrapper for Random Forest programmed in C (M. Wright & Ziegler, 2015). Both are readily documented and published through the CRAN repository. The technique is easily understandable and only two extra options need to be ticked, when predicting a quantile should be provided and the leafs need to keep all response values. Kriging is a well-established method that was developed 50 years ago, therefore its availability is very high with R packages such as gstat (Pebesma, 2004) and geoR (Ribeiro Jr et al., 2001) and its inclusion in proprietary software such as

ArcGIS and SAGA. Kriging is not very usable straight away, a spatial correlation model (with variogram) needs to be fitted which might differ from person to person. The jackknifing method has gained quite some traction and since the original paper was released implementations in python and R were available (Polimis et al., 2017); the method is also embedded in the beta of the Ranger package (M. Wright & Ziegler, 2015). The technique is easily usable and requires (like QRF) only one extra true or false options to check when predicting on unknown points. The U-statistics approach is very recent (from 2016) and has not been implemented in a package yet.

3.10.3 Rigor

With QRF the complete conditional distribution is estimated and thus prediction intervals can be estimated. If the training set is large enough QRF is also robust (Meinshausen, 2016). It is unclear what number of points the prediction interval estimates stabilizes, accuracy will depend on some assumptions and might vary. Validation studies are still lacking. Regression kriging is dependent on the dispersion of observations (Brus & Heuvelink, 2007); for its variogram close and far distances need to be included in the sampling scheme. Kriging has a couple of assumptions (e.g. stationarity) that need to be fulfilled for it to be a correct model. When there is good reason to believe that kriging will satisfy its underlying assumptions it is fair to assume it is an accurate model. However, regression kriging can only be used in an auto-correlation context (i.e. space or time). Jackknifing and U-statistics are both techniques for quantifying the reducible error, hence they can only be used for estimating confidence intervals of the expected prediction. They have currently not been researched well on their accuracy beside some mathematical simulations which were successful. Jackknifing has only recently been proven to be mathematically consistent under a set of assumptions (Wager & Athey, 2017). U-statitistics is also proven to be mathematically consistent and can be used for making formal hypothesis tests.

4 SPATIAL EVALUATION METHODS

The two most viable methods Random Forest uncertainty quantification methods, regression kriging (RK) and quantile regression forest (QRF) both model the complete uncertainty that is suitable for constructing prediction intervals on any probability level. However, an important question that needs to answered is how reliable the methods are when subjected to a spatial context. Therefore, QRF and RK were evaluated to a certain methodology. This chapter outlines the evaluation approach. First, maps were made to visualize the predictions and their upper and lower boundaries and map the widths of the prediction intervals. Second, an overall validation assessment on the Random Forest model's performance was conducted to provide a framework for interpretation. Third, the local uncertainty was assessed through a rigorous cross-validation approach. The evaluation consisted of comparing metrics of the performance, visual inspection and interpretation of anomalies with geographic knowledge (i.e. why does the model perform bad at certain locations in the landscape?).

4.1 Mapping prediction intervals

Random Forest parameters were chosen with standard settings used by the Ranger package (M. Wright & Ziegler, 2015): 1000 trees, mtry of $\sqrt{N_{\text{predictors}}}$, a minimum nodesize of 5 training observations and the splitting criterion is the local maximum decrease in variance of the target value (see Section 2.4 for a discussion of these parameters).
Consider an arbitrary unsampled location u_0. Now suppose that the target value z is a realization of the random variable Z at location u_0, where the probability distribution of Z is conditioned to the available local information represented by $\mathcal{I}$. Then the conditional cumulative probability distribution function (CCDF) F at location u_0 can be written as:

$$F(u_0; z \mid \mathcal{I}) = Prob\{Z(u_0) \leq z \mid \mathcal{I}\} \tag{43}$$

The ccdf (Equation 43) gives the probability that a value is smaller than a given threshold z. Hence, the inverse of the ccdf F^{-1} gives the boundary values of a predefined probability interval. In the equation of the ccdf above (Eq. 43), the local information $\mathcal{I}$ either represents the multivariate (say of size c) observation vector $\mathbf{x} \in \mathcal{X} \subset \mathbb{R}^c$ at point u_0 that is used for the QRF model, or $\mathcal{I}$ represents information of neighbor locations $\{z(u_\alpha) : \ \alpha \in U \subseteq \mathcal{D}_n$ that was used in regression kriging. Where in this case U stands for the set of neighbors with influence (i.e. the complete study site) and $\mathcal{D}_n$ the set of all training points in the study site.

Information obtained with QRF and RK on the ccdf was used to predict values at two different quantiles to represent the 0.9 prediction interval. First, the 0.05 quantile was estimated from the inverse of the empirical or assumed CCDF ($\hat{F}^{-1}$) given by QRF and RK respectively. Second, the 0.95 quantile was calculated through submitting in $\hat{F}^{-1}$. Additionally, the regular Random Forest prediction was computed. For the mapping of the prediction interval a width metric was used defined as W. This width is the total

distance of the prediction interval in the original unit of the response. Let, u_0 be an unknown location that was not included during the modeling process, then the width $W(u_0; p)$ at probability level $p \in [0, 1]$ at this location is defined as:

$$W(u_0; p) = \left[\hat{F}^{-1}\left(u_0; \frac{(1+p)}{2}\right) - \hat{F}^{-1}\left(u_0; \frac{(1-p)}{2}\right)\right] \tag{44}$$

The function above (Equation 44) was applied for $p = 0.9$ on the complete study site both for QRF and RK to estimate the width of the 0.9 probability prediction interval for every pixel. The resulting maps were then plotted side-by-side and differences between QRF and RK were visually assessed and compared with the other results.

4.2 Cross validation

Cross validation was done in a k-fold manner. k-fold cross-validation means the partitioning a dataset into k partitions and each partition consists out of two, complementary, subsets. In this study k was set to 10. The first part of this partitioned subset, the test subset (size 10%), and the second part of the subset, also called the training subset (size $100 - 10 = 90\%$). With the RF model callibrated on this training set, the test set was predicted. This process was repeated 100 times. The subsections below describe in further detail what quantity was validated.

4.2.1 Random Forest model performance

The overall validation was conducted using pair-wise validation scatterplots of the model's prediction versus the actual observations and two quantities that numerically summarize the validation of the goodness of fit below. The scatter plots visualize the performance of the model on 100 iterations of 10-fold tests. The end results of the three methods are visually presented in the form of a scatter plot comparing the models in one figure. The overall statistics Root Mean Squared Error (RMSE) shown in Equation 45 and the coefficient of determination, R^2 calculated as in equation 46, of the observed values versus the predicted accompany these graphs and provide a metric of the overall performance and guided the interpretation of the results.

Consider the set of N cross validation locations, represented by the coordinate vectors $\{u_i, i = 1, 2 \cdots N\}$. Let $z(u_i)$ stand for the real value of the soil property at location u_i and $\hat{z}(u_i))$ stand for the predicted value respectively. Then, the RMSE is defined as:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{z}(u_i) - z(u_i))^2} \tag{45}$$

And the R^2 of the k-fold predictions:

$$R^2_{pred} = 1 - \sum_{i=1}^{n}\frac{(\hat{z}(u_i) - z(u_i))^2}{(z(u_i) - \bar{z})^2} \tag{46}$$

Where $\bar{z}$ denotes the average response over all locations. Both measures were averaged over all 100 iterations.

4.2.2 Uncertainty assessment

For a measure of the quality of the uncertainty quantification at a local level, a slight alteration of the method outlined by Goovaerts (2001) was followed that was originally set out by Wadoux et al. (2018).

The overall uncertainty assessment was conducted in three major steps. Let, p-probability intervals (p-PI) be intervals that are confined by $(1-p)/2$ and $(1+p)/2$ quantiles. First, all occurrences of the observed target value within a p-PI were counted. Second, this number of observed values within p-PI was compared to the corresponding proportion of these occurrences at 19 p values. Third, the absolute deviation was used to summarize the correctness of the uncertainty quantification models QRF and RK.

The fraction of occurrences within these quantiles is expressed as Equation 47 for the set of N locations u_i that have a validation measure of z and an associated ccdf estimate $\{\langle z(u_i), \hat{F}(u_i; z \mid I)\rangle,\ i = 1, 2, \ldots, N\}$:

$$\bar{\xi}(p) = \frac{1}{N}\sum_{i=1}^{N} \xi(u_i; p) \quad p \in [0, 1] \tag{47}$$

where the indicator function $\xi(u_i; p)$ is defined as:

$$\xi(u_i; p) = \begin{cases} 1 & \text{if} \quad \hat{F}^{-1}(u_i; \frac{(1-p)}{2}) < z(u_i) \leq \hat{F}^{-1}(u_i; \frac{(1+p)}{2}) \\ 0 & \text{otherwise} \end{cases} \tag{48}$$

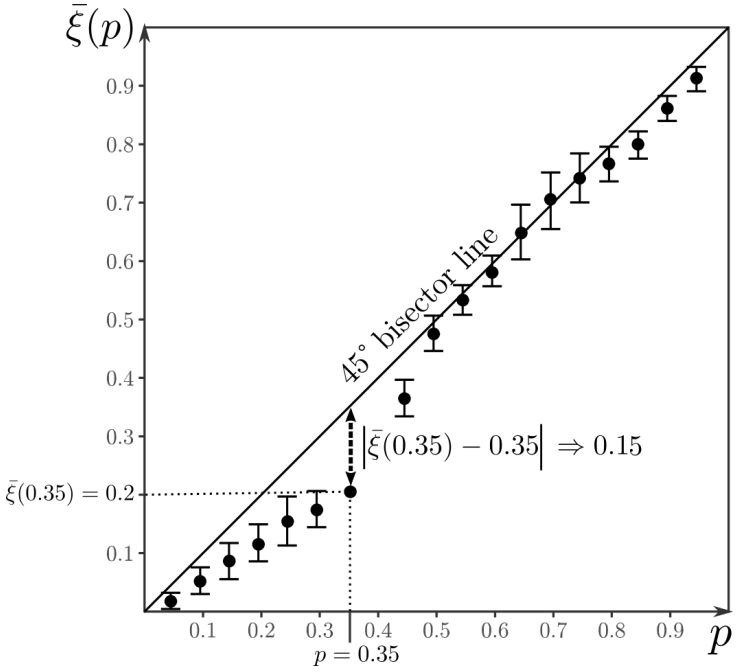


Figure 4.1. Example of reading from an accuracy plot and its components.
On the x-axis the expected proportion of observations in the interval (= p) and on the y-axis the actually observed proportion in the interval (= $\xi(p)$). The black dot denotes the average of the 100 10-fold validation and the error bars the 95% confidence interval of the mean being the true value. The 1 : 1 line in the middle, also named 45 deg bisector, displays the ideal situation where the expected uncertainty is the same as the observed uncertainty. The plot gives an example for a hypothetical situation where a validation of the 0.35-PI ($p = 0.35$) was tested but only a proportion of $\bar{\xi}(0.35) = 0.2$ was observed, a difference of 0.15.

For every $p \in [0.1]$ the value of $\bar{\xi}(p)$ was plotted in an accuracy plot (see figure 4.1) against the true value,

with p on the x-axis and $\bar{\xi}(p)$ on the y-axis. The accuracy plot displays the predicted versus observed fractions of cross validation observations within a prediction interval for different p values. The $\bar{\xi}$ was averaged over the 100 10-fold cross validation and plotted as a single dot. The variance within these iterations was used to construct 95% confidence intervals, indicated with error bars above and below the points, on where the true value of the accuracy lies. A correct uncertainty model demonstrates a relationship of 1 : 1, i.e. the closer to the 45° bisector line ($\bar{\xi}(p) \approx p$) the better the quality of the uncertainty quantification. Falling above the 1 : 1 line indicates overestimation of the PI widths and falling below an underestimation; too pessimistic versus too optimistic. Additionally, a threefold of numerical measures for the deviation of the points from the 45° bisector line were derived, following Wadoux et al. (2018). First, the absolute deviation A_d was measured that determines the total area of PI over- and underestimation together. In an ideal scenario A_d should equal 0. A_d is defined according to the following equation (Eq. 49):

$$A_d = \int_0^1 \left| \bar{\xi}(p_k) - p_k \right| \, \mathrm{d}p \tag{49}$$

Two additional over- and underestimation metrics were also used, these metrics provide the percentage of under- (P_u) or overestimation(P_o) of the total absolute deviation A_d. These are given in the following equations (Eqs.50 & 51):

$$P_u = \frac{1}{A_d} \int_0^1 \left| \bar{\xi}(p) - p \right| \cdot \mathbb{1}_{\bar{\xi} > p} \, \mathrm{d}p \tag{50}$$

$$P_o = \frac{1}{A_d} \int_0^1 \left| \bar{\xi}(p) - p \right| \cdot \mathbb{1}_{\bar{\xi} < p} \, \mathrm{d}p \tag{51}$$

Where the indicator functions $\mathbb{1}_{\bar{\xi}>p}$ and $\mathbb{1}_{\bar{\xi}<p}$ are defined as giving a 1 when the logical condition in the subscript does holds and a 0 if the logical condition in the subscript does not hold.

Only a finite number of p values was studied, hence the over- and underestimation equation above (Eqs. 51 & 50) were not integrated but summed to approach them numerically. All metrics above were averaged over the total 100 replications of 10-fold cross validation.

Suppose that two models achieve similar accuracy and both $\xi(\bar{p}) \approx p$, then the model with the narrowest width of the probability intervals should be regarded as the best model for that probability. As a rule-of-thumb: the lower the width while staying above the 1 : 1 line, the better the model. Therefore, the models were also assessed on their average width using a similar notation as in Goovaerts (2001). All sites were assessed: $\{u_i : i \in 1, \ldots, n\}$ the set of all locations. For any p-PI, the average width $\bar{W}(p)$ is defined as the difference between the lower and upper boundaries for a given probability level p:

$$\bar{W}(p) = \frac{1}{n} \sum_{i=1}^{n} W(u_i; p) \tag{52}$$

Where the width is defined as in Eq. 44. The mean widths at probability levels $\{p_k : k \in 1, \ldots, K\}$ were used in the equation above (Eq. 52). The results were subsequently averaged over all 100 10-fold cross validations. This average width was then plotted against p and interpreted visually.

4.2.3 Mapping spatial outliers

Some point locations or groups of locations can be more prone to falling outside PI width estimates. A measure was developed to identify such locations.

Let $u_{t,i}$ denote an arbitrary location in the i^{th} test fold sample and let $z(u_{t,i})$ be its observed target value. Let W (without a bar) be defined as in Eq. 44 as the prediction interval width of indexed location $u_{t,i}$ at probability level p. Then, the lower boundary is defined as $Q_l(u_{t,i};p) = \hat{F}^{-1}\left(u_{t,i}; \frac{(1-p)}{2}\right)$ and $Q_u(u_{t,i};p) = \hat{F}^{-1}\left(u_{t,i}; \frac{(1+p)}{2}\right)$ denotes the upper boundary. Then, let $R = 100$ denote all iterations of 10-fold cross validation and the root mean squared normalized distance from PI (RMSND_{PI}) (Eq. 53) can be defined as:

$$\text{RMSND}_{\text{PI}}(u_{t,i};p) = \sqrt{\frac{1}{R}\sum_{i=1}^{R}\Bigg[\underbrace{\left(\frac{z(u_{t,i}) - Q_l(u_{t,i})}{W(u_{t,i};p)}\right)^2 \cdot \mathbb{1}_{z(u_{t,i})<Q_l(u_{t,i})}}_{\text{Lower boundary test}} + \underbrace{\left(\frac{z(u_{t,i}) - Q_u(u_{t,i})}{W(u_{t,i};p)}\right)^2 \cdot \mathbb{1}_{z(u_{t,i})>Q_u(u_{t,i})}}_{\text{Upper boundary test}}\Bigg]} \tag{53}$$

where indicator functions $\mathbb{1}$ returns a 1 if the conditions are met and a 0 otherwise. Note that only one boundary test fraction within the sum can return a value other than 0 at a time; and both tests always return a 0 if the target value at the respective location falls within the prediction interval. The normalized deviation from PI was calculated only for a probability level of p =90% (i.e. the 0.9-PI).

An additional measure of under- or over estimation at specific sites was also performed that looks for consistently bad geographic regions of the uncertainty quantification methods. Over 100 iterations it was counted how many times an observed target value fell above or below the respective upper and lower boundaries of the prediction interval according to the following formula:

$$B_{PI}(u_{t,i};p) = \frac{1}{R}\sum_{i=1}^{R}\left(\mathbb{1}_{z(u_{t,i})>Q_u(u_{t,i})} - \mathbb{1}_{z(u_{t,i})<Q_l(u_{t,i})}\right) \tag{54}$$

where the same boundary tests are defined by the indicator functions as in Equation 53 that return a 0 when the target value falls within the prediction interval, a 1 if it falls above and -1 when it falls below the PI. Again, this was done for all locations and mapped on a blank canvas to look for spatial patterns of over- and underestimation.

Ultimately, the spatial outliers analysis was used to cross-check the validation plots and to look for underlying erratic behaviors the uncertainty quantification models QRF and RK. Locations that fall consistently outside of a prediction interval score high on RMSND_{PI} were depicted with a relative higher circle size. Whether it fell, on average, below or above the prediction interval was depicted in red and green, respectively.

4.3 Geographic interpretation

For the geographical interpretation a measure of importance was used. This importance measure is defined as the decrease in node purity (see section 2.2 for more information) when leaving out the covariates one-by-one. The covariates that lead to the largest decrease in node purity were considered to be most important for the modeling process. These covariates were ordered by importance and summarized in a bar plot. Performance of the techniques was evaluated and regions that fell consistenly outside the prediction intervals were investigated and interpreted by comparing them visually with their underlying covariates. The covariate importance was implicitly linked with this interpretation process. Initially, local hotspots that share high uncertainty overlap in the methods were analyzed as these hotspots could be related to uncertainties in the input data. By overlaying several covariate layers it may be determined where this uncertainty originates from. Next, a detailed comparison was done on differing

areas aimed to identify possible model related uncertainties, identifying where QRF outperforms RK and vice versa.

4.4 Scalability assessment

After the determination of how the model performed on a small case study the relation between computation time and number of covariates and number of points was studied and compared between the methods. First, the cell size was changed to increase the number of points to predict. Second, the amount of covariates was increased. Total processing time was then plotted against these two measures for both QRF and RK. The model lowest in processing times was regarded as more scalable.

4.5 Materials

4.5.1 Hardware

The modelling process was run on a Intel(R) Dual Core(TM) i7-3520M CPU @ 2.90GHz processor with a total of 4 threads; the total RAM was 8GB with at least 5GB of free RAM before the processes were run.

4.5.2 Software

R 64-bit version 3.3.3 (2017-03-06) was used as programming interface. For the Random Forest modeling process the "Ranger" package (M. Wright & Ziegler, 2015) was used that optimally uses all threads of the processor. This package includes an option to keep the information in the end nodes (leaves) that was used in the QRF uncertainty modelling procedure. For the RF model used by kriging a separate model was trained on the same random seed as the QRF forest but without the option to conserve all target values in the end leaves. This was mainly done in order to get optimal performance times during the benchmarks. Interpolation with ordinary kriging of the RF residuals was done using the package "gstat" (Pebesma, 2004).

5 SOIL PROPERTY CASE STUDY

The selected study site is located in Australia in the region of Edgeroi, about 500 km north-east from Sydney (see Figure 5.1). The study site lies in a valley of of the Namoi River and has a surface area of around 1500 km^2. Most of its land use is agriculture with cotton and wheat as main crops and pasture. An elevated zone in the east of the study site serves corresponding to a vegetation dominated by native plants. These elevated zones form the lower foothills of the Nandewar Range (Malone et al., 2009). The chosen case study has an extensive dataset of sand, clay, silt, soil organic carbon (SOC) and pH observations. The dataset has a total of 359 sampling locations, consisting of non-harmonized horizon measurements of non-uniform depth intervals. The soil profile at each location was classified according to the Australian soil taxonomy (see Isbell (2016) for more details on the taxonomy).

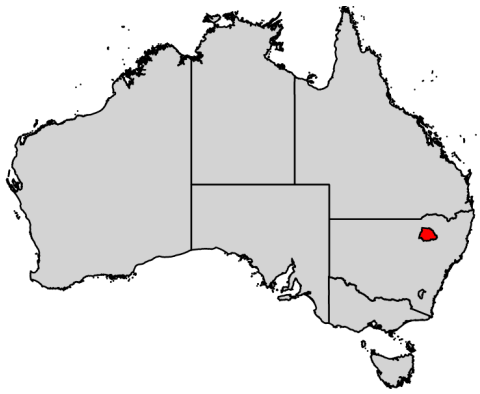


Figure 5.1. Narrabri (red), Australia, where the study site is located.

5.1 Soil property and covariate selection

Two continuous soil properties were chosen: soil pH and soil organic carbon content (SOC). pH was measured in a 1:5 suspension with H2O and SOC was measured in mg/g dry soil. pH was chosen because it is a frequently measured soil property, hence it is easy to compare with other study results and SOC because it is closely related to climate change functional ecosystems, hence crucial for food, soil and water security (Stockmann et al., 2015).

Hengl, et al. (2017) list many of the covariates derived from remote sensing products as layers to supply the algorithm with training information on the soil forming factors. These covariates were acquired through ISRIC's WorldGrids covariates database at a resolution of 1000m. Table 5.1 below gives an overview of all selected input layers that are candidates as predictor variables in the random forest algorithm. One Random Forest model was built on 4 randomly selected covariates and one on all 14 covariates. These two covariate numbers provide the RF models with high or low amounts of variation explanation. Testing different settings supported the investigation on how QRF and RK react to different circumstances and test whether it affected their uncertainty modeling.

Table 5.1. Environmental covariates with sources, grouped by soil forming factor (long term average).

Forming factor	Derived Covariates	Code	Source data
Location	Topographic openness	OPISPRE	SRTMGL3 & ETOPO DEM
	Topographic Wetness Index	TWISRE	
Relief	Digital elevation model	DEMSRE	SRTDMGL3 DEM
	Slope	SLPSRT	
Vegetation	Mean monthly Enhanced Vegetation Index (EVI)	EVMMOD	MOD13A3
	Monthly standard deviation EVI	EVSMOD	
	Standard deviation 8-day Leaf Area Index (LAI)	LASMOD	MOD15A2
Climate	Mean potential solar radiation	INMSRE	SRTMGL3
	Standard deviation potential solar radiation	INSSRE	
	Long term estimated evapotranspiration	ETMNTS	MOD16
	Mean 8-day day surface temperature	TDMMOD	MOD11A2 LST
	Mean 8-day night surface temperature	TNMMOD	MOD11A2 LST
	Monthly mean precipitation	PREGSM	WorldClim & GPCP2.2
Parent material & Soil properties	Soil survey point data	PHIHO5 & ORCDRC	GSIF Edgeroi dataset
	Lithological age	GEAISG	USGS Surface Geology Map

5.2 Preprocessing

5.2.1 Top soil profile harmonization

A two dimensional model of the top soil, defined as 0-30 cm measured from the top, was made by fitting a mass preserving spline. This harmonization was needed as the Edgeroi dataset contains measurements of non-uniform depth profiles that corresponding to horizons. A mass preserving spline (2009) is generalization of the quadratic spline, popularized in soil profile context by Bishop et al (1999), and can deal with missing values in between measurements of soil depth intervals. In essence, the mass preserving spline tries to fit an as smooth as possible curve through the centers of the horizons within a soil profile without the loss of mass. In other words, mass preserving means that the area under the spline between two different depth boundaries should be equal to the area of its corresponding measurement at this depth interval. The Figure below (5.2) shows an example of a disjoint soil profile and a mass preserving spline fitted through the centers of its horizons.

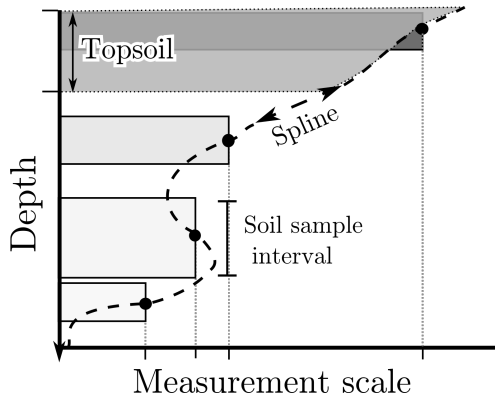


Figure 5.2. Example of a mass preserving spline

The fitting of a mass preserving spline depends on balancing two different quantities: the first term is called the fidelity, the closeness of the spline to the observations and the second term represents the roughness where a lower value will be deemed more realistic as it smooths the curve. The smoothness is controlled by a parameter named λ that controls the trade-off between the two terms. Malone et al. (2009) already reported a tuned λ for the Edgeroi study site and therefore, this $lambda = 0.1$ was used for the soil profile harmonization process. Subsequently the depth interval could be filled in using the GSIF package (Hengl, Kempen, et al., 2017) and a harmonized value of the top soil was obtained.

Figures 5.3 and 5.4 show the spatial distribution of top soil pH and SOC respectively.

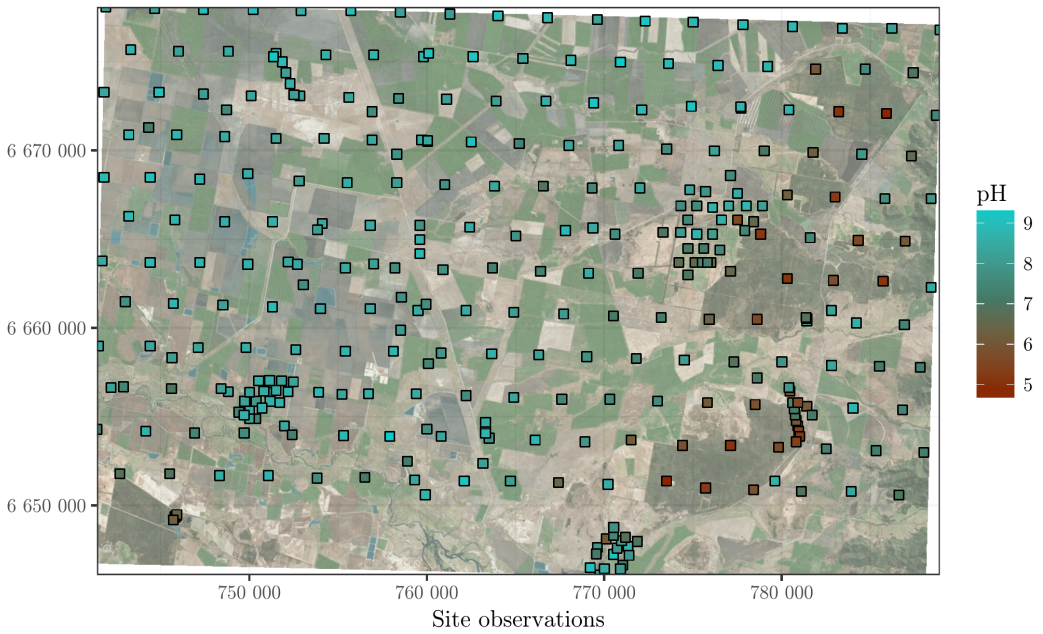


Figure 5.3. Edgeroi top soil pH observations (for satellite background photo see Schwartz (2009))

5.2.2 Variogram modeling

Regression kriging requires a variogram model to model the spatial correlation of the residuals. First, all covariate data and site data were projected to the Australian Albers GDA94 national coordinate system (datum close to WGS84) with distance units of meters. Then, a variogram cloud was build on the Random Forest residuals. For the estimation of the variogram model parameters a Restricted Maximum Likelihood (REML) was applied, the parameters are further specified in Table 5.2 below. The modeling was done for both pH and SOC on 4 and 14 randomly selected covariate RF modeling residuals.

Table 5.2. Variogram parameters

Covariates	Soil property	Variogram model	Nugget	Sill	Range
14	pH	Pure nugget	0.48	N/A	N/A
	SOC	Pure nugget	17.8	N/A	N/A
4	pH	Exponential	0.56	0.01	8000m
	SOC	Exponential	20.9	4.5	8000m

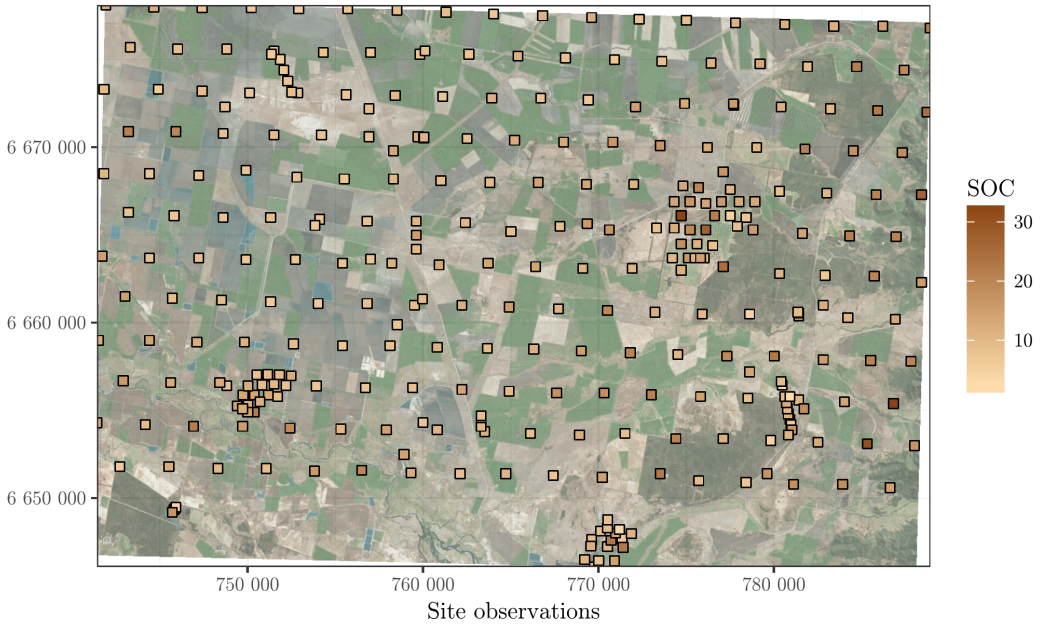


Figure 5.4. Edgeroi site observations of soil organic carbon content (SOC in mg/g) (for satellite background photo see Schwartz (2009))

5.3 Results

5.3.1 Soil pH

General performance Random Forest

The Random Forest model for pH with 14 covariates had a k-fold average predictive R^2 of 0.50 against an R^2 of 0.41 for 4 covariates. The average root mean squared error is also smaller for 14 covariates, 0.67 versus 0.78. From now the 14 covariate model will be called RF14 and the 4 covariate model RF4.

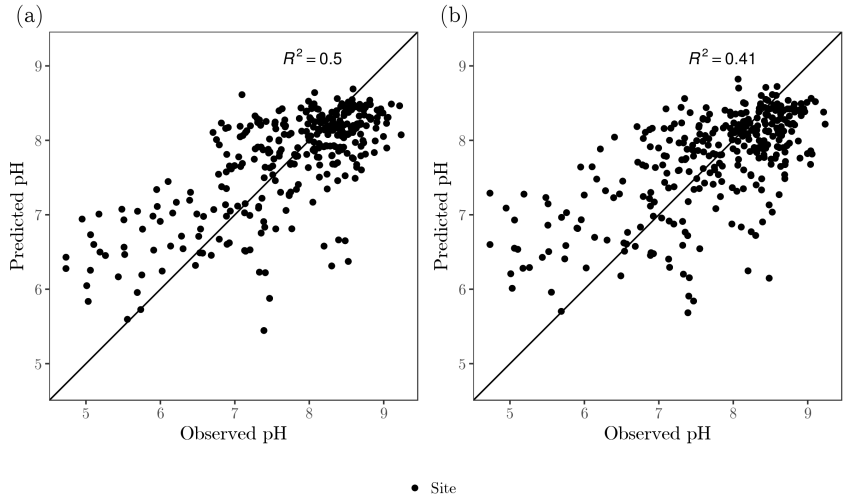


Figure 5.5. Overall performance of the Random Forest soil pH predictions. (a) Observed versus predicted pH (14 covariates); (b) Observed versus predicted pH (4 covariates).

Figure 5.5 summarizes the performance of the Random Forest model for soil pH; in (a) the observed pH versus the prediction is plotted for RF14, is clustered more around the $1:1$ line than RF4 (b). Especially low pH values show higher variation for 4 covariates.

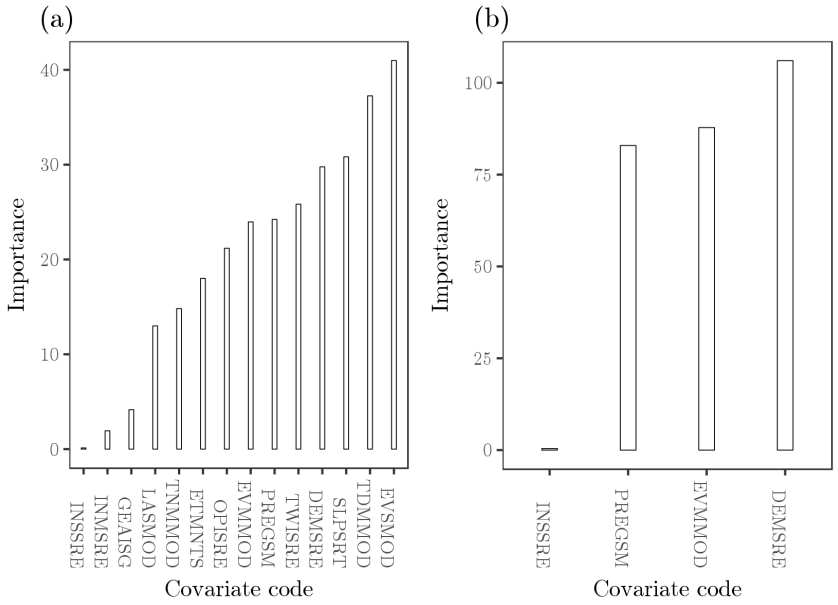


Figure 5.6. Variable importance of the Random Forest model for soil pH.
Variable importance defined as the increase in error when leaving out the covariate.
(a) 14 covariates (RF14); (b) 4 covariates (RF4).

With 14 covariates, the variable importance is unevenly distributed (Figure 5.6). The standard deviation of the monthly vegetation index (EVSMOD) and the 8-day average land surface temperature (TDMMOD) are on the high side of the spectrum. Geological age (GEAISG) and potential incoming solar radiation (INMSRE, INSSRE) were on the low side of the spectrum. For 4 covariates, all selected covariates are important with the exception of the potential incoming solar radiation (INSSRE).

Uncertainty mapping

Figure 5.7 shows lower and upper boundaries for pH of the 0.9 p-PI for the RF14 model. The pH 0.05 and 0.95 quantiles range from a minimum of 4.5 pH to over 10 pH. The lower quantile map (a) of QRF predicts that the eastern part of the study area could be low in pH, i.e. the south-east is acidic. In comparison, RK predicts that the east is slightly higher in pH (alkaline) looking at lower boundary map, with values differing up to 1.5 pH difference locally between the two methods. The lower quantile in the north-east is noticeably higher in pH for quantile regression forest than for regression kriging, i.e. higher acidity. The prediction maps (b) in Figure 5.7 for QRF and RK resemble each other closely, there are no trends visible. This similarity is expected because only the kriged mean of the residual is added to the kriging model which should be close to 0 as residuals displayed a normal distribution centered around 0. The upper boundary map (c) of the 0.9 p-PI for regression kriging estimates the center and north east of the map to be much higher in pH than its QRF counterpart (more alkaline). In the south-east, however, RK predicts that some regions are lower in pH than the QRF estimates (brighter red); for QRF the complete range of values falls within 3 pH units, while regression kriging has differences totaling up to 4 units. Nevertheless, over all three maps in general, QRF expects the area to be lower in pH than RK, i.e. lower alkalinity and higher acidity.

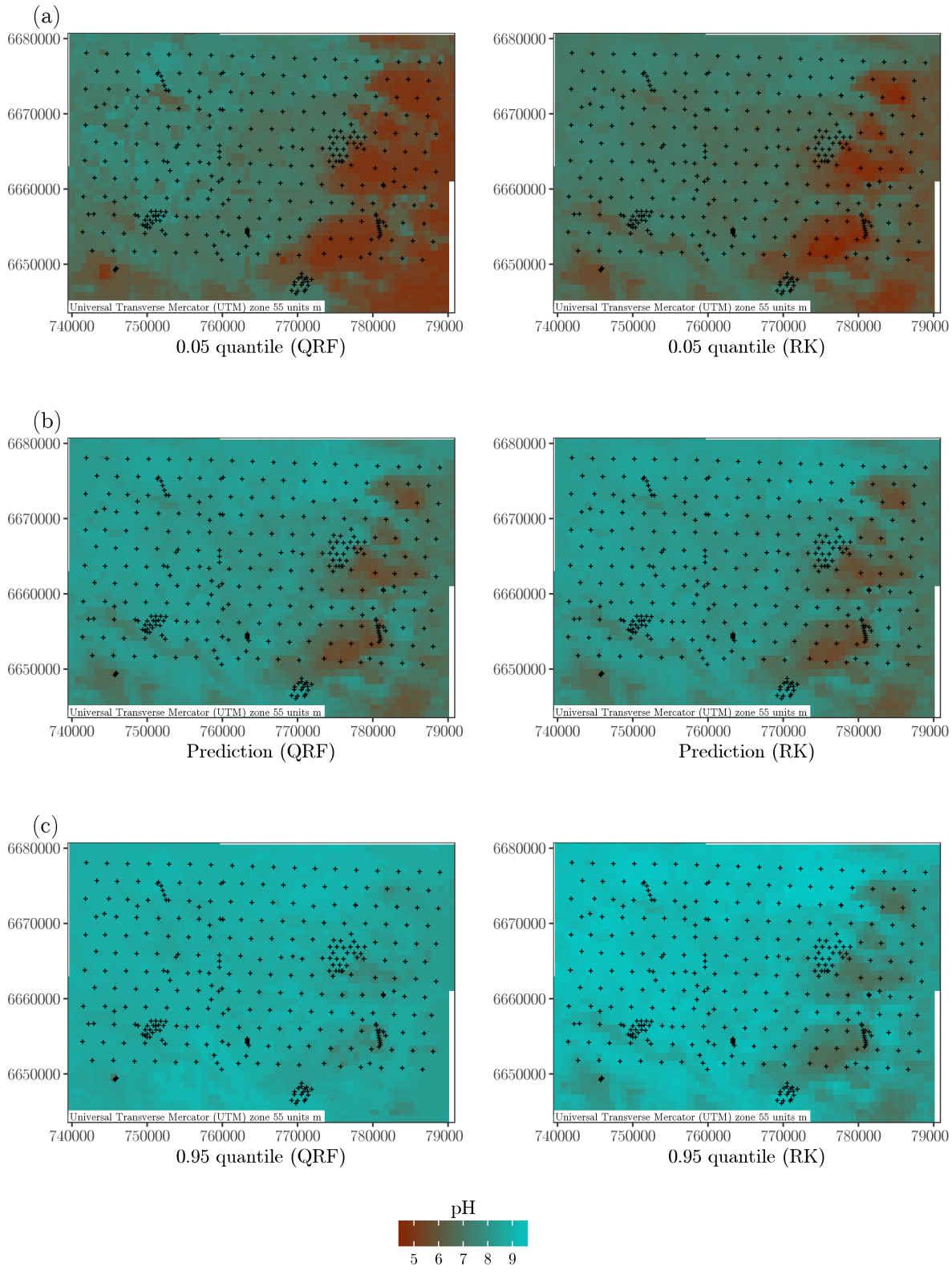


Figure 5.7. Maps of the 0.9 prediction interval boundaries for pH (14 covariates).
Quantile regression forest (left) and regression kriging (right). Values below 7 pH, indicating acidity, are red; values above 7 pH, indicating alkalinity, are blue.
(a) The 0.05 quantile depicts the predicted lower boundary of the 0.9 p-PI, on average 1 out of every 20 predictions should fall below these pH values; (b) The expected value of the RF prediction that minimizes the squared error; (c) The 0.95 quantile depicts the predicted upper boundary of the 0.9 p-PI, on average 1 out of every 20 predictions should fall above these pH values.

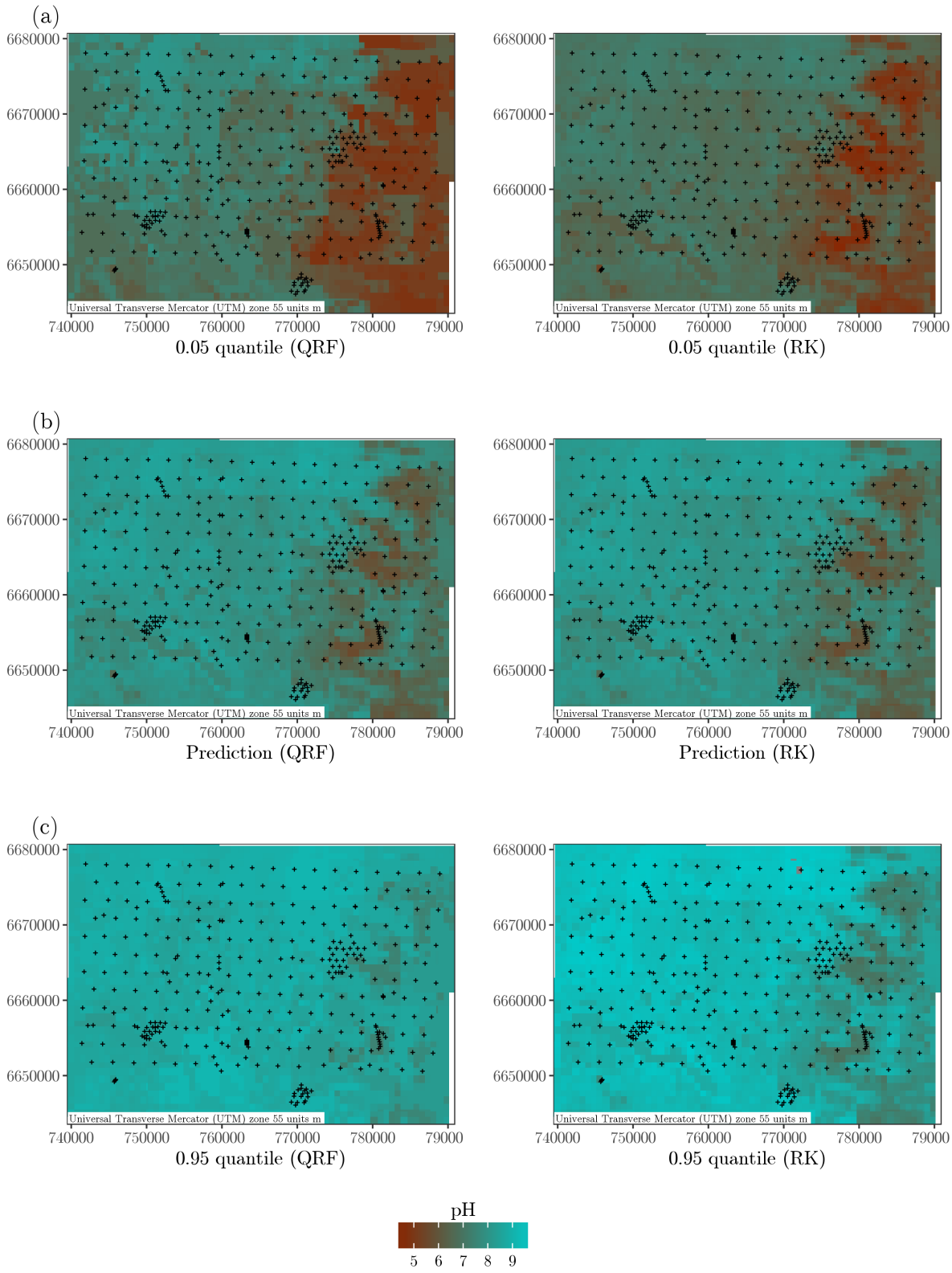


Figure 5.8. Maps of the 0.9 prediction interval boundaries for pH (4 covariates).
Left shows quantile regression forest (QRF) and right shows regression kriging (RK). Values below 7 pH, indicating acidity, are red; values above 7 pH, indicating alkalinity, are blue.
(a) The 0.05 quantile depicts the predicted lower boundary of the 0.9 p-PI, on average 1 out of every 20 predictions should fall below these pH values; (b) The expected value of the RF prediction that minimizes the squared error; (c) The 0.95 quantile depicts the predicted upper boundary of the 0.9 p-PI, on average 1 out of every 20 predictions should fall above these pH values.

Figure 5.8 shows the lower and upper quantiles for pH at the 0.9 p-PI for RF4. The main difference between RF14 and RF4 is in the prediction map, with RF4 being being lower in pH than RF14; the red regions are more abundant. The expected pH values of northern part of the study area are also different; RF14 predicts this area to be more alkaline (both RK and QRF) instead. The boundary maps (a) and (c) are similar to the RF14 maps of Figure 5.8, apart from the more coarse patterns (pixel effect). Both quantile regression forest and regression kriging in figure 5.8 show lower pH values (i.e. less alkalic) in the north-west. QRF seems to estimate more extreme differences than RK for the lower boundary. The upper boundary maps (c) show close resemblance between QRF and RK.

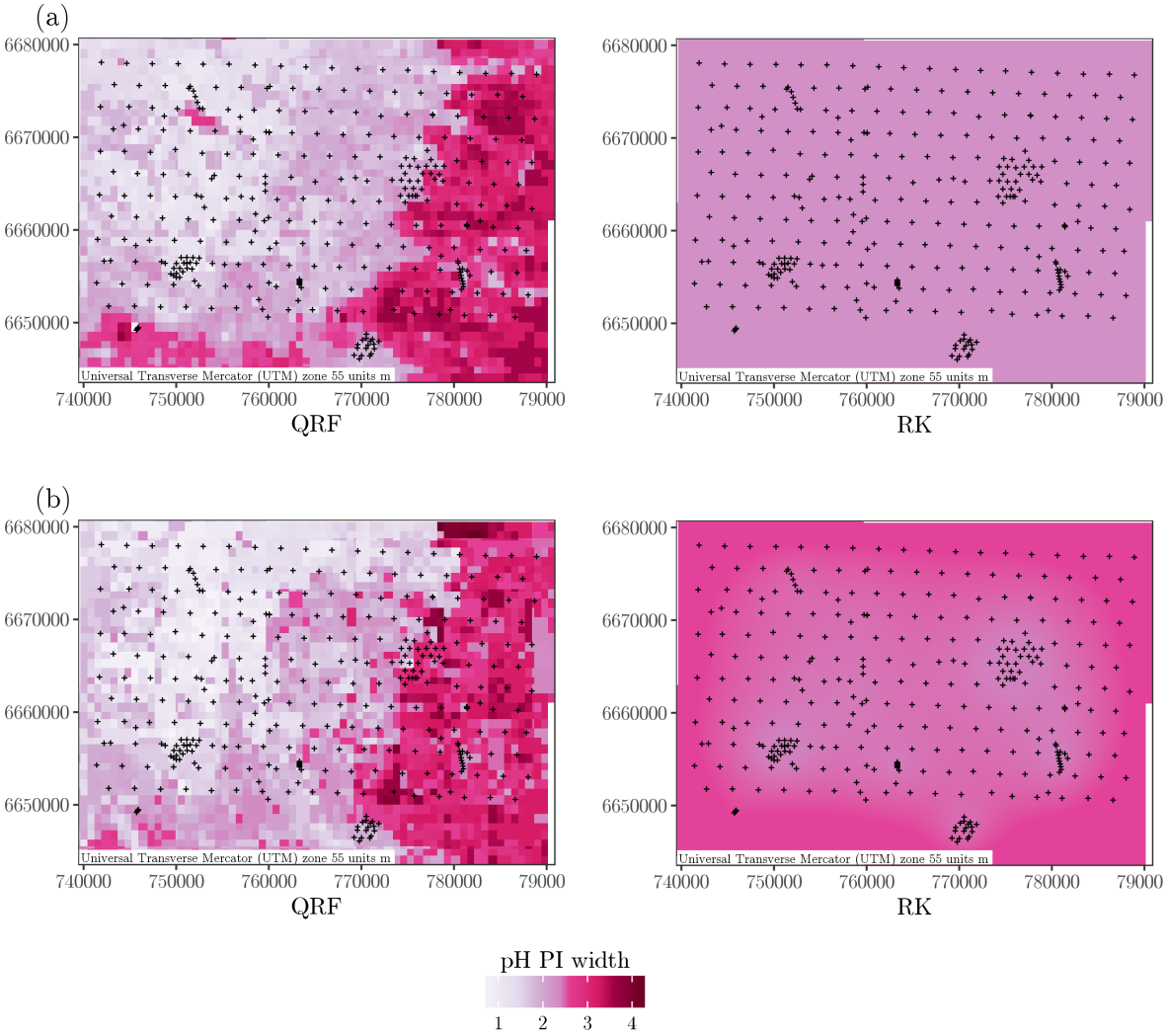


Figure 5.9. Prediction interval width maps of the 0.9 prediction interval for pH.
On the left quantile regression forest and on the right regression kriging. Color scale from light purple to dark red on a quantile based scale. (a) presents the respective widths for 14 covariates and (b) for 4 covariates.

Figure 5.9 above shows the total width of the 0.9 p-PI per pixel in the study area by the RF14 (a) and RF4 (b) models respectively. QRF shows a noticeable pattern of uncertainty for both RF14 and RF4, while RK shows an equal PI width map in figure 5.9 (a) and almost equal PI width map in (b). The boundary transitions in QRF have prediction interval widths that sometimes differ up to 2 pH within 1000 meters, this coincides with a pH transition from alkaline to acidic. The structure of PI width

patterns of QRF within the two figures is very similar, with the most noticeable difference in the south, where the RF14 predicts PI widths of almost 1 pH more than its RF4 counterpart. This means that more explanatory information led to higher uncertainty. Comparing the PI width figures with the uncertainty boundary maps above (Figures 5.7 & 5.8), the most wide PI widths are found where the RF models predict high acidicity and lowest where the model predicts high alkalinity, at least for QRF. RK predicts that, based on the spatial dispersion of the residual, PI width will stay more or less the same throughout the study area. RF4 uncertainty decreases marginally (<0.2 pH locally) when approaching the site observations. Note that the number of site observations that had acidic measurements were less abundant than alkaline measurements. Geographically most of the uncertainty seems to stem from EVSMOD (standard deviation of EVI), the covariate values of EVSMOD is an important predictor (Figure 5.6 (a) left) and shows very low values in the east that coincide visually with the width map (compare Figures 5.9 with A.1 in appendix A). As both width maps (a) and (b) resemble each other it seems reasonable to assume EVSMOD was a major culprit for the RF14 uncertainty as well (being the most important covariate).

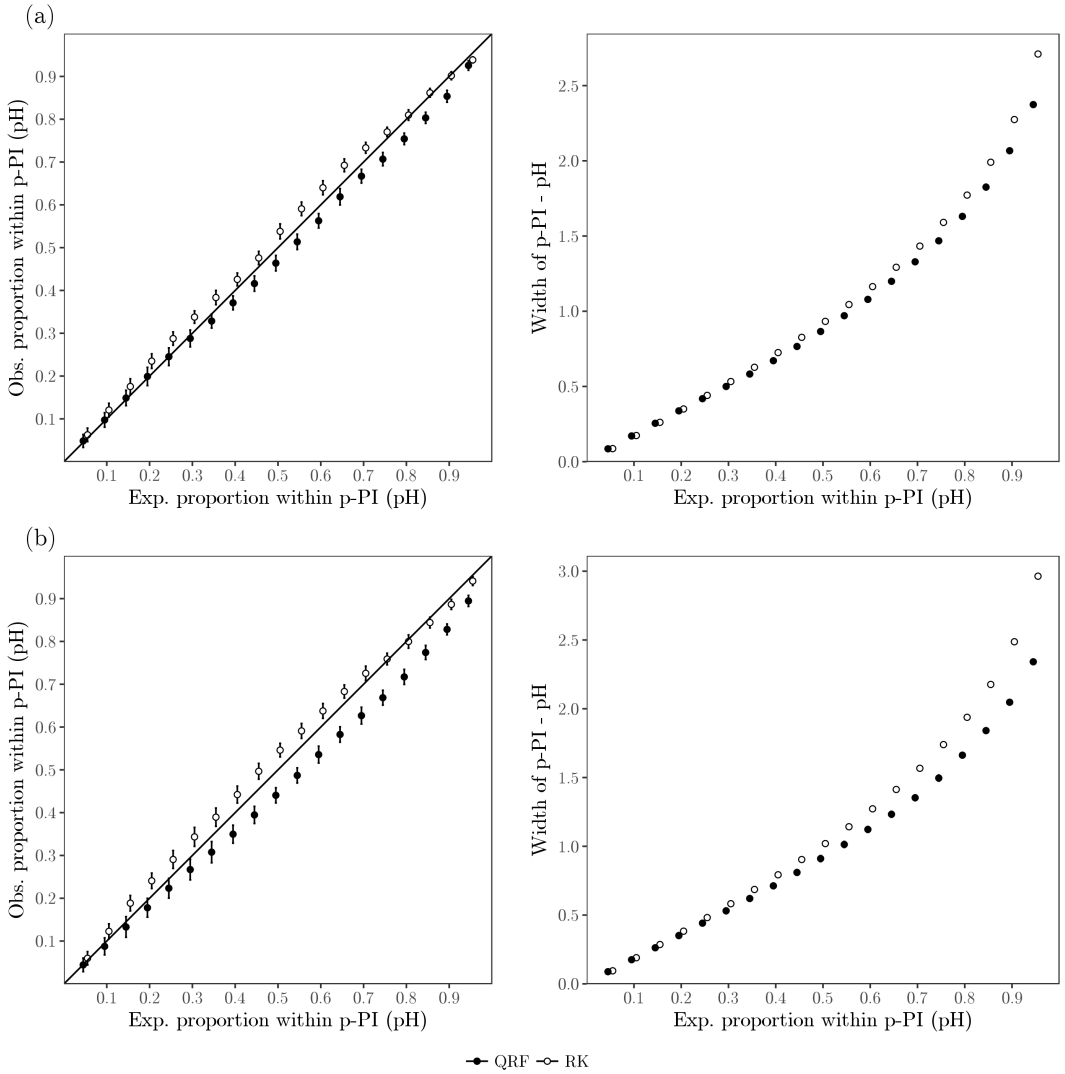


Figure 5.10. Validation plots for all p-PIs of soil pH (10 k-fold, 100 iterations).
(a) Validation plots for 14 covariates; (b) Validation plots for 4 covariates. Accuracy plot (left); PI width (right). Accuracy plot shows expected proportions within interval versus observed predictions within interval; Prediction interval width plot shows the probability level versus the PI width

Prediction interval validation

Figure 5.10 (a) RK is consistently slightly above the 45 deg bisector line and QRF often falls slightly below it for RF14 and for RF4 this effect is increased for QRF (not RK). The error bars (95% confidence interval) of QRF are wider on average than RK, especially for RF14 which means that the deviations from the 1 : 1 line were more dependent on both the validation fold that determined what data the model was calibrated on. The accuracies for both the lower p-PIs as the higher p-PIs of QRF and RK lie close to each other; most of the differences in accuracy are found around the 0.3 - 0.7 p-PIs. In general, RK seems to be more optimistic than QRF (see P_0 Table 5.3), but at least for RF14 equally apart from the 1 : 1 line - judging from the absolute deviation A_d. This over- versus understimation is supported by the width plots (Figure 5.10 right): the widths of RK are consistently higher for RK than for QRF. RK widths increase faster for higher p-PIs reaching a maximum difference with QRF of 0.25 pH for RF14. A noticeable effect for both RF14 and RF4 is seen on the higher probability levels, where both accuracies are equal for QRF and RK but the width of QRF is (substantially) lower.

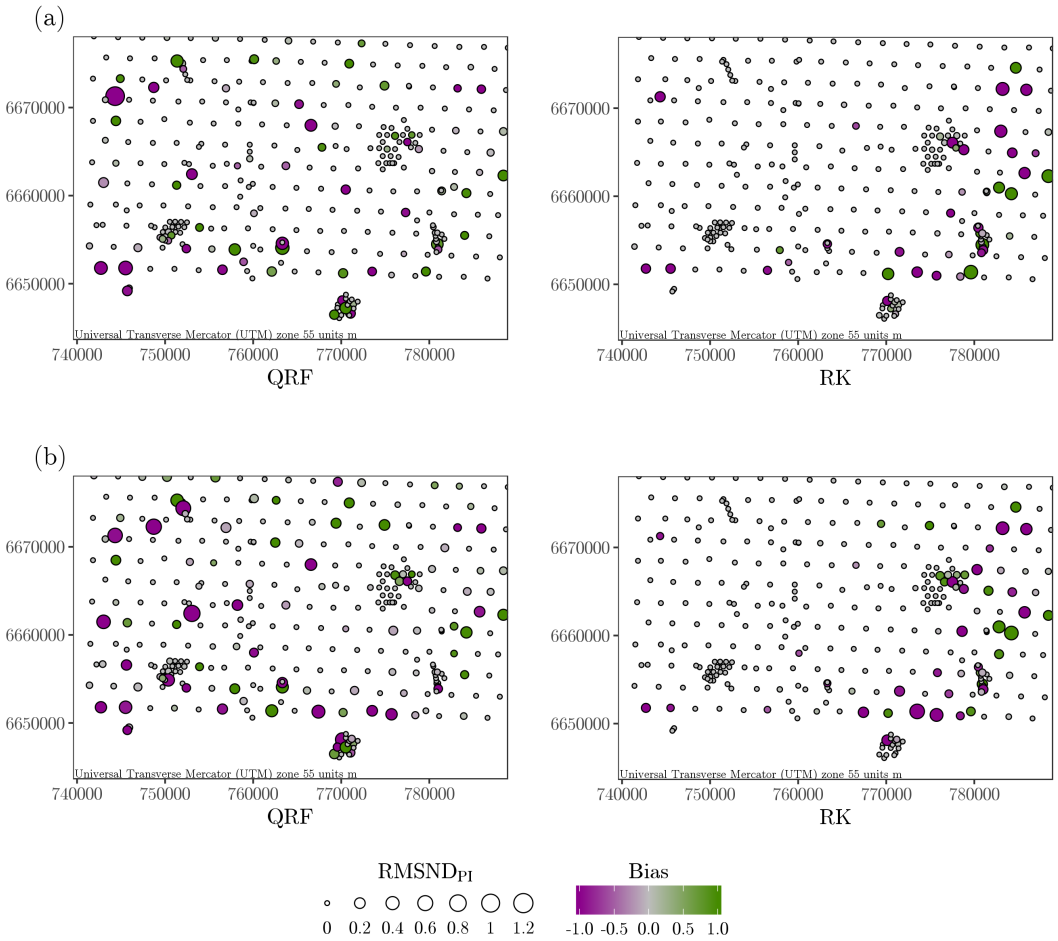


Figure 5.11. Spatial outliers for every site observation of the 0.9-PI for pH.
(a) Spatial outliers for 14 covariates; (b) Spatial outliers for 4 covariates. Quantile regression forest (left); regression kriging (right). Circle size indicates the RMSND$_{PI}$ for 0.9-PI of all 100 10-fold cross validation results. Bias shows the average proportion of under- or overestimation: -1 for all observations smaller than 0.9-PI (green); +1 for all observations larger than 0.9-PI (pink).

Zooming in on a local level, Figure 5.11, shows a dispersal of locations with a high RMSND$_{PI}$ for QRF, whereas the RK consistently high RMSND$_{PI}$ locations are located mainly in the south-east, coinciding

with the acidic predictions. Regression kriging underestimates the pH-value of the sites in the center of the acidic regions, while the boundaries of these acidic regions are often overestimated (positive bias). Apart from these technique specific outcomes, the differences between the RF14 (a) and RF4 (b) are minimal. The regional differences cannot be found on the accuracy plots (Figure 5.10).

Table 5.3. PI estimate validation summary for pH

Covariates	Method	Absolute deviation (A_d)	Proportion overestimation (P_o)	Proportion underestimation (P_u)
14	**QRF**	2.4%	0	1
	RK	2.5%	0.98	0.02
4	**QRF**	4.8%	0	1
	RK	2.7%	0.95	0.05

5.3.2 Soil organic carbon

General performance Random Forest

The Random Forest model for soil organic carbon content predictions with 14 covariates has a R^2 of 0.32 against an R^2 of 0.08 for 4 covariates; both are considered low. The average root mean squared error of the for 14 covariates 3.88 is slightly smaller than the 4 covariate RF model 4.17. Furher on these covariate models will be denoted by RF14 and RF4. Figure 5.12 summarizes the performance of the Random Forest model for soil SOC; in (a) the observed SOC versus the prediction is plotted. As expected, RF14 is clustered more around the 45 deg bisector line than RF covariates (b). Especially lower pH values had higher variation for RF4 covariates.

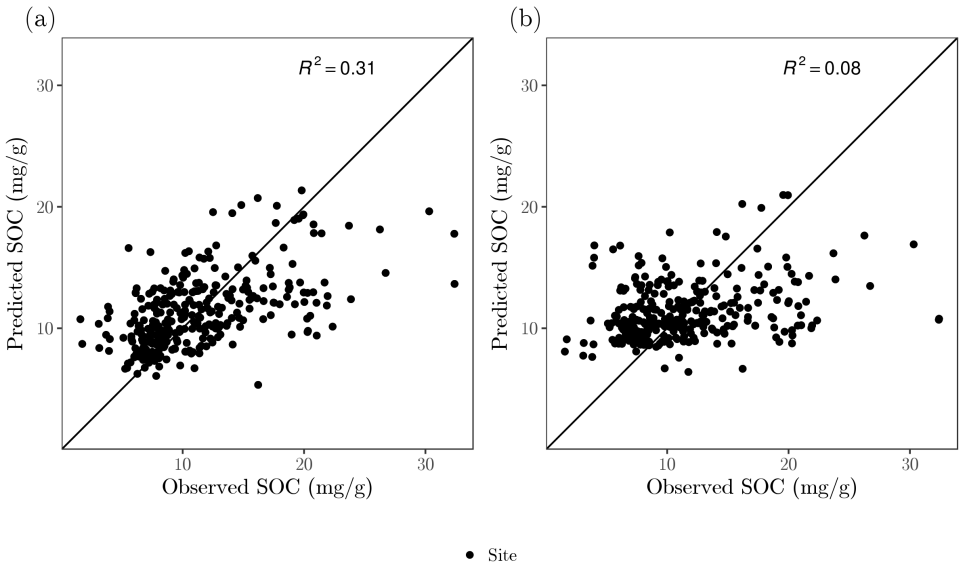


Figure 5.12. Overall performance of the Random Forest soil organic carbon predictions (14 covariates).
(a) Observed versus predicted SOC (14 covariates); (b) Observed versus predicted SOC (4 covariates).

With 14 covariates, the variables of importance seem stratified into 3 classes of importance. On the top of error increase when leaving out the covariate are the elevation (DEMSRE), long-term evapotranspiration

(ETMNTS3) and mean monthly precipitation (PREGSM1). On the lower end of the spectrum are the potential surface are, just as for pH, eological age (GEAISG) and potential incoming solar radiation covariates (INMSRE, INSSRE). For RF14, the mean potential incoming solar radiation seems to be ineffective for predicting organic carbon content of the soil and especially the slope seems an important covariate that mainly depends on how well the regression does.

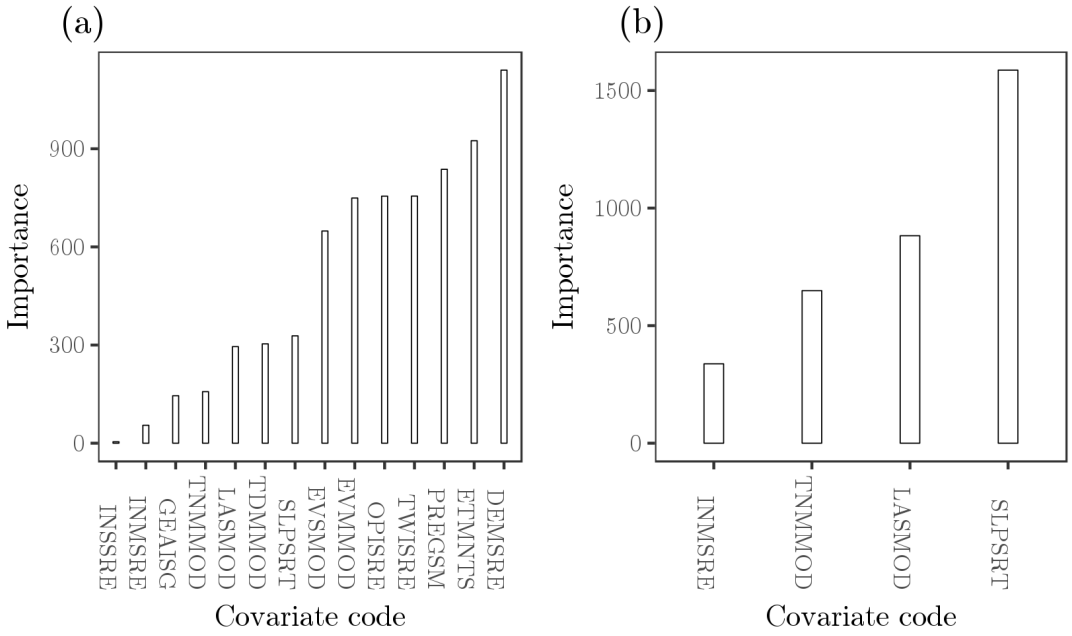


Figure 5.13. Variable importance of the Random Forest model for soil organic carbon (SOC). Variable importance defined as the increase in error when leaving out the covariate. (a) 14 covariates; (b) 4 covariates.

Uncertainty mapping

Figure 5.14 shows the estimated lower and upper quantiles for SOC of the 0.9-PI. The random forest model underlying these maps is trained on a combination of 14 different covariates The SOC boundary predictions range from a minimum of 0 to around 30 mg/g SOC.

The lower quantile map (a) shows most difference between the two methods for SOC. RK predicts that the western part of the study area has a much lower soil organic carbon content than QRF predicts. Differences between the two maps range up to 10 mg/g. QRF estimates that the lower boundary of SOC does not often fall below 10, while RK does. The prediction maps (b) in Figure 5.14 for QRF and RK resemble each other closely, there are no trends visible mainly because the mean of the predicted residual is close to 0. The upper boundary map of the 0.9 p-PI (c) displays a couple of regional differences, but the overall patterns are quite similar. The elevation ridge in the east is estimated to be much higher in soil organic carbon content according to QRF than to RK. Some typical block patterns around the center become visible for QRF, around the darkened pixels west of the center, which seem absent for RK. Moreover, quantile regression forest predicts some small regions in the north west and center of the study area much lower (lighter) than regression kriging. Overall the RK maps seem to be smoother than the pixelated maps of QRF, much more than for pH.

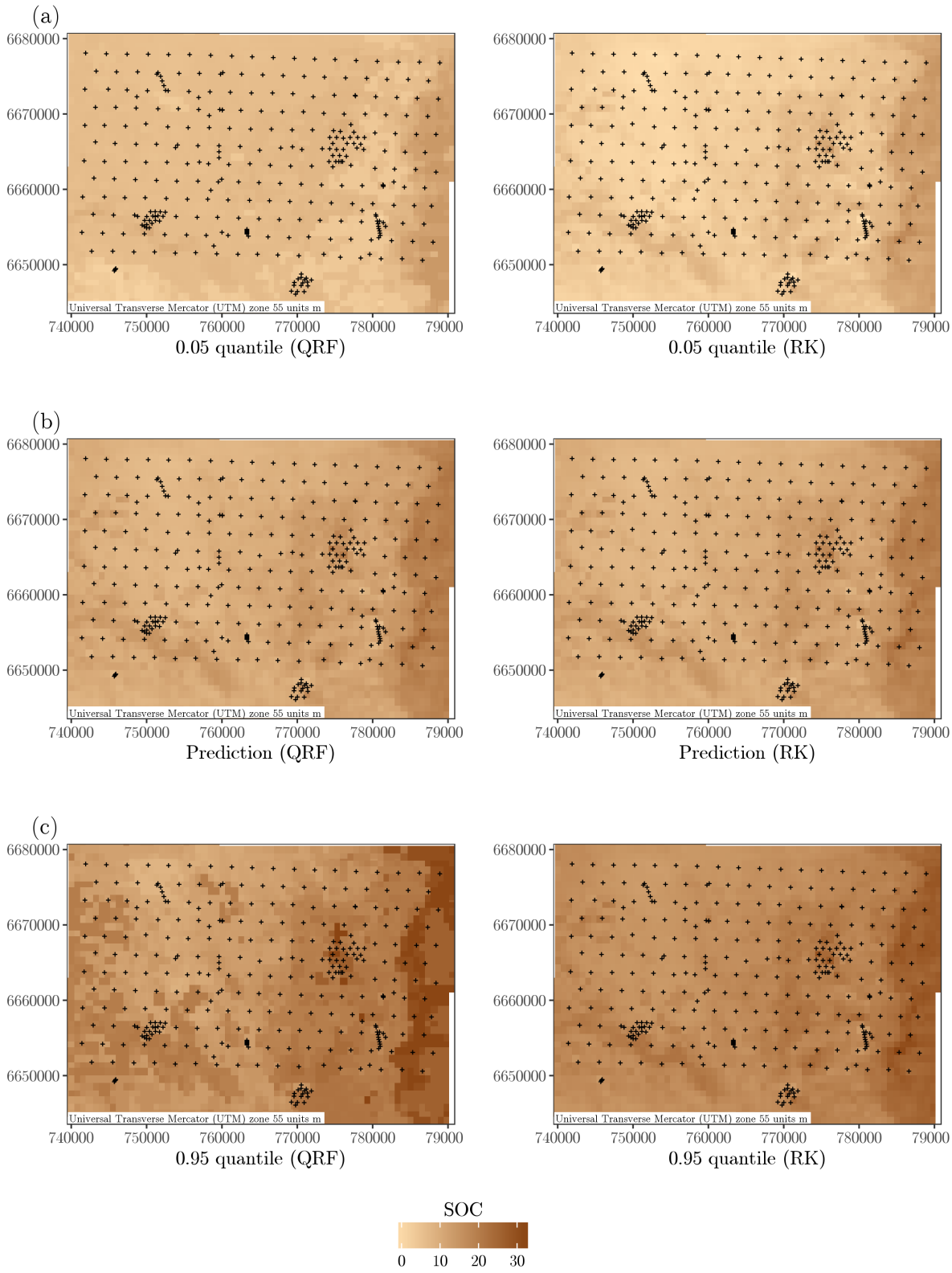


Figure 5.14. Maps of the 0.9 prediction interval boundaries for SOC (14 covariates).
Left shows quantile regression forest (QRF) and right shows regression kriging (RK). Values in mg/g of soil organic carbon content.
(a) The 0.05 quantile depicts the estimated lower boundary of the 0.9 p-PI, on average 1 out of every 20 predictions should fall below these SOC values; (b) The expected value of the RF prediction that minimizes the squared error; (c) The 0.95 quantile depicts the estimated lower boundary of the 0.9 p-PI, on average 1 out of every 20 predictions should fall above these SOC values.

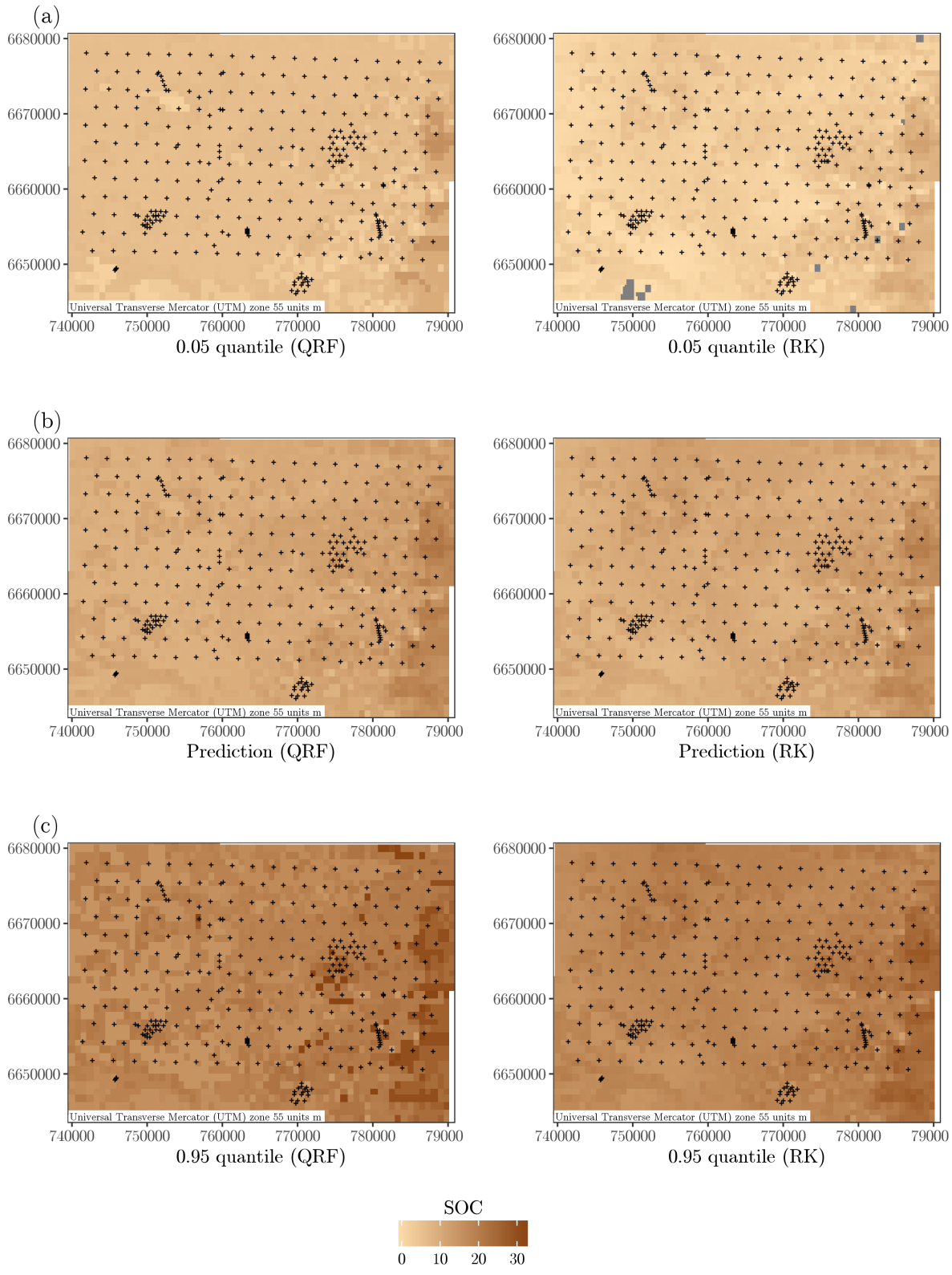


Figure 5.15. Maps of the 0.9 prediction interval boundaries for SOC (4 covariates).
Left shows quantile regression forest (QRF) and right shows regression kriging (RK). Values in mg/g of soil organic carbon content.
(a) The 0.05 quantile depicts the estimated lower boundary of the 0.9 p-PI, on average 1 out of every 20 predictions should fall below these SOC values; (b) The expected value of the RF prediction that minimizes the squared error; (c) The 0.95 quantile depicts the estimated lower boundary of the 0.9 p-PI, on average 1 out of every 20 predictions should fall above these SOC values.

Figure 5.15 shows the estimated lower and upper boundaries for SOC of the 0.9 p-PI for RF4. The maps and the differences between SOC and RK appear to be very similar to RF14. In general, both maps appear to be more pixelated, with an emphasis on QRF. The major differences between 14 covariates and 4 covariates can be found in the upper boundary maps (c). Especially QRF exhibits a certain speckle effect with differences locally (within 1000m) that can differ up to 10 mg/g SOC.

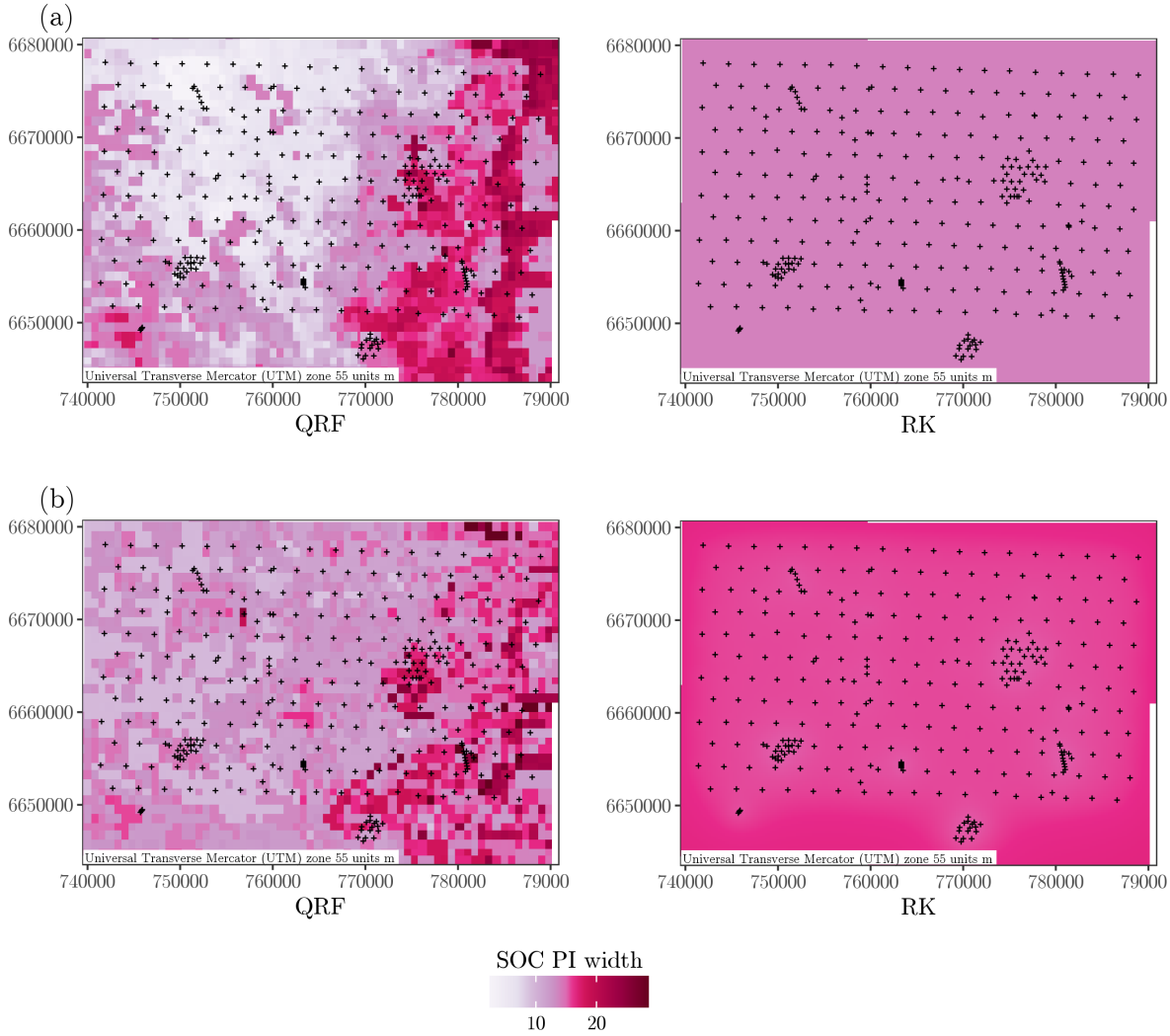


Figure 5.16. Prediction interval width maps of the 0.9 prediction interval for SOC.
On the left quantile regression forest and on the right regression kriging. Color scale from light purple to dark red on a quantile based scale. (a) presents the respective widths for 14 covariates and (b) for 4 covariates.

Figure 5.16 above maps the total width of the 0.9 p-PI per pixel in the study area for RF14 (a) and RF4 (b) respectively. QRF shows a noticeable pattern for both figures, while RK shows an equal PI width map for all locations for RF14 (a) and a near uniform width in the center for RF4 (b) with differences <2 mg/g at the edge of the study. In contrast, QRF estimates prediction interval widths that sometimes differ up to 10 mg/g SOC within 1000 meters. The structure of PI width patterns of QRF within the two figures is similar for the edges of the study area and some clear differences appear in the mid-center to mid-west. RF14 (a) displays a somewhat clear and smooth PI width transition pattern (5 mg/g), and RF4 shows some abrupt transitions in PI widths (10 mg/g). Geographically there does not seem to be

one major covariate that determines the uncertainty (compare Figure 5.16 with Appendix A) as seemed the case with pH. SOC uncertainty is more likely related to specific combinations between the covariates.

Uncertainty validation

Figure 5.10 shows the 10 k-fold, 100 iteration validation results for RF14 and RF4 SOC models respectively. In comparison with pH similar patterns of over- and underestimation appear. QRF seems to be too optimistic and has too small widths, while RK has the opposite. Table 5.4 confirms this, judging from comparing P_0 with P_u. Again, the absolute (A_d) deviations follow each other closely (e.g. 3.4% vs 4% for RF14). Although the R^2 was almost 4 times for RF14 as for RF4, the PI width (right) did not decrease substantially both for QRF and RK (2 mg/g for the higher p-PIs). Furthermore, the accuracy even decreased for both RK and QRF (e.g. from %3.4 to %2.5 A_d for QRF). Again, QRF performs well in terms of accuracy and for the higher probability levels, the accuracy is equal to RK but the width is 10% smaller.

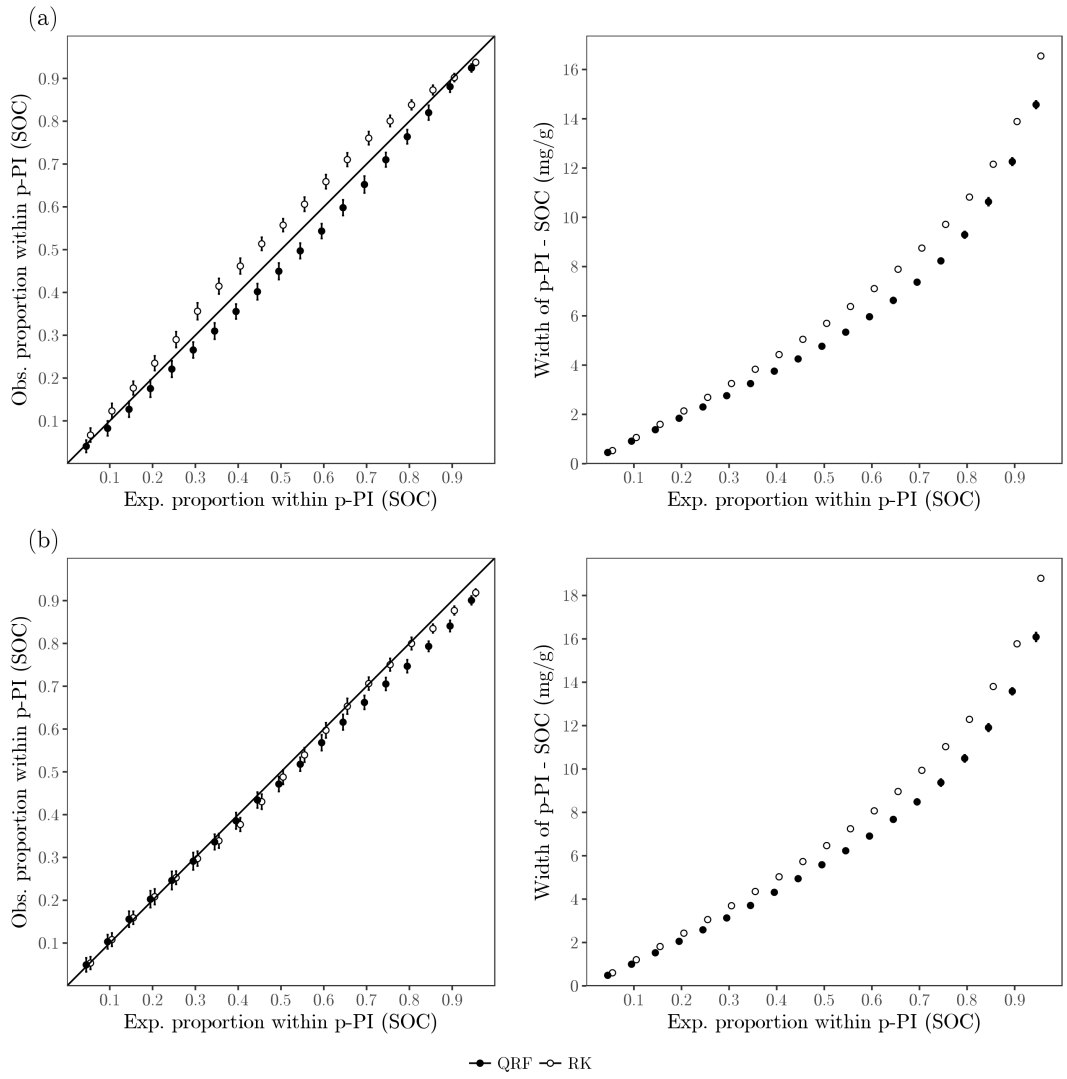


Figure 5.17. Validation plots for all p-PIs of SOC (10 k-fold, 100 iterations).
(a) Validation plots for 14 covariates; (b) Validation plots for 4 covariates. Accuracy plot (left); PI width (right). Accuracy plot shows expected proportions within interval versus observed predictions within interval; Prediction interval width plot shows the probability level versus the PI width

Table 5.4. PI estimate validation summary for SOC

Covariates	Method	Absolute deviation (A_d)	Proportion overestimation (P_o)	Proportion underestimation (P_u)
14	**QRF**	3.4%	0	1
	RK	4.0%	0.98	0.02
4	**QRF**	2.5%	0.03	0.97
	RK	1.0%	0.20	0.80

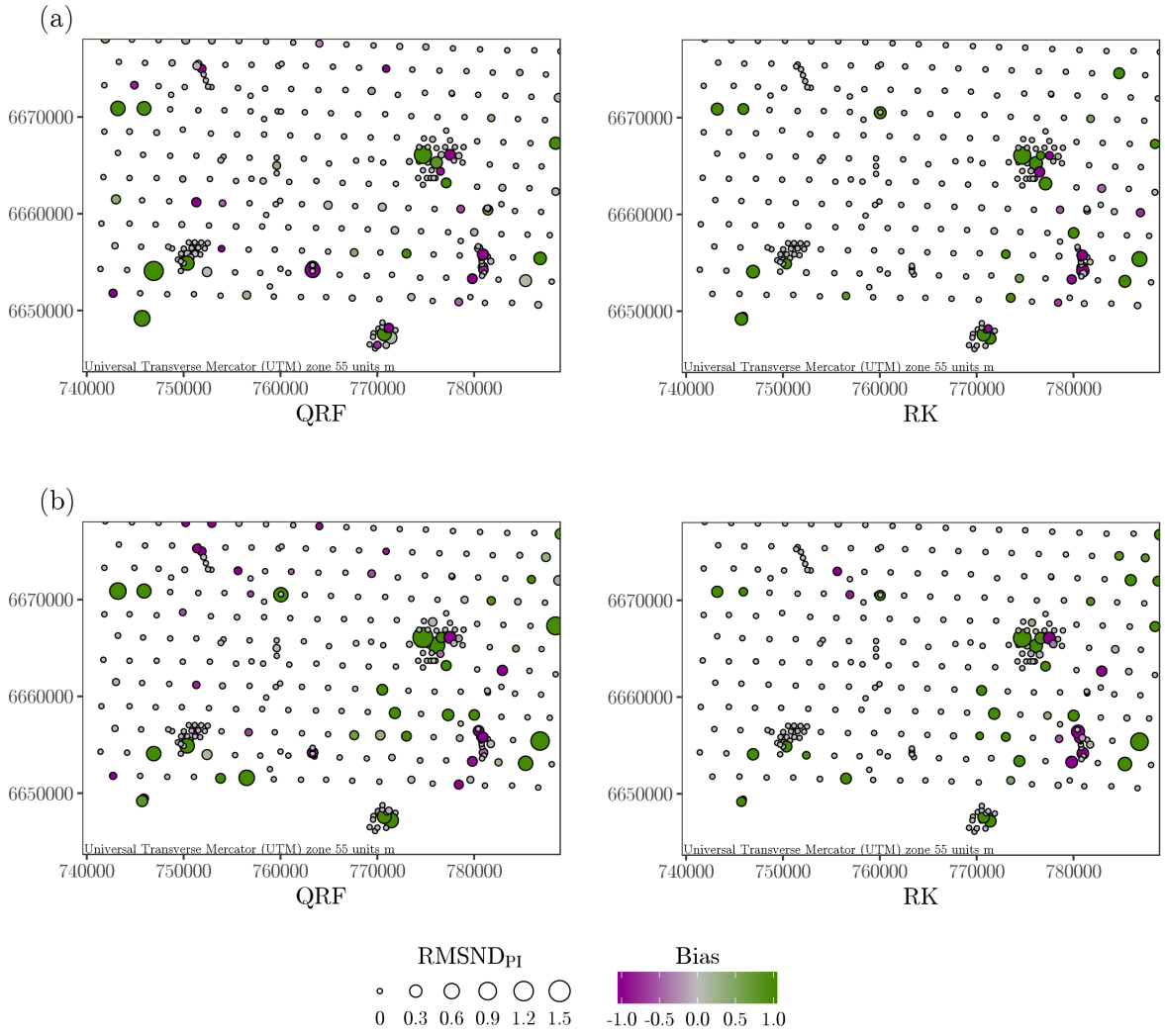


Figure 5.18. Spatial outliers for every site observation of the 0.9-PI for SOC.
(a) Spatial outliers for 14 covariates; (b) Spatial outliers for 4 covariates. Quantile regression forest (left); regression kriging (right). Circle size indicates the $RMSND_{PI}$ for 0.9-PI of all 100 10-fold cross validation results. Bias shows the average proportion of under- or overestimation: -1 for all observations smaller than 0.9-PI (green); +1 for all observations larger than 0.9-PI (pink).

Figure 5.18 above shows the root mean squared distance ($RMSND_{PI}$). In this figure, QRF shows almost the same spatial pattern for $RMSND_{PI}$ as RK, for both covariate models. This differs completely from pH where regional differences between QRF and RK were spotted 5.11. The spatial outlier patterns for SOC do not seem to be directly related to their geographical location. The only difference between RF14 and RF4 is that the $RMSND_{PI}$ is higher for RF4, for both QRF and RK.

5.3.3 Scalability assessment

In the Figure below (5.19) an overview can be found of the effect of (a) resolution on the total processing time; and (b) the number of covariates on processing timetime. The number of points seems highly dependent on the processing time, judging from (a). A maximum time of 4 s was achieved by RK on $3E4$ points (1000m resolution) and QRF only took 1 s more. When the number of points increased to $3E6$ (100m resolution), the difference between QRF and RK doubled. Hence, RK seems more scalable than QRF. In (b) it is seen that the number of covariates does not seem to affect the computation time much both for the RF modelling and uncertainty modelling as only a slight linear increase can be witnessed. Quantile regression forest seems slightly more affected when the number of covariates increase. Nevertheless, the absolute deviation metrics and goodness metrics do variate. RK seems to be more resistant to the number of covariates as the computation difficulty does not necessarily increase for the kriging step. Therefore, it seems that the training of a regular Random Forest is not affected much by number of predictors/covariates.

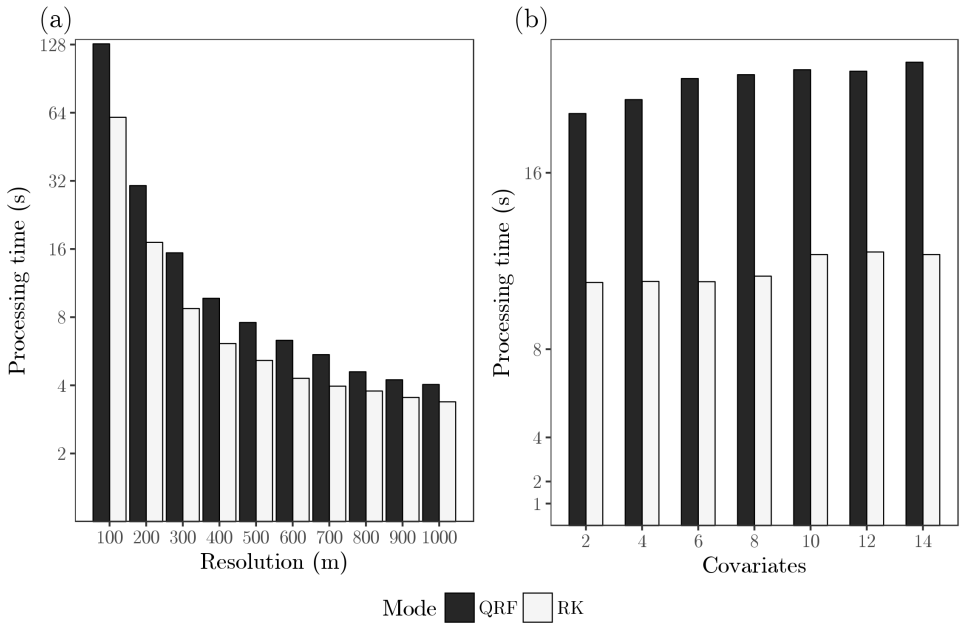


Figure 5.19. Total processing time of QRF versus RK.
Black colored bars represent QRF (dodged left) and white colored bars represent RK (dodged right); (a) The effect of resolution, at logarithmic scale (log2), on computation time; (b) The effect of the number of covariates on computation at regular scale in steps of powers of 2.

Figure 5.20 shows the effect of covariates on the absolute deviation. RK has a local minimum on 6 covariates and QRF on 8 covariates. The expectation was that both lines should monotonously decrease with an increase in covariates because more information would in theory lead to better uncertainty quantification as RF is resistant to noise present within the inputs so that the reducible error should further decrease.

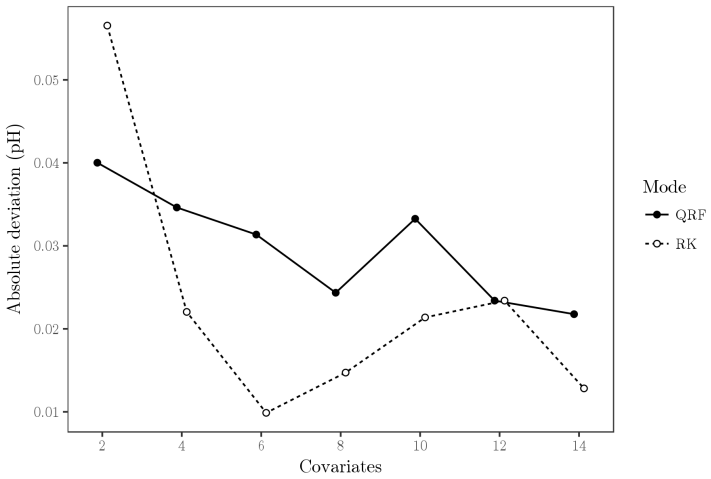


Figure 5.20. Effect covariates on absolute deviation (A_d).

6 GENERAL DISCUSSION

Judging from the included case study both quantile regression forest and regression kriging show promising uncertainty quantification potential for Random Forest predictions of soil properties. Even though the correctness of RK was high in the conducted Edgeroi case study, the spatial correlation of the residual of Random Forest regression was, in general, limited: the variograms almost existed out of pure nuggets (see table 5.2). Other studies that combined regression kriging with random forest also found limited spatial correlation in the RF residuals (Vaysse & Lagacherie, 2015, Fig. 3; Hengl et al., 2015, Fig. 7; Guo et al., 2015, Fig. 6). Surprisingly, in this case study this also occured for the 4 covariate RF model where more spatial correlation in the residuals was expected due to the presence of more information. Therefore, Random Forest looks very promising for modeling all soil relationships. The result is that the uncertainty maps, or 0.9-PI width maps, of RK show a near to uniform uncertainty estimate for all of the study site (Figures 5.9, 5.16). In contrast with RK, QRF provides more structured uncertainty maps with high differences throughout the map as can also be seen from the same figures also reported in the French QRF research paper (Vaysse & Lagacherie, 2015). It is likely that where most contrast can be seen on the QRF uncertainty maps for pH was dominated by one covariate (EVSMOD), for SOC it was probably the combination of different satellite products changes altogether or where the end leafs of RF find a differentiating characteristic, e.g. when looking at the clear partitions in uncertainty of the pH map (figure 5.9) that is much more uncertain for acidic observations than alkaline observations and coincides with patterns observed in EVSMOD (Appendix A.1). The uncertainty might be exacerbated by the limiting number of acidic observations and their distribution see Figure 5.3.

A potential useful uncertainty modeling extension would be to incorporate a covariate importance measure on a local level to quantify these effects because information on the spatial uncertainty given by QRF seems useful in understanding the underlying patterns of satellite products that lead to uncertainty that RK does not give. Currently, this is done subjectively by the researcher but a quantitative approach would be more objective. Note that in other Random Forest case studies that model (different) target variables under other conditions, RK might provide an extra layer of extra information that misses when using only quantile regression forest. The absence of such an extra spatial correlated information layer must be carefully considered when asking the question (such as in Hengl et al. (2018)) whether regression kriging is still needed in predictive soil modeling. Moreover, Nussbaum et al. (2017) note that RF models could simply include the coordinates as an extra covariates to provide such an extra layer, but they remark that this might lead to checkerboard artifacts due to recursive splitting that might be difficult to interpret.

6.1 Validity of the uncertainty quantification models.

In the included case study, an uncertainty estimation bias on the accuracy plots can be seen for both QRF and RK (see figures 5.10 & 5.17). QRF consistently underestimated the width of the p-PI , which leads to a lower than expected accuracy, i.e. QRF is said to be too optimistic. In contrast, RK tends to

be too pessimistic and kept the widths of the p-PI too wide, so that the observed proportion of predictions that fall within the prediction interval falls above the 45-degree correctness line. In general the accuracy plots very closely approximated the $1:1$ line so the over- and underestimation are relative. Furthermore, on average QRF consistently has smaller p-PI prediction interval widths than RK, because QRF limits the range the observed values. Interestingly, at higher p-PIs QRF almost consistently displays lower p-PI widths as RK, with a comparable correctness. This means that for this case study QRF can be regarded as a better model for these probability levels. A possible explanation is that QRF empirically estimates the ccdf (note the word cumulative) from node information, which implies more explanatory information is included at higher p-PIs, that in turn might mitigate the effect of error. Nonetheless, the increase in accuracy for QRF for higher probabilities seems to be absent in the Vaysse and Lagacherie (2015) study that did not expose an optimistic bias for QRF. This stresses the need for experimenting with more case studies to derive some general rules-of-thumb.

6.1.1 Effect of covariates

The validation of the quantile regression forest uncertainty quantification technique seems to be slightly dependent on the number of covariates (e.g. Figure 5.20 for pH). Higher number of covariates increases the predictive R^2, and seems to lead to higher validity of QRF as uncertainty model. Comparing Figure 5.10 (a) with (b) for instance, which seems to be performing worse for a smaller number of covariates judging from the accuracy plots. RK seems to be more resistant to a lower number of covariates, as its underlying uncertainty model is based on spatial correlation, which adds unexplained spatial uncertainty information on top of the existent RF prediction model. In theory, a higher number of covariates and training points should lead to more explanatory information within the leafs of the RF model, such that it becomes increasingly sufficient for QRF to correctly model the uncertainty of the RF prediction. This is in line with studies on including a large number of covariates in Random Forest soil modeling (Behrens et al., 2010, 2014). Meinshausen's (2016) theory also supports this feature as the the empirical ccdf should asymptotically converge to the real ccdf when providing the RF model with more information such as covariates or samples. In general, non-paramatric regression such as Random Forest should perform better with more training points (Friedman et al., 2001). More research is needed to look whether severe input uncertainties could impair this effect in DSM context such as is noted by X. Zhu et al. (2012). An observation that supports this uncertainty propagation of input uncertainties is the increases in absolute deviation A_d from the $1:1$ line in Figure 5.20, which could not be explained. The increases in A_d at 6 and 8 covariates for RK and QRF respectively were quite en expected. Therefore the addition of extra covariates should be carefully investigated.

Research could provide answers on how uncertainty quantification correctness of QRF correlates with the coefficient of determination or the RMSE of RF predictions. Under what conditions does the addition of covariates lead to an increase or decrease in uncertainty? An interesting case study for comparison between RK and QRF could be to pick soil properties that are difficult to model with satellite products, but do show high auto-correlation on small to medium distances ($<$1km) and look whether the differences in average p-PI width of QRF and RK will stay close to zero, even for the higher probability values under such conditions.

In addition, logarithmic transformations could be performed on the SOC concentrations to mitigate the effects of outliers on the relative error increase minimization problem that is fundamental to regression. For pH this is not the case at it is already a logarithmic scale of the H_3O^+ concentration in itself. In theory, Random Forests training is resistant to skewed distributions (Kuhn & Johnson, 2013), therefore it would be expected to look whether this skewedness resistance translates to QRF uncertainty estimate

correctness. If this is the case, then QRF does not need any transformations of the datasets.

6.2 Spatial patterns of uncertainty.

In general, QRF uncertainty is related to explanatory data and RK uncertainty to the point configuration. Moreover, QRF seems to partition the uncertainty mapping into distinct regions of small and high uncertainties, with abrupt changes at their boundaries, e.g. figures 5.16 and 5.9 for QRF have clear boundaries. Within these partitions there seems to be a degree of smoothness which becomes very visible for more covariates (Figure 5.9 for example). This raises the idea that within the uncertainty partitions, the uncertainty still has a degree of spatial correlation. Interestingly, the variogram could not find this spatial correlation, probably because it only considered the map as a whole instead of stratifying the space. For example, regression kriging seems to base its spatial correlation mainly on the sites with high pH (high alkalinity) that is likely related to low values of EVSMOD (Appendix A); this also explains the highly biased $RMSND_{PI}$ values in the parts where there is high acidity (see Figure 5.11). The assumption of stationarity of RK failed to hold for pH. Note that Wadoux et al. (2018) gained huge improvements in uncertainty modeling when taking the stationarity assumption into play, indicating that the current case study was not ideal for RK. This points to the hypothesis that if two separate models were trained on the alkaline and acidic partitions separately, the consistent over- and underestimations RK makes can be mitigated. Future research can focus on exploring such a hypothesis by constructing multiple validation plots on partitions of the landscape to summarize the local uncertainty assessment.

6.2.1 Spatial outliers

For pH, QRF made random errors (especially with less covariates), whereas RK seems to cluster them in certain regions 5.11. These spatial outlier clusters seem to be present in the parts where QRF estimates the RF uncertainty to be high. For SOC similar dispersal of spatial outliers was observed between QRF and RK. Differences in spatial outlier clusters do not seem to show up in the accuracy plot, e.g. comparing figure 5.9 with figure 5.10. On the contrary, the accuracy plot can be interpreted as quite positive for both techniques. Therefore, uncertainty evaluation with the accuracy plot alone might give a false impression, especially when only small regions are investigated. For example, if a a farmer or a policy maker would solely base his decision for a specific region on the accuracy plot, this region might happen to fall within a cluster of spatial outliers, leading to bad decision making. QRFs results on the other hand, show a more random dispersal of errors and are thus safer to use for projects of a larger scale. Note that the GlobalSoilMap predicament currently only asks for an overall validation of the uncertainty on the 0.9-PI (Arrouays et al., 2014), not for an extra local uncertainty assessment.

6.2.2 Sampling scheme dependence

Kriging needs a proper sampling design to estimate both the spatial correlation needed to improve its accuracy, but also in order to interpolate the target value based on neighbours close by (Brus & Heuvelink, 2007). In the case of the Edgeroi dataset sampling design was a mix of systematic and clustered sampling which can be considered good for variogram modeling. Vaysse and Lagacherie (2015) dealt with a more sparse sampling design and RK uncertainty modeling accuracy was very poor. Furthermore, the resolution of the satellite products was, with the exception of the DEM, standardized to $1km^2$ which caused some of the clustered sample locations to be intersected with the same pixels, whereas in reality circumstances might differ. Constructing a variogram from this can lead a decrease in detection of autocorrelation on close distances.

As kriging always demands a strict sampling scheme, an important question that remains is on how QRF is affected by sampling design. The need for including clusters into the that aim to estimate the spatial correlation better during the variogram modeling might prove to be unnecessary for quantile regression forest and this can, subsequently, reduce both the cost of time and financial budget of a soil survey.

6.2.3 Variogram modeling dependence

RK is very dependent on its variogram fitting and there are multiple ways to conduct the variogram modeling process (Calder & Cressie, 2009). The case study in France that also tested the uncertainty of pH and SOC of RF reported similar findings with very small spatial correlations of the RF residuals (Vaysse & Lagacherie, 2017); high nugget patterns. Peculiarly, their study showed opposite results to this case study, especially for RK the accuracy plots were deviating largely from the 45-degree bisector line. An explanation is that their sampling design was more irregular than the Edgeroi study site and no clusters of sampling points were used. In other words, the sampling scheme is decisive for making a proper choice for the variogram (Z. Zhu and Stein (2006); Brus and Heuvelink (2007). Another reason might be the choice of covariates, where Vaysse and Lagacherie (2017) chose more relief based covariates, this study chose more climate based covariates. Further research could expand on searching the limitations for variogram fitting in a RF context for example for more irregular sampling schemes.

6.3 Computation time

The ranger package has an option for threading that allows for computation in parallel for both the training of the model and predicting with the model, all cores are utilized (M. Wright & Ziegler, 2015). Moreover, the ranger package is an R package that functions like a wrapper for the low level programming language C. The advantage of C is that much faster computation times can be achieved as the model is compiled; hereafter it can be executed in machine code without further inferences. The gstat R package for kriging predictions only uses 1 core, but it does use multi-threading (Pebesma, 2004). It also acts as a wrapper for C, but to what degree has not been investigated during this research.

During the experiments with the different resolutions (increases in number of points) the ranger package seemed to have a certain threshold (<100m resolution) where it will crash R because it consumes too much memory. This means that QRF with the current ranger package implementation could make improvements regarding memory garbage collection in C for it to become applicable on large scale mapping projects. The gstat package did not have any memory issues. A solution to the memory problem can be implementation of such studies on a Big Data architecture (Marz & Warren, 2015). Even with a large number of processing cores, the computation times become impractical when the number of points increases.

Although the amount of covariates barely seemed to effect processing time, the performance of the uncertainty modeling measured in A_d did not monotonously decrease when the number of covariates increased. For RK an increase in A_d was observed for 6 covariates, for QRF an A_d increase on 8 covariates. The reason for this might be the inclusion of an environmental layer that seems decisive for RF for PI estimation but in reality is not as important. The effect of uncertainty propagation of input layers with QRF could provide answers.

6.4 Other methods

The other two uncertainty quantification methods, U-statistic based RF and jackknifing-after-bootstrap RF, could also provide more information on what the effect of the correctness of the regression fit is on the

total uncertainty; their derivations of confidence intervals could be compared to the the prediction interval estimates of QRF in future case studies. Especially for the purpose of measuring the effects uncertainty propagation in the input data they might become useful. Note that Meinshausen (2016) measures the complete uncertainty that includes the conditional variance of the target variable. However, the relation to the expected prediction (which USI and JKIJ quantify) might be a stronger indicator of sensitivities to the input uncertainty. Although jackknifing-after-bootstrap for RK uncertainty quantification is probably unfeasible for additional functionalities, it still remains a question on whether U-statistics can be for other research goals. For example, U-statistics has a possibility for constructing hypothesis test to look for the presence of algebraic structures in Random Forest models, e.g. additive structure (Mentch & Hooker, 2017). Therefore, U-statistics shows potential for using it to better understand on how Random Forest models complex soil formation processes which is is beyond mere uncertainty quantification.

7 CONCLUSION

7.1 Uncertainty quantification methods

Four uncertainty quantification methods for random forest predictions were found in literature. Two of these methods were aimed at quantifying the prediction uncertainty related to reducible error: U-statistic based random forest and jackknife-after-RF & infinitesimal jackknife. These techniques can be used for confidence interval estimation. The other two techniques focus on quantifying the complete uncertainty of reducible and irreducible error together: QRF and RK. These techniques are used for constructing prediction intervals. QRF aims to quantify the uncertainty related to the complete prediction error through the construction of an empirical ccdf. RK only works in a context of auto-correlation and it builds a spatial model on top of the regression model. RK minimizes the prediction error and through this process it automatically quantifies the uncertainty.

7.2 Viable methods

RK, QRF and the jackknifing approach are implemented in a variety of packages. Jackknifing as it is a a posteriori approach is very fast and scalable. RK is computation heavy, but should be more scalable than QRF. All techniques are proven to be mathematically conistent, but practical rules-of-thumb on how many trees, how many training points and how many prediction points are lacking. Information is especially in a spatial context where the number of points to predict are often high. The two most viable techniques in a spatial context are RK and QRF, as they both are suitable for estimating a prediction interval. Prediction intervals provides information directly on the uncertainty of the prediction. The other two methods are used for estimating confidence intervals instead which is aimed on information on the uncertainty of the model itself.

7.3 Validation of uncertainty quantification on soil case study

For the overall uncertainty assessment the results are consistent for set-aside test sets. Both QRF and RK demonstrated highly accurate plots with similar deviations from the $1:1$ line ($<5\%$). In general, QRF had smaller PI widths than RK that ranged up to differences of $>10\%$ for the higher probability levels ($p>0.8$). These results showed that QRF is slightly more favorable than RK on the Edgeroi case study, characterized by its systematic sample design with some high density cluster sampling. However, the spatial outlier analysis for soil pH showed that the outliers that fall outside of the predicted quantiles were clustered within non-overlapping regions and it can be argued that these results are more inconsistent because the RK assumptions did not hold for pH. QRF seemed to produce a more dispersed pattern of spatial outliers with high RMSND_{PI} values than RK. The spatial outlier patterns for SOC did not show this behavior, the overlap between the outliers was very similar between RK and QRF. This example proves that a spatial outlier assessment might show useful local validation information that the overall accuracy plot cannot show.

The uncertainty maps created with QRF were very detailed with high contrasts (boundaries) in p-PI widths throughout the study site, indicating conditional information based on the underlying environmental attributes. This extra information directly from the regression uncertainty might be useful for future applications such as sampling design optimization and choosing covariate combinations. In conclusion, when taking the Global Soil Map predicament into account − 9 out of 10 predictions should fall within the 0.9-PI boundaries − the current study cannot reject either RK or QRF as they both performed well.

7.4 Scalability assessment

The number of covariates did not seem to affect computation time considerably. However, when increasing the number of points to predict, computation time for QRF seems to have a more than double increase when comparing with RK with the currently used packages. Therefore QRF is less suitable for fine resolution, large scale digital soil mapping use cases. Further parallelization distributed over a computer network can provide an answer to these difficulties. Ultimately, the number of covariates did not lead to an expected monotonously decreasing absolute deviation from the accuracy plot. In general a trend was observed, but this trend displayed a local discrepancy on 6 covariates for RK and 8 for QRF. More research is needed on how the type and number of covariates influence the uncertainty quantification modeling.

8 RECOMMENDATIONS

The use of QRF in soil science is still limited. Therefore, this recommendation chapter will pitch some particular useful research ideas that can help to provide rules-of-thumb for uncertainty quantification modeling choice. Rules-of-thumb regarding sampling design, choice of covariates and geographical location. What are the limitations of QRF and could RK be useful in a Random Forest context?

The uncertainty conditioned on the environmental predictors shows clear patterns on the map into different parts with steep transitions in uncertainty along the borders. Therefore, it seems logical to use this information for studying the effect of underlying covariates on uncertainty and see whether this can be linked to the current covariate importance metric. A possibility could also be to develop a new importance metric on a local level that can be computed and visualized geographically. Another recommendation is to test whether this partitioning that QRF exposes (perhaps also the jackknifing and U-statistic approaches) can be used in conjunction with a RK approach, such that the kriging stationarity assumption holds within these partitions and see whether better variograms can be fitted and how this affects the performance of RK with regards to reductions in PI width and in the prevention of spatial outlier effects.

A different recommendation is to look whether QRF can replace RK. This can be done by including the coordinates as predictors, as suggested by (Nussbaum et al., 2017). Maybe QRF can completely replace the potential extra information a variogram can provide by doing so. The question remains whether this will lead to a loss of detail on the predicted PI width maps.

Currently both an overall (accuracy plots) and local assessment (spatial outliers) was done which did not always seem to match with each other. It would be useful to introduce more case studies to investigate whether the spatial outlier regions will occur under different sets of environmental conditions. The question is whether QRF stays more randomly dispersed than RK, or that this was specific for this case only. This research proposed a metric (RMSND_{PI}) for outlier detection and it is likely that improvements can be made to this metric to better visualize the spatial outliers.

Ultimately, experimentation with statistical transformations, such as a logarithm or a square root should be done for QRF and see what its effect is on its uncertainty model. Random Forest is resistant to the distributions of the response and predictors so it needs to be identified whether this translates to QRF.

REFERENCES

Addiscott, T. M., & Tuck, G. (2001). Non-linearity and error in modelling soil processes. *European Journal of Soil Science*, *52*(1), 129-138.

Arrouays, D., Grundy, M. G., Hartemink, A., Hempel, J. W., Heuvelink, G. B. M., Hong, S. Y., ... Zhang, G. (2014). Chapter three - globalsoilmap: Toward a fine-resolution global grid of soil properties. In D. L. Sparks (Ed.), *Advances in agronomy* (Vol. 125, p. 93-134). Academic Press.

Bazaglia, O., Rizzo, R., Lepsch, I., Prado, H., Gomes, F., Mazza, J. A., & DemattÃł, J. (2013). Comparison between detailed digital and conventional soil maps of an area with complex geology. *Revista Brasileira de ciência do solo*, *37*(5), 1136-1148.

Beckett, P., & Burrough, P. (1971). The relation between cost and utility in soil survey. *European Journal of Soil Science*, *22*(4), 466-480.

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

Behrens, T., Zhu, A., Schmidt, K., & Scholten, T. (2010). Multi-scale digital terrain analysis & feature selection for digital soil mapping. *Geoderma*, *155*(3-4), 175–185.

Biau, G., & Scornet, E. (2016). A random forest guided tour. *Test*, *25*(2), 197–227.

Bishop, T., McBratney, A., & Laslett, G. (1999). Modelling soil attribute depth functions with equal-area quadratic smoothing splines. *Geoderma*, *91*(1-2), 27–45.

Breiman, L. (1996). Bagging predictors. *Machine learning*, *24*(2), 123–140.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.

Brus, D. J., & Heuvelink, G. B. M. (2007). Optimization of sample patterns for universal kriging of environmental variables. *Geoderma*, *138*(1), 86-95.

Burgess, T. M., & Webster, R. (1980). Optimal interpolation and isarithmic mapping of soil properties. *Journal of Soil Science*, *31*(2), 315-331.

Calder, C., & Cressie, N. A. (2009). Kriging and variogram models.

Campbell, J. B., & Edmonds, W. J. (1984). The missing geographic dimension to soil taxonomy. *Annals of the Association of American Geographers*, *74*(1), 83-97.

Cressie, N. (1985). Fitting variogram models by weighted least squares. *Journal of the International Association for Mathematical Geology*, *17*(5), 563–586.

Cressie, N. A. (1993). *Statistics for spatial data*. Wiley Online Library.

Declercq, F. A. N. (1996). Interpolation methods for scattered sample data: accuracy, spatial patterns, processing time. *Cartography and Geographic Information Systems*, *23*(3), 128–144.

Dijkerman, J. C. (1974). Pedology as a science: The role of data, models and theories in the study of natural soil systems. *Geoderma*, *11*(2), 73-93.

Dubois, P. C., Van Zyl, J., & Engman, T. (1995). Measuring soil moisture with imaging radars. *IEEE Transactions on Geoscience and Remote Sensing*, *33*(4), 915-926.

Efron, B. (1981). Nonparametric estimates of standard error: the jackknife, the bootstrap and other methods. *Biometrika*, *68*(3), 589–599.

Efron, B. (1992a). Bootstrap methods: another look at the jackknife. In *Breakthroughs in statistics* (pp. 569–593). Springer.

Efron, B. (1992b). Jackknife-after-bootstrap standard errors and influence functions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 83–127.

Fox, J. (1997). *Applied regression analysis, linear models, and related methods.* Sage Publications, Inc.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning* (Vol. 1). Springer series in statistics New York.

Goovaerts, P. (2001). Geostatistical modelling of uncertainty in soil science. *Geoderma*, *103*(1), 3-26.

Guo, P.-T., Li, M.-F., Luo, W., Tang, Q.-F., Liu, Z.-W., & Lin, Z.-M. (2015). Digital mapping of soil organic matter for rubber plantation at regional scale: An application of random forest plus residuals kriging approach. *Geoderma*, *237-238*, 49 - 59. Retrieved from `http://www.sciencedirect.com/science/article/pii/S0016706114003164` doi: https://doi.org/10.1016/j.geoderma.2014.08.009

Hartemink, A., Hempel, J. W., Lagacherie, P., McBratney, A. B., McKenzie, N. J., MacMillan, R. A., ... Sanchez, P. A. (2010). Globalsoilmap. netâĂŞa new digital soil map of the world. *Digital soil mapping*, 423-428.

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

Hengl, T., Heuvelink, G., Kempen, B., Leenaars, J., Walsh, M., Shepherd, K., ... others (2015). Mapping soil properties of africa at 250 m resolution: random forests significantly improve current predictions. *PloS one*, *10*(6), e0125814.

Hengl, T., Heuvelink, G., & Rossiter, D. (2007). About regression-kriging: From equations to case studies. *Computers & Geosciences*, *33*(10), 1301 - 1315. Retrieved from `http://www.sciencedirect.com/science/article/pii/S0098300407001008` (Spatial Analysis) doi: https://doi.org/10.1016/j.cageo.2007.05.001

Hengl, T., Kempen, B., G.B.M., & Malone, B. (2017). Package 'gsif'. *R Package*.

Hengl, T., Mendes de Jesus, J., Heuvelink, G. B. M., Ruiperez Gonzalez, M., Kilibarda, M., Blagotic, A., ... Kempen, B. (2017). Soilgrids250m: Global gridded soil information based on machine learning. *PLoS One*, *12*(2), e0169748.

Hengl, T., Nussbaum, M., Wright, M., & Heuvelink, G. (2018). Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. *PeerJ PrePrints*.

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

Hillis, D. M., & Bull, J. J. (1993). An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. *Systematic biology*, *42*(2), 182–192.

Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. *The annals of mathematical statistics*, 293–325.

Hudson, B. (1992). The soil survey as paradigm-based science. *Soil Science Society of America Journal*, *56*(3), 836-841.

Isbell, R. (2016). *The australian soil classification*. CSIRO publishing.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112). Springer.

Janson, S. (1984). The asymptotic distributions of incomplete u-statistics. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, *66*(4), 495–505.

Jenny, H. (1941). *Factors of soil formation: A system of quantitative pedology, 281 pp*.

Kerry, R., & Oliver, M. (2007). Comparing sampling needs for variograms of soil properties computed by the method of moments and residual maximum likelihood. *Geoderma*, *140*(4), 383 - 396. Retrieved from `http://www.sciencedirect.com/science/article/pii/S0016706107001176` (Pedometrics 2005) doi: https://doi.org/10.1016/j.geoderma.2007.04.019

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling* (Vol. 810). Springer.

Lee, J. (1990). *U-statistics: Theory and practice*. Citeseer.

Lichtenstern, A. (2013). Kriging methods in spatial statistics. *Technische Universität München*.

Lorenzetti, R., Barbetti, R., Fantappié, M., L'Abate, G., & Costantini, E. A. C. (2015). Comparing data mining and deterministic pedology to assess the frequency of wrb reference soil groups in the legend of small scale maps. *Geoderma*, *237*(Supplement C), 237-245.

Louppe, G. (2014). Understanding random forests: From theory to practice. *arXiv preprint arXiv:1407.7502*.

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

Malone, B. P., Minasny, B., & McBratney, A. B. (2016). *Using r for digital soil mapping.*

Marz, N., & Warren, J. (2015). *Big data: Principles and best practices of scalable realtime data systems.* Manning Publications Co.

Matheron, G. (1971). Theory of regionalized variables and its applications. *Les Cahiers du Centre de Morphologie Mathèmatique de Fontaineblea.*, *5*, 211.

McBratney, A. B., Mendonça Santos, M. L., & Minasny, B. (2003). On digital soil mapping. *Geoderma*, *117*(1), 3-52.

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

Meinshausen, N. (2016). quantregforest: Quantile regression forests. r package version 1.3-5 [Computer software manual].

Mentch, L., & Hooker, G. (2016). Quantifying uncertainty in random forests via confidence intervals and hypothesis tests. *The Journal of Machine Learning Research*, *17*(1), 841–881.

Mentch, L., & Hooker, G. (2017). Formal hypothesis tests for additive structure in random forests. *Journal of Computational and Graphical Statistics*, *26*(3), 589-597. doi: 10.1080/10618600.2016.1256817

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

Minasny, B., McBratney, A. B., & Lark, R. M. (2008). Digital soil mapping technologies for countries with sparse data infrastructures. *Digital soil mapping with limited data*, 15-30.

Moore, I. D., Gessler, P. E., Nielsen, G. A., & Peterson, G. A. (1993). Soil attribute prediction using terrain analysis. *Soil Science Society of America Journal*, *57*(2), 443-452.

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

Nussbaum, M., Spiess, K., Baltensweiler, A., Grob, U., Keller, A., Greiner, L., ... Papritz, A. (2017). Evaluation of digital soil mapping approaches with large sets of environmental covariates. *SOIL Discuss.*, *2017*, 1-32.

Paoli, J. N., Tisseyre, B., Strauss, O., & Roger, J.-M. (2003). Methods to define confidence intervals for kriged values: Application on precision viticulture data. *Proceedings of 4th ECPA, edited by J. Stafford, A. Werner, Wageningen Academic Publishers, The Netherlands*, 521–526.

Pebesma, E. J. (2004). Multivariable geostatistics in s: the gstat package. *Computers & Geosciences*, *30*, 683-691.

Polimis, K., Rokem, A., & Hazelton, B. (2017). Confidence intervals for random forests in python. *The Journal of Open Source Software*, *2*, 124.

Ribeiro Jr, P. J., Diggle, P. J., et al. (2001). geor: a package for geostatistical analysis. *R news*, *1*(2), 14–18.

Rowell, D. L. (2014). *Soil science: Methods & applications*. Routledge.

Schwartz, J. (2009). Bing maps tile system. *Microsoft Developer network Available: http://msdn. microsoft. com/en-us/library/bb259689. aspx*.

Scornet, E. (2015). *Learning with random forests* (Theses, Université Pierre et Marie Curie - Paris VI). Retrieved from `https://tel.archives-ouvertes.fr/tel-01250221`

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

Sexton, J., & Laake, P. (2009). Standard errors for bagged and random forest estimators. *Computational Statistics & Data Analysis*, *53*(3), 801 - 811. Retrieved from `http://www.sciencedirect.com/science/article/pii/S0167947308003988` (Computational Statistics within Clinical Research) doi: https://doi.org/10.1016/j.csda.2008.08.007

Shih, Y.-S. (1999). Families of splitting criteria for classification trees. *Statistics and Computing*, *9*(4), 309–315.

Solomon, J., & Rock, B. (1985). Imaging spectrometry for earth remote sensing. *Science*, *228*(4704), 1147-1152.

Stockmann, U., Padarian, J., McBratney, A. B., Minasny, B., de Brogniez, D., Montanarella, L., ... Field, D. J. (2015). Global soil organic carbon assessment. *Global Food Security*, *6*(Supplement C), 9-16.

Van Beers, W. C., & Kleijnen, J. P. C. (2003). Kriging for interpolation in random simulation. *Journal of the Operational Research Society*, *54*(3), 255–262.

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*(Supplement C), 20-30.

Vaysse, K., & Lagacherie, P. (2017). Using quantile regression forest to estimate uncertainty of digital soil mapping products. *Geoderma*, *291*(Supplement C), 55-64.

Wackernagel, H. (2003). *Multivariate geostatistics, 387 pp*. Springer, New York.

Wadoux, A. M. C., Brus, D. J., & Heuvelink, G. B. (2018). Accounting for non-stationary variance in geostatistical mapping of soil properties. *Geoderma*.

Wager, S., & Athey, S. (2017). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, *0*(ja), 0-0. Retrieved from `https://doi.org/10.1080/01621459.2017.1319839` doi: 10.1080/01621459.2017.1319839

Wager, S., Hastie, T., & Efron, B. (2014). Confidence intervals for random forests: the jackknife and the infinitesimal jackknife. *Journal of Machine Learning Research*,

15(1), 1625-1651.

Wagner, W., Blöschl, G., Pampaloni, P., Calvet, J., Bizzarri, B., Wigneron, J., & Kerr, Y. (2007). Operational readiness of microwave remote sensing of soil moisture for hydrologic applications. *Hydrology Research*, *38*(1), 1-20.

Webster, R. (2000). *Is soil variation random?* (Vol. 97).

Were, K., Bui, D., Øystein, B., & Bal Ram, S. (2015). A comparative assessment of support vector regression, artificial neural networks, and random forests for predicting and mapping soil organic carbon stocks across an afromontane landscape. *Ecological Indicators*, *52*, 394 - 403. Retrieved from `http://www.sciencedirect.com/science/article/pii/S1470160X14006049` doi: https://doi.org/10.1016/j.ecolind.2014.12.028

Wilding, L. (1985). Spatial variability: Its documentation, accommodation and implication to soil surveys. *Soil Spatial Variability*, 166-194.

Wright, M., & Ziegler, A. (2015). ranger: A fast implementation of random forests for high dimensional data in c++ and r. *arXiv preprint arXiv:1508.04409*.

Wright, R., & Wilson, S. (1979). On the analysis of soil variability, with an example from spain. *Geoderma*, *22*(4), 297-313.

Zaouche, M., Bel, L., & Vaudour, E. (2017). Geostatistical mapping of topsoil organic carbon and uncertainty assessment in western paris croplands (france). *Geoderma Regional*, *10*(Supplement C), 126-137.

Zhu, X., Vondrick, C., Ramanan, D., & Fowlkes, C. (2012). Do we need more training data or better models for object detection?. In *Bmvc* (Vol. 3, p. 5).

Zhu, Z., & Stein, M. L. (2006, Mar 01). Spatial sampling design for prediction with estimated parameters. *Journal of Agricultural, Biological, and Environmental Statistics*, *11*(1), 24. Retrieved from `https://doi.org/10.1198/108571106X99751` doi: 10.1198/108571106X99751

Appendix A: COVARIATES

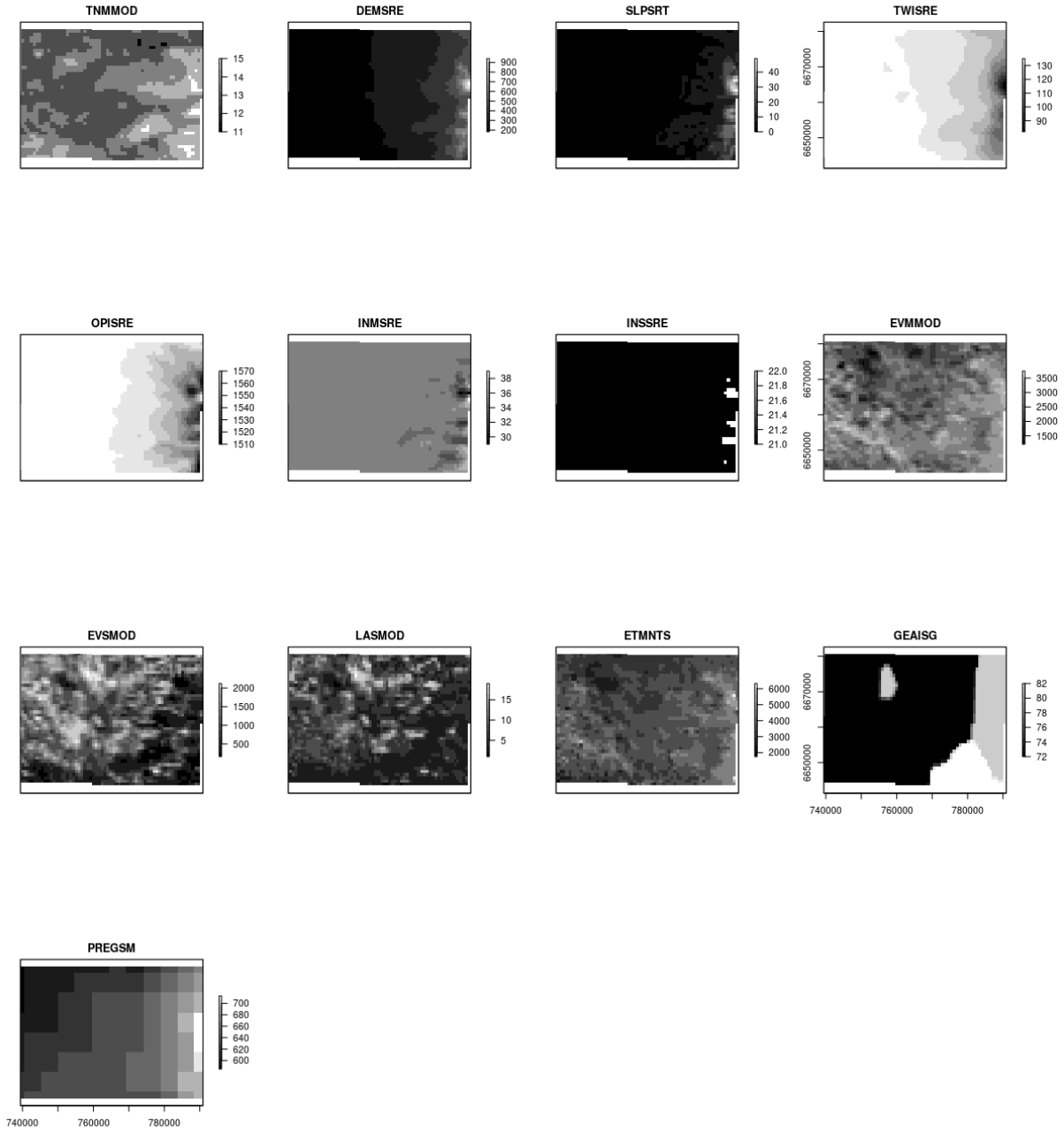


Figure A.1. Maps of the 14 covariates.