High-Resolution Digital Soil Mapping for Managed Forests using Airborne LiDAR Data

by

Babak Kasraei

B.Tech., British Columbia Institute of Technology (BCIT), 2014

B.Sc., Shahid Chamran University, 1998

Thesis Submitted in Partial Fulfillment of the

Requirements of the Degree of

Master of Science

in the

Department of Geography

Faculty of Environment

SIMON FRASER UNIVERSITY

Summer 2020

Approval

Name:	Babak Kasraei
Degree:	Master of Science (Geography)
Title:	High-Resolution Digital Soil Mapping for Managed Forests using Airborne LiDAR Data
Examining Committee:	**Chair:** William Jasse Hahm Assistant Professor
	Margaret Schmidt Senior Supervisor Associate Professor
	Chuck Bulmer Supervisor Soil Scientist (PhD) Provincial Government of British Columbia Ministry of Forests, Land, Natural Resource Operations and Rural Development
	Brandon Heung Supervisor Assistant Professor Faculty of Agriculture Dalhousie University
	Angela Bedard-Haughn External Examiner Professor College of Agriculture and Bioresources University of Saskatchewan
Date Defended/Approved:	July 14, 2020

Abstract

A goal of sustainable forest management using digital soil mapping (DSM) is to ensure that current and future generations have the best soil information so they can use forest resources wisely. This goal can be achieved using new technologies of generating digital soil maps and high-resolution light detection and ranging (LiDAR) data. Uncertainty in digital soil maps can be quantified using quantile regression (QR). The overall objective of this study is to generate several digital soil maps using different machine learning (ML) methods for forest management purposes and use a QR method to estimate their uncertainty. The study area is the Eagle Hill Forest (95 km^2), located west of Kamloops, BC, Canada. Five soil properties were mapped and locations with soil erosion, displacement, and compaction and puddling hazards were displayed on maps and discussed. 90% prediction interval (PI) maps were produced and the performance of the QR method in uncertainty quantification of different ML models was illustrated by producing Prediction Interval Coverage Probability (PICP) plots.

Keywords: Digital soil mapping; LiDAR; Quantile regression; Machine learning; Prediction interval

Acknowledgements

I would like to thank the Province of British Columbia: Ministry of Forest, Lands, and Natural Resources for their financial support of this study awarded through Dr. Margaret Schmidt.

I would like to thank Dr. Margaret Schmidt for her constant and enduring support. She allowed me the flexibility to pursue my degree even through many hardships which I am incredibly grateful for.

To Dr. Chuck Bulmer for the many long encouraging conversations to continue pursuing my degree with confidence.

To Dr. Brandon Heung for his guidance in all the stages of my studies through holding Skype sessions, introducing useful papers and adding excellent comments to my works.

To Daniel D. Saurette in the Ontario Ministry of Agriculture, Food and Rural Affairs who has helped me with many online conversations regarding R coding and modeling.

I would also like to thank my parents, Manoochehr Kasraei and Fatemeh Bibi Adli. My ability to pursue my dreams in Canada is because of my father's magnanimity and my mother's cordiality and encouragement. My father provided me with enough immigration fund and my mother provided me with enough education fund for my second bachelor in Canada at BCIT.

Table of Contents

List of Tables

List of Figures

List of Acronyms

BC	British Columbia
BEC	Biogeoclimatic ecosystem
cLH	Conditioned Latin Hypercube
CL	confidence levels
Cubist	Cubist decision tree
DEM	Digital elevation models
DSM	Digital soil mapping
DSA	Digital Soil Assessment
DSRA	Digital Soil Risk Assessment
GPS	geographical positioning system
IDF	Interior Douglas-fir
K	Soil erodibility
kNN	k nearest neighbour
LiDAR	Light detection and ranging
ML	Machine learning
MPI	Mean prediction interval
MSE	Mean square error
OP	Opportunistic point
PI	Prediction interval
PICP	Prediction interval coverage probability
QR	Quantile regression
QRF	Quantile regression forest
RC	Road cuts
RF	Random forest
RMSE	Root mean square error
SCORPAN-SSPFe	Soil spatial prediction function with spatially autocorrelated errors
SL	Slope length
SVM	Support vector machine
UNCED	United Nation Conference on Environment and Development

Chapter 1.

Introduction

Sustainable forest management can be defined as the practice of preserving and improving forest health, while creating environmental, economic, social and cultural opportunities for present and future generations (Franc et al., 2001). Regarding this definition, considerable agreement has been made to ensure that forests of all types are well managed in ways that are environmentally sensitive, socially aware and economically viable (Wallis et al., 1997). Another objective of sustainable forest management is in sustaining spatial complexity or variability across a range of spatial scales (Franklin and Forman, 1987). Spatial complexity in forested systems refers to the range of forest age classes, the size of patches in each class and the variation in overstory and understory structure and floristics. These patterns of spatial variation are directly related to environmental changes in terrain morphometry, aspect, elevation and soil type (Austin et al., 1990). With respect to the goals of forest management, in the last few decades process based forest growth models have been developed; however, they have not been very successful regarding the complexity of a forest system and difficulties in experimenting with large, long-lived plants such as trees (Battaglia and Sands, 1998). The principal goal for making the forest growth models have been to predict the volume growth yield of the forests. In other words the goal has been to estimate the forest productivity (Vanclay, 1988).

Forest productivity can be defined as the merchantable yield of an individual stand, and for a group of stands that comprises the forest. Forest productivity is the result of integration of environmental factors which include soil, climate, species composition and stocking, and stand history. Stand history includes disturbances such as fire, logging, insects or disease (Nyland, 1992). Productive forests depend on the soil, whereby soil disturbance or the depletion of soil nutrients can be associated with declines in productivity; hence information on the soils will facilitate sustainable soil management. (YuSheng et al., 2000). Forest soil productivity can be affected by factors such as wind and water erosion (Burger, 2009), and soil properties such as soil texture, pH, and thickness (Bontemps and Bouriaud, 2014; Tan et al., 2005) and thus, forest soil

information is very important for sustainable forest management. However, in many areas soil information at local scale is not readily available because conventional soil maps have been produced at small and mostly national scales (Ghaderi et al., 2019; Kempen et al., 2015).

To solve the problems of conventional soil maps having high production costs, limited accuracy and precision, and problems related to map scale, DSM appeared in 1978 (Kempen et al., 2012; McBratney et al., 2003a; Minasny and McBratney, 2016). DSM is based on the SCORPAN-SSPFe (soil spatial prediction function with spatially autocorrelated errors) concept. The SCORPAN model is a reformulation of Jenny's soil formation factors (Jenny, 1941) that describes the relationships between soil and other environmental factors and is used to establish a soil spatial prediction function (McBratney et al., 2003a). In DSM, quantitative prediction methods, such as machine learning (ML) methods, are used to correlate ancillary variables and soil properties. Examples of ML techniques may include random forest (RF), artificial neural network, and Cubist decision tree (McBratney et al., 2003a). The ML techniques sometimes are used as independent platforms for modelling, and sometimes they are hybridized with geostatistical modelling methods such as regression kriging (Odeha et al., 1994).

Within the SCORPAN model, topographic data is most commonly used (McBratney et al., 2003a) due to its wide availability and strong correlation with soil properties. This data comes in the form of digital elevation models (DEMs), which can be derived from satellite or other remotely sensed data such as Light Detection and Ranging (LiDAR) data (Boettinger et al., 2008; Shi et al., 2012a). Demands for LiDAR-based DEMs have increased recently because the LiDAR data is acquired at extremely fine spatial resolutions and with a high level of accuracy (Shi et al., 2012a). High-resolution LiDAR DEMs can provide an opportunity to produce accurate, local scale maps. Using LiDAR data in precision agriculture is common (Hämmerle and Höfle, 2014; Höfle, 2014; Koenig et al., 2015); however, it has less commonly been used for DSM in forest systems over large areas. One reason can be the high soil variability and complex topography of forested areas that makes it often difficult to obtain calibration and validation data in forests.

Digital soil maps are associated with substantial uncertainty that should be quantified (Vaysse and Lagacherie, 2017). The errors may happen in measurements

that will contribute to uncertainty in predictions (Arrouays et al., 2014b). Uncertainty in model output can be due to three sources including uncertainties in the model structure, model parameters, and model inputs (Minasny and McBratney, 2002). Even though international standards for DSM products necessitate the inclusion of uncertainty assessment for each soil property prediction (Arrouays et al., 2014a), such evaluations are not always applied (Arrouays et al., 2017). Different methods are used for uncertainty quantification. Some of these methods include empirical uncertainty quantification methods (Malone et al., 2017) bootstrapping (Malone et al., 2017), Monte Carlo method, Bayesian method (Solomatine and Shrestha, 2009), and quantile regression forest (QRF) (Meinshausen, 2006). QR which selects the quantiles from model outputs and finds a linear relationship among them is a novel method in DSM and will be tested in this thesis (Koenker and Hallock, 2001).

Due to the need for improved soil information to support forest productivity assessments and the increasing availability of LiDAR elevation data in forested areas, the application of DSM techniques using LiDAR and uncertainty assessments of the results should be further evaluated in forested areas. Therefore, the overarching goal of this thesis is to investigate the usefulness of LiDAR-derived DEMs for DSM and digital soil maps for sustainable forest management and to quantify uncertainty in digital soil maps using a novel method in DSM using QR.

1.1. Theoretical Background and Methods

1.1.1. DSM and SCORPAN

In the late 20th century, McBratney et al. (2003) introduced the generic framework of digital sol mapping, called the SCORPAN-SSPFe (soil spatial prediction function with spatially autocorrelated errors) model to predict soil properties and measure errors. The SCORPAN model is particularly relevant for those places where soil resource information is limited. It is based on the seven predictive factors described in Equation 1 as follows:

$$S_c = f(s,c,o,r,p,a,n) \; or \; S_a = f(s,c,o,r,p,a,n) \qquad (1)$$

In Equation 1 S_c is soil classes, and S_a is soil attribute. The factor s refers to soil conventional information such as legacy data or expert knowledge. The factor n stands

for space, spatial or geographic position. The other factors are a generalization of Jenny's (1994) five factors. Jenny's famous equation which was intended as a mechanistic model for soil development is: *S= f(c,o,r,p,t,…).* In this equation, S- stands for soil, c- (sometimes cl) represents climate, o- organisms including humans, r- relief or topography, including terrain attributes and classes, p- parent material including lithology and t- time factor (Jenny, 1994). DSM, based on the SCORPAN theoretical concept, involves the following steps:

1) Defining soil attribute(s) of interest and deciding on the resolution and block size
2) Assembling data layers
3) Specifying spatial decomposition of data layers
4) Sampling data to obtain sampling sites
5) Conducting GPS field sampling and laboratory analysis to obtain soil class or property data
6) Fitting quantitative relationships with autocorrelated errors
7) Predicting a digital map
8) Conducting field sampling and laboratory analysis for corroboration and quality testing
9) If necessary, simplifying the legend or decreasing resolution (McBratney et al., 2003b)

1.1.2. Light Detection and Ranging (LiDAR)

LiDAR is an active, remote sensing technology that uses a laser beam to measure distances. In this method a laser beam is emitted to a target object and the reflection of the beam is recorded. Then the distance is measured based on the product of the speed of light and the travel time of the reflected beam (Wehr and Lohr, 1999). By developing and advancing global positioning systems (GPS) in the late 20th century, the application of LiDAR increased (Lim et al., 2003). Since then it has been used in flood risk mapping (McArdle et al., 1999), terrain modelling (Kraus and Pfeifer, 1998), and land cover classification (Schreier et al., 1985).

A DEM is a raster dataset that consists of a matrix of pixels and represents the surface of an area (ESRI, 2019). Exceptional accuracy, and spatial resolution and the capability of scanning high density patterns provides the opportunity of producing high quality DEMs with LiDAR (Liu, 2008; Lohr, 1998). A great number of covariates that can be used as inputs to DSM can be derived from DEMs (Lagacherie, 2008). LiDAR DEMs have been used in modelling soils. For instance, Greve et al. (2012) used DSM methods to quantify the relationship between soil texture and different environmental covariates in Denmark. They derived their topographic indices from LiDAR data. Fink and Drohan (2016) used LiDAR derived terrain indices to predict hydric soils in Pennsylvania, USA and to improve their soil survey mapping. Although LiDAR derived DEMs have been used for modelling in hydrology, geology and ecology in forested areas and in precision agriculture (Bässler et al., 2011; Galzki et al., 2011; James et al., 2007), only a few studies have investigated the capability of LiDAR derived DEMs in modelling forest soils.

1.1.3. Machine Learning (ML)

Machine learning uses the theory of statistics to learn from training data or past experience to make a model. The model can be predictive to make predictions in the future or descriptive to gain knowledge from data, or both. Learning refers to the execution of a computer program to optimize the parameters of the model using the training data or past experience (Alpaydin, 2020). RF, cubist decision tree, k Nearest Neighbors (kNN) and support vector machine (SVM) are examples of ML methods and are used in this study.

RF is a non-parametric technique in which many ensembles of trees and classifiers are generated. Each tree in an ensemble grows based on the realization of a random vector. RF employs bagging which is a popular classification tree and ML method in which the trees are constructed independently using a bootstrap sample of the dataset. Bagging or bootstrap aggregation is used to reduce the variance of an estimated prediction function (Breiman 1996). After constructing the trees, RF predicts on new data by combining the predictions of the trees (Liaw and Wiener, 2014). For example, Grimm et al. (2008) used RF to predict the spatial distribution of soil organic carbon on Barro Colorado Island. Another example for using RF in DSM is in a study by Wiesmeier et al. (2011) in which they used RF to model the spatial distribution of soil

organic carbon, total carbon, total nitrogen and total sulphur in a semi-arid catchment in Inner Mongolia, Northern China.

Cubist decision tree is a model tree which is used to generate rule-based predictive models. It has been developed from C4.5 and M5 model trees (Quinlan, 2014). A Cubist tree is grown where there are intermediate linear models in leaves, and terminal nodes of the tree and can capture both linear and hierarchical relationships between the variables. The tree and the linear model is finally adjusted to reduce the absolute error (Minasny and McBratney, 2008). It is a data partitioning algorithm that mines non-linear relationships in data (Malone et al., 2017). For example, Ma et al. (2017) used the Cubist decision tree algorithm to model and map various soil properties such as soil texture and pH in eastern China using legacy data and available covariates. In another example Pouladi et al. (2019) used some modelling approaches including Cubist decision tree to predict soil organic matter in Denmark.

kNN is a ML method that classifies training data points based on closest training data in the environmental covariate space (Subburayalu and Slater, 2013). In this method a dataset is explored for k closest soil attributes based on similarity in feature space. To implement this, the similarity distance to the target soil is measured using Euclidean distance after normalization and rescaling of the soil attribute data in the dataset. Normalization and rescaling are done to ensure that the soil attribute values receive equal weights (Taghizadeh-Mehrjardi et al., 2016). kNN have been used in many DSM studies; for example, Mansuy et al. (2014) used kNN to generate continuous national maps for selected soil variables such as carbon, nitrogen and soil texture for the Canadian managed forest landbase. In another example, Taghizadeh-Mehrjardi et al. (2016) used several data mining techniques including kNN to map soil organic carbon and vertical variations down to 1 m depth in a semi-arid region in Kurdistan Province in Iran.

Support vector machine (SVM) is a non-parametric learning algorithm that is mostly used in pattern recognition and classification problems in remote sensing. SVM approximates a function that assigns a value to each input sample by constructing a hyperplane (Mountrakis et al., 2011). The hyperplane separates the dataset into discrete predefined classes that are consistent with the training dataset. The optimal separation hyperplane is used as a decision boundary to minimize misclassifications. Then the

model separates the simulation data under the same configurations (Zhu and Blumberg, 2002). In one example Ballabio (2009) used SVM to map several soil properties such as organic carbon content and extractable Al concentration in the B horizon in mountainous areas of Northern Italy.

1.1.4. *k*-fold Cross Validation

In *k*-fold cross validation the training data is split into k smaller mutually exclusive subsets. Then the model is trained using k-1 of the folds as training data. Following that, the model is validated on the remaining fold of the data, and the process is repeated for all k folds. The performance measure then is reported for the average of the values computed across all folds (Kohavi, 2001; Yadav and Shukla, 2016).

1.1.5. Uncertainty Estimation

Existing Methods

DSM uses statistical methods to relate soil observations to environmental covariates, and if there are errors in the sampling or the sample locations, these errors will be incorporated into the model (Cressie and Kornak, 2003). Another source of uncertainty is the quality of environmental covariates used in DSM. The covariates from various sources can contribute errors to DSM because of their different acquisition scales, resolutions, or age (Lagacherie and Holmes, 1997). These errors should be quantified otherwise they will lead to poor DSM results (Arrouays et al., 2014b). International standards require 90% PI uncertainty quantification (Arrouays et al., 2014a), but even with the increased use of ML methods in DSM, quantification of uncertainty is relatively uncommon among digital soil mappers (Arrouays et al., 2017; Minasny and McBratney, 2002).

Although geostatistical methods have been used for uncertainty quantification in DSM (Mueller and Pierce, 2003; Wu et al., 2009; Zhao and Shi, 2010), methods used for measuring uncertainty in ML methods require further development and their use is quite novel (Vaysse and Lagacherie, 2017). Methods that have been used to quantify uncertainty in DSM may include: bootstrapping, empirical uncertainty quantification through data partitioning and cross validation, empirical uncertainty quantification through fuzzy clustering and cross validation, Bayesian and Monte Carlo methods, and

QRF (Malone et al., 2017; Meinshausen, 2006; Solomatine and Shrestha, 2009). All these methods have limitations. Bootstrapping depends on computational capabilities when this method is applied to big datasets because in this method many map realizations need to be predicted and stored in a database. Empirical uncertainty quantification methods are calculated from distributions of model errors and such predictions are not spatially uniform while also varying for different landscape situations (Malone et al., 2017). Bayesian and Monte Carlo methods (Solomatine and Shrestha, 2009) measure only certain sources of error and QRF can be used for RF only (Meinshausen, 2006).

Quantile Regression

Currently, there is not a comprehensive uncertainty assessment method that could be used with all ML methods; however, this knowledge gap could possibly be filled using a QR method (Koenker and Hallock, 2001). Unlike an ordinary linear regression, QR estimates the quantiles of a data distribution. While an ordinary linear regression model gives a picture of the central mean, the QR gives us a more complete picture of the conditional data distribution. An advantage of QR is that it is not sensitive to outliers (Hunter and Lange, 2000). Chamberlain (1996) developed two empirical applications of QR techniques. The first application was about the changes in the returns to schooling from 1979 to 1987 and the second application was about a union relative wage effect in 1987. In both cases the goal was to provide a more detailed description of the conditional distribution of wages. Other studies were conducted related to problems in labour markets such as studies by Fitzenberger (2012) in Germany and Schultz and Mwabu (1998) in South Africa. Rahmati et al. (2019) used QR along with ML methods in hydrology. They sought detailed descriptions of ML outputs to estimate uncertainty in ML models by calculating quantiles from certain portions of the model output. We believe the QR method can help soil scientists quantify uncertainty easily and extensively in their model predictions by the generic DSM framework to produce more reliable digital soil maps.

The QR method is further elaborated here. A linear quantile regression is similar to a simple linear regression function. In a simple regression function, to find the regression equation the square residual is minimized to construct the least square residual line (McGrew and Monroe, 2009; Schneider et al., 2010). Therefore, the goal in

a simple linear regression is to solve Equation 2 when the random sample dependent variables are $\{y_1, y_2, \ldots, y_n\}$.

$$\underset{\mu \in R}{min} \sum_{i=1}^{n}(y_i - \mu)^2 \quad (2)$$

In Equation 1, y is the response variable. i is the number of the variable. The character μ is the mean of response variables or in statistics it is referred to as unconditional population mean that is an element of the real values (R), and finally n is the last number of the variable (Koenker and Hallock, 2001; McGrew and Monroe, 2009). In Equation 1, μ can be replaced by a parametric function $\mu(\chi, \beta)$ that predicts a value y using a covariate x in equation 3.

$$\underset{\beta \in Rp}{min} \sum_{i=1}^{n}(y_i - \mu(x_i, \beta))^2 \quad (3)$$

In Equation 3, the parametric function $\mu(\chi, \beta)$ uses the independent variables to predict the dependent variables (Koenker and Hallock, 2001). This is called an unconditional regression function.

Unlike an unconditional regression function, in QR, a conditional relationship is sought between predictors and dependent variables $\mathrm{E}(Y|x)$. In QR, if dependent variables are called y and independent variables are called x, only desired quantiles of x are selected, and the linear model is established between y and x. For every quantile τ, a linear relationship between the predicted value, y, and the real observed value, x is assumed (Equation 4).

$$y = \alpha_\tau x + b_\tau , \quad (4)$$

where α_τ and b_τ are the parameters of the linear regression. These parameters are found by minimizing the sum of residuals of the portion that we are looking for. This is quite similar to a simple linear regression except for the condition that is enforced (Equation 5) (Dogulu et al., 2015b).

$$min \sum_{j=1}^{J} \rho_\tau(y_j - (\alpha_\tau x_j + b_\tau)) \quad (5)$$

The letter j is the number of variables, and J is the total number of variables. Therefore, the variables are x_1, x_2, x_3, …, x_J. The parameter ρ_τ is called the quantile

regression function for the quantile τ. The quantile regression function is defined as in Equation 7 (Dogulu et al., 2015b).

A name for the residual is first selected:

$$\varepsilon_j = y_j - (\alpha_\tau \chi_j + b_\tau) \quad (6)$$

Then, the QR function is written as follows:

$$\rho_\tau(\varepsilon_j) = \begin{cases} (\tau - 1).\varepsilon_j & \varepsilon_j \leq 0 \\ \tau.\varepsilon_j & \varepsilon_j > 0 \end{cases} \quad (7)$$

1.2. Research Rationale and Objectives

Three pivotal topics will be investigated and discussed in this thesis: using digital soil maps for forest management, using LiDAR to generate fine-resolution digital soil maps, and estimating uncertainty using QR to evaluate the reliability of digital soil maps. A primary purpose of producing digital soil maps was to reduce costs and increase accuracy of soil maps compared to conventional methods (Yang et al., 2011; Zhu et al., 2001). We believe digital soil maps not only can serve as a new approach to reduce cost and increase accuracy over large areas, but also can be used at local scales to help forest managers utilize and preserve valuable forest soils much more effectively. One question that many forest managers ask is how digital soil maps can contribute to improved forest management. Secondly, forest managers would like to know how accurate and reliable digital soil maps are. It would be very beneficial to develop moderately fast methods to measure prediction accuracy and uncertainty of digital soil maps. I discuss how digital soil maps can be used for sustainable forest management in Chapter 2 and then in Chapter 3, I present a new and novel approach in DSM called QR for measuring uncertainty easily and extensively. Therefore, this thesis is composed of two phases:

In the first phase of the study RF, as a regression method, will be used to map 5 soil properties: soil thickness, depth to carbonates, soil pH, coarse fragment content and clay content. The covariates used in this phase were derived from a LiDAR DEM originally prepared at one-meter resolution. The total covariate numbers were 16, three of which were categorical covariates. The other 13 covariates have been turned into

different resolutions and the final covariates have been transferred to 3 m resolution rasters. Objectives of phase one of this study are:

1. to produce soil maps using RF for the five soil properties using LiDAR derived covariates;
2. to validate soil property predictions using the *k*-fold cross validation method;
3. to discuss and illustrate how these maps are useful for forest management.

In phase two of this study, a new uncertainty quantification method in soil science called QR will be used. To do so, first, four ML methods will be used to map three soil properties: soil thickness, depth to carbonates, and soil pH. Then, the model prediction output will be used as QR input to generate uncertainty maps. Finally, uncertainty quantification and model performance will be assessed using PICP graphs and mean prediction interval (MPI) bar-charts. The objectives of phase two of this study are:

1. To develop a framework for producing local estimates of uncertainty by coupling ML models with quantile regression
2. To demonstrate the coupling using a variety of ML techniques for a case study
3. To evaluate the uncertainty estimations using metrics such as MPI and PICP

In this study several soil properties are predicted by fitting quantitative relationships. The statistical quantitative relationships used in this study are ML methods including RF, Cubist decision tree, kNN and SVM. The errors and model accuracy are measured using the *k*-fold cross validation method and uncertainty of the four ML methods are quantified using the QR method.

1.3. Thesis Structure

The thesis is divided into four chapters. Chapter 1 provides an: introduction, including a description of the context of the study, theoretical background and methods and research rationale and objectives. The methods section describes DSM and SCORPAN, LiDAR, ML methods, cross validation, and QR as an uncertainty estimation method.

In Chapter 2, soil properties were predicted using RF and their prediction results were validated using a nested 10-fold cross validation with 20 repeats. The soil properties that have been mapped are, soil thickness, depth to carbonates, soil pH, coarse fragment content and clay content. Chapter 2 presents a generic mapping framework in which the usefulness of digital soil maps in forest management are discussed. First, the digital soil maps produced for the soil properties are discussed in terms of data distribution and validation. Then, the usefulness of the high-resolution maps produced using high-resolution LiDAR data for soil hazard assessment is discussed.

Chapter 3 focuses on uncertainty estimation. In this chapter soil properties were modeled using RF, Cubist decision tree, kNN and SVM. The modelling results were validated using a nested 10-fold cross validation with 20 repeats. Property maps were produced using the four ML methods for 3 soil properties including soil thickness, depth to carbonates, and soil pH. The validation results of models produced using the 4 ML methods were compared. Then, to quantify the uncertainty, QR uncertainty methods were used to generate uncertainty maps and then they were compared in the uncertainty maps. Moreover, uncertainty predictions in QR using four ML methods were assessed using prediction interval coverage probability (PICP) plots and mean prediction intervals (MPI).

Finally, in Chapter 4 the overall thesis conclusions are discussed including a brief description of the background knowledge and the results obtained in Chapters 2 and 3. In the last part of the conclusion chapter, the challenges in this research and future possible research have been described.

Chapter 2.

Exploring the Novel Use of Soil and LiDAR-derived Terrain Information to Support Forest Management

2.1. Abstract

The goal of sustainable forest management is to conserve biological diversity and maintain forest ecosystem productivity. LiDAR can be used to collect information for forest inventories, biomass monitoring and ecosystem modelling. Development of new digital mapping technologies has increased our ability to monitor soil properties and changes within them. To generate digital soil maps, a numerical model is used to relate field soil observations and environmental variables to make new predictions for all areas to be mapped. High-resolution topographic data derived from LiDAR have been used for mapping non-forested regions. However, there are only a few instances in which LiDAR derived DEMs have been used for mapping forest soils over large areas. The objectives of this study are 1) to produce digital soil maps for five soil properties: soil thickness, depth to carbonates, soil pH, coarse fragment content and soil clay content; 2) to validate model predictions using the *k*-fold cross validation method; and 3) to discuss and illustrate how these maps can be useful for forest management and soil degradation prevention. The study area is the Eagle Hill Forest located west of Kamloops, British Columbia (BC), Canada (95 km^2). Covariates were derived from 1 m resolution LiDAR data. RF model was used to predict five soil property and maps were produced. A nested 10-fold cross validation with 20 repeats was conducted to estimate the accuracy of maps. The best validation results were obtained for modelling soil thickness with R^2 of 0.35 and concordance of 0.47. The soil maps for individual properties can be used directly in forest management or can be used to prepare interpretive maps such as maps of compaction and puddling hazards.[1]

[1] A version of the following chapter will be submitted to a peer reviewed journal for publication under the co-authorship of Chuck E. Bulmer, Margaret G. Schmidt, Brandon Heung, and William Bethel

2.2. Introduction

Depletion of forest resources in the early years of the 20th century caused the forestry community to focus on management activities starting in the 1960s (Nyland, 1992). World leaders met at the United Nation Conference on Environment and Development (UNCED) in 1992 in Rio de Janeiro to develop a nonbinding statement of forest principles. National policymakers in many countries now are committed to conserve biodiversity, forest productivity and the growth of forests in the long-term. Sustainable forest management can be defined as the efforts to conserve biological diversity, and to maintain the health and productive capacity of forest ecosystems and their role in watersheds and the global carbon cycle. The goal is to ensure that forest resources will continue to exist at some acceptable levels for the benefit of current and future generations (Szaro et al., 2000).

Technology and information systems have been important tools for sustainable forest management. To conduct successful forest management, scientists need to adopt a multidisciplinary approach that comprises the human research capability both to use the knowledge of the field that technology deals with and to improve analytical and decision making skills (Szaro et al., 2000). Moreover, an integration of training programs, networking, technology transfer and information management is necessary to build a significant research capacity. For this goal, forest simulation models that describe growth, succession, mortality, reproduction, and associated stand changes have been used (Peng, 2000; Vanclay, 1994). Remote sensing and airborne data can provide spatial information for forest management and ecosystem modelling. The data collected can be used to classify forested land cover and to track forest health, structure, biomass and natural disturbances (Wulder et al., 2004). The remote sensing and airborne data can also be used in agriculture, erosion monitoring and risk assessment, geomorphology and hydrology, and land use, and land cover mapping (Bahrawi et al., 2016; Henderson and Lewis, 1998; Natural Resources Canada, 2013).

LiDAR is an airborne remote sensing technique that can be used in forest management (Dubayah and Drake, 2000). LiDAR works like radar; however, it uses a laser beam. It is an active remote sensing technology that uses laser pulses to measure distance between objects. LiDAR can be incorporated into an airborne scanning system that produces image-like coverage of surface height (Asner et al., 2012). LiDAR

technology has extensive applications in forestry. It can be used to collect information for forest inventories and biomass monitoring such as tree location within plots and tree height (Wulder et al., 2012). It also can be used to collect information for ecosystem modelling such as vertical forest stratification, gas exchange, and canopy carbon content (Dassot et al., 2011). LiDAR systems can provide direct measurement of the height of the canopy, the topography of the subcanopy and the vertical distribution of intercepted surfaces between the top of the canopy and the ground. From these direct measurements, other forest structural features, such as above-ground biomass, are modelled or inferred (Dubayah et al., 2000). Like other technologies, LiDAR technology also has some advantages and disadvantages. The limitations of LiDAR remote sensing are that it has a small footprint and like other remote sensing techniques, LiDAR is restricted by clouds and dense atmosphere haze. The other disadvantage of LiDAR technology is that few LiDAR datasets are available, and they are costly. The big strength of LiDAR remote sensing compared to satellite remote sensing is the ability to directly measure canopy height, subcanopy topography, and vertical distribution of intercepted surfaces (Dubayah and Drake, 2000).

Successful forest management practices depend on the maintenance or the enhancement of forest productivity. One of the important variables in controlling above ground biomass productivity is forest soil (Ayma-Romay and Bown, 2019; Schoenholtz et al., 2000). Soil is a medium for growth of trees and a healthy soil can contribute to forest productivity (Weil and Brady, 2017). Foresters rely on knowledge of soil chemical and physical properties to assess the capacity of sites to support productive forests. Soil quality can be related to concepts such as the capacity of soil water retention, carbon sequestration, plant productivity, waste remediation and the capability to produce biomass (Schoenholtz et al., 2000). Furthermore, soil may provide an immediate sink of atmospheric CO_2 with proper forest management (Bruce et al., 1999). Therefore, soil quality and productivity can be used as an indicator of sustainable forest management (Burger and Kelting, 1999). Burger and Kelting (1999) have suggested 10-steps for soil quality monitoring and in step 7 they suggest evaluating the soil quality by using geostatistical techniques or some other type of spatial extrapolation to produce soil quality maps.

A soil map is a representation of a soil attribute distribution that is used to convey soil information (Yaalon, 1989). Early soil maps were produced for the purpose of land

valuation and taxation and agronomic planning (Brevik and Hartemink, 2010). Conventional methods of soil map production continued until the end of the 20th century. Conventional soil mapping was expensive and legacy soil maps were based on soil surveyor's conceptual mental models of landscapes. Moreover, legacy soil maps suffer from two major problems: lack of consistency and unknown accuracy (Yang et al., 2011; Zhu et al., 2001). Development of new technologies such as GPS, GIS, and remote sensing, and development of statistical and geostatistical techniques increased our ability to collect, analyze, and predict soil spatial information and properties. From the late 20th century a new era of soil mapping called DSM appeared (Brevik et al., 2016; McBratney et al., 2003a; Minasny and McBratney, 2016). DSM is a subdiscipline of soil science in which a numerical model relates field soil observations and environmental variables to make new predictions for a mapping area (Minasny and McBratney, 2016). Since the emergence of DSM many national soil maps have been produced for forest soils (Baritz et al., 2010; Morisada et al., 2004; Yang et al., 2011) but most of them are coarse and would not be applicable for site scale forest management.

In topographically varying areas, the success and accuracy of producing a good digital soil map depends on finding suitable digital elevation models (DEM) (Cavazzi et al., 2013). Many important model covariates such as hydrologic units, hillslope segments, slope gradient, aspect, flow networks, hillshade illumination and catchment boundaries can be derived from DEM rasters using GIS software. These covariates are used as independent variables in statistical models to predict soil properties in a DSM framework (MacMillan et al., 2004). DEMs can be derived from LiDAR data in forested and non-forested areas. The advantage of LiDAR derived DEMs is their accuracy and the small pixel size that is known as raster resolution (Haneberg et al., 2009; Liu, 2008).

High-resolution topographic data derived from LiDAR has been used for soil mapping purposes in non-forested regions and in precision agriculture. For instance, Shi et al. (2012) conducted a comparison between high-resolution LiDAR based data and satellite data in a non-forested area in the northern Vermont, USA. Their study showed that LiDAR based data showed significantly better performance than a USGS-sourced DEM. Campbell et al. (2013) studied soil resistance to penetration for two study areas in Alberta, Canada. In their study DEM related covariates were derived from LiDAR data. They used best-fitted regression between cone index and depth-to-water index and elevation to map cone index. They showed that cone index increased with increasing

depth-to-water index. There are only a few instances in which LiDAR derived DEMs have been used in DSM in forested systems. In one example, Kristensen et al. (2015) combined LiDAR data with fine-scale spatial carbon data relating to vegetation and the soil surface to describe the spatial distribution of carbon pools within spruce stands in Norway. Other studies conducted using LiDAR derived DEMs in forested areas have been carried out by Lidberg et al. (2020), Li et al. (2016), and Niemi et al. (2017). The potential of LiDAR in assisting in mapping of forest soil properties and forested systems has not yet been extensively explored. One of the goals of this study is to fill the knowledge gap of forest soil mapping using LiDAR data.

2.2.1. Soil Degradation Processes and DSM

Forest soil quality is an important indicator of forest productivity (Schoenholtz et al. 2000; Ayma-Romay and Bown 2019) and better forest management entails providing support and information to prevent soils from being degraded or eroded. In BC, Canada, the sensitivity of soils in terms of degradation and erosion have been described as the result of three soil degrading processes including, soil erosion, soil displacement, and soil compaction and puddling (Lewis and Carr, 1993). Each of these soil degrading processes have their own definitions, controlling site factors, management considerations and hazard assessment keys (Lewis and Carr, 1993). Soil erosion can be defined as the removal of the productive soil surface by water and wind (Ratta and Lal, 1998). Soil displacement can be defined as mechanical movement of soil that causes the exposure of unfavorable subsoils (Naghdi et al., 2009). Soil compaction is defined as the increase in soil bulk density because of rearrangement of soil particles caused by external forces (Sparks, 2012). Soil puddling is defined as the destruction of soil structure and reorientation of soil particles by running machinery on the soil when it is wet (Grigal, 2000)

In forested areas in BC where logging is planned, hazards need to be assessed so that the logging plan can consider any special concerns for soil conservation and adjust accordingly. For example, forest managers can avoid running machinery on sensitive sites, or can operate in the winter when the ground is frozen and will not be compacted by heavy equipment. Currently, field observations of every site are required to evaluate these hazards, and this is expensive. DSM can potentially simplify the process and make it more efficient. Data required for performing a soil disturbance

hazard assessment include climate information, slope and terrain information, site hydrology information and soil information (Lewis and Carr, 1993). Climatic information includes biogeoclimatic subzone/variant information. Slope and terrain information includes slope gradient, slope length, presence of slope instability indicators, and presence of hummocky terrain. Site hydrology information includes gully spacing, soil moisture regime and depth to seepage. Soil information includes, forest floor depth, soil thickness, depth to carbonates, soil texture, coarse fragment content, depth to unfavorable subsoil, type of unfavorable subsoil and depth to water-restricting layer (Lewis and Carr, 1993).

Digital soil assessment (DSA) and digital soil risk assessment (DSRA) are two new branches in DSM. The goal of DSA is to make quantitative models for soil attributes that are difficult to measure such as soil erosion, salinization and landslide susceptibility (Carré et al., 2007). Finke (2012) believes that DSM has reached its scientific maturity and that a shift can be made from DSM to digital function assessment. We believe that for each soil hazard assessment, a set of information is necessary; for example, for soil erosion, climate information, slope, depth to water restricting layer and coarse fragment content information is needed. For soil displacement hazard assessment, slope gradient, slope complexity, soil thickness, depth to carbonates and soil pH information is needed. For soil compaction assessment, clay content, coarse fragment content and moisture regime information is needed. The data for hazard assessment can be obtained from DEMs and from soil attribute maps produced using DSM methods. One of the objectives of this research is to introduce the advantages of using DSM and DEM-derived data for assessing soil degradation.

Five soil attributes that can be used in soil degradation assessment are, soil thickness, depth to carbonates, soil pH, coarse fragment content and clay content. Soil thickness is defined as the thickness of the soil material from the top of the mineral soil to underlying bedrock. It includes the developed soil layers as well as unconsolidated C horizon material and represents all soil material that could potentially support plant roots, whether developed or not (Bonfatti et al., 2018; Weil and Brady, 2017). In many landscapes, soil thickness depends on the balance between soil formation and soil erosion (Dosseto et al., 2011). Soil erosion is mainly caused by water erosion and mass movement that is governed by topography. Therefore, soil thickness is highly correlated with slope angle, relative height, curvature, and compound topographic index (Gessler et

al., 2000). Soil inorganic carbon is a big part of the large terrestrial carbon pool in soil in dry climates and where soils develop in calcareous parent materials. The soil carbon pool contains organic and inorganic carbon (Zamanian et al., 2016). Soils containing appreciable amounts of calcium carbonate can make an unfavorable subsoil condition that limits the growing conditions which in turn will increase the erosion rate and hazardous condition of the soil. High pH in carbonate rich soils is an indicator of the unfavorable condition (Lewis and Carr, 1993) because it influences the availability of many soil nutrients such as calcium (Thomas, 1996; van den Driessche, 1984). It also influences the abundance of soil microbial taxa, fungi and bacteria (Rousk et al., 2010; Urbanová et al., 2015).

Soil separates are classified based on their diameters. They are ordered in six levels of magnitude from boulders (1 m) to submicroscopic clays ($<10^{-6}$ m). Particles with diameter over 2 mm are considered as coarse fragments and they are not part of the fine earth fraction. Clay particles are smaller than 0.002 mm. They have a great surface area which gives them the capacity to absorb water and other substances such as soil nutrients (Weil and Brady, 2017). Coarse fragment content and clay content are directly related to soil productivity, and it is one of the forest management challenges to identify forest lands that are most productive for tree growth in terms of coarse fragment content and clay content (Carmean, 1996). Coarse fragment content and texture of the upper 30 cm of mineral soil can be used to assess soil compaction and puddling hazard (Curran et al., 2007). Soil with coarse fragment content of less than 70% and clay content of over 20% can be considered as soil with very high risk of compaction and puddling (Lewis and Carr, 1993).

2.2.2. Objectives of the Study

The objectives of this study are: 1) to produce soil maps using RF for five soil properties: soil thickness, depth to carbonates, soil pH, coarse fragment content and soil clay content using LiDAR derived covariates; 2) to validate soil property predictions using the *k*-fold cross validation method; and 3) to discuss and illustrate how these maps are useful for assessment of soil degrading processes in forest management.

2.3. Methods

2.3.1. Study Area

The study area is the Eagle Hill Forest located west of Kamloops, BC, on the northern side of Kamloops Lake. The forest covers approximately 95 km^2 between $50^{\circ}46'\ 49.9''$, $-120^{\circ}57'\ 02.5''$ and $50^{\circ}55'\ 05''$, $-120^{\circ}46'\ 36.7''$ geographic coordinates (Figure 2.1). The study area is delineated by Criss Creek in the northwest, Sedge Creek in the west, Kamloops Lake and Thompson River in the south, Sparks Creek and Lake in the north, and Red Lake in the northeast. Sabiston Creek and Sabiston Lake separate the area into two distinct uplands in the northeast and southwest with maximum elevation of 1452 m and 1389 m respectively (Figure 2.1). The elevation of the study area is between 545 m and 1455 m above sea level. The climate is semi-arid due to being located in a rain shadow with annual precipitation of about 380 mm. The annual temperature in Kamloops is between “-10” in winter and about “30” degrees Celsius in summer (Canada, 2013).

The study area is mostly in the interior Douglas-fir biogeoclimatic (BEC) zone, dry cool subzone and Thompson variant (IDFdk1) and very dry hot and Thompson variant (IDFxh2). There is also a small portion in the southern part of the study area that belongs to the ponderosa pine zone and very hot dry subzone (PPxh2) (Figure 2.2). The Interior Douglas fir zone (IDF) covers 5.5% of southern BC at mid-low elevations. The lower elevation ranges from 130 m to 900 m and the upper elevation ranges from 1200 m to 1600 m. The most common tree species in this zone is Douglas fir (*Pseudotsuga menziesii*). However, at higher elevation, lodgepole pine (*Pinus contorta*) is widespread. In drier and hotter subzones such as IDFxh2 ponderosa pine (*Pinus ponderosa*) occurs. The minor species in this zone is trembling aspen (*Populus tremuloides*), which is also common in the IDFdk. Many other minor species restricted to specific areas include grand fir (*Abies grandis*), western white pine (*Pinus monticola*), Rocky Mountain juniper (*Juniperus scopulorum*), balsam popular (*Populus balsamifera*), choke cherry (*Prunus virginiana*), alders (*Alnus*) and willows (*Salix*). PP zone is the driest forest zone In BC and occurs below the moister and cooler IDF zone. It is close to the drier treeless grasslands of BG zone at lower elevation. Douglas-fir (*Pseudotsuga menziesii*), trembling aspen (*Populus tremuloides*), and lodgepole pine (*Pinus contorta*) tree species can be found in this zone (Faculty of Forestry, UBC, 2009; Roberta, 1948).

Several soil types occur in this study area; some of the most important ones are Tunkwa, Gisborne, Timber, and Glossey (Figure 2.2). Tunkwa soil covers most of IDFdk1 BEC subzone in this study area, and in Thompson and Fraser plateau physiographic regions that can be found between Kamloops Lake and Merritt, and north of Kamloops Lake and Thompson River. Tunkwa is described as a silt loam or loam and its pH is slightly alkaline. Parent material is composed of morainal deposits (till) that are associated with volcanic bedrock and generally are stony. This soil can be found at elevations ranging from 900 m to 1455 m in this study area (Figure 2.2) The common subgroup of soil is Orthic Gray Luvisol. Other less common soils such as degraded Eutric Brunisol, Brunisolic Gray Luvisol, Lithic Gray Luvisol, Orthic Black Chernozem, Orthic Dystric Brunisol and Orthic Sombric Brunisol can be found in this soil type (Young et al. 1992).

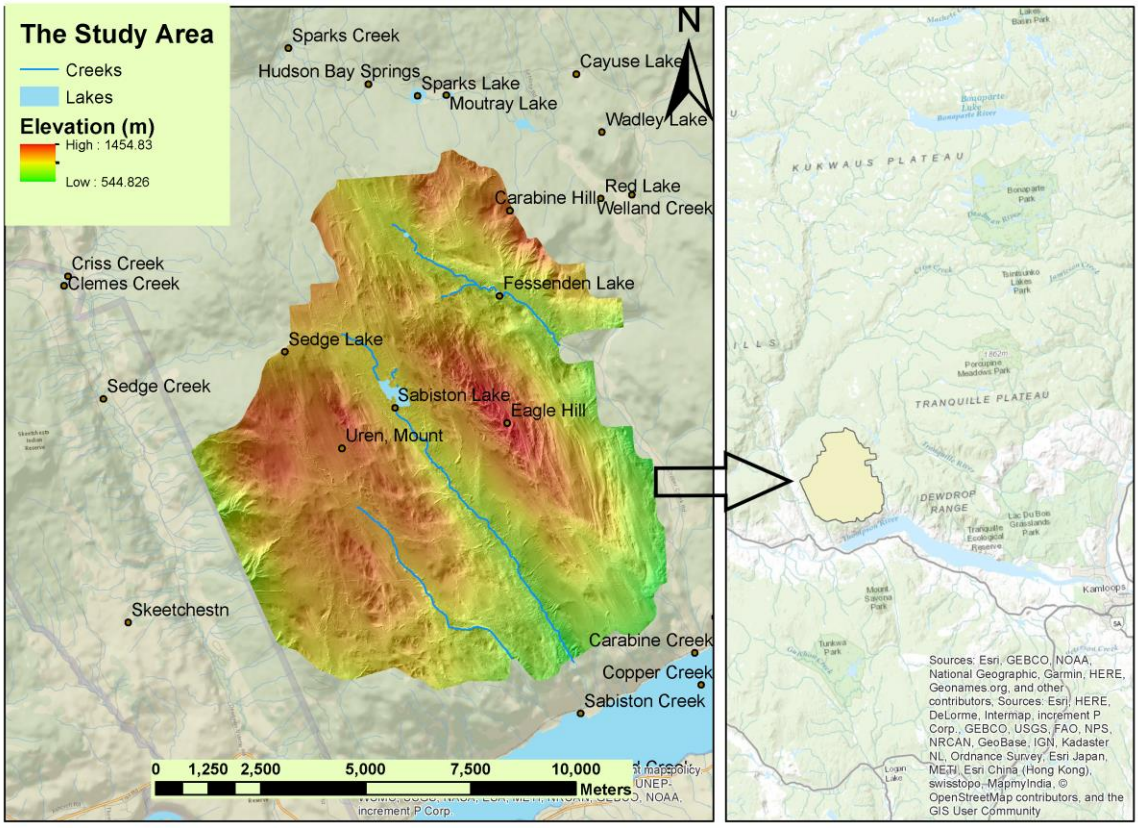

Figure 2.1. The study area and water bodies and rivers.

The IDFxh2 BEC subzone in the study area is covered mostly by the three other soils including Gisborne, Timber, and Glossey. All those three soils can be found in the Thompson and Fraser plateau physiographic region in the dry interior Douglas fir zone.

Gisborne covers the lands with elevation ranging from 800 m to 1200 m around the central Sabiston Creek and Lake (Figure 2.2) and one strip of land on the northern side of the study area with the same elevation range. The soil texture of Gisborne is gravelly to very gravelly sandy loam to sand, and the parent material of this soil is fluvioglacial. The most common soil subgroup is degraded Eutric Brunisol and less common soil subgroups are Orthic Humo-Ferric Podzol, degraded Dystric Brunisol, Orthic Dark Gray Chernozem and carbonated Black and Melanic Brunisol (Canada, 2013).

Timber soils can be found mostly on the southeast, and north corners and partly on the west corner of the study area (Figure 2.2). The range of elevations are from 544 m to 1200 m in this study area. Timber soil is moderately alkaline, and the texture of this soil can be silt loam and silty clay loam. It is slightly to moderately stony soil. The parent material of this soil is morainal deposits (till) that is associated with volcanic bedrock. The most common soil subgroups are degraded Eutric Brunisol and Lithic Eutric Brunisol. The less common soil subgroups are Orthic Brown, Orthic Gray Luvisol, Orthic Dark Gray Chernozem, Orthic Dark Brown Chernozem, Eutric Brunisol and Orthic Regosol (Young et al. 1992).

Glossey soil can be found in some patches on the south and west sides at elevations ranging from 544 m to about 1000 m in the study area (Figure 2.2). This gravelly soil has soil texture of sandy loam and slit loam with overlaying sand or loamy sand. In terms of acidity it is neutral to basic and it occurs on fluvioglacial deposits derived mainly from volcanic bedrock. The most common soil subgroup found in this soil is degraded Eutric Brunisol, and less common soils that occur in this soil are Orthic Brown Chernozem, degraded Eutric Brunisol, Orthic Dark Gray Chernozem, carbonated Dark Brown Chernozem and Melanic Brunisol (Young et al., 1992). There are also three other soil types that cover very small areas and are not discussed here including Cavanaugh, Commonage, and Trapp lake.

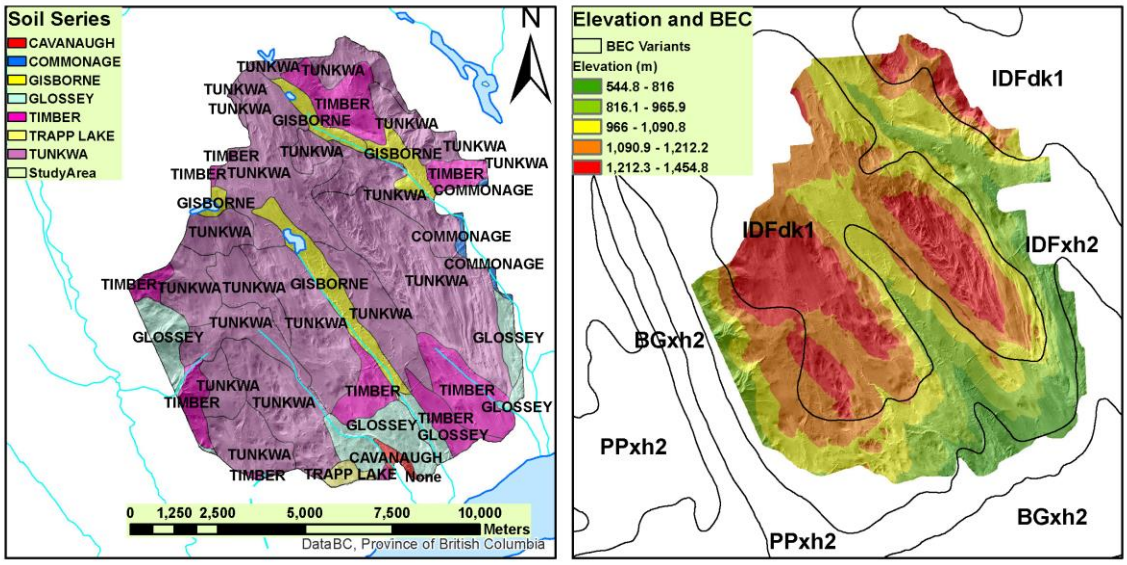

Figure 2.2. Soil types and BEC subzones in the Eagle Hill Forest study area.

2.3.2. DSM workflow

The workflow for DSM in this study starts with integration of observation data values (dependent variables) collected at sample locations and their associated values from environmental covariates derived mainly from LiDAR data (independent variables) into a data table, referred to here as the full dataset in Figure 2.3. The full dataset consisting of all observation points for each attribute was used as input data into a statistical model, and parameter optimization was carried out using a 10-fold cross validation method. Then the fitted model was used to generate spatial predictions of soil patterns using the stack of environmental covariates as shown in step 2 of Figure 2.3.

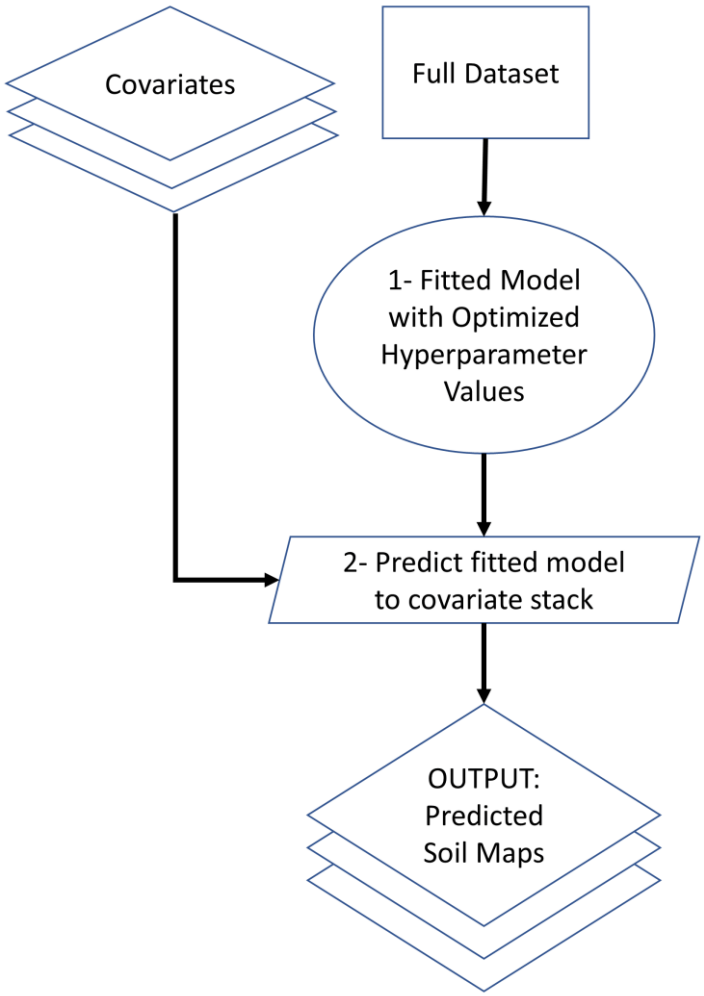

Figure 2.3. DSM workflow: the workflow was embedded in a cross-validation process where 10-folds of the data were created to tune the hyperparameters of the model, and the tuned model was used to generate a soil map.

2.3.3. Soil Sampling in the Study Area and Data Acquisition

In this study, five soil properties have been modelled: soil thickness, depth to carbonates, soil pH, coarse fragment content and clay content. Three different methods were used to identify sampling and observation points for the evaluation of these soil properties. The three approaches used were conditioned Latin Hypercube (cLH), random sampling on road cuts (RC) and opportunistic point (OP) locations. For depth to carbonates and soil thickness, points located using all three methods were used. Samples for determination of soil pH were only collected from the cLH and RC plots, as were observations of coarse fragments and collection of samples for clay content

determination. For soil thickness, depth to carbonates, and soil pH, data were collected at 410, 171, and 230 sites respectively. For coarse fragment and clay content 231 and 233 sample points were used (Figure 2.4).

Conditioned Latin Hypercube (cLH) Sampling

The cLH sampling method, implemented in the R package “clhs”, was used to specify sampling locations. Sixteen environmental covariates were generated using a LiDAR DEM originally prepared at 1 m resolution and these were used to inform the sampling scheme. A constraint was added so that the locations for sampling were within 200 m distance from roads. Within each of the two BEC subzones (Faculty of Forestry, UBC, 2009), 100 site locations were identified, and therefore, the total cLH sampling sites were 200. A 30 cm deep pit was dug in each cLH plot and 5 samples were collected: two samples from forest floor and 3 mineral soil samples from the depths of 0 to 5 cm, 5 to 15 cm and 15 to 30 cm.

Randomized Road Cut (RC) Sampling

A spatial layer of the road network was converted to a polygon using a 5 m buffer, and 50 random RC sites were selected from the buffering zone. In each road cut plot, a 100 cm deep pit was dug, or a face was exposed, and 7 samples were collected from the pit or road cut: two samples from forest floor and 5 mineral soil samples from depths of 0-5, 5-15, 15-30, 30-60, and 60-100 cm. With the sampling constraints, 43 of the RC sampling sites were visited.

Locations Recorded on a GPS device: Opportunistic Points (OP)

Additional observations were recorded at sample points without previous plans based on surveyors’ decisions in the field. These were called opportunistic points (OP) and there were 237 of these sites. The coordinates of locations of those points were recorded on a GPS device. Most of the GPS sample points were on road cuts. Some other sample points at which depth to carbonates or soil thickness could be measured were also recorded.

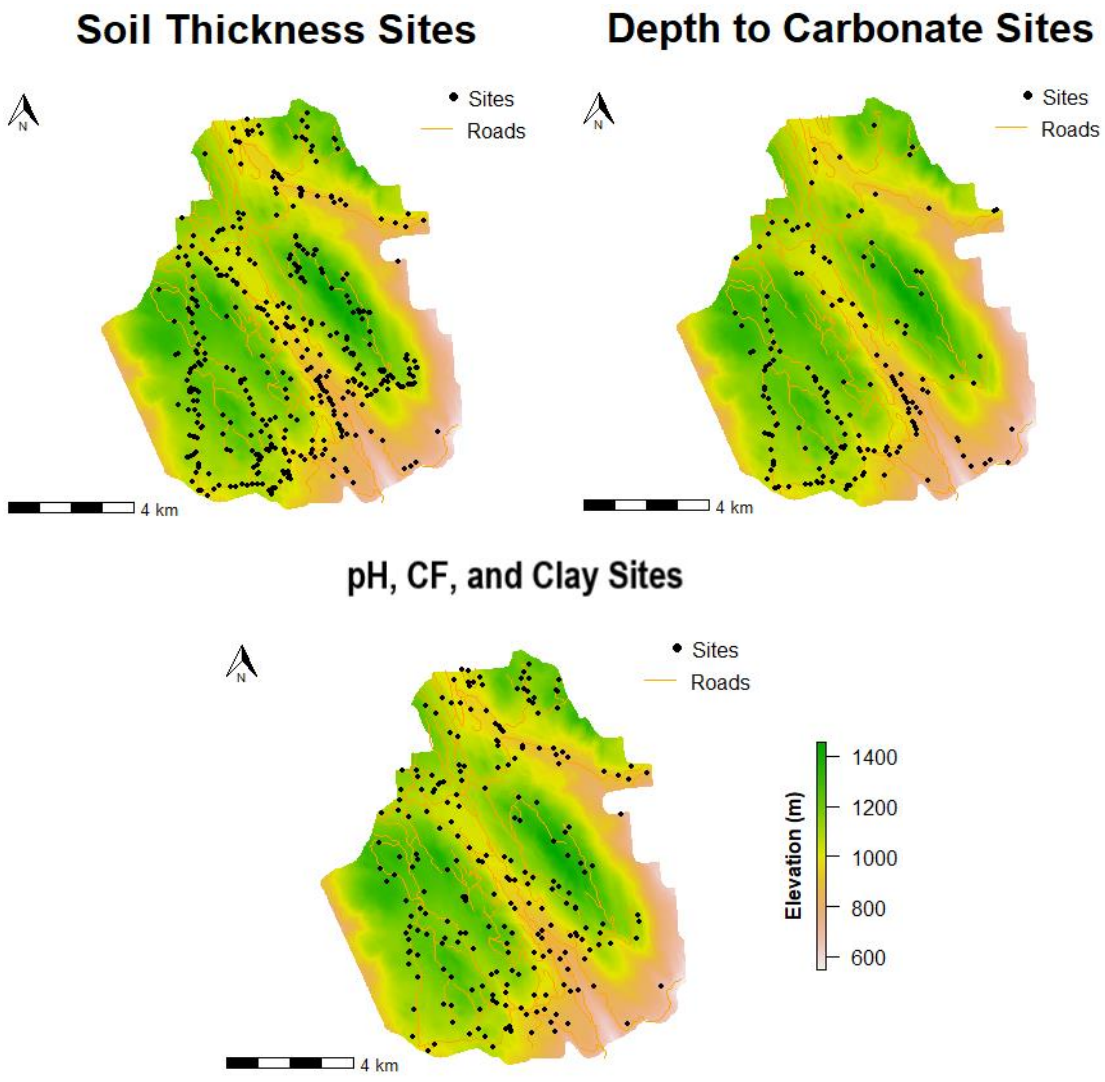

Figure 2.4. Sampling sites for soil thickness, depth to carbonates and lab measured properties (soil pH, coarse fragment content and clay content) in the study area.

2.3.4. Dependent Variables

Soil Thickness

Soil thickness was directly measured on RCs (Figure 2.6) where bedrock was visible and could be measured (Figure 2.5). At RC and OP sites where soil thickness exceeded the depth of visible soil in the road cut, an estimation of soil thickness was made based on the presence or absence of exposed bedrock and indicators of a shallow soil in the surrounding area, as well as elevation, slope percentage and position. Lack of exposed bedrock, lower elevation, and flat locations were signs of deeper soils. From

410 sites measured for thickness, 155 sites were cLH sites; 230 sites were OP sites; and 25 were RC sites. Bedrock outcrops were recorded at 56 of the OP sites based on visual inspection and assigned a depth of 0 to 20 cm.

Figure 2.5. Soil thickness observed on a road cut.

Depth to Carbonates

Estimates of depth to carbonates in this study were mostly obtained from OP and RC sites, along with a small number from cLH sampling sites. From 171 points, 26 points were cLH sites; 105 points were OP sites; and 40 points were RC sites. There were relatively few cLHS sites where depth to carbonates was recorded, because carbonated soils were not close to the soil surface (i.e. within 30 cm). In some places, considerable time was required to excavate and observe the presence of carbonates as deep as 200 cm of soil depth. At sample sites, while soils were excavated, 10% HCl was dropped down the soil profile and the depth where effervescence occurred was recorded. Where effervescence did not occur, we assumed that either the carbonated horizon did not exist, or it was deeper. To make the soil digital map for depth to carbonates we only used the sites in which we were certain about the depth to carbonates (171 sites).

Soil pH

Soil pH data was collected from 230 sites. It was measured in water for samples from each designated depth: 0-5, 5-15, and 15-30 (Carter, 1993). From the total of 230 points, 191 plots were cLHs sites and 39 were RC sites. The three pH values collected from each sampling site were weight averaged according to their sampling depth (0-5, 5-15, 15-30 cm).

Coarse Fragment Content

Coarse fragment data were collected from a total of 231 sites including 192 cLH sites and 39 RC sites. Coarse fragment content measurement was conducted in two stages. In the first stage, volumetric content of coarse fragments larger than 10 mm in diameter was assessed in the field for the 0-30 cm depth by visual comparison to area percentage charts. Then samples were collected from 0-5 cm, 5-15 cm, and 15-30 cm. In the field, samples were passed through a 10 mm sieve with the <10 mm fraction being placed in plastic bags. The collected samples were transferred to the soil science lab and the second stage of coarse fragment content measurement was carried out by passing the samples through a 2 mm sieve. Following that, the coarse fragments on the sieve were weighed and coarse fragment content (2 mm – 10 mm) was calculated gravimetrically.

To calculate the total coarse fragment content, the coarse fragment content measured in the field (field >10 mm) was summed with the coarse fragment content measured in the lab. For this goal, it was necessary to convert the lab coarse fragment measurements from a weight to a volume basis. If we name coarse fragments larger than 10 mm as CF>10, and coarse fragment content from the lab as fine coarse fragment (FCF) and consider bulk density as Db and particle density as Dp, we then can calculate the total coarse fragment content from equation 1:

$$CF > 10 + (1 - CF > 10) \times \frac{FCF/Dp}{(\frac{FCF}{Dp} + (1 - FCF))/Db} \qquad (1)$$

In this equation Dp was 2.65 and Db was estimated for each site. FCF consists of the coarse fragments less than 10 mm in diameter. The complexity of this equation can be explained if we consider equation 2. In Equation 2, (1-FCF) is the volume of fine fraction passed from the 2 mm sieve.

$$\frac{FCF/Dp}{(\frac{FCF}{Dp}+(1-FCF))/Db} = \frac{Volume\ of\ FCF}{Total\ Volume\ of\ a\ Sample\ Collected\ for\ Lab\ Analysis} \quad (2)$$

Since direct measurement of Db for each site was not conducted in this study, an indirect measurement method was used by making a RF model for Db using a regional subset of the BCSIS data (Sondheim and Suttie, 1983). Four variables were extracted from the BCSIS data subset including bulk density, organic carbon content, sand content and clay content. Then a RF model was trained using bulk density as a dependent variable and organic carbon content, sand content and clay contents as covariates. After making the model it was used to estimate Db for our own sites using our dataset. The results of the three depths were weight averaged to make only one value for the 0- 30 cm layer at each site. Lastly, a new variable was defined in the dataset for total coarse fragment content, and it was used in RF modelling analysis and map production.

Clay Content

Clay content data collected from 233 sites were used in this study including 194 cLH sites and 39 RC sites. Clay content was measured using the hydrometer method (Carter, 1993). The three depth values (0-5, 5-15, 15-30 cm) of clay content for each site were weight averaged and one value for each plot was calculated to be used in the modelling.

2.3.5. Environmental Covariates

A 1 m spatial resolution LiDAR DEM was provided for this project by the BC Ministry of Forests, Land and Natural Resource Operations. Before generating other covariates, the DEM was passed through an adaptive filter in WhiteBox GIS (Lindsay, 2015). The adaptive filter algorithm passes a window over all cells and calculates the average value centered on each cell. If the difference of the absolute average value and the central cell value in that window was more than a threshold, the filter will assign the mean value to that cell value (Lindsay, 2015). The filtered DEM was further processed to reduce the perturbing effects of the road network on topographic derivatives derived from hydrologic flow. Roads were identified from access network datasets that were manually checked for completeness and accuracy. Then a 4 m buffer was applied to the road lines and the underlying DEM cut, and then gap filled to recreate a smooth slope where the ditches and road fill were previously visible in the LiDAR dataset. For the next

step,15, 3 m resolution covariates were derived from the pre-processed DEM in SAGA GIS. All covariates were upscaled to 3 m, 9 m, 15 m, 27 m, and 30 m resolution grids. Then all variables at different resolutions were downscaled to 3 m resolution (Behrens et al., 2018). Forty-two varieties of the first 15 variables listed in Table 2.1, plus three categorical covariates (#16-18 in Table 2.1) were used to make the models.

Table 2.1. List of covariates.

No	Variable Name	Description
1	Full hillshade illumination	Shaded relief or hillshade can be calculated from a DEM. The azimuth angle of the hillshade illumination can be adjusted. To increase the influence of this variables, hillshade illuminations have been calculated from different angles (0, 120 and 240) and the results were averaged to make one variable.
2	Diurnal anisotropic heating: an adjusted aspect	The highest values are found on sloping hillsides facing southwest. The anisotropic adjustment is used to incorporate the fact that hot temperature and intense sunshine in the afternoon (i.e. when the sun is west of south) causes more plant moisture stress than the site exposures which are still relatively cool in the morning (i.e. sun is east of south)
3	Digital elevation model (DEM)	A 3D representation of the earth surface created from terrain elevation data
4	Standardized height	A measure of relative elevation
5	Multiresolution index of valley bottom flatness	An identification of valley bottoms from DEM; SAGA GIS uses slope and elevation to classify valley bottoms as flat, low areas.
6	Multiresolution ridge top flatness	Unlike multiresolution index of valley bottom flatness that is used to identify the areas of deposited material, multiresolution ridge top flatness is used to identify high flat areas at the range of scales. (Gallant et al., 2000).
7	Plan curvature	A type of curvature that emphasizes different aspects of the slope; A curvature can be defined as profile, planform and standard. The planform curvature (usually called plan curvature) is perpendicular to the direction of the maximum slope.
8	Profile curvature	Parallel to the direction of the maximum slope
9-10	Openness positive and openness negative	These terms refer to the 'exposure' of a point on the earth surface. High degrees of positive openness occur on convex topographic highs with high exposure to the atmosphere. Concavities on the lower parts of the landscape have low values of positive openness (Yokoyama, 2002).
11	Slope	The ratio of vertical change to horizontal change between to distinct points
12	Slope height	The vertical distance from slope toe to crest
13	Topographic Wetness Index (TWI)	Ln of Local upslope area draining through a certain point per unit contour length (α) divided by local slope (tan β) [ln(α/tanβ)] (Sørensen et al., 2006)
14	Valley depth	Vertical distance from crest of a slope to a channel network base level
15	Vegetation canopy height model	Also called canopy height model and is the distance between ground and top of the trees. It is obtained from LiDAR data by subtracting digital terrain model from digital surface model (Wasser, 2017).
16	BEC subzone	A system of ecological classification widely used in British Columbia, Canada (Pojar et al., 1987). This is meant to provide the model with information similar to many climate variables

No	Variable Name	Description
17	Geonut	An index derived from the bedrock geology map to reflect light and dark coloured rocks and minerals. Soils developed from dark coloured rocks are generally thought to contain more plant nutrients than those from light coloured rocks and minerals.
18	Geotex	Very similar to geonut, derived from the same bedrock geology dataset. Rocks with coarse grains such as igneous intrusive rocks have large size mineral grains and are thought to weather to sandy parent materials. This is very common in the coast mountains in almost the whole western part of BC, Canada. Fine grain size rocks such as basalt and shale tend to produce soils with finer texture like in many parts of the interior of BC, Canada.

2.3.6. Machine Learning Model

The process of using a ML model can be described as training a statistical model using predictor and response variables to make the model and using it to make new predictions in a study area (Heung et al. 2016; Witten et al. 2016). In this study, the RF model was used, and its output prediction results were validated. RF is a non-parametric ML technique in which many trees are trained, and the results are obtained from the predictions from an ensemble of trees. The RF algorithm incorporates bagging, where numerous trees are constructed independently using a bootstrap sample of the dataset. Bagging or bootstrap aggregation is used to reduce the variance of an estimated prediction function. The node splitting rules are randomly selected by using a subset of independent variables. RF randomizes the partitioning procedure by considering a parameter called *Mtry* which is the number of variables that are tested at each split. In this way RF reduces prediction errors, measured from variance reduction resulting from averaging. Final predictions are made based on average weights over the ensemble (Breiman, 2001; Heung et al., 2016).

2.3.7. Variable Importance Plots

RF model made with Caret package (Kuhn, 2019) in RStudio application has a built in variable importance score. The function automatically scales the importance of variables to be between 0 to 100. The most important variable receives score 100 and the least important one receives score 0. The scores can be used as a list or a plot. The plot method visualizes the results (Kuhn, 2019). In this study variable importance plots were generated for RF models.

2.3.8. *k*-fold Cross Validation

Model validation shows us the proportion of the variance for a dependent variable that can be explained by the independent variables. There are different methods of validation in DSM such as random holdback, leave one out cross validation, and *k*-fold cross validation (Malone et al., 2017). In this study we used *k*-fold cross validation, with k = 10.

k-fold cross validation sometimes is named rotation estimation. In *k*-fold cross validation, the dataset D is split into K mutually exclusive subsets of $D_1, D_2, \dots, D_k$ which are roughly the same size. A model is trained and tested k times, each time $t \in \{1, 2, \dots, K\}$. The model is trained on $D \backslash D_t$ and tested on D_t by calculating the accuracy metrics: R^2, concordance correlation coefficient and root mean squared error (RMSE). The summation of the number of correct classifications divided by the number of instances in the dataset will create the cross validation estimate of accuracy.

$$acc_{cv} = \frac{1}{n} \sum_{(v_i, y_i) \in D} \delta(I(D \backslash D_{(i)}, v_i), y_i) \qquad (1)$$

In this equation $D = \{x_1, x_2, \dots, x_n\}$ is a dataset. This dataset consists of n labelled instances and $x_i = (v_i \in V, y_i \in Y)$. A classifier is a function that maps an unlabelled instance to a label using internal data structures. An inducer I, or an induction algorithm such as random forest, builds a classifier from a given dataset. As mentioned previously, a classifier C maps an unlabeled instance $v \in V$ to a label $y \in Y$. A RF model maps D into C. To obtain a complete cross-validation, all $\binom{m}{m/k}$ possibilities should be averaged for selecting m/k instances out of m. δ refers to the cross validation function (Kohavi, 2001).

Three important statistics that are measured by a *k*-fold cross validation method are R^2, concordance, and RMSE. R^2 is the square sample correlation coefficient (Pearson's) between the observation and their corresponding predictions. Lin's concordance correlation coefficient or simply concordance evaluates the accuracy and precision of the relationship. RMSE is the standard deviation of the residuals and is called prediction error (Malone et al., 2017).

To generate the most unbiased validation results using a 10-fold cross validation, a nested 10-fold cross validation approach with 20 repeats was used. In the nested 10-fold cross validation the best hyperparameter of the RF is selected in the inner loop and the trained model with the best hyperparameter is tested 10 times on unseen randomly selected subsets of the dataset (Wainer and Cawley, 2018). The validation results of the 10 times accuracy measurement are saved and the whole process is repeated 20 times. All validation results finally are averaged to obtain the most accurate validation metrics.

2.4. Results and Discussion

2.4.1. Modelling and Validation

Training Data in BEC subzones

Before starting to model, an analysis was conducted to see how training data for each soil attribute are distributed in BEC subzones in the study area. For this goal, training datapoints in the two BEC subzones were compared using boxplots in Figure 2.6 and values in Table 2.2. According to Figure 2.6 and Table 2.2, soil thickness data in IDFdk1 compared to IDFxh2 have been collected from shallower soils as the lowest quartile is 20 cm for IDFdk1 and is 100 cm for IDFxh2 BEC subzones. However, in both subzones the median value (200 cm) and highest quartile (250 cm) are the same, although there are some potential outliers in IDFxh2 data that show very deep soils in IDFxh2 (Figure 2.6 & Table 2.2). The values for very deep soil were not removed from the dataset because they represent true determinations of soil thickness and were not errors.

Comparison of depth to carbonate data collected in the two BEC subzones shows that carbonated soils can be found in deeper soils in IDFdk1, and carbonates are closer to the surface soil in IDFxh2 as the lowest quartile of the boxplot shows a depth of 58 cm compared to IDFxh2 that shows 30 cm depth (Figure 2.6 & Table 2.2). Comparison of soil pH in the two BEC subzones shows more acidic soils in IDFdk1 which is in accordance with the results of depth to carbonates in the two BEC subzones. In both the lowest quartile and the highest quartile, the soil pH values in IDFdk1 are lower with values of 5.14 and 5.96 compared to those related to IDFxh2 that are 5.61 and 6.31 (Figure 2.6 & Table 2.2). This means that in the IDFdk1, carbonated soil can be

found deeper in soils and soil pH is more acidic, and in IDFxh2 carbonated soil can be found closer to the soil surface and the soil pH is more basic. Comparison of coarse fragment content in the two BEC subzones shows that soils with both less coarse fragment content with value of 9.14% and coarser fragment content with value of 13.88% for lowest and highest quartiles respectively can be found in IDFdk1. Clay content comparison in the two BEC subzones shows less clay content in IDFdk1 (Figure 2.6 & Table 2.2).

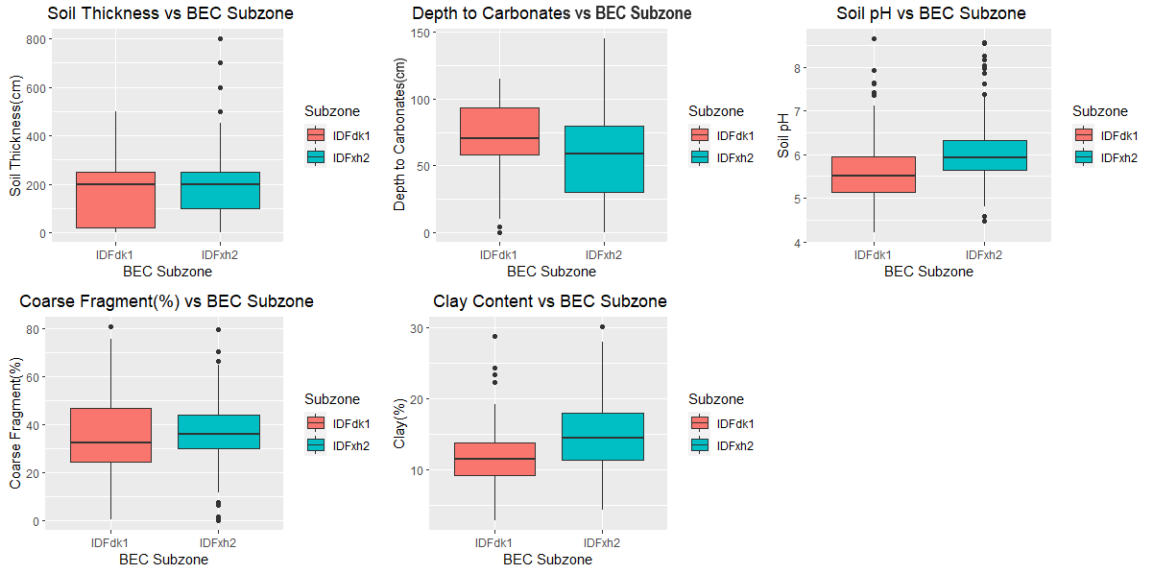

Figure 2.6. Comparison of 5 soil properties from training data between the two BEC subzones

Table 2.2. Statistics for the training data in the two BEC subzones.

Soil property	BEC Subzone	Number of Points in each Subzone	Data statistics						
			Median	Mean	Min	Max	Range	Lowest Quartile	Highest Quartile
Soil Thickness (cm)	IDFdk1	190	200	162.8	0	500	500	20	250
	IDFxh2	220	200	194.7	0	800	800	100	250
Depth to Carbonates (cm)	IDFdk1	61	70	69.9	0	115	115	58	93
	IDFxh2	110	59	58.09	0	145	145	30	80
Soil pH	IDFdk1	116	5.51	5.64	4.21	8.65	4.44	5.14	5.96
	IDFxh2	114	5.93	6.06	4.48	8.56	4.08	5.61	6.31
Coarse Fragment Content (%)	IDFdk1	117	32.25	35.59	0.6	80.7	80.1	24.3	47.33
	IDFxh2	114	36.15	36	0.06	79.35	79.29	29.67	43.98
Clay Content (%)	IDFdk1	103	11.5	11.81	2.85	28.83	25.98	9.14	13.88
	IDFxh2	97	14.5	14.94	4.28	30.17	25.89	11.32	18

Covariates for BEC subzones and Data Points

An elevation model of the study area depicted in Figure 2.9; map 1, shows the elevation range from 544.8 m to 1454.8 m. The three BEC subzones also are shown on all maps (Figure 2.9). The elevation map shows that IDFdk1 BEC subzone is associated with high elevations with mean elevation value of 1177 m (Table 2.3), and IDFxh2 is associated with mid elevations with mean elevation value of 971 m (Table 2.3 & Figure 2.7). The lowest elevations can be found in PPxh2 subzone which is 544 m (Table 2.3). The distribution of cell values for three other covariates also was investigated (Table 2.3 & Figure 2.7). Boxplots of slope raster cell values show that steeper slopes can be found at lower elevation BEC subzones and the steepest slopes are in PPxh2. The statistics of cell values of TWI and negative openness covariates show that there is not a significant difference between them in different BEC subzones (Figure 2.7).

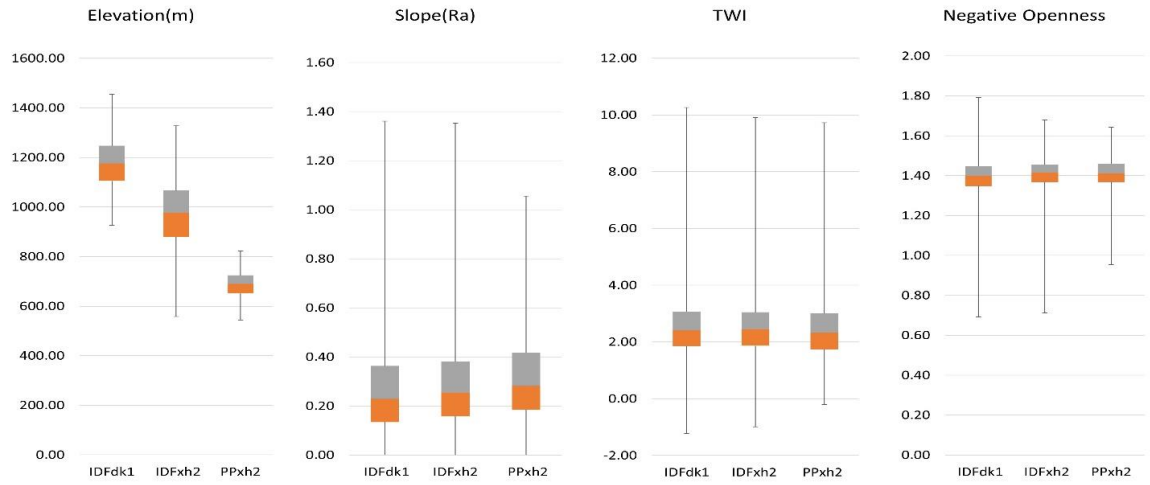

Figure 2.7. Data distribution of four covariates in the three BEC subzones.

Table 2.3. Statistics for four covariates used in modelling.

Covariate	BEC Subzone	Covariate Statistics							
		Mean	Median	Min	Max	Range	SD	Lowest Quartile	Highest Quartile
Elevation(m)	IDFdk1	1177	1176	926	1455	529	96	1105	1247
	IDFxh2	971	977	558	1328	769	130	880	1066
	PPxh2	688	690	545	822	277	53	651	725
Slope (Ra)	IDFdk1	0.26	0.23	0	1.36	1.36	0.17	0.13	0.36
	IDFxh2	0.28	0.25	0	1.35	1.35	0.16	0.16	0.38
	PPxh2	0.3	0.28	0	1.06	1.05	0.15	0.18	0.42
Wetness index (TWI)	IDFdk1	2.51	2.41	-1.23	10.26	11.5	0.98	1.85	3.06
	IDFxh2	2.54	2.45	-0.99	9.9	10.89	1.03	1.86	3.05
	PPxh2	2.47	2.32	-0.21	9.72	9.93	1.04	1.74	3.01
Openness negative	IDFdk1	1.39	1.4	0.69	1.79	1.1	0.08	1.35	1.45
	IDFxh2	1.41	1.42	0.71	1.68	0.97	0.07	1.37	1.46
	PPxh2	1.41	1.41	0.95	1.64	0.69	0.06	1.37	1.46

Statistics for 4 important covariates for data points show that the sampling coverage was best for soil thickness (Figure 2.8). However, the sampling for depth to carbonates is biased and needs improvement. Depth to carbonate points are taken from slightly lower elevations overall, with fewer samples on the steep slopes, in slightly lower landscape positions, and in drier and more convex areas (Figure 2.8).

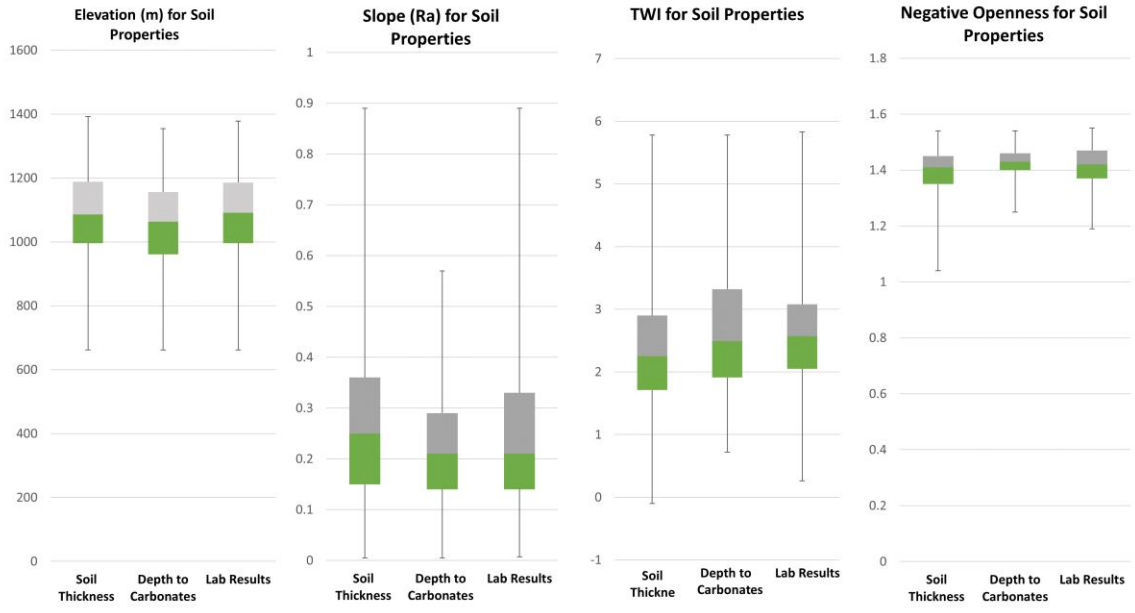

Figure 2.8. Boxplots showing values for 4 topographic covariates (elevation, slope, TWI, and negative openness) for soil thickness, depth to carbonates and soil properties measured in the lab (soil pH, coarse fragment content, clay content).

The predicted soil maps

Five soil properties that can be used for improving forest management have been mapped in this study area using a RF model and the data collected. They are soil thickness, depth to carbonates, soil pH, coarse fragment content and clay content. The first map in Figure 2.9 shows the elevation map (made from DEM) and has been added for comparison purposes. Moreover, the BEC subzones have been outlined on the maps (Figure 2.9).

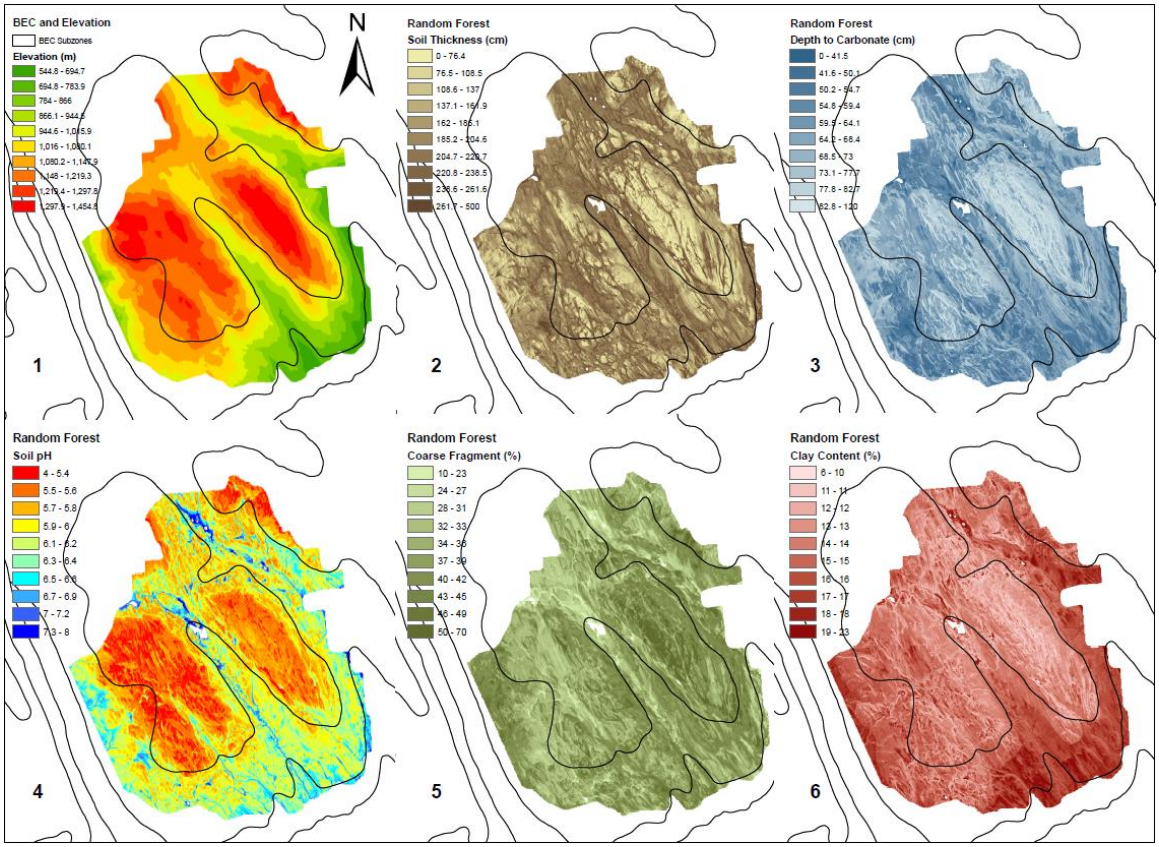

Figure 2.9. Five maps generated using RF to predict soil properties (soil thickness, depth to carbonates, soil pH, coarse fragment, and clay content). BEC subzones are outlined on the maps and an elevation map is also presented.

Soil properties for BEC subzones

Most of the predicted soil properties (soil thickness, depth to carbonates, soil pH and clay) tend to follow the BEC subzone pattern, but coarse fragment content does not. Most of the shallowest soils are associated with IDFdk1 BEC subzone which covers the top of the hills (Figure 2.9). The minimum value of soil thickness in IDFdk1 is 18 cm and

the maximum value is 387 cm. The mean value of soil thickness in IDFdk1 is 175 cm (Table 2.4). The mean value of soil thickness in IDFdk1 is less than that in IDFxh2 (188 cm) and PPxh2 (205 cm). In IDFdk1 BEC subzone, 50% of soil thickness values were between 140 cm and 216 cm. Fifty percent of cell values in IDFxh2 BEC subzone were between 162 cm and 218 cm. The range of values for PPxh2 was between 180 cm and 236 cm. Therefore, the deepest soil has been found in PPxh2. The relationship between soil thickness and elevation raster cell values has an R^2 of 6.27% (Figure 2.10). A slight negative relationship between soil thickness and elevation indicates that by increasing the elevation, soil thickness decreases.

Depth to carbonates is greatest at higher elevation (Figure 2.9). The average depth to carbonates is 69.7 cm in the IDFdk1 subzone. However, in the IDFxh2 and PPxh2 subzones the average depths are 63.2 cm and 61.9 cm respectively (Table 2.4 & Figure 2.9). Moreover, 50% of cells in the IDFdk1 subzone have depth to carbonates ranging from 62 cm to 77.8, whereas the ranges for the IDFxh2 and PPxh2 are 56.4 cm-70.1 cm and 56.4 cm-67.9 cm, respectively. This means that soil carbonates are closer to the soil surface in the PPxh2 subzone (Table 2.4). The relationship between depth to carbonate map cell values and the elevation model cell values is depicted in Figure 2.10.

Table 2.4. Statistics for predicted soil properties for each BEC subzone.

Soil Property	BEC Subzone	Map statistics							
		Mean	Median	Min	Max	Range	SD	Lowest Quartile	Highest Quartile
Soil Thickness(cm)	IDFdk1	175	190	18	387	369	54	140	216
	IDFxh2	188	197	23	375	352	45	162	218
	PPxh2	205	212	70	329	259	44	180	236
Depth to Carbonates (cm)	IDFdk1	69.7	70.2	22.7	97.9	75.2	10.1	62	77.8
	IDFxh2	63.2	63.5	21.3	106.3	84.9	9.3	56.3	70.1
	PPxh2	61.9	63.4	26.9	83.7	56.9	7.7	56.4	67.9
Soil pH	IDFdk1	5.72	5.67	5.04	7.68	2.64	0.3	5.52	5.87
	IDFxh2	6.07	6.03	5.08	7.68	2.6	0.32	5.88	6.22
	PPxh2	6.41	6.37	5.79	7.68	1.9	0.31	6.18	6.59
Coarse Fragment Content (%)	IDFdk1	36.9	37.1	13.7	62.8	49	6.3	32.2	41.2
	IDFxh2	37	38.3	14.8	57.4	42.6	5.6	33	41.2
	PPxh2	37	38.9	15	51.1	36.1	6.4	32.8	41.9
Clay Content (%)	IDFdk1	12.5	12.3	6.1	21.6	15.5	1.7	11.3	13.6
	IDFxh2	15	15.1	8.4	23.2	14.8	1.6	14	16
	PPxh2	16.2	16.1	12.5	21	8	1.1	15.5	16.8

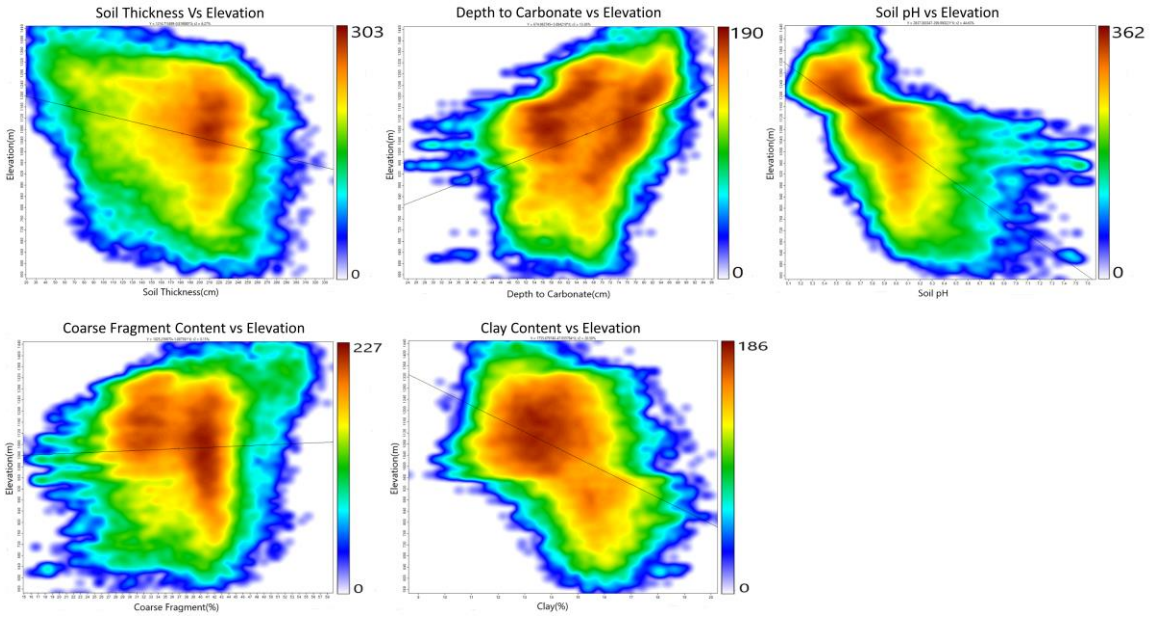

Figure 2.10. Scatter plots of cell values of soil property maps vs elevation; the plots show high cell density in brown and low cell density in blue.

The R^2 value for the relationship between depth to carbonates and elevation was 13.36% which is a higher value than that for the relationship between the soil thickness and the elevation model (Figure 2.10). The positive relationship between depth to carbonates and the elevation in Figure 2.10 indicates that by increasing the elevation, the depth to carbonates increases.

The most acidic soils are found in the IDFdk1 subzone (Figure 2.9) with a mean value of 5.72. The mean pH value in IDFxh2 is 6.07 and in PPxh2 is 6.41. The basic soils can be found at lower elevations in IDFdk1 and PPxh2, which are the same areas with shallower depths to carbonates (Table 2.4). Fifty percent of pH values in the IDFdk1 subzone are between 5.52 and 5.87 while the same percentage of values in IDFxh2 and PPxh2 are in the range of 5.88-6.22 and 6.18-6.59 respectively (Table 2.4). This analysis indicates that soil in the IDFdk1 subzone is more acidic. The relationship of the soil pH raster and the elevation model cell values shows a relatively strong relationship with an R^2 of 44.63% (Figure 2.10). The negative relationship means that by increasing the elevation the soil pH decreases.

A visual inspection of the map in Figure 2.9, map 5 shows that the coarse fragment content in IDFdk1, especially on the northwest hill is higher than other places on the map. However, the mean coarse fragment content in the IDFxh2 subzone is

37.04% which is higher than the other subzones (Table 2.4). Fifty percent of coarse fragment content values in IDFdk1 are between 32.2% and 41.2% and for the IDFxh2 and PPxh2 subzones the ranges are 33%-41.2% and 32.8%-41.9% respectively (Table 2.4). Coarse fragment content has a very weak relationship with elevation (Figure 2.10).

A visual inspection of the clay percentage map shows that the IDFdk1 subzone has lower clay content than the IDFxh2 and PPxh2 subzones (Figure 2.9). The mean clay content in IDFdk1 (12.5%), is lower than the clay content in IDFxh2 (15%) and PPxh2 (16.2%) (Table 2.4). Fifty percent of clay content values in IDFdk1 are between 11.3 and 13.6 % and in the IDFxh2 and PPxh2 subzones the ranges are 14%-16% and 15.5%-16.8% clay respectively. This suggests that the lower the elevation; the higher the clay content is (Table 2.4). According to Figure 2.10, the cell values of the clay content map shows a relatively strong relationship with elevation with R^2 of 23.26%. Since the relationship is negative, at increased elevation, the clay content decreases.

Validation Results

A nested 10-fold cross validation method with 20 repeats was used to validate the RF predictions. The R^2 and concordance values for the RF model for soil thickness were 0.35 and 0.47 respectively which were the highest validation results amongst all models. The second-best validation results were for soil pH with R^2 of 0.26 and concordance of 0.37 (Table 2.5). The relatively poor validation results for depth to carbonates (Table 2.5) may be due to a shortage of training data or a lack of an intrinsic relationship between the soil properties and the DEM. Collecting more training data can solve the first problem. To investigate the second possible problem the relationship between the soil property predictions and the DEM has been sought (Figure 2.10). As can be seen in Figure 2.10, predicted depth to carbonates is relatively highly correlated with elevation (R^2 of 13.36%). This suggests that the poor validation results for depth to carbonates may be due to a lack of enough training data for this property.

Table 2.5. Accuracy metrics for 5 soil property maps produced for the study area using RF.

Dataset	R^2	concordance	MSE	RMSE	bias
Soil Thickness	0.35	0.47	10997.74	103.18	1.58
Depth to Carbonates	0.07	0.14	898.8	29.98	-0.24
Soil pH	0.26	0.37	0.47	0.68	0.02
Coarse Fragment Content	0.11	0.2	213.07	14.48	0.08
Clay Content	0.13	0.20	21.58	4.58	0.25

Covariate Importance

A total of 45 covariates were used in modelling five soil attributes in this study. The importance of variables was different for modelling each soil property (Figure 2.11). The most important covariates for modelling soil thickness were open negative and parallel curvature. For modelling depth to carbonates, elevation and slope were the most important variables. For modelling soil pH, vegetation height model (canopy height model), elevation and total wetness index were the most influential variables. For modelling coarse fragment content, multiresolution index of valley bottom flatness, slope, and multiresolution ridge top flatness were the most important variables. For modelling clay content, elevation and topography were the most important variables (Figure 2.11 & Tables 2.1 & 2.6).

As was explained in section 2.3.5 all covariates used in modelling were downscaled to 3 m resolution. However, the covariates were originally upscaled to different resolutions. Covariates originated from low resolutions such as 9, 27 and 30 m are often high in the rankings for importance (Figure 2.11). For example, 9 out of 10 of the most important covariates for modelling soil thickness were from 9 m resolution and lower. For modelling depth to carbonates, 6 out of 10 important covariates were from 9 m and lower resolutions. For modelling soil pH, 5 out of 10 covariates were from 9 m and lower resolution covariates. Likewise, for coarse fragment content and clay content, 7 covariates out of 10 were from 9 m and lower resolutions respectively. One possible reason for this observation is that much of the fine scale originated covariates may contain a lot of noise, and thus the RF model can make better models with covariates originated from lower resolutions.

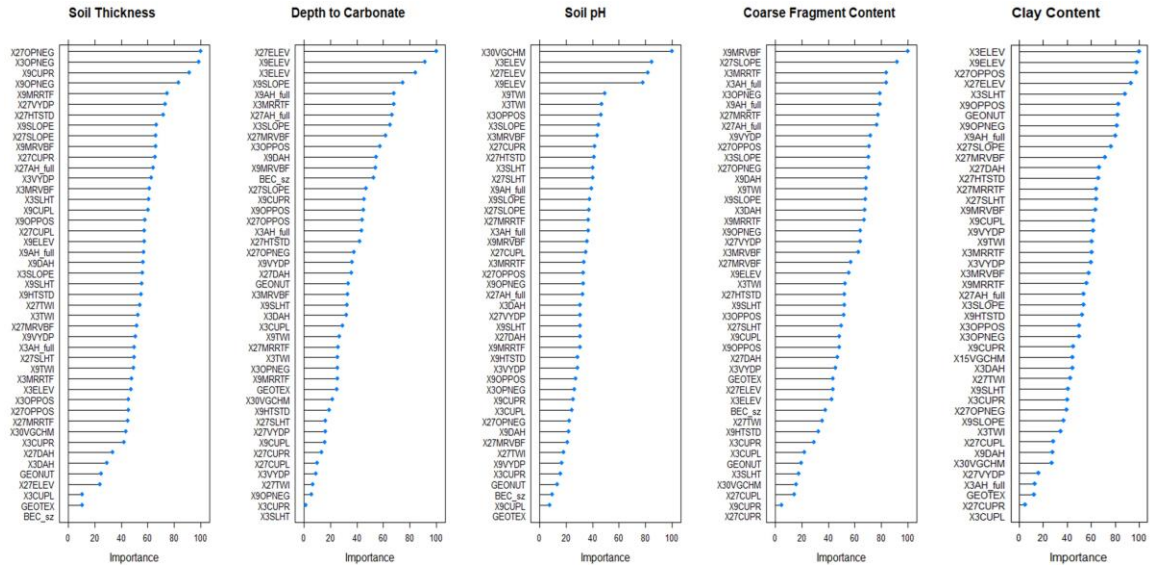

Figure 2.11. Importance plots for modelling 5 soil properties using RF demonstrated using percentage of importance level.

Table 2.6. List of covariates and acronyms.

No.	Covariate Acronym	Original Reso (m)	Covariate Name	No.	Covariate Acronym	Original Reso (m)	Covariate Name
1	X3AH_full	3	Full hillshade illumination	24	X3OPEPOS	3	Openness positive
2	X9AH_full	9		25	X9OPEPOS	9	
3	X27AH_full	27		26	X27OPEPOS	27	
4	X3DAH	3	Diurnal anisotropic heating	27	X3OPNEG	3	Openness negative
5	X9DAH	9		28	X9OPNEG	9	
6	X27DAH	27		29	X27OPNEG	27	
7	X3ELEV	3	DEM	30	X30OPENEG	30	
8	X9ELEV	9		31	X3SLOPE	3	Slope
9	X27ELEV	27		32	X9SLOPE	9	
10	X9HTSTD	9	Standardized height	33	X27SLOPE	27	
11	X27HTSTD	27		34	X3SLHT	3	Slope height
12	X3MRVBF	3	Multiresolution index of valley bottom flatness	35	X9SLHT	9	
13	X9MRVBF	9		36	X3TWI	3	Topographic wetness index
14	X27MRVBF	27		37	X9TWI	9	
15	X3MRRTF	3	Multiresolution ridge top flatness	38	X27TWI	27	
16	X9MRRTF	9		39	X3VYDP	3	Valley depth
17	X27MRRTF	27		40	X9VYDP	9	
18	X3CUPL	3	Plan curvature	41	X27VYDP	27	
19	X9CUPL	9		42	X30VGCHM	30	Canopy height model
20	X27CUPL	27		43	BEC_sz	3	BEC subzone
21	X3CUPR	3	Profile curvature	44	GEONUT	3	Bedrock (darkness)
22	X9CUPR	9		45	GEOTEX	3	Bedrock(texture)
23	X27CUPR	27					

2.4.2. Maps for Forest Management

Digital soil maps can be used to assess the sensitivity of soils in cutblock sites to degradation processes and help forest managers decide how to preserve soil productivity and prevent soil-related site degradation. In this study, cutblocks have been depicted with respect to predicted soil properties (Figure 2.12). Availability of digital soil information provided at the time of authorizing the harvest could assist forest managers of these cutblocks to ensure soil preservation and protection. As can be seen in Figure 2.12; map 1, which shows the relationship between cutblock location and slope, cutblocks are mostly positioned in areas of relatively gentle slope within the generally steep terrain, although certain portions of cutblocks located at high elevations have steep slopes. Also, on the high hills of the northeast side of the study area the soils were generally predicted to be very shallow ranging from 0-76 cm, but the cutblocks tend to be located in areas with somewhat thicker soils (Figure 2.12; map 2) except for certain portions of cutblocks. At the same time, with soils (e.g. clay < 10%) that partially mitigate the erosion risks associated with steep slopes and thin soils on the northeast (Figure 2.12; map 1, 2 & 6), careful management could occur and some harvesting can be carried out without significant soil degradation and diminished soil productivity.

The forest soil displacement hazard refers to the risk of exposing unfavourable subsoil such as carbonated soils (Lewis and Carr, 1993) and forcing tree and plant roots to grow in them instead of the more favorable topsoil materials. In the southwest of the study area and along the edge of the study area highlighted with a square in maps 3 and 4 of Figure 2.12, there are cutblocks in which the carbonated soils and those that have pH over 6 are close to the surface (Figure 2.12; map 4). In this cutblock although soil is deep, and slope is minimal, the unfavorable subsoil condition represents a risk that exposure of such materials can lead to reduced plant growth compared to other forest locations. High-resolution digital soil maps made for depth to carbonates can easily show forest managers where in the forest the unfavorable carbonated soils are expected to be close to the soil surface. Therefore, the maps can provide them with an effective decision-making tool to ensure sustainable forest management.

Lastly, there are also some cutblocks in this study area that are at risk of soil compaction and puddling hazard. As previously described, soils with coarse fragment content < 70% and clay content > 20% are susceptible to soil compaction and puddling

hazard (Lewis and Carr, 1993). On the southern region of the study area (highlighted cutblock in Figure 2.12; maps 5 and 6 & Figure 2.13), some soils were predicted to have low coarse fragment content and high clay contents; hence, harvesting activity carried out under wet soil conditions could lead to an increased risk of soil compaction and puddling on these sites.

Soil hazard assessment maps can be produced to assist forest managers. An example of a hazard map is produced in this study (Figure 2.13). The soil compaction and puddling hazard map demonstrated in Figure 2.13 shows areas with high level of hazardous condition in dark blue. Areas shown in red depict areas with very high hazardous condition (Figure 2.13). For this area, there do not appear to be any contiguous areas of soils with very high compaction hazards. By considering a hazardous threshold value and combining slope gradient, slope complexity, slope curvature, depth to bedrock, depth to carbonates and soil chemistry (pH), a displacement hazard map can be produced. In the same fashion an erosion hazard assessment map also could be produced. To generate a soil erosion hazard assessment map, the following data could be used: climate, slope, depth to water restricting layer from depth to bedrock (soil thickness), clay content and coarse fragment content.

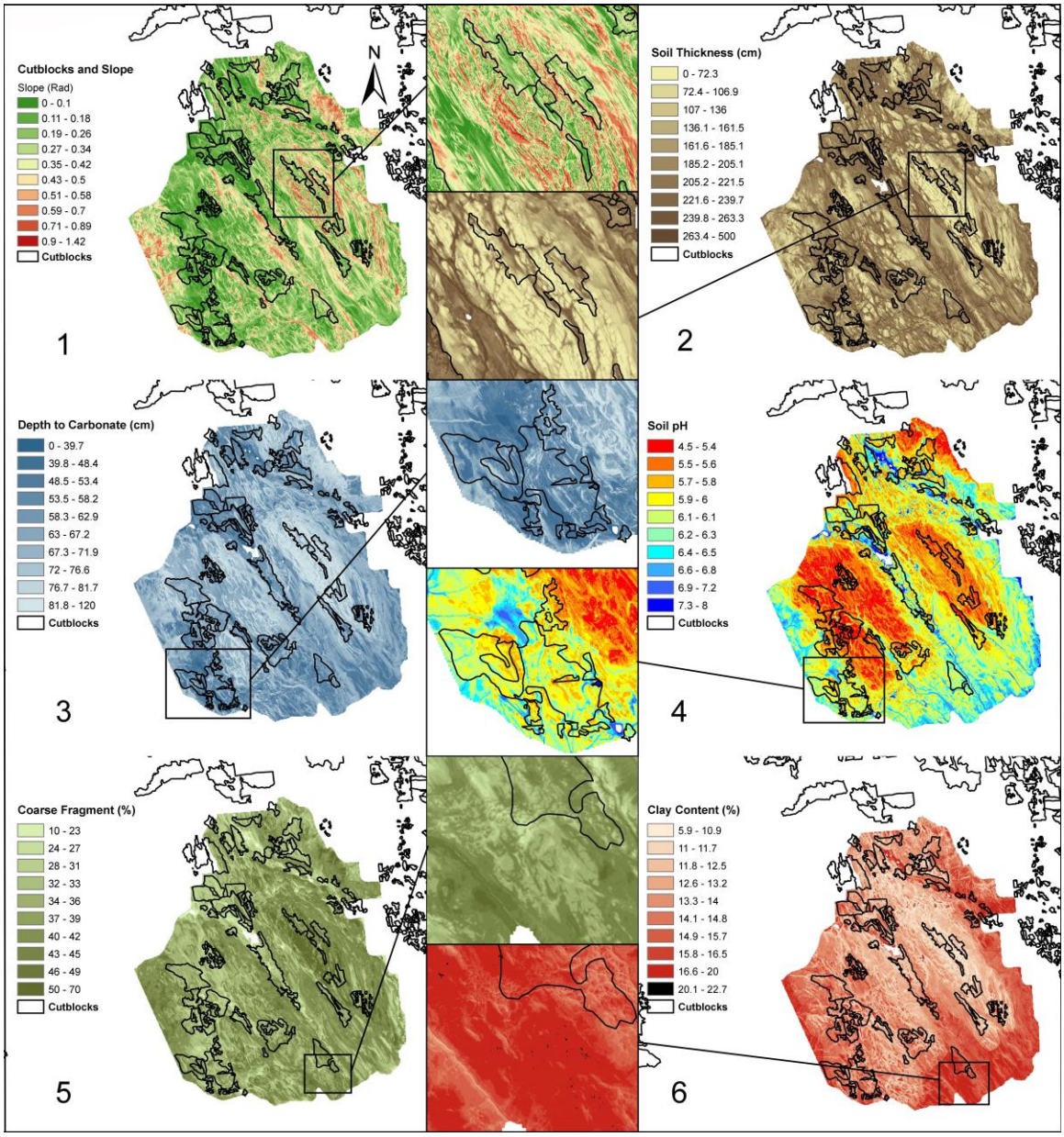

Figure 2.12. Cutblocks in the study area susceptible to mismanagement and high risk of soil degradation.

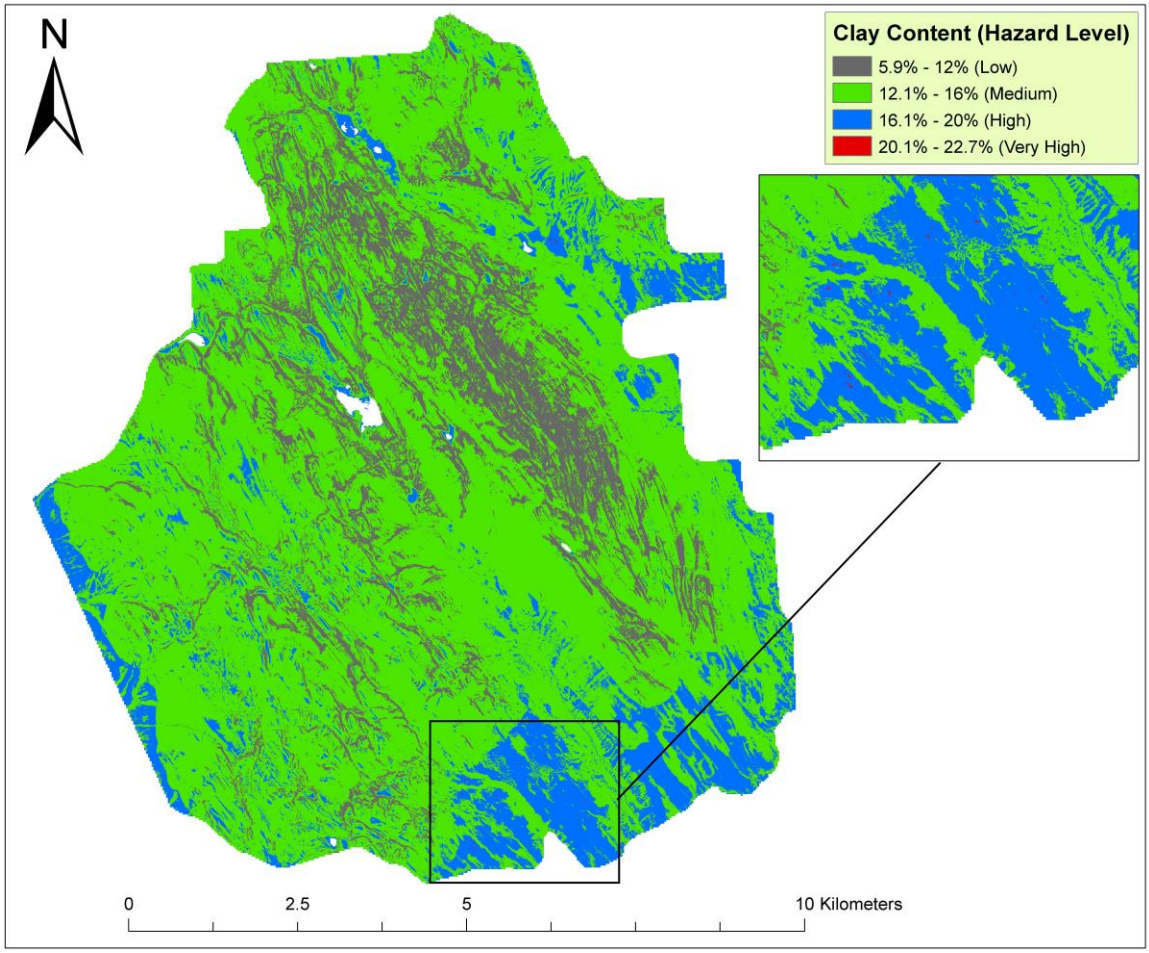

Figure 2.13. Compaction and puddling hazard assessment map.

2.4.3. General Discussion and Future Work

In this study we used RF to model five soil properties and produce digital soil maps. Other ML models could be selected such as Cubist decision tree, kNN and SVM. We selected RF because it has several advantages compared to most of the modeling techniques mentioned above. First, it can model high dimensional non-linear relationships. Second, it is resistant to overfitting. Overfitting occurs when a ML model fits closely to the training data and the accuracy results are too optimistic; however, the performance of the model on an unseen dataset is poor (Hawkins, 2004). Third, RF performance is relatively robust with respect to noisy covariates. Fourth, RF implements an unbiased measure of the error rate; and fifth, RF can measure variable importance (Grimm et al., 2008). Moreover, the results in Chapter 3 confirm the superiority of RF models compared to the other three ML models as RF has shown the lowest local uncertainty in the 90% PI range.

In this study we used a multiscale approach for covariate generation. The reason for selecting this method is that we wanted to regionalize soil information as accurately as possible and cover all relevant landscape characteristics in our modelling (Behrens et al., 2010; McBratney et al., 2003b). The landscape characteristics such as topography, rivers, waterbodies and vegetation are soil formation forces and can induce soil formation and pedogenesis at different scales (Kerry and Oliver, 2011). This approach provides measures of the entire landscape that is called the geomorphic signature (Pike, 1988) and increases prediction accuracy (Behrens et al., 2014).

The goal of producing digital soil maps in this chapter was to help forest managers by providing them with useful information about soil. This is a new branch in DSM introduced by Carré et al. (2007) and is called digital soil assessment (DSA). Their intention of producing digital soil maps was to translate the information gained from DSM into something practical and useful that can help foresters assess risks and make spatial decision-making surfaces (McBratney et al., 2012). This branch is still very young and is still in the framework development phase. To develop an initial framework McBratney et al. (2012) discussed that digital soil assessment can be driven by soil scientists or stakeholders. Global issues such as food, water and energy security and climate change mitigation can motivate stakeholders and soil protection can motivate soil scientists to develop DSA methods. Soil scientists can provide useful information about soil for stakeholders. This chapter was just a start in developing DSRA maps. This topic will be further developed and discussed and new methods will be investigated for mapping soil properties that are difficult to be mapped such as soil erosion, unfavorable subsoil layers and salinization (Carré et al., 2007).

2.5. Conclusions

The first objective of this study was to produce soil digital maps for five soil properties that provide valuable information for forest managers. The soil properties that have been selected for this study were soil thickness, depth to carbonates, soil pH, coarse fragment content and soil clay content. The model that has been used in this study was RF. The second objective of this study was to validate the soil property predictions using a nested 10-folds cross validation with 20 repeats. The last objective of this study was to discuss and illustrate how the digital maps produced in this study can help forest managers.

Five digital soil maps were produced in the Eagle Hill Forest study area located west of Kamloops, BC with the area of 95 km^2. The maps produced in this study shows that the harvesting sites on the northeast slopes are susceptible to erosion, and more care should be considered when forest harvest is authorized on these slopes. Moreover, the depth to carbonates digital soil map highlights some cutblocks on the southwest corner of the study area with unfavorable subsoil condition in which tree growth may be delayed. Forest harvesting protocols should be carefully reviewed in these cutblocks to prevent soil displacement and subsoil unfavorable hazard. By looking at digital soil maps of coarse fragment content and clay content it is evident that in the south part of the study area there is a cutblock in which in some places the clay content is over 20% and coarse fragment content is less than 70%. This condition is a sign of puddling and compaction hazard in this cutblock that should be considered.

This study was designed to show how digital soil maps can be produced using high-resolution LiDAR data and how they can be used for forest management. In future work the improvement of the map validation results will be sought by increasing the number of training data points. Making further digital soil maps using high-resolution LiDAR data for forest in soils susceptible to hazards and a closer look and comparison between cutblocks that are susceptible to mismanagement will be conducted in future work as well.

Chapter 3.

Quantile Regression as a Generic Approach for Estimating Uncertainty for Machine-Learning Techniques

3.1. Abstract

The word uncertainty in DSM refers to both model error and spatially explicit uncertainty. Spatially explicit uncertainty is known as quantification of confidence intervals for model output and is called local error. Based on importance of local errors, international standards require 90% prediction interval (PI) measurements for quantification of uncertainty in DSM. Regarding the limitation of uncertainty quantification methods, this study proposes a new framework for uncertainty estimation using QR. The objectives of this study are 1) to produce soil attribute maps using 4 ML models for three soil properties: soil thickness, depth to carbonates, and soil pH and to validate the prediction results; 2) to produce 90% PI maps using the QR method; and 3) to assess those uncertainty estimations using metrics such as mean prediction intervals (MPI) and prediction interval coverage probability (PICP) analysis. We demonstrated the integration of ML and QR using a case study from a dry-forest ecosystem in the Kamloops region of British Columbia, Canada. Within the QR framework, model residuals from predictions using ML were obtained using a nested cross-validation procedure, which were then used as inputs into the QR model. In QR, the conditional distribution of a response variable was described as a linear function between the predicted and observed values of a soil variable. Uncertainty estimates were provided for every pixel in a predicted map and then were evaluated using mean prediction intervals (MPI) and prediction interval coverage probability (PICP) analyses. The results showed that RF performance was the best among the ML models and quantification of uncertainty using QR was accurate for modelling all soil properties in all four ML methods. This illustrates the capability of QR in quantification of uncertainty in DSM. [2]

[2] A version of the following chapter has been submitted to Geoderma under the co-authorship of Brandon Heung, Daniel D. Saurette, Margaret G. Schmidt, Chuck E.Bulmer and William Bethel.

3.2. Introduction

Like other kinds of maps, digital soil maps are representations of reality, and prediction in DSM is not error free. In other words, we are uncertain about the true properties and processes in soil because they are highly variable in space and time (Arrouays et al., 2014b; Malone et al., 2017). Some common forms of error are errors in measurements, digitization, typing, interpretation, classification, generalization and interpolation (Arrouays et al., 2014b). Uncertainty in the output data can be the result of bias in modelling, uncertainty in parameters, or even errors in measurements of the input data (McBratney et al., 2002). Since these errors can lead to poor decision making that may sometimes lead to serious consequences, we should be aware of them and quantify them (Arrouays et al., 2014b).

It is important to consider the difference between model error and spatially explicit uncertainty. Model error is called mean square error and is the average squared difference between the estimated value and the actual value (Malone et al., 2011; Wang and Bovik, 2009). However, pixel-based estimates of uncertainty (i.e. local uncertainty) may be generated to provide an understanding of the spatial distribution of uncertainty (Feizizadeh et al., 2014; Malone et al., 2011; Vaysse and Lagacherie, 2017). Based on the importance of spatially explicit uncertainty quantification in DSM, international standards require 90% prediction interval (PI) measurements for quantification of uncertainty in DSM (Arrouays et al., 2014a). Nevertheless, measuring uncertainty in DSM seldom has been carried out (Minasny and McBratney, 2002; Vaysse and Lagacherie, 2017).

Several uncertainty assessment methods have been used in DSM. Each of the uncertainty assessment methods has its own advantages and disadvantages. Geostatistical modelling approaches, for example, can produce uncertainty estimations using the kriging prediction variance where spatial representations of the prediction intervals may be calculated. However, the uncertainty estimations are specific to the modelling approach itself and not applicable for ML techniques (Fouedjio and Klump, 2019; Malone et al., 2017). Moreover, geostatistical techniques are computationally intensive (Mckay et al., 2000). When using ML techniques, other studies have used a bootstrapping approach (Ackerson et al., 2015; Padarian et al., 2017; Stumpf et al., 2017; Thomas et al., 2015), where a model is trained on a random subset of the full

training data and multiple realizations of a soil map are produced through the prediction process. By generating these realizations, spatial representations of uncertainty are produced by calculating the average MSE from the realizations and summing it with the bootstrap prediction variance that is estimated for each pixel (Malone et al., 2017). However, computational capabilities may limit the use of this approach when applied to large datasets since each map realization needs to be predicted and stored in order to estimate the prediction variance.

Two other methods for measuring uncertainty are, empirical uncertainty quantification through data partitioning and cross validation, and empirical uncertainty quantification through fuzzy clustering and cross validation (Malone et al., 2017). Both these methods determine prediction intervals from the distribution of model errors. Model errors are calculated from the deviation between observation and model predictions. The prediction limits calculated in both these methods are not spatially uniform, and they are a function of the landscape. It means that the accuracy of the prediction depends on the areas of the landscape and particular landscape situations (Malone et al., 2017). Some uncertainty assessment methods such as Bayesian and Monte Carlo methods deal only with certain sources of uncertainty; for example, the Bayesian method only measures uncertainty associated with input data while the Monte Carlo method measures uncertainty in parameters (Solomatine and Shrestha, 2009).

ML are computer algorithms that perform a specific task and improve through experience. They use dependent variables and environmental data as independent variables to predict soil attribute values for the whole study area (Dietterich, 2000; Heung et al., 2016). Within the DSM literature, the use of ML techniques has become increasingly popular because of their computational power availability (Rossiter, 2018). For example, Grimm et al. (2008) used RF to model soil organic carbon for different depth intervals on Barro Colorado island. Adhikari et al. (2019) developed a spatially explicit prediction model between soil organic carbon observations and 17 covariates using Cubist decision trees in Wisconsin, USA. Mancini et al. (2019) used RF, support vector machine (SVM), and linear discriminant analysis models to predict parent materials from A and B horizon samples in Brazil. Merchant et al. (2018) evaluated multiple remotely sensed datasets to map wetlands in the subarctic, boreal cordillera in Yukon, Canada. For their analysis they used RF, support vector machine (SVM) and k-nearest neighbor (kNN) using various data combinations. These are only a few

examples of the use of ML methods in DSM and numerous studies have used ML methods.

Quantification of uncertainty within a ML framework is quite novel, and many ML techniques are not capable of generating localized uncertainty maps (Szatmári and Pásztor 2019; Velronesi and Schillaci 2019; Vaysse and Lagacherie 2017). Among all ML techniques used in DSM, quantile regression forest (QRF) is one of the ML techniques that has a built-in mechanism for uncertainty quantification. QRF (Meinshausen, 2006) is an extension of RF (Breiman, 2001) in which for every leaf of every decision tree all observations in this leaf are retained, instead of only the average. Then QRF models calculate the full conditional distribution, rather than only the conditional mean (Fouedjio and Klump, 2019; Meinshausen, 2006).

To assess uncertainty in the other ML methods, either bootstrapping or quantification through data partitioning and cross validation should be used. Regarding the bootstrapping limitations, Hengl et al. (2017) mention that although tools for modelling uncertainty in ML methods already exist, they are computationally intensive and development of other robust statistical frameworks such as quantile regression forest in future should be considered. Generating many map realizations is time consuming, expensive, and has high memory requirements. Moreover, in bootstrapping, by increasing the number of realizations, using the approach described in Malone et al. (2017), may result in the artificial decrease of the PI width. In data partitioning and cross validation, the residuals are computed after cross validation performance to develop an uncertainty map for the ML algorithms. The problem of this method is that it is too optimistic because it uses the results of cross validation and is not suitable for constructing an uncertainty map. A problem with the results of cross validation is that a selection of random points is used and this does not account for spatial autocorrelation and assumes samples to be independent (Veronesi and Schillaci, 2019).

Considering the shortcomings of uncertainty estimation methods and ML limitations in quantifying uncertainty, a knowledge gap is apparent with regards to ML uncertainty estimations. Geostatistical techniques produce uncertainty estimations using kriging prediction variance; bootstrapping is computationally intensive especially for big datasets; QRF works when we use RF only; and finally, data partitioning and cross validation does not account for autocorrelation and is too optimistic.

Given the limitations of existing approaches, this study proposes a new framework for uncertainty estimation using QR and demonstrates its flexibility and integration with ML techniques. QR is an approach that has originated from the field of quantitative economics but has been extended to other applications. QR is a linear statistical method that is used to estimate quantile conditional functions, of prediction and distribution. The estimation is based on causal relationships in the dataset. It uses data observation and ML prediction output and calculates the residuals. Residuals are the difference between observed and prediction values. QR finds a linear relationship between observation and prediction values of the quantile requested (Koenker and Hallock, 2001). Within the soil science literature (outside of DSM), QR has been used for estimating the uncertainty of models used for modelling nitrate contamination of groundwater using different ML techniques (Rahmati et al., 2019). In terms of the DSM literature however, relatively few studies have explored this approach with the exception of Lombardo et al. (2018), which demonstrated the coupling of a generalized linear model approach with QR for predicting SOC; however, such couplings may be potentially extended to other ML techniques as part of a generic framework.

Hence, the objectives of this study are 1) to develop a framework for producing local estimates of uncertainty by coupling ML models with quantile regression. 2) to demonstrate the coupling using a variety of ML techniques for a case study: a dry-forest ecosystem in the Kamloops region of British Columbia, Canada, and 3) to evaluate the uncertainty estimations using metrics such as MPI and PICP.

3.3. Methods

3.3.1. Study Area, Soil Sampling and Data Acquisition

The study area is the Eagle Hill Forest located west of Kamloops, BC, on the north side of Kamloops Lake (Figure 3.1). A detailed description of the study area can be found in section 2.3.1 of Chapter 2. Different sampling methods were used to sample these soil properties. The sampling methods and data acquisition have been described in section 2.3.3 of Chapter 2.

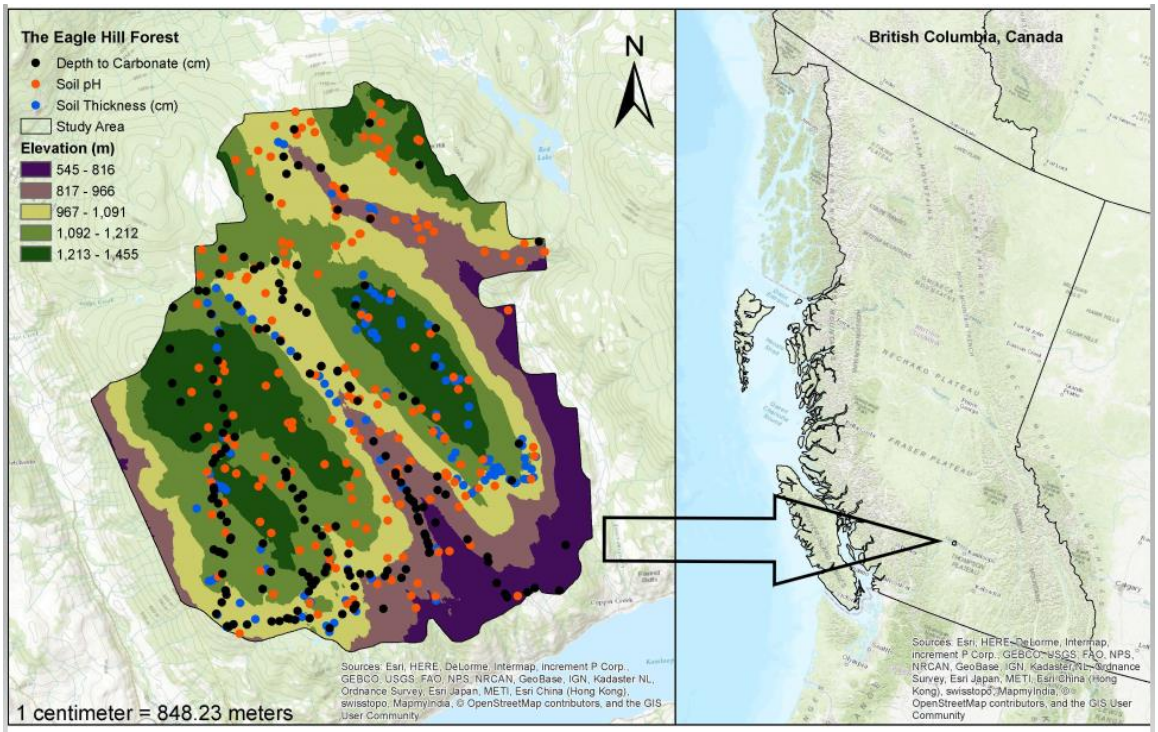

Figure 3.1. Study area and sampling points.

3.3.2. Dependent Variables and Covariates

In this study three soil properties have been modelled: soil thickness, depth to carbonates, and soil pH (Figure 3.1). Soil thickness, depth to carbonates, and soil pH, data were collected at 410, 171, and 230 sites respectively (Figure 3.1). Forty-five covariates at 3 m resolution were used. Description of dependent variables and covariates can be found in sections 2.3.4 and 2.3.5.

3.3.3. Machine Learning Models

The process of using a ML model can be described as training a statistical model using predictor and response variables to make the model and using it to make new predictions in a study area (Heung et al. 2016; Witten et al. 2016). In this study, four ML models were used, and their output prediction results were validated. Then, the performance of the models was assessed and compared using the QR uncertainty quantification method. The ML models used in this study are RF, Cubist decision tree, kNN, and SVM. A detailed description of the ML methods used in this study can be found in Chapter 1; section 1.1.3.

3.3.4. Model Validation

Model validation shows us the proportion of the variance for a dependent variable that can be explained by the independent variables. In this study we used multiple replicates of *k*-fold cross validation. By applying the iterative *k*-fold cross validation, R^2 and concordance metrics were calculated in each iteration and the final results were averaged. Accuracy metrics measured in this study are R^2, concordance, and RMSE. R^2 is the square of sample correlation coefficient (Pearson's) between the observation and their corresponding predictions. Lin's concordance correlation coefficient or simply concordance evaluates the accuracy and precision of the relationship. RMSE is the standard deviation of the residuals and is called prediction error (Malone et al., 2017).

3.3.5. Uncertainty

QR method

QR was first introduced by Koenker and Bassett (1978) originally in the field of quantitative economics but its use has since been extended to other applications. Within the soil science literature (outside of DSM), QR has been used for estimating the uncertainty of models used for modelling nitrate contamination of groundwater using different ML techniques (Rahmati et al., 2019). In terms of the DSM literature however, relatively few studies have explored this approach with the exception of Lombardo et al. (2018), who demonstrated the coupling of a generalized linear model approach with QR for predicting soil organic carbon (SOC); however, such couplings may be potentially extended to other ML techniques as part of a generic framework.

When applying QR for uncertainty estimation we assume that there is a linear relationship between a soil variable's observed value and its model predicted value. QR consists of a set of linear regression models, where the response variable is the selected quantile of the variable's conditional distribution. Within the hydrological modelling literature (Dogulu et al., 2015a; Lopez et al., 2014; Rahmati et al., 2019), QR is applied as a post-processing technique whereby the prediction of the response variable is dissociated from the uncertainty estimation process. Because of this dissociation, there is the added flexibility in the choice of the predictive model and uncertainty estimation may be generated using model residuals. QR estimates the value for a soil property for

any quantile that is needed (Koenker and Hallock, 2001) which may then be used to calculate prediction intervals (e.g. 90%) using the upper (95%) and lower (5%) quantile maps. Here, a brief description of QR is provided; however, detailed descriptions may be found in Lopez et al. (2014).

Similar to Lopez et al. (2014), Dogulu et al. (2015) and Rahmati et al. (2019), for each quantile (τ), QR assumes a linear relationship between observed values (x) and predicted values (y):

$$y = \alpha_\tau x + b_\tau \quad (2)$$

Where, a_τ is the slope and b_τ is the intercept of the linear regression. Here, a_τ and b_τ are both determined by minimizing the sum of residuals in the following loss function:

$$min \sum_{j=1}^{J} \rho_\tau(y_j - (\alpha_\tau x_j + b_\tau)) \quad (3)$$

Where, y_j and x_j are jth paired samples (i.e. soil measurements), with a total of *J* samples, and $\rho_\tau$is the QR function for the τ-th quantile:

$$\rho_\tau(\varepsilon_j) = \begin{cases} (\tau - 1).\varepsilon_j & \varepsilon_j \leq 0 \\ \tau.\varepsilon_j & \varepsilon_j > 0 \end{cases} \quad (4)$$

Where, the model residuals, ε_j are the difference between the observed and predicted values, acquired from Equation 2, for the τ-th quantile. The QR function is applied for the residual, ε_j, in Equation 4 for the desired quantile τ (Dogulu et al., 2015a; Lopez et al., 2014; Rahmati et al., 2019).

QR limitations

The quantile crossing problem and the assumption of linear model for a non-linear data distribution are two limitations of QR. The crossing problem may occur when the predicted soil value for a lower percentile is greater than that of its corresponding higher percentile. For example, a predicted 95th percentile of the response variable may become smaller than the 90th percentile which is impossible (Bondell et al., 2010). To avoid the crossing problem He (1997) suggested forcing proper ordering of percentile curves if we use nonlinear QR method. Koenker (1984) considered parallel quantile

planes for linear models. Cole (1988) and Cole and Green (1992) suggested a suitable transformation that would yield normality of the response variable to fully determine the quantile functions.

This approach assumes that there is a linear relationship between the predicted and observed values and therefore a linear QR function is applied. However, in some cases, there may not be a linear relationship and hence, the QR uncertainty estimations have the potential to produce non-realistic uncertainty estimates (Lopez et al., 2014): To solve this problem Van Steenbergen et al. (2012) applied linear model to different parts of the predictor and were able to achieve more reliable results. Koenker (2005) has referred to this problem as a faulty notion and has suggested the segmenting of the response variable into subsets.

Prediction Interval Coverage Probability (PICP) and Mean Prediction Intervals (MPI)

PICP graphs, called accuracy plots, are used to assess the performance of QR in terms of uncertainty quantification by evaluating the encapsulation of observation values into an associated prediction interval. For a particular confidence level (CL), we should expect that the same percentage of observations, equal to the associated CL, is encapsulated by the PI. For example, it is expected that about 90% of observations fall within the 90% PI. This percentage is defined as the Prediction Interval Coverage Probability (PICP). Therefore, to assess the sensitivity of QR uncertainty quantification, PIs at a number of CLs are defined and then the PICP is assessed. Ideally, the observed fractions are equal to the expected fraction and a 1:1 relationship should be found (Malone et al., 2017). If observation fractions are lower than the prediction, then the uncertainty has been underestimated, and if observation fractions are higher than the prediction, then the uncertainty has been overestimated (Szatmári and Pásztor, 2019). PICP can be calculated by equation 5.

$$PICP = \frac{1}{n}\sum_{t=1}^{n} C, C = \begin{cases} 1, PL_t^{lower} \leq y_t \leq PL_t^{upper} \\ 0, otherwise \end{cases} \quad (5)$$

In equation 5, y_t is the observed value, PL_t^{lower} is the lower limit and $PL_t^{upper}$is the upper limit (Rahmati et al., 2019).

MPI shows the average of the widths of the prediction intervals. MPI can be used to quantify the level of predicted uncertainty. A wider PI represents a higher uncertainty of modelling prediction, and a narrower PI represents a lower uncertainty of the model used (Ding et al., 2018; Rahmati et al., 2019). MPI can be calculated using equation 6.

$$MPI = \frac{1}{n}\sum_{\tau=1}^{n}(PL_{\tau}^{Upper\ Limit} - PL_{\tau}^{Lower\ Limit}) \quad (6)$$

In this equation PL_t^{upper} is the upper limit and PL_t^{lower} is the lower limit of the PI. Between two models with the same PICP, the model with the lower MPI is regarded as the better model (Muthusamy et al., 2016).

3.3.6. Integration of QR for DSM

The proposed framework consists of two phases: (1) testing the predictive model and uncertainty estimates (Figure 3.2) and (2) generating digital soil maps and uncertainty maps (Figure 3.3).

Testing the Predictive Model and Uncertainty Estimations

In Phase 1, the objective is to ascertain the accuracy of the ML model using goodness-of-fit metrics and to quantitatively evaluate the uncertainty estimates using PICP (Equation 5) and MPI (Equation 6) graphs. Here, the process consists of seven steps, where the only input is a matrix that consists of the observed soil attribute value and the corresponding covariate values for each sample location (Figure 3.2). This is acquired by spatially intersecting the geographical position of each sample point with a suite of environmental covariates representing the SCORPAN factors (McBratney et al., 2003a).

To generate estimates of model accuracy, PICP, and MPI, a nested cross-validation procedure is applied, consisting of an ‘outer loop’ and an ‘inner loop’. The ‘inner loop’ (Steps 2-5) calibrates and selects the predictive model with the optimal combination of model hyperparameters using the validation data. The ‘outer loop’ (Steps 1-7) assesses the ML model’s accuracy and estimates the uncertainty. In Step 1, the matrix with the soil-environmental data is randomly partitioned into k_{outer} folds, whereby $k_{outer} - 1$ folds are used as the input data to the ‘inner loop’. The remaining fold is reserved for testing the predictive model.

In Step 2, the $k_{outer} - 1$ training fold is further partitioned into k_{inner} folds, whereby $k_{inner} - 1$ folds are used to build the predictive model. Here, the predictive models are fitted (Step 3) using different combinations of hyperparameters (e.g. *mtry* for RF and *sigma* and *cost* for SVM) and the model is predicted on the validation data, which then allows the calculation of the model's accuracy. The accuracy values are used to select the optimal hyperparameter values (Step 4). This process is reiterated k_{inner} times so that each fold is used to validate the model once and the optimized model is selected in Step 5. It is important to note that the optimized model is calibrated using all the data in the inner loop (i.e. non-partitioned). Lastly, residual distribution (i.e. observed vs. predicted) values are retained from each validation fold and combined. Steps 2-5 may be carried out entirely using the *caret* package (Kuhn, 2018) in the *R* statistical language (Kleinman and Horton, 2015), which includes the model parameterization and selection functions.

Step 5 will produce a matrix of the residual distribution for the validation data (inner loop), which is compiled from each validation fold. This matrix is then used to fit the QR function in the *quantreg* package (Koenker, 2019) in Step 6. Because an independent test fold was retained in Step 1, it is now possible in Step 7 to produce an estimate of model accuracy by predicting the optimized model on the test data and generating the residual distribution for the test fold (outer loop). Secondly, the fitted QR model is applied to the test fold at the desired percentiles from which the PICP and MPI are later calculated. To complete the 'outer loop', the model testing process is also reiterated k_{outer} times so that each test fold is used to test the model once and generate the percentiles for each sample location. In Step 7, a matrix consisting of the residual distribution values and their percentile values are retained from each test fold and combined. This final matrix is used to calculate goodness-of-fit statistics (e.g. Lin's concordance correlation coefficient (CCC), root mean square error); calculate PICP for a range of confidence levels using Equation 5 and generate the corresponding PICP plot; and calculate the MPI using Equation 6.

To assess the reliability of the accuracy, PICP, and MPI metrics, Phase 1 may be repeated over multiple iterations so the mean of the metrics may be reported as well as their standard deviations. To generate the final uncertainty estimation maps using QR, the residual distribution values for the test data are all compiled (including the repeats) and used as the input to Step 9 in Phase 2.

Generating Digital Soil Maps and Uncertainty Maps

To follow standard DSM production practices in Phase 2, the final digital soil maps and uncertainty maps are produced using all soil sample locations (Figure 3.3). In Step 8, the soil-environmental matrix is used to calibrate the predictive model and hyperparameter values are selected using cross-validation in the *caret* package. The optimized model is then used to make the spatial predictions using the covariate stack to generate the final maps in Step 9 using the *raster* package (Hijmans, 2019). To generate the uncertainty maps, the residual distribution values for the test data from Step 7 are used to fit the QR function (Step 10), which is then applied to the final maps produced in Step 9 to generate the upper and lower limit percentile maps as well as the PI map. Step 11 uses the *quantreg* and *caret* packages.

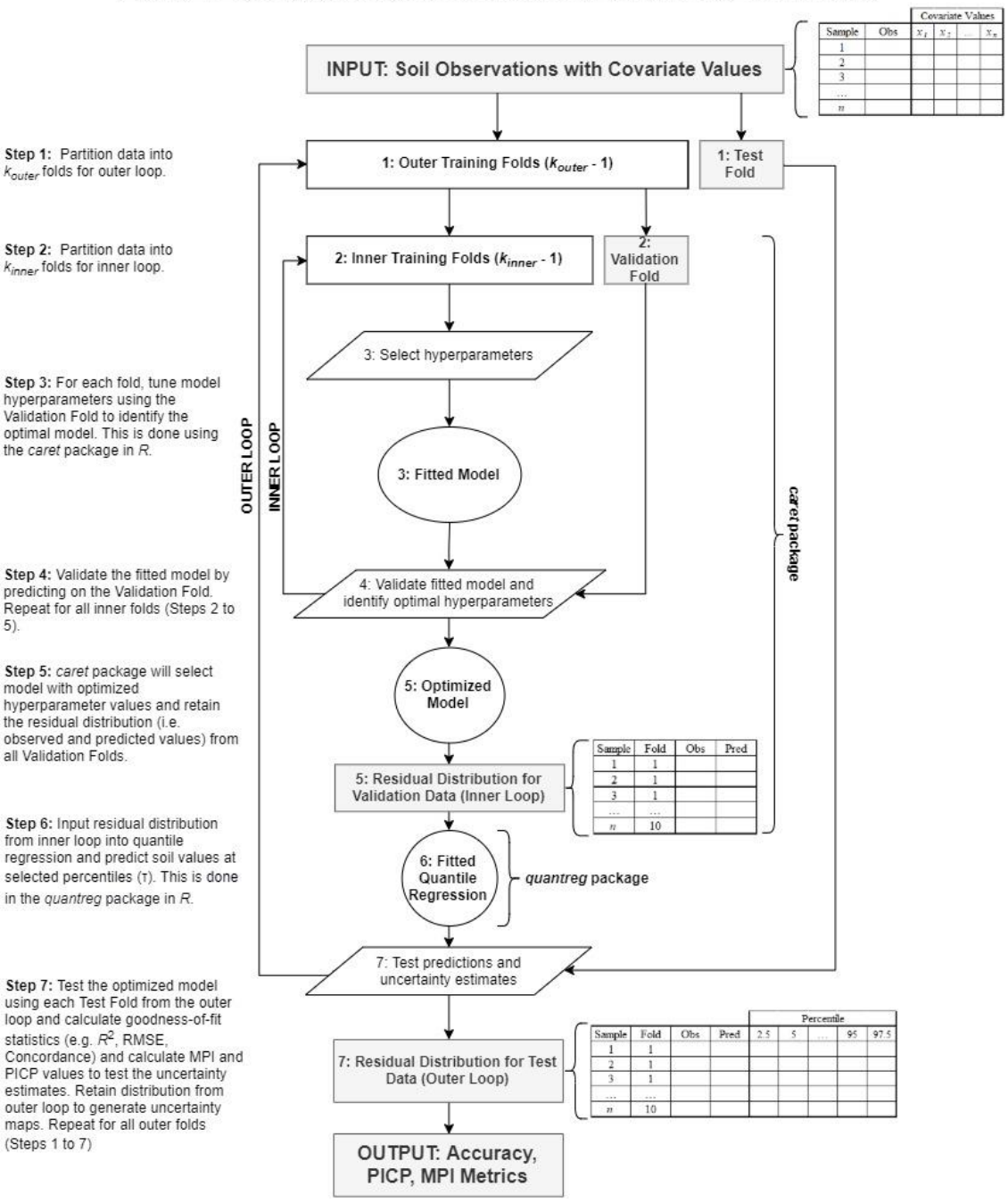

Figure 3.2. Generic framework for evaluating accuracy and uncertainty estimations of digital soil maps using ML and QR.

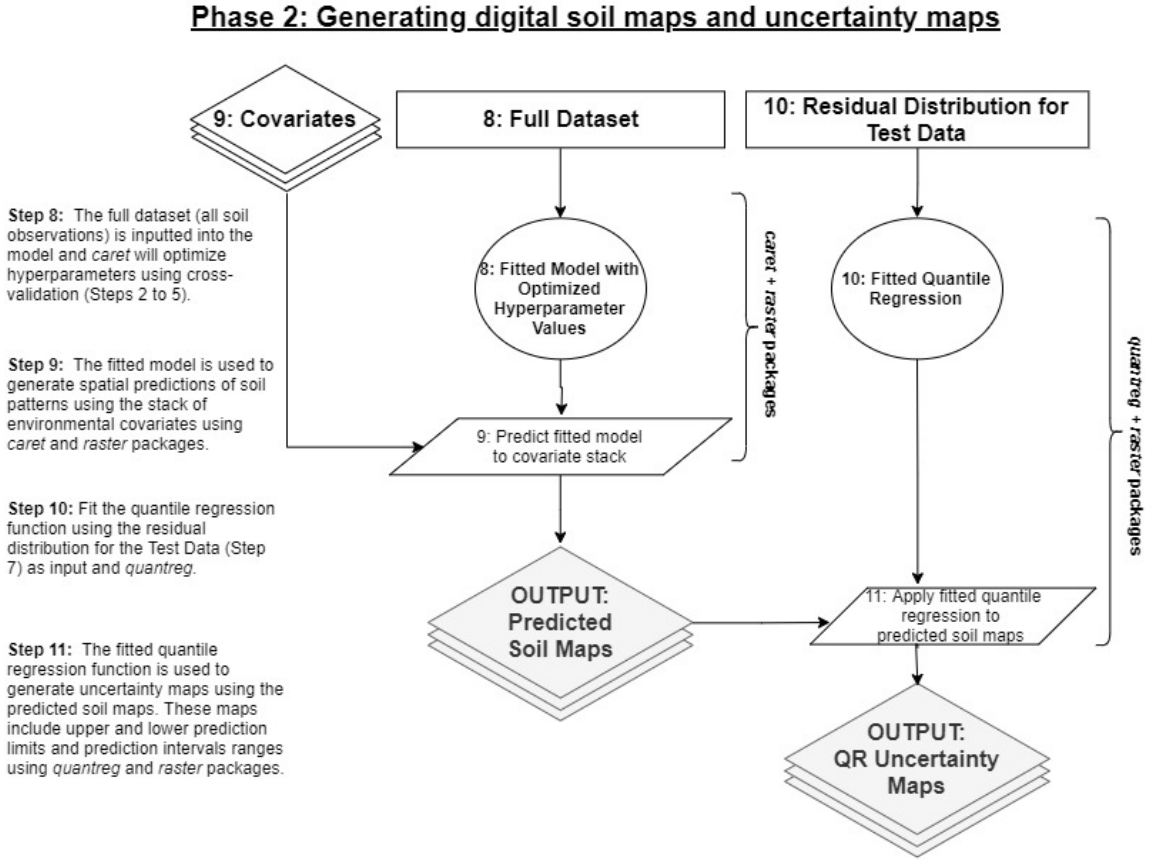

Figure 3.3. Generic framework for producing digital soil maps and uncertainty maps using ML and QR.

3.4. Results and Discussion

3.4.1. Modelling and Validation

In this study, 4 ML methods were used to predict 3 soil properties. The four ML models were RF, Cubist decision tree, k nearest neighbors (kNN) and support vector machine (SVM). The soil properties used for modelling were soil thickness, depth to carbonates and soil pH. The validation method used to validate the prediction results, as described in section 3.3.4 and 3.3.6, was a nested 10-fold cross validation with 20 repeats.

Maps Produced Using Different ML Models

A total of 12 maps were produced using the 4 ML models and the 3 soil properties using a 10-fold cross validation method with the optimized hyperparameter

values (Figure 3.4). As can be seen for all 3 soil properties in Figure 3.4 and Table 3.2, Cubist decision tree prediction range of values in the maps are larger than the other ML methods. If Figure 3.4 is compared with the elevation map (Figure 3.1), it can be seen that on the top of the hills the soil is shallow and depth to carbonates shown in the maps is deep. Depth to carbonates on the top of the hills is conceptual, and it means that if there were deep soils in that area, the depth to carbonates would be that deep. In the soil pH maps, acidic soils have been shown in red and are generally found at the higher elevations. Visual inspection of the maps in Figure 3.4 confirms that in this study area deeper soils are at lower elevations. Moreover, carbonated soils are found at lower elevations and basic soils are also at the lower elevations.

Statistics for soil predictions were calculated (Table 3.2). In Table 3.2 negative values have been predicted using the SVM method when predicting soil thickness. The negative values are likely a result of SVM extrapolating values beyond the range of the training data. Similarly, the Cubist model extrapolated large positive numbers for soil thickness and depth to carbonates. In both cases, extreme values were very limited in distribution. Furthermore, we chose to not bind the predicted values to the range of values found in the training data to ensure that QR estimated the quantile values from the actual predicted values.

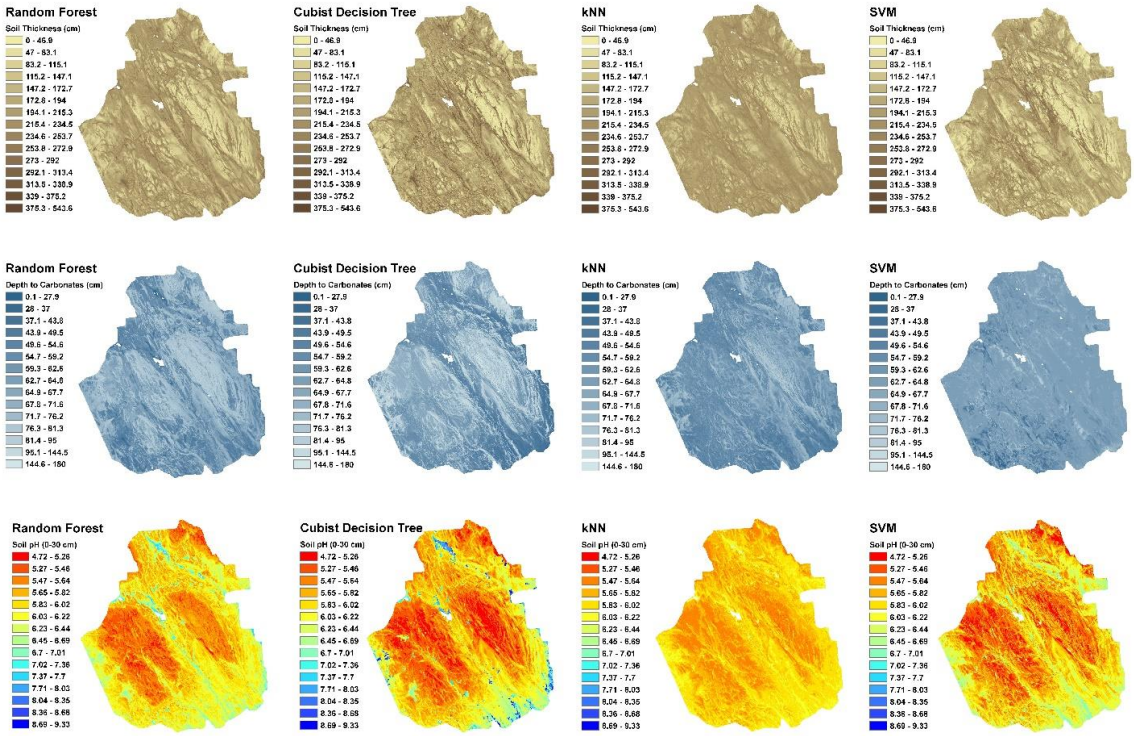

Figure 3.4. Maps of soil thickness, depth to carbonates, and soil pH generated using 4 ML models.

Validation Results

To validate the prediction results for the 3 soil properties and 4 ML models, a 10-fold nested cross validation method with 20 repeats, described in section 3.3.4 and 3.3.6, was used. Validation results were collected for prediction of soil thickness, depth to carbonates and soil pH using the 4 ML models (Table 3.1). The results in Table 3.1 show that RF had the highest R^2 and concordance and the lowest root mean squared error (RMSE) in modelling soil thickness. The R^2 and concordance for modelling soil thickness using RF were 0.35 and 0.47 respectively and RMSE was 103.19. The lowest validation results for the same soil property were found for kNN for which R^2 is 0.30 and concordance is 0.37. The highest error rate was found for kNN for modelling soil thickness (Table 3.1).

For depth to carbonates, kNN had the highest R^2 (0.13). The lowest R^2 results were found for Cubist decision tree in which R^2 was 0.05. However, modelling depth to carbonates using Cubist decision tree generated the highest concordance value which is 0.16. All the other three models generated the same concordance value which was 0.14.

RMSE results for modelling depth to carbonates were very close in all models; however, the lowest RMSE was found for kNN which was 29.44 (Table 3.1).

The validation results of modelling soil pH using the 4 ML models show that the highest R^2 and concordance and the lowest RMSE were found for RF. For RF modelling soil pH, R^2 is 0.26 and concordance is 0.37. The RMSE value for modelling soil pH using RF was 0.68. The lowest validation results and the highest RMSE result for modelling soil pH were found for kNN. The R^2 for modelling soil pH in kNN is 0.18 and concordance is 0.16. The RMSE value for modelling soil pH using kNN model was 0.73 (Table 3.1).

Table 3.1. Validation results (R2, concordance and RMSE) of ML models for 3 soil properties.

	Model Validation								
	Soil Thickness			Depth to Carbonates			Soil pH		
	R^2	Con	RMSE	R^2	Con	RMSE	R^2	Con	RMSE
RF	0.35	0.47	103.19	0.07	0.14	29.98	0.26	0.37	0.68
Cubist	0.31	0.49	106.15	0.05	0.16	30.56	0.2	0.34	0.72
kNN	0.3	0.37	108.36	0.13	0.14	29.44	0.18	0.16	0.73
SVM	0.32	0.48	105.36	0.08	0.14	29.76	0.22	0.34	0.7

3.4.2. QR 90% Prediction Interval (PI) Maps

Ninety (90%) percent PI range maps are known as uncertainty maps. The process for generating 90% PI maps was described in section 3.3.6. The importance of 90% PI maps is that they show how certain or uncertain a modelling method has been in its prediction at different locations on the map. In the uncertainty maps produced in this study (Figure 3.5) the highest uncertainty in all models has been depicted in red. It is important to mention that these uncertainty maps only show the level of uncertainty in modelling and the map cell values show this attribute.

According to uncertainty maps in Figure 3.5, different models have shown different levels of uncertainty for soil properties. For modelling soil thickness, all models have shown some levels of uncertainty in their prediction at lower elevations (Figure 3.1). For modelling depth to carbonates however, the Cubist decision tree model had the highest uncertainty at lower elevations while for the other ML methods the uncertainty

has been shown to be greater at higher elevations. For modelling soil pH uncertainty in all four model predictions was shown at lower elevations.

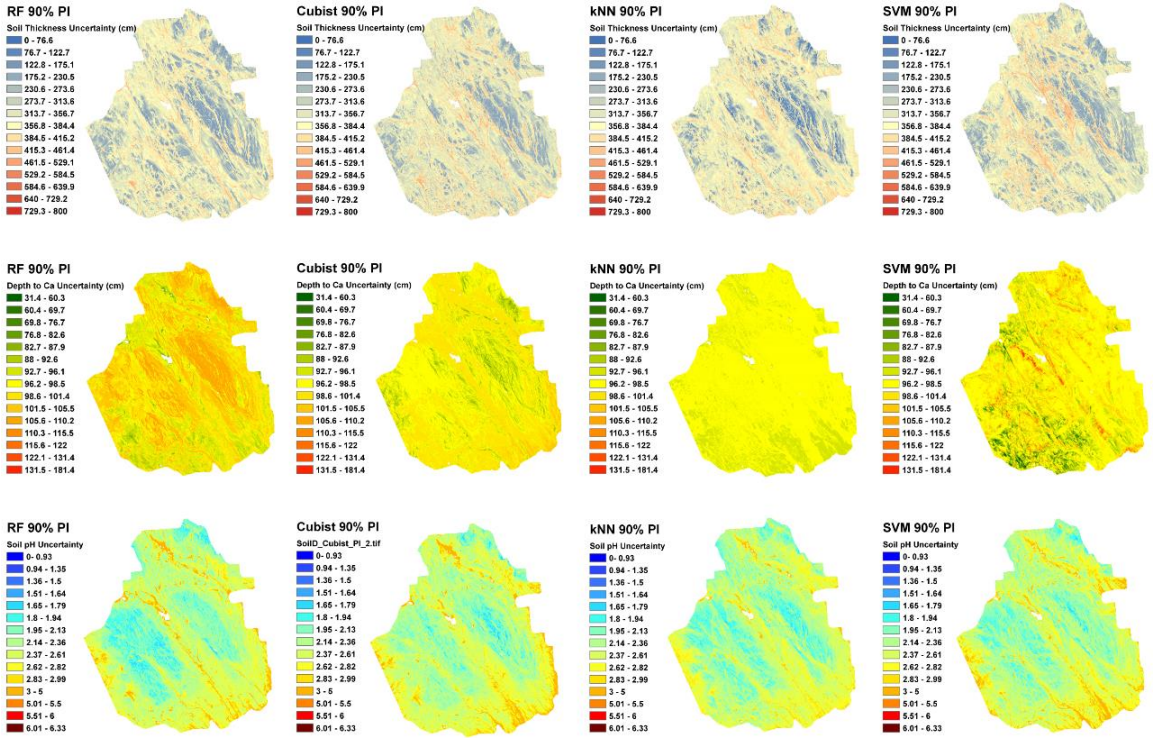

Figure 3.5. 90% prediction interval uncertainty maps generated using QR method for 3 soil properties (soil thickness, depth to carbonates, soil pH) for each of 4 ML models

Moreover, descriptive statistics of 90% PI maps show that the mean uncertainty values between the ML methods were similar. However, the range and variability in uncertainty values differed considerably (Table 3.2), which led to the differences in the appearance of the maps.

Table 3.2. Descriptive statiatics for soil prediction and 90% prediction interval maps

Soil Property	Depth	Model	Soil Prediction Maps					90% Prediction Interval Maps				
			Min	Max	Range	Mean	SD	Min	Max	Range	Mean	SD
Soil Thickness (cm)		RF	10.5	474.4	463.9	182.1	50.3	104	710.8	606.8	328.5	65.8
		Cubist	0	543.6	543.6	183.9	61.8	151.1	696.1	545	335.5	61.9
		kNN	21.3	291.8	270.5	195.2	36	-1.3	525.6	526.9	337.4	70.2
		SVM	-38.9	343	381.9	167	57.8	99	535	436	334.1	65.9
Depth to Carbonates (cm)		RF	19	111.4	92.4	66.4	10.2	79.8	119.3	39.5	100	4.4
		Cubist	32.3	180	147.7	71.2	20.4	82.1	103	20.9	97.5	2.9
		kNN	33.8	98.8	65	62.1	7.7	95.4	97.3	1.9	96.3	0.3
		SVM	0.1	144.9	144.8	63.9	5.4	31.5	181.4	149.9	97.5	5.6
Soil pH	0-30 cm	RF	4.9	7.8	2.9	5.9	0.4	1.5	4.1	2.6	2.4	0.3
		Cubist	4.7	9.3	4.6	5.8	0.5	1.4	4.4	3	2.4	0.3
		kNN	5.3	6.5	1.2	5.8	0.2	1.5	3.5	2	2.3	0.3
		SVM	4.8	7.5	2.7	5.8	0.4	1.6	3.8	2.2	2.4	0.3

3.4.3. QR crossing problem

In this study a small crossing problem occurred producing 90% PI map for soil thickness prediction using kNN. In the 90% PI map produced for kNN modeling soil thickness the minimum value is negative (Table 3.2). By generating the scatterplot of observed versus predicted values and delineating 5%, 50% and 95% quantile lines using QR, it can be seen that the left tail of the regression lines cross each other that causes the negative values in 90% PI maps (Figure 3.6).

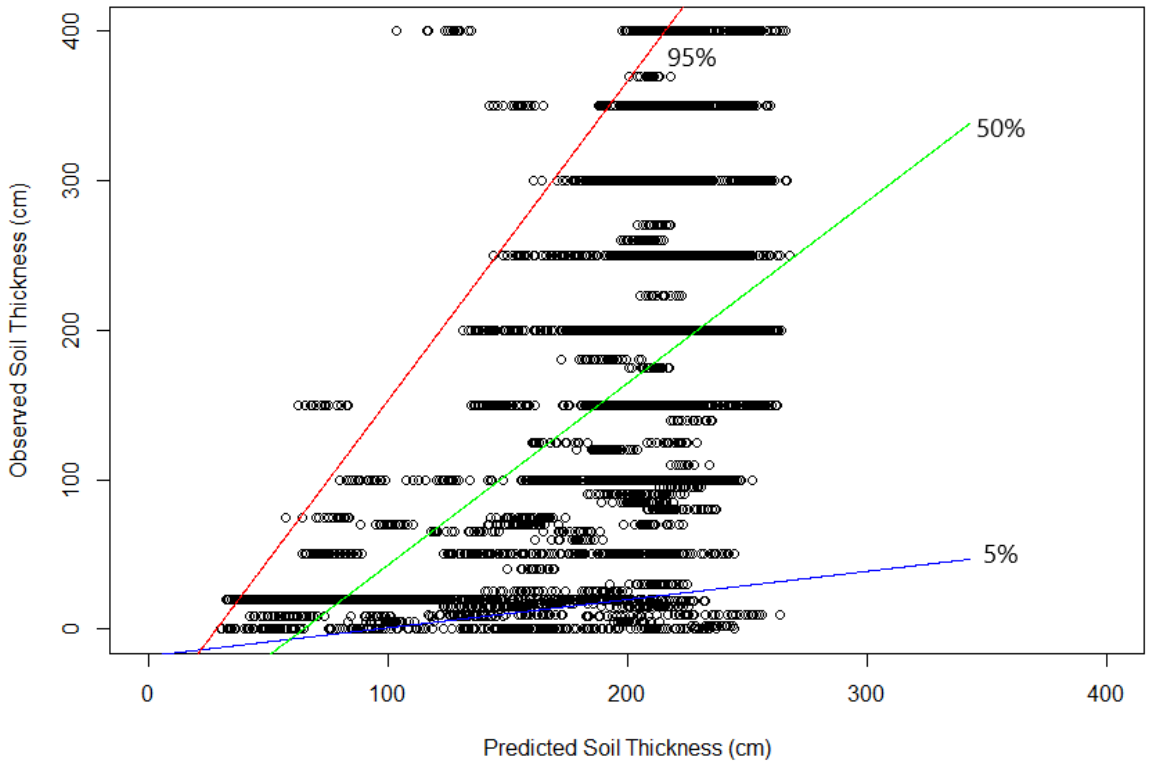

Figure 3.6. Scatterplot of predicted and observed values for soil thickness with linear regressions showing 5, 50 and 95% quantile predictions using kNN.

3.4.4. Mean Prediction Interval (MPI) results

Mean prediction interval (MPI) (Equation 6) is an indicator that is used to measure the quality of model performance. The lower the value of MPI, the better the model prediction performance is in terms of uncertainty. Every pair of the 20 quantiles on opposite sides, such as 5% and 95% quantiles that make a 90% PI, specified a range of each PI. The PIs used in this study were: 5%, 10%, 20%, 40%, 60%, 80%, 90%, 95%, 97.5%, and 99%. Upper quantile and lower quantile values were collected in different columns of the table produced in step 7 of phase 1 (Figure 3.2). To generate the MPI, the lower quantiles were subtracted from the upper quantiles and the total results of subtractions were averaged (equation 6) (Table 3.3).

In this section MPIs produced from different model outputs using QR were compared to determine the best model performance for modelling each soil property. The comparison in Table 3.2 and Figure 3.7 between MPIs of the four model outputs show that the performance of RF at 90% CL has been better than the other models for

modelling soil thickness (Figure 3.6: a). For modelling depth to carbonates, comparison of MPI values in Table 3.2 and Figure 3.7(b) shows that RF outperformed the other models at 90% CL and over. In the MPIs of the soil pH model, RF also outperformed the other models at 90% CL and over (Table 3.3 & Figure 3.7:c). Therefore, if we consider 90% confidence level as an indicator of model performance in modelling all three soil properties, RF is better than the other models because MPI values of the RF model predictions at 90% PI are less than for the other models.

Table 3.3. Mean prediction interval results for soil thickness, depth to carbonates and soil pH.

		Mean Prediction Intervals (MPI) for the Different Confidence Levels									
Dataset	Model	5%	10%	20%	40%	60%	80%	90%	95%	97.50%	99%
Soil Thickness	RF	9.26	18.52	38.73	83.73	142.70	236.44	330.31	404.21	458.12	556.86
	Cubist	9.29	19.02	40.18	88.23	150.59	246.64	336.89	419.73	462.91	553.52
	kNN	8.27	16.74	35.19	91.9	162.98	251.08	331.48	401.30	471.54	573.24
	SVM	8.62	17.34	36.67	89.82	157.25	247.67	333.34	401.28	479.29	612.68
Depth to Carbonates	RF	3.4	7.21	14.42	32.89	56.48	80.09	95.96	101.61	105.13	112.46
	Cubist	3.45	6.89	15	33.86	57.58	81.30	99.09	111.24	120.57	135.59
	kNN	3.65	7.44	16.02	33.58	57.10	82.26	99.76	116.78	124.95	139.12
	SVM	3.64	7.35	15.62	33.17	56.71	81.06	99.13	116.48	125.83	138.07
Soil pH	RF	0.06	0.12	0.25	0.54	0.91	1.59	2.33	2.86	3.4	4.11
	Cubist	0.06	0.12	0.25	0.6	0.98	1.67	2.4	2.98	3.58	4.14
	kNN	0.06	0.14	0.27	0.56	0..94	1.76	2.37	3.02	3.73	4.12
	SVM	0.06	0.12	0.24	0.53	0.91	1.71	2.38	2.98	3.66	4.12

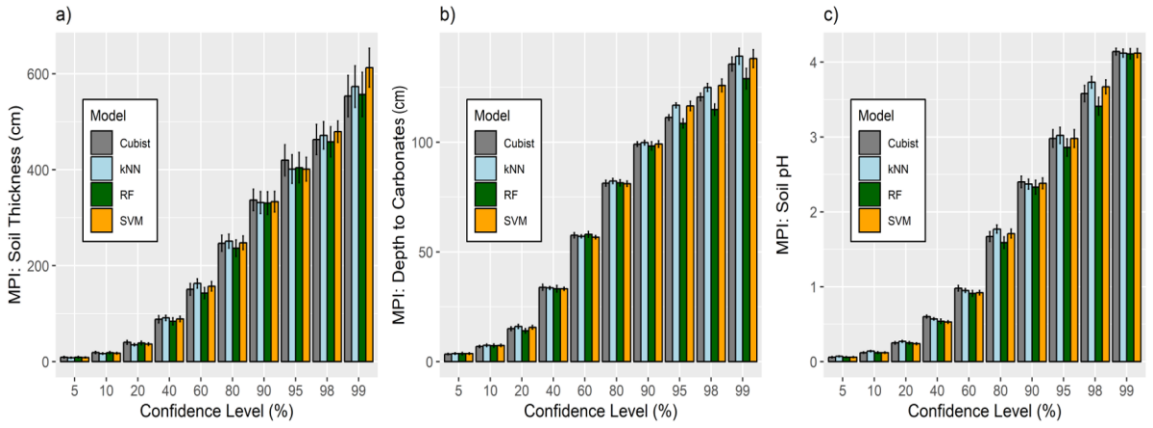

Figure 3.7. Mean prediction intervals for soil thickness (a), depth to carbonates (b), and soil pH (c). Error bars were generated using 20 repeats of nested cross-validation.

3.4.5. Prediction Interval Coverage Percentage (PICP)

PICP graphs are used to assess the performance of QR in terms of uncertainty quantification while MPIs assess the performance of models. In other words, PICPs tell us how accurate the QR uncertainty quantification has been by encapsulating observation data into 10 PI ranges. The process of producing 20 quantiles using QR methods was explained in section 3.3.6. To generate a 90% PI range, 95% quantile and 5% quantile predictions were collected in a table in step 7 (Figure 3.2) to estimate the upper limit and lower limit of the 90% PI range. Following that, the observation values associated with model predictions of training folds in step 1 of Figure 3.2 for all 10 folds were placed into the PI ranges and the number of the values that fell in each range were counted. Then the average of counted values was calculated by dividing it by the number of the rows. The final counting results averaged were collected in a table as one row. In the next repeat of the whole process of phase 1 in Figure 3.2, a new row was added to the PICP counting table until a total of 20 rows for 20 repeats of phase 1 (Figure 3.2) made the PICP table (Table 3.4). Table 3.4 is an example of a PICP table in which the observation values encapsulated into PI ranges are counted, averaged and collected for modelling soil thickness using a RF model. For the next step, each row of the table was plotted on a graph and compared with a 1:1 relationship line in a PICP graph (Figure 3.8).

Table 3.4. Prediction interval coverage probability table with 20 rows and 10 prediction intervals.

Replicate	5%	10%	20%	40%	60%	80%	90%	95%	97.5%	99%
1	4.88	10.49	19.51	41.46	61.46	79.76	90.00	94.39	96.59	99.02
2	3.66	8.54	19.27	40.24	61.46	78.54	88.54	94.63	97.07	99.27
3	4.63	9.76	18.78	40.98	59.51	80.24	89.02	95.61	97.32	98.54
4	4.63	9.51	20.49	40.98	60.73	79.76	90.24	94.63	97.07	99.02
5	5.12	9.51	19.51	39.02	60.49	77.56	90.24	94.39	96.83	99.27
6	5.85	8.29	19.76	40.00	59.51	79.27	89.27	94.88	97.32	99.02
7	4.63	10.00	20.49	40.73	60.49	80.73	89.76	94.63	96.83	99.27
8	6.34	9.27	22.44	40.24	59.02	80.49	88.78	94.15	96.83	99.02
9	6.10	11.22	17.07	40.49	59.76	81.46	90.00	94.63	96.59	99.02
10	4.39	8.78	20.49	40.00	60.49	78.05	89.27	94.39	97.07	99.27
11	5.37	10.73	19.27	40.24	59.02	79.51	88.78	94.63	97.07	99.02
12	5.12	8.78	15.85	37.56	60.24	80.24	90.98	94.88	96.83	98.78
13	5.12	10.49	20.00	40.24	60.24	79.27	90.49	94.39	97.32	99.27
14	4.39	8.54	18.54	41.46	61.22	79.02	90.24	94.15	96.83	99.02
15	7.32	11.71	19.02	40.24	58.78	79.76	90.73	94.88	96.83	99.27
16	5.85	10.98	18.54	40.98	59.27	80.00	89.27	94.88	97.32	99.27
17	4.39	10.49	19.51	41.71	59.27	79.51	89.02	94.15	96.83	99.27
18	6.34	10.73	20.00	40.24	59.27	80.00	89.76	94.15	96.10	98.78
19	5.61	9.27	20.73	39.02	59.51	80.24	89.27	94.39	96.83	99.02
20	3.66	9.27	19.02	40.00	59.51	78.78	89.76	94.39	97.56	99.27

In PICP graphs there are two important factors. The first factor is the closeness of the points to a 1:1 relationship on the graphs. The closer the points to the bisector line, the more accurate the quantification of uncertainty. On the other hand, if the points are far from the 1:1 line on the graphs, it will mean that the quantification of uncertainty was less accurate. Moreover, the distance from the bisector line shows the level of overestimation or underestimation of the uncertainty method whether the number of observation fractions are over or under the bisector line respectively. The second factor is the length of the boxplots. Boxplots are used in the graphs of Figure 3.8 to show the distribution of PICP values in 20 replicates for each PI range. Since 20 replicates of PICPs were plotted on each PICP graph, for each replicate there might have been some variation in uncertainty quantification. This variation is known as error range. The shorter the error range, the more stable and accurate the uncertainty quantification.

Assessment of QR uncertainty quantification was conducted for 4 ML models predicting 3 soil properties: soil thickness, depth to carbonates and soil pH. The PICP

graphs of soil thickness predictions for all models are almost the same. The PICP graphs for depth to carbonates and soil pH depict similar results. The PICP graphs counting observation in the PIs of 3 different soil properties are very accurate and in all cases the boxplots are very close to the bisector line.

As can be seen in Figure 3.8, the best uncertainty quantification using QR belongs to the soil thickness dataset in which the boxplots are close to the bisector line and the length of boxplots is very short. The second most accurate assessment of uncertainty quantification according to Figure 3.8 belongs to soil pH. The lengths of boxplots are small in the soil pH PICP graph; however, the length of boxplots shows that the variation in the 20 PICP replicates is more than soil thickness. The least accurate uncertainty quantification belongs to depth to carbonates in which the length of boxplots is longer than for soil thickness and soil pH PICPs. The main reason for these results is likely the lower number of datapoints for the depth to carbonates dataset compared to soil thickness and soil pH. The closeness of points to the bisector line infers that the quantification of uncertainty in all models using QR method is accurate, and QR is capable of measuring the uncertainty accurately.

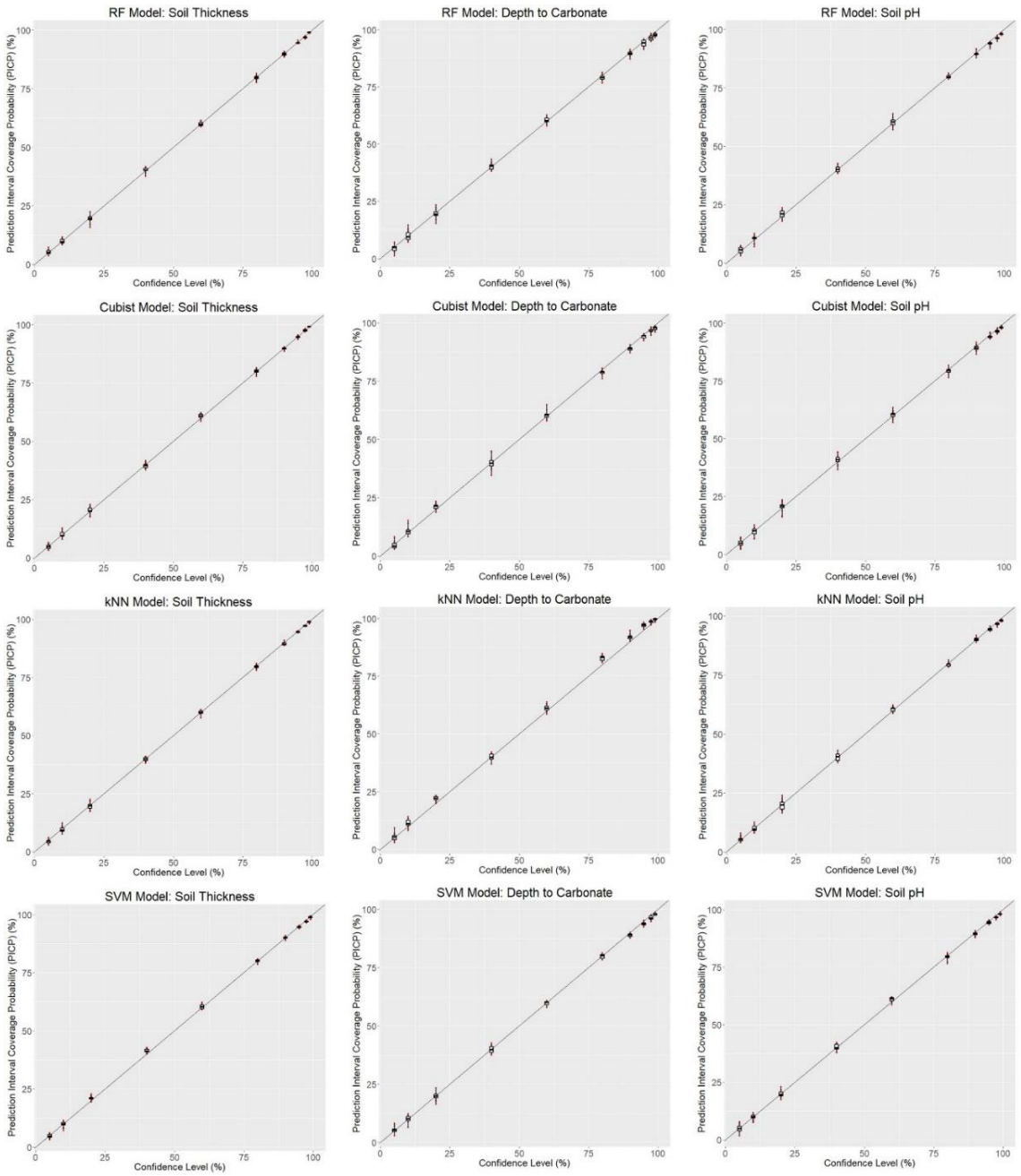

Figure 3.8. Prediction interval coverage probability plots using quantile regression and machine learning for the study area. Boxplots were generated using 20 repeats of nested cross-validation.

3.4.6. General Discussion and Future Work

The resulting PICP plots were optimal, where the confidence level and PICP showed a 1:1 relationship regardless of soil attribute, and machine learner. Although we followed the standard approach for assessing uncertainty estimates in the ML, QR, and

DSM literature (e.g. Ding et al., 2018; Malone et al., 2011; Muthusamy et al., 2016; Shrestha and Solomatine, 2006) by using PICP and MPI, recent studies such as Szatmári and Pásztor (2019) have suggested the use of the G statistic, which quantifies how close the PICP values are to their associated confidence levels.

Although it is recognized that the use of opportunistic sampling may not seem like an optimal approach by some, we do not anticipate this to be an issue when the core objective is to test an approach for estimating uncertainty. Future research may investigate the use of different sampling approaches and combinations of environmental covariates to minimize the uncertainty estimations while maximizing accuracy.

In modelling soil thickness using the kNN method, we raised a potential issue related to QR generating negative estimates of uncertainty and we had attributed it to crossing QR lines (Bondell et al., 2010), which may be related to the distribution of soil attribute values and the model residuals. Furthermore, we also recognize that the assumption of a linear relationship between the observed and predicted values may be a weakness of this approach; however, nonparametric quantile curves could be a potential solution (Bondell et al., 2010). These issues warrant further investigation as applied to a DSM context.

Given that this is the first implementation of the QR approach for DSM, there are several areas for future research. First, the generalizability of the approach should be further investigated for other study areas, soil properties, spatial scales, and machine learners. Although this study specifically tested the integration of QR and ML, ML may be substituted with other types of predictive models such as a purely geostatistical model (e.g. ordinary kriging) or a hybrid model (e.g. regression-kriging), given that the main inputs for QR are the model residuals. Such a comparison would be useful especially since kriging approaches provide an alternative for uncertainty estimation using the kriging variance.

In terms of existing ML-based approaches for uncertainty estimation, a comparison between the QR and bootstrapping approaches would be warranted. For example, Szatmári and Pásztor (2019) compared the uncertainty estimates produced from the kriging variance from universal kriging and RF regression kriging, with sequential Gaussian simulation (SGS), QRF, and bootstrapping of RF regression

kriging. There, it was shown that QRF and SGS were more optimal in estimating uncertainty; however, they also indicated that QRF and SGS were computationally demanding.

It is also necessary to note that our integration of RF and QR is fundamentally different from the QRF approach. In QRF, all predictions made at the terminal node across all individual trees are retained. Following this, the residual distribution produced by the individual trees for each terminal node is then used to calculate and predict the quantiles (Meinshausen, 2006). Although QRF is the primary example of a ML technique that generates uncertainty estimates in the DSM literature (e.g. Rudiyanto et al., 2018; Szatmári and Pásztor, 2019; Vaysse and Lagacherie, 2017), other types of ML techniques such as quantile regression neural networks (Cannon, 2011) could be tested and compared in future work.

A potential issue with the bootstrapping approach (Malone et al., 2017) is that increasing the number of bootstrap predictions would lead to a decrease in the prediction variance and thereby cause the narrowing of the prediction interval maps. Through a comparative study, if QR yields similar or better results than bootstrapping, this framework will overcome the major computational demands of bootstrapping, which requires the spatial prediction of each model iteration. Although we did not carry out a direct comparison with bootstrapping with respect to computational time, our personal experience was that the computationally demanding part of the bootstrapping process was in generating multiple (e.g. 100) realizations of a soil property map (i.e. applying a model to a covariate stack) and having to store each realization. In this proposed framework, only one realization of a soil property map is generated, as well as one realization of each quantile map (e.g. 95% and 5%). The uncertainty maps are generated by applying the QR function, fitted using only the residual distribution from the observed sites, and applied to the predicted soil map; and hence, not requiring multiple map realizations to be generated.

This solution would be particularly valuable when applied over large spatial extents (e.g. national and global mapping initiatives), when using ultra-high-resolution datasets (e.g. LiDAR), and when there are other Big Data challenges to overcome. For example, the most recent global-scale mapping effort was the SoilGrids250m (Hengl et al., 2017), whereby the authors avoided the modelling of uncertainty for continuous soil

variables and have indicated that QRF was a computationally intensive process—especially with increasing data volumes. By May 4, 2020, SoilGrids250m was updated to include uncertainty estimations using QRF.

3.5. Conclusion

The first objective of this study was to generate attribute maps using four ML models: RF, Cubist decision tree, kNN and SVM, modelling three soil properties: soil thickness, depth to carbonates and soil pH, and to validate the prediction results using *k*-fold cross validation method. The second objective of this study was to produce 90% PI maps using QR method. The third goal of this study was to assess model performance and uncertainty estimation using MPIs and PICPs. The case studies that have been used to demonstrate the integration of ML methods and QR, was located in a dry-forest ecosystem in the Kamloops region of British Columbia, Canada.

In the study area in Kamloops, the soil attribute maps generated for soil thickness showed the deepest soils in the lower elevations in all four maps. Moreover, depth to carbonates was shown to be close to the soil surface at lower elevations in all four maps. For soil pH the most basic soils also are located at lower elevations. The model validation results using a 10-fold nested cross validation method with 20 repeats were shown to be the highest for RF modelling soil thickness and soil pH. For modelling depth to carbonates the highest validation results belonged to kNN.

In the 90% PI maps it is evident that different ML methods show different levels of uncertainty modelling the three soil properties. The quality of model prediction was measured using MPIs at 90% prediction levels. The MPI results showed that RF was the most certain ML method modelling all three soil properties. The quantification of uncertainty also was assessed in this study using PICP graphs. The results showed that quantification of uncertainty using QR was the most accurate for the soil thickness dataset. According to the obtained results the best model made in this study was RF modelling soil thickness because of high validation results, and low MPI value compared to the other ML methods. Moreover, quantification of uncertainty was the most accurate for soil thickness using the QR method.

The results including 90% PI maps, assessment of model performance and quantification of uncertainty using PICPs, have been used to demonstrate the capability of the QR method in quantification of uncertainty regardless of ML methods used in digital soil mapping. Further study will be designed to improve the mapping and uncertainty quantification by increasing the numbers of datapoints. Increasing the number of datapoints should improve the model prediction quality which in turn will improve the QR prediction results. Moreover, the quality of uncertainty quantification using QR can be compared with the other uncertainty quantification methods such as bootstrapping and QRF to illustrate the performance of the QR method in measuring uncertainty.

Chapter 4.

Thesis Conclusions

The goal of sustainable forest management is to ensure that forest resources will continue to exist for the benefit of current and future generations (Szaro et al., 2000). Successful forest management practices depend on enhancement of forest productivity which is in turn directly related to forest soil productivity (Ayma-Romay and Bown, 2019; Schoenholtz et al., 2000). Burger and Kelting (1999) have suggested a10-step soil quality monitoring approach and in step 7 they suggested evaluating the soil quality by using geostatistical techniques or some other type of spatial extrapolation to produce soil quality maps. Early soil maps referred to as conventional soil maps, were expensive to produce and they suffered from limited accuracy and precision. Moreover, they were at small scale and were not suitable for local use purposes (Yang et al., 2011; Zhu et al., 2001). In late 20^{th} century a new subdiscipline of soil science was born called DSM (Brevik et al., 2016; McBratney et al., 2003a; Minasny and McBratney, 2016).

In DSM, a quantitative model relates field soil observations and environmental variables to make new predictions for all mapping areas (Minasny and McBratney, 2016). Most environmental variables are derived from digital elevation models (DEM) (Cavazzi et al., 2013). The best DEMs can be derived from LiDAR data because they are accurate products, and they have high-resolutions that make them suitable for forest management (Haneberg et al., 2009; Liu, 2008). The numerical models can be ML methods. MLs are statistical algorithms that a computer uses to perform a specific task without explicit instruction (Dietterich, 2000; Heung et al., 2016). Examples of ML methods are RF, Cubist decision tree and multi linear regression (Heung et al., 2016). Since foresters require accurate maps, the errors in digital soil maps should be assessed. According to international standards for producing digital soil maps the uncertainty of the final product of DSM should be quantified with a 90% PI map (Arrouays et al., 2014a). QR is a new novel approach in DSM that can be used for all ML methods to quantify uncertainty in digital soil maps (Koenker and Bassett, 1978; Rahmati et al., 2019).

In the first component (Chapter 2) of this thesis, RF was used to map soil thickness, depth to carbonates, soil pH, coarse fragment content, and clay content, using LiDAR derived covariates. Then the prediction results were validated using a nested 10-fold cross validation with 20 repeats. Lastly, how these maps can be useful for forest management was discussed. The mapping results of the five soil properties showed that the patterns of digital soil maps are associated with BEC subzones. The validation results showed the best results for soil thickness with concordance of 0.47, and the worst validation results for depth to carbonates with concordance of 0.13. In the results section of Chapter 2 it was discussed that digital soil maps can be used as useful tools to preserve productive forest soils. In that chapter it was discussed that clear cutting in some cutblocks has exposed soil to erosion hazard, displacement and exposed to unfavorable hazard, and puddling and compaction hazard, and digital soil maps can help select better places or management protocols for forest harvesting.

In the second component (Chapter 3) of this thesis, different ML methods including RF, Cubist, kNN and SVM were used to map three soil properties: soil thickness, depth to carbonates and soil pH. Then the prediction results were validated using a nested 10-fold cross validation with 20 repeats. Following that, 90% PI maps were produced for each of those property maps using a QR method. Lastly, the uncertainty estimations and model performance were assessed using metrics such as PICP and MPI. The mapping results for soil thickness in all ML methods showed deeper soils at lower elevations. The maps produced for depth to carbonates also showed more carbonated soils at lower elevations. The soil pH maps showed soils with lower pH at higher elevations. The validation results in different models for one soil property were different. For modelling soil thickness the best validation results belonged to RF with concordance of 0.47 and the lowest validation results belonged to kNN with concordance of 0.37. The validation results for modelling depth to carbonates showed the highest results for Cubist model with concordance of 0.16 although R^2 was lowest for the Cubist model. The validation results for soil pH was the highest for RF with concordance of 0.37 and the lowest for kNN with concordance of 0.16.

The 90% PI maps produced using the QR method demonstrated uncertainties in the study area. Uncertainty in soil thickness predictions for all models showed the highest values at lower elevations. Uncertainty in depth to carbonate predictions showed the highest values at higher elevations except Cubist decision tree. Uncertainty in soil

pH predictions showed the highest values at lower elevations. The performance of the models was checked using MPIs using QR method and showed that RF had the best performance for 90% PI. The results of assessment of QR uncertainty quantification using PICPs were promising and in all cases showed a bisector relationship between observed and predicted values. The results suggest that QR can be used as a powerful method to quantify uncertainty in all ML models. The QR method described in this research has potential to provide a computationally efficient approach for estimating local uncertainty that may have applications for a variety of different spatial modelling applications.

4.1. Challenges in this Research and Future Research

Challenges in this research included challenges in sampling, modelling and uncertainty measurements. Challenges in sampling related to the logistics of sampling in the forest. Since the forest was remote, two trucks with radios were necessary due to safety concerns. Forest roads were narrow, sometimes steep and some were blocked with fallen tree trunks that needed to be removed to reach sampling locations. Hiking 5 hours per day to reach some sampling locations was quite common since roads were often not passable. Some sampling sites were located on extremely steep slopes. Environmental conditions in the forest were quite variable and a sunny day could turn to hail in less than an hour. Bugs were abundant and sampling soils with their presence was almost impossible. These factors increased the time required to collect samples and therefore reduced the number of sites where observations could be made within the project budget.

Challenges in modelling were mostly related to data selection and preprocessing. Creating environmental variables needed a lot of analyses and preparation. Noisy LiDAR data was necessary to be passed through WhiteBox filtering tools and roads needed to be fuzzed out to reduce the perturbing effects of the road network on topographic derivatives derived from hydrologic flow. Roads were identified from access network datasets that were manually checked for completeness and accuracy. Then a 4m buffer was applied to the road lines and the underlying DEM cut, and then gap filled to recreate a smooth slope where the ditches and road fill were previously visible in the LiDAR dataset. The continuous covariates needed to be scaled and centered and categorical covariates needed to be encoded before using them in kNN and SVM models. Encoding

is used to turn the categorical variables into numerical variables that can be used in modelling (Graves, 2017).

Challenges in uncertainty measurements were mostly related to incorporation and connection between validation measurement and uncertainty quantification at the same looping process. Although uncertainty quantification using QR is quite easy, in order to calculate uncertainty assessment metrics, we needed to collect prediction results from the model that was run for 10 iterations of 10-fold cross validation, 20 times. The validation results were collected and averaged in separate tables for 200 times (20 repeats X10 iterations) of running 10-fold cross validation and uncertainty quantile results were collected for each of 20 repeats to calculate 20 PICPs. All quantile prediction results were used to measure MPIs. The other challenge was with regard to uncertainty map production. Since the QR package is a statistical package and not a spatial analysis package, it does not have a built-in method for producing uncertainty maps. The first approach was to reclassify the property maps and then we came up with a novel method to recalculate the map property values using the new 5% and 95% QR line equations created by QR.

In our future work we plan to improve the sampling method by collecting much more data in sampling sites for one variable only and without spending a lot of time on lab analysis. More samples will be collected for some of the current property maps with low validation results such as depth to carbonates. We plan to attempt to use the QR method in different stages of sampling to reduce the sampling costs by going to the locations in which the models show the highest uncertainty. More digital soil maps for other soil properties such as sand content, organic carbon content, total nitrogen content, and forest floor will be produced for the same study area. Interpretive maps for other soil degradation hazards such as soil erosion and soil displacement will be produced. We also plan to conduct a comparison between the QR method and other uncertainty quantification methods such as quantile regression forest to assess the uncertainty quantification in both methods using PICPs and other statistical tests. One statistical approach would be to study the distribution of residuals in confidence levels produced by QR.

References

Ackerson, J.P., Demattê, J.A.M., Morgan, C.L.S., 2015. Predicting clay content on field-moist intact tropical soils using a dried, ground VisNIR library with external parameter orthogonalization. Geoderma 259, 196–204. https://doi.org/10.1016/j.geoderma.2015.06.002

Adhikari, K., Owens, P.R., Libohova, Z., Miller, D.M., Wills, S.A., Nemecek, J., 2019. Assessing soil organic carbon stock of Wisconsin, USA and its fate under future land use and climate change. Sci. Total Environ. 667, 833–845. https://doi.org/10.1016/j.scitotenv.2019.02.420

Alpaydin, E., 2020. Introduction to Machine Learning. MIT Press.

Arrouays, D., Lagacherie, P., Hartemink, A.E., 2017. Digital soil mapping across the globe. Geoderma Reg., Digital soil mapping across the globe 9, 1–4. https://doi.org/10.1016/j.geodrs.2017.03.002

Arrouays, D., Mcbratney, A., Minasny, B., Hempel, J., Heuvelink, G., Macmillan, R.A., Hartemink, A., Lagacherie, P., McKenzie, N., 2014a. The GlobalSoilMap project specifications. Glob. Basis Glob. Spat. Soil Inf. Syst. - Proc. 1st Glob. Conf. 9–12. https://doi.org/10.1201/b16500-4

Arrouays, D., McKenzie, N., Hempel, J., Forges, A.R. de, McBratney, A.B., 2014b. GlobalSoilMap: Basis of the global spatial soil information system. CRC Press.

Asner, G.P., Mascaro, J., Muller-Landau, H.C., Vieilledent, G., Vaudry, R., Rasamoelina, M., Hall, J.S., van Breugel, M., 2012. A universal airborne LiDAR approach for tropical forest carbon mapping. Oecologia 168, 1147–1160. https://doi.org/10.1007/s00442-011-2165-z

Austin, M.P., Nicholls, A.O., Margules, C.R., 1990. Measurement of the Realized Qualitative Niche: Environmental Niches of Five Eucalyptus Species. Ecol. Monogr. 60, 161–177. https://doi.org/10.2307/1943043

Ayma-Romay, A.I., Bown, H.E., 2019. Biomass and dominance of conservative species drive above-ground biomass productivity in a mediterranean-type forest of Chile. For. Ecosyst. 6, 47. https://doi.org/10.1186/s40663-019-0205-z

Bahrawi, J.A., Elhag, M., Aldhebiani, A.Y., Galal, H.K., Hegazy, A.K., Alghailani, E., 2016. Soil Erosion Estimation Using Remote Sensing Techniques in Wadi Yalamlam Basin, Saudi Arabia [WWW Document]. Adv. Mater. Sci. Eng. https://doi.org/10.1155/2016/9585962

Ballabio, C., 2009. Spatial prediction of soil properties in temperate mountain regions using support vector regression. Geoderma 151, 338–350. https://doi.org/10.1016/j.geoderma.2009.04.022

Baritz, R., Seufert, G., Montanarella, L., Van Ranst, E., 2010. Carbon concentrations and stocks in forest soils of Europe. For. Ecol. Manag. 260, 262–277. https://doi.org/10.1016/j.foreco.2010.03.025

Bässler, C., Stadler, J., Müller, J., Förster, B., Göttlein, A., Brandl, R., 2011. LiDAR as a rapid tool to predict forest habitat types in Natura 2000 networks. Biodivers. Conserv. 20, 465–481. https://doi.org/10.1007/s10531-010-9959-x

Battaglia, M., Sands, P.J., 1998. Process-based forest productivity models and their application in forest management. For. Ecol. Manag. 102, 13–32. https://doi.org/10.1016/S0378-1127(97)00112-6

Behrens, T., Schmidt, K., MacMillan, R.A., Viscarra Rossel, R.A., 2018. Multiscale contextual spatial modelling with the Gaussian scale space. Geoderma 310, 128–137. https://doi.org/10.1016/j.geoderma.2017.09.015

Behrens, T., Schmidt, K., Ramirez-Lopez, L., Gallant, J., Zhu, A.-X., Scholten, T., 2014. Hyper-scale digital soil mapping and soil formation analysis. Geoderma 213, 578–588. https://doi.org/10.1016/j.geoderma.2013.07.031

Behrens, T., Zhu, A.-X., Schmidt, K., Scholten, T., 2010. Multi-scale digital terrain analysis and feature selection for digital soil mapping. Geoderma 155, 175–185. https://doi.org/10.1016/j.geoderma.2009.07.010

Boettinger, J.L., Ramsey, R.D., Bodily, J.M., Cole, N.J., Kienast-Brown, S., Nield, S.J., Saunders, A.M., Stum, A.K., 2008. Landsat Spectral Data for Digital Soil Mapping, in: Digital Soil Mapping with Limited Data. Springer, Dordrecht, pp. 193–202. https://doi.org/10.1007/978-1-4020-8592-5_16

Bondell, H.D., Reich, B.J., Wang, H., 2010. Noncrossing quantile regression curve estimation. Biometrika 97, 825–838. https://doi.org/10.1093/biomet/asq048

Bonfatti, B.R., Hartemink, A.E., Vanwalleghem, T., Minasny, B., Giasson, E., 2018. A mechanistic model to predict soil thickness in a valley area of Rio Grande do Sul, Brazil. Geoderma 309, 17–31. https://doi.org/10.1016/j.geoderma.2017.08.036

Bontemps, J.-D., Bouriaud, O., 2014. Predictive approaches to forest site productivity: recent trends, challenges and future perspectives. For. Int. J. For. Res. 87, 109–128. https://doi.org/10.1093/forestry/cpt034

Breiman, L., 2001. Random Forests. Mach. Learn. 45, 5–32. https://doi.org/10.1023/A:1010933404324

Brevik, E.C., Calzolari, C., Miller, B.A., Pereira, P., Kabala, C., Baumgarten, A., Jordán, A., 2016. Soil mapping, classification, and pedologic modeling: History and future directions. Geoderma, Soil mapping, classification, and modelling: history and future directions 264, 256–274. https://doi.org/10.1016/j.geoderma.2015.05.017

Brevik, E.C., Hartemink, A.E., 2010. Early soil knowledge and the birth and development of soil science. CATENA 83, 23–33. https://doi.org/10.1016/j.catena.2010.06.011

Bruce, J.P., Frome, M., Haites, E., Janzen, H., Lal, R., Paustian, K., 1999. Carbon sequestration in soils. J. Soil Water Conserv. 54, 382–389.

Burger, J.A., 2009. Management effects on growth, production and sustainability of managed forest ecosystems: Past trends and future directions. For. Ecol. Manag., Forest Soil Science: Celebrating 50 Years of Research on Properties, Processes and Management of Forest Soils Forest Soil Science: Celebrating 50 Years of Research on Properties, Processes and Management of Forest Soils Forest Soil Science: Celebrating 50 Years of Research on Properties, Processes and Management of Forest Soils Forest Soil Science: Celebrating 50 Years of Research on Properties, Processes and Management of Forest Soils 258, 2335–2346. https://doi.org/10.1016/j.foreco.2009.03.015

Burger, J.A., Kelting, D.L., 1999. Using soil quality indicators to assess forest stand management. For. Ecol. Manag. 122, 155–166. https://doi.org/10.1016/S0378-1127(99)00039-0

Campbell, D.M.H., White, B., Arp, P.A., 2013. Modeling and mapping soil resistance to penetration and rutting using LiDAR-derived digital elevation data. J. Soil Water Conserv. 68, 460–473. https://doi.org/10.2489/jswc.68.6.460

Canada, G. of C. and A.-F., 2013. Canadian Soil Information Service [WWW Document]. URL http://sis.agr.gc.ca/cansis/nsdb/slc/v3.2/index.html (accessed 10.15.17).

Cannon, A.J., 2011. Quantile regression neural networks: Implementation in R and application to precipitation downscaling. Comput. Geosci. 37, 1277–1284. https://doi.org/10.1016/j.cageo.2010.07.005

Carmean, W.H., 1996. Forest Site-Quality Estimation Using Forest Ecosystem Classification in Northwestern Ontario, in: Sims, R.A., Corns, I.G.W., Klinka, K. (Eds.), Global to Local: Ecological Land Classification: Thunderbay, Ontario, Canada, August 14–17, 1994. Springer Netherlands, Dordrecht, pp. 493–508. https://doi.org/10.1007/978-94-009-1653-1_35

Carré, F., McBratney, A.B., Mayr, T., Montanarella, L., 2007. Digital soil assessments: Beyond DSM. Geoderma 142, 69–79. https://doi.org/10.1016/j.geoderma.2007.08.015

Carter, M.R., 1993. Soil Sampling and Methods of Analysis. CRC Press.

Cavazzi, S., Corstanje, R., Mayr, T., Hannam, J., Fealy, R., 2013. Are fine resolution digital elevation models always the best choice in digital soil mapping? Geoderma 195–196, 111–121. https://doi.org/10.1016/j.geoderma.2012.11.020

Chamberlain, G., 1996. Quantile regression, censoring, and the structure of Wages, in: Advances in Econometrics: Volume 1: Sixth World Congress. Cambridge University Press, pp. 171–207.

Cole, T.J., 1988. Fitting Smoothed Centile Curves to Reference Data. J. R. Stat. Soc. Ser. A Stat. Soc. 151, 385–406. https://doi.org/10.2307/2982992

Cole, T.J., Green, P.J., 1992. Smoothing reference centile curves: The lms method and penalized likelihood. Stat. Med. 11, 1305–1319. https://doi.org/10.1002/sim.4780111005

Cressie, N., Kornak, J., 2003. Spatial Statistics in the Presence of Location Error with an Application to Remote Sensing of the Environment. Stat. Sci. 18, 436–456.

Curran, M., Maynard, D., Heninger, R., Terry, T., Howes, S., Stone, D., Niemann, T., Miller, R.E., 2007. Elements and rationale for a common approach to assess and report soil disturbance. For. Chron. 83, 852–866. https://doi.org/10.5558/tfc83852-6

Dassot, M., Constant, T., Fournier, M., 2011. The use of terrestrial LiDAR technology in forest science: application fields, benefits and challenges. Ann. For. Sci. 68, 959–974. https://doi.org/10.1007/s13595-011-0102-2

Dietterich, T.G., 2000. Ensemble Methods in Machine Learning, in: Multiple Classifier Systems, Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 1–15.

Ding, C., Duan, J., Zhang, Y., Wu, X., Yu, G., 2018. Using an ARIMA-GARCH Modeling Approach to Improve Subway Short-Term Ridership Forecasting Accounting for Dynamic Volatility. IEEE Trans. Intell. Transp. Syst. 19, 1054–1064. https://doi.org/10.1109/TITS.2017.2711046

Dogulu, N., López López, P., Solomatine, D.P., Weerts, A.H., Shrestha, D.L., 2015a. Estimation of predictive hydrologic uncertainty using the quantile regression and UNEEC methods and their comparison on contrasting catchments. Hydrol. Earth Syst. Sci. 19, 3181–3201. https://doi.org/10.5194/hess-19-3181-2015

Dogulu, N., Lopez, P.L., Solomatine, D.P., Weerts, A.H., Shrestha, D.L., 2015b. Estimation of predictive hydrologic uncertainty using the quantile regression and UNEEC methods and their comparison on contrasting catchments [WWW Document]. Hydrol. Earth Syst. Sci. https://doi.org/10.5194/hess-19-3181-2015

Dosseto, A., Buss, H., Puthiyaveetil Othayoth, S., 2011. The delicate balance between soil production and erosion, and its role on landscape evolution. Appl. Geochem. - APPL GEOCHEM. https://doi.org/10.1016/j.apgeochem.2011.03.020

Dubayah, R.O., Drake, J.B., 2000. Lidar Remote Sensing for Forestry. J. For. 98, 44–46. https://doi.org/10.1093/jof/98.6.44

Dubayah, R.O., Knox, R., Hofton, M., Blair, J.B., Drake, J., 2000. Land surface characterization using lidar remote sensing, in: Spatial Information for Land Use Management.

ESRI, 2019. Digital Elevation Models. URL https://learn.arcgis.com/en/related-concepts/digital-elevation-models.htm

Faculty of Forestry, UBC, 2009. IDF zone [WWW Document]. CFCG. URL https://cfcg.forestry.ubc.ca/resources/cataloguing-in-situ-genetic-resources/idf-zone/ (accessed 4.6.20).

Feizizadeh, B., Jankowski, P., Blaschke, T., 2014. A GIS based spatially-explicit sensitivity and uncertainty analysis approach for multi-criteria decision analysis. Comput. Geosci. 64, 81–95. https://doi.org/10.1016/j.cageo.2013.11.009

Fink, C.M., Drohan, P.J., 2016. High Resolution Hydric Soil Mapping using LiDAR Digital Terrain Modeling. Soil Sci. Soc. Am. J. 80, 355–363. https://doi.org/10.2136/sssaj2015.07.0270

Finke, P.A., 2012. On digital soil assessment with models and the Pedometrics agenda. Geoderma, Entering the Digital Era: Special Issue of Pedometrics 2009, Beijing 171–172, 3–15. https://doi.org/10.1016/j.geoderma.2011.01.001

Fitzenberger, B., 2012. Wages and Employment Across Skill Groups: An Analysis for West Germany. Springer Science & Business Media.

Fouedjio, F., Klump, J., 2019. Exploring prediction uncertainty of spatial data in geostatistical and machine learning approaches. Environ. Earth Sci. 78, 38. https://doi.org/10.1007/s12665-018-8032-z

Franc, A., Laroussinie, O., Karjalainen, T., 2001. Criteria and indicators for sustainable forest management at the forest management unit level. European Forest Institute, Joensuu.

Franklin, J.F., Forman, R.T.T., 1987. Creating landscape patterns by forest cutting: Ecological consequences and principles. Landsc. Ecol. 1, 5–18. https://doi.org/10.1007/BF02275261

Galzki, J.C., Birr, A.S., Mulla, D.J., 2011. Identifying critical agricultural areas with three-meter LiDAR elevation data for precision conservation. J. Soil Water Conserv. 66, 423–430. https://doi.org/10.2489/jswc.66.6.423

Gessler, P.E., Chadwick, O.A., Chamran, F., Althouse, L., Holmes, K., 2000. Modeling Soil–Landscape and Ecosystem Properties Using Terrain Attributes. Soil Sci. Soc. Am. J. 64, 2046–2056. https://doi.org/10.2136/sssaj2000.6462046x

Ghaderi, A., Abbaszadeh Shahri, A., Larsson, S., 2019. An artificial neural network based model to predict spatial soil type distribution using piezocone penetration test data (CPTu). Bull. Eng. Geol. Environ. 78, 4579–4588. https://doi.org/10.1007/s10064-018-1400-9

Graves, E.E., 2017. Package 'onehot.'

Greve, M.H., Kheir, R.B., Greve, M.B., Bøcher, P.K., 2012. Quantifying the ability of environmental parameters to predict soil texture fractions using regression-tree model with GIS and LIDAR data: The case study of Denmark. Ecol. Indic. 18, 1–10. https://doi.org/10.1016/j.ecolind.2011.10.006

Grigal, D.F., 2000. Effects of extensive forest management on soil productivity. For. Ecol. Manag. 138, 167–185. https://doi.org/10.1016/S0378-1127(00)00395-9

Grimm, R., Behrens, T., Märker, M., Elsenbeer, H., 2008. Soil organic carbon concentrations and stocks on Barro Colorado Island — Digital soil mapping using Random Forests analysis. Geoderma 146, 102–113. https://doi.org/10.1016/j.geoderma.2008.05.008

Hämmerle, M., Höfle, B., 2014. Effects of Reduced Terrestrial LiDAR Point Density on High-Resolution Grain Crop Surface Models in Precision Agriculture. Sensors 14, 24212–24230. https://doi.org/10.3390/s141224212

Haneberg, W.C., Cole, W.F., Kasali, G., 2009. High-resolution lidar-based landslide hazard mapping and modeling, UCSF Parnassus Campus, San Francisco, USA. Bull. Eng. Geol. Environ. 68, 263–276. https://doi.org/10.1007/s10064-009-0204-3

Hastie, T., Tibshirani, R., Friedman, J., 2009. Random Forests, in: Hastie, T., Tibshirani, R., Friedman, J. (Eds.), The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer Series in Statistics. Springer New York, New York, NY, pp. 587–604. https://doi.org/10.1007/978-0-387-84858-7_15

Hawkins, D.M., 2004. The Problem of Overfitting. J. Chem. Inf. Comput. Sci. 44, 1–12. https://doi.org/10.1021/ci0342472

He, X., 1997. Quantile Curves without Crossing. Am. Stat. 51, 186–192. https://doi.org/10.1080/00031305.1997.10473959

Henderson, F.M., Lewis, A.J., 1998. Principles and applications of imaging radar. Manual of remote sensing: Third edition, Volume 2.

Hengl, T., Mendes de Jesus, J., Heuvelink, G.B.M., Ruiperez Gonzalez, M., Kilibarda, M., Blagotić, A., Shangguan, W., Wright, M.N., Geng, X., Bauer-Marschallinger, B., Guevara, M.A., Vargas, R., MacMillan, R.A., Batjes, N.H., Leenaars, J.G.B., Ribeiro, E., Wheeler, I., Mantel, S., Kempen, B., 2017. SoilGrids250m: Global gridded soil information based on machine learning. PLoS ONE 12. https://doi.org/10.1371/journal.pone.0169748

Heung, B., Ho, H.C., Zhang, J., Knudby, A., Bulmer, C.E., Schmidt, M.G., 2016. An overview and comparison of machine-learning techniques for classification purposes in digital soil mapping. Geoderma 265, 62–77. https://doi.org/10.1016/j.geoderma.2015.11.014

Hijmans, R.J., 2019. The raster package.

Höfle, B., 2014. Radiometric Correction of Terrestrial LiDAR Point Cloud Data for Individual Maize Plant Detection. IEEE Geosci. Remote Sens. Lett. 11, 94–98. https://doi.org/10.1109/LGRS.2013.2247022

Hunter, D.R., Lange, K., 2000. Quantile Regression via an MM Algorithm. J. Comput. Graph. Stat. 9, 60–77. https://doi.org/10.1080/10618600.2000.10474866

James, L.A., Watson, D.G., Hansen, W.F., 2007. Using LiDAR data to map gullies and headwater streams under forest canopy: South Carolina, USA. CATENA, Soil erosion and sediment transport under different land use/land cover scenarios 71, 132–144. https://doi.org/10.1016/j.catena.2006.10.010

Jenny, H., 1994. Factors of Soil Formation: A System of Quantitative Pedology. Courier Corporation.

Jenny, H., 1941. Factors of Soil Formation; A System of Quantitative Pedology. McGraw-Hill N. Y. 281.

Kempen, B., Brus, D.J., de Vries, F., 2015. Operationalizing digital soil mapping for nationwide updating of the 1:50,000 soil map of the Netherlands. Geoderma 241–242, 313–329. https://doi.org/10.1016/j.geoderma.2014.11.030

Kempen, B., Brus, D.J., Stoorvogel, J.J., Heuvelink, G.B.M., de Vries, F., 2012. Efficiency Comparison of Conventional and Digital Soil Mapping for Updating Soil Maps. Soil Sci. Soc. Am. J. 76, 2097–2115. https://doi.org/10.2136/sssaj2011.0424

Kerry, R., Oliver, M.A., 2011. Soil geomorphology: Identifying relations between the scale of spatial variation and soil processes using the variogram. Geomorphology 130, 40–54. https://doi.org/10.1016/j.geomorph.2010.10.002

Kleinman, K., Horton, N.J., 2015. Using R and RStudio for Data Management, Statistical Analysis, and Graphics, Second Edition. ed.

Koenig, K., Höfle, B., Hämmerle, M., Jarmer, T., Siegmann, B., Lilienthal, H., 2015. Comparative classification analysis of post-harvest growth detection from terrestrial LiDAR point clouds in precision agriculture. ISPRS J. Photogramm. Remote Sens. 104, 112–125. https://doi.org/10.1016/j.isprsjprs.2015.03.003

Koenker, R., 2019. quantreg: Quantile Regression.

Koenker, R., 2005. Quantile Regression. Cambridge University Press.

Koenker, R., 1984. A note on L-estimators for linear models. Statist. Prob. Lett. 323–5.

Koenker, R., Bassett, G., 1978. Regression Quantiles. Econometrica 46, 33–50. https://doi.org/10.2307/1913643

Koenker, R., Hallock, K.F., 2001. Quantile Regression. J. Econ. Perspect. 15, 143–156. https://doi.org/10.1257/jep.15.4.143

Kohavi, R., 2001. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection.

Kraus, K., Pfeifer, N., 1998. Determination of terrain models in wooded areas with airborne laser scanner data. ISPRS J. Photogramm. Remote Sens. 53, 193–203. https://doi.org/10.1016/S0924-2716(98)00009-4

Kristensen, T., Næsset, E., Ohlson, M., Bolstad, P.V., Kolka, R., 2015. Mapping Above- and Below-Ground Carbon Pools in Boreal Forests: The Case for Airborne Lidar. PLOS ONE 10, e0138450. https://doi.org/10.1371/journal.pone.0138450

Kuhn, M., 2019. The caret Package.

Kuhn, M., 2018. caret: Classification and Regression Training.

Lagacherie, P., 2008. Digital Soil Mapping: A State of the Art, in: Hartemink, A.E., McBratney, A., Mendonça-Santos, M. de L. (Eds.), Digital Soil Mapping with Limited Data. Springer Netherlands, Dordrecht, pp. 3–14. https://doi.org/10.1007/978-1-4020-8592-5_1

Lagacherie, P., Holmes, S., 1997. Addressing geographical data errors in a classification tree for soil unit prediction. Int. J. Geogr. Inf. Sci. 11, 183–198. https://doi.org/10.1080/136588197242455

Lewis, T., Carr, W., 1993. Hazard assessment keys for evaluating site sensitivity to soil-degrading processes–Interior sites. BC Min Vic. BC Land Manage Handb Field Guide Insert.

Li, C., Xu, Y., Liu, Z., Tao, S., Li, F., Fang, J., 2016. Estimation of Forest Topsoil Properties Using Airborne LiDAR-Derived Intensity and Topographic Factors. Remote Sens. 8, 561. https://doi.org/10.3390/rs8070561

Liaw, A., Wiener, M., 2014. Classification and Regression by RandomForest.

Lidberg, W., Nilsson, M., Ågren, A., 2020. Using machine learning to generate high-resolution wet area maps for planning forest management: A study in a boreal forest landscape. Ambio 49, 475–486. https://doi.org/10.1007/s13280-019-01196-9

Lim, K., Treitz, P., Wulder, M., St-Onge, B., Flood, M., 2003. LiDAR remote sensing of forest structure. Prog. Phys. Geogr. Earth Environ. 27, 88–106. https://doi.org/10.1191/0309133303pp360ra

Lindsay, J., 2015. Whitebox GAT | Help [WWW Document]. URL https://www.uoguelph.ca/~hydrogeo/Whitebox/getting_help.html (accessed 12.13.18).

Liu, X., 2008. Airborne LiDAR for DEM generation: some critical issues. Prog. Phys. Geogr. Earth Environ. 32, 31–49. https://doi.org/10.1177/0309133308089496

Lohr, U., 1998. Digital Elevation Models By Laser Scanning. Photogramm. Rec. 16, 105–109. https://doi.org/10.1111/0031-868X.00117

Lombardo, L., Saia, S., Schillaci, C., Mai, P.M., Huser, R., 2018. Modeling soil organic carbon with Quantile Regression: Dissecting predictors' effects on carbon stocks. Geoderma 318, 148–159. https://doi.org/10.1016/j.geoderma.2017.12.011

Lopez, P.L., Verkade, J.S., Weerts, A.H., Solomatine, D.P., 2014. Alternative configurations of Quantile Regression for estimating predictive uncertainty in water level forecasts for the Upper Severn River: a comparison. 22.

Ma, Y., Minasny, B., Wu, C., 2017. Mapping key soil properties to support agricultural production in Eastern China. Geoderma Reg. 10, 144–153. https://doi.org/10.1016/j.geodrs.2017.06.002

MacMillan, R.A., Jones, R.K., McNabb, D.H., 2004. Defining a hierarchy of spatial entities for environmental analysis and modeling using digital elevation models (DEMs). Comput. Environ. Urban Syst., GIS for Environmental Modeling 28, 175–200. https://doi.org/10.1016/S0198-9715(03)00019-X

Malone, B.P., de Gruijter, J.J., McBratney, A.B., Minasny, B., Brus, D.J., 2011. Using Additional Criteria for Measuring the Quality of Predictions and Their Uncertainties in a Digital Soil Mapping Framework. Soil Sci. Soc. Am. J. 75, 1032–1043. https://doi.org/10.2136/sssaj2010.0280

Malone, B.P., Minasny, B., McBratney, A.B., 2017. Using R for Digital Soil Mapping. Springer International Publishing Switzerland.

Mancini, M., Weindorf, D.C., Silva, S.H.G., Chakraborty, S., Teixeira, A.F. dos S., Guilherme, L.R.G., Curi, N., 2019. Parent material distribution mapping from tropical soils data via machine learning and portable X-ray fluorescence (pXRF) spectrometry in Brazil. Geoderma 354, 113885. https://doi.org/10.1016/j.geoderma.2019.113885

Mansuy, N., Thiffault, E., Paré, D., Bernier, P., Guindon, L., Villemaire, P., Poirier, V., Beaudoin, A., 2014. Digital mapping of soil properties in Canadian managed forests at 250m of resolution using the k-nearest neighbor method. Geoderma 235–236, 59–73. https://doi.org/10.1016/j.geoderma.2014.06.032

McArdle, S.S., Farrington, G., Rubinstein, I., 1999. A preliminary comparison of flood risk mapping using integrated remote sensing technology to aerial photography. In Proceedings, Fourth International Airborne Remote Sensing Conference and Exhibition. Ann Arbor MI ERIM Int. 616–23.

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

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

McBratney, A.B., Minasny, B., Cattle, S.R., Vervoort, R.W., 2002. From pedotransfer functions to soil inference systems. Geoderma 109, 41–73. https://doi.org/10.1016/S0016-7061(02)00139-8

McBratney, A.B., Minasny, B., Wheeler, I., Linden, B.P.M.& D. van der, 2012. Frameworks for digital soil assessment [WWW Document]. Digit. Soil Assess. Beyond. https://doi.org/10.1201/b12728-6

McGrew, J.C., Monroe, C.B., 2009. An Introduction to Statistical Problem Solving in Geography: Second Edition. Waveland Press.

Mckay, M.D., Beckman, R.J., Conover, W.J., 2000. A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output From a Computer Code. Technometrics 42, 55–61. https://doi.org/10.1080/00401706.2000.10485979

Meinshausen, N., 2006. Quantile Regression Forests. J. Mach. Learn. Res. 7, 983–999.

Merchant, M.A., Warren, R.K., Edwards, R., Kenyon, J.K., 2018. An Object-Based Assessment of Multi-Wavelength SAR, Optical Imagery and Topographical Datasets for Operational Wetland Mapping in Boreal Yukon, Canada: Canadian Journal of Remote Sensing: Vol 45, No 3-4 [WWW Document]. URL https://www-tandfonline-com.proxy.lib.sfu.ca/doi/full/10.1080/07038992.2019.1605500 (accessed 11.24.19).

Minasny, B., McBratney, A.B., 2008. Regression rules as a tool for predicting soil properties from infrared reflectance spectroscopy. Chemom. Intell. Lab. Syst. 94, 72–79. https://doi.org/10.1016/j.chemolab.2008.06.003

Minasny, B., McBratney, A.B., 2002. Uncertainty analysis for pedotransfer functions. Eur. J. Soil Sci. 53, 417–429. https://doi.org/10.1046/j.1365-2389.2002.00452.x

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

Morisada, K., Ono, K., Kanomata, H., 2004. Organic carbon stock in forest soils in Japan. Geoderma 119, 21–32. https://doi.org/10.1016/S0016-7061(03)00220-9

Mountrakis, G., Im, J., Ogole, C., 2011. Support vector machines in remote sensing: A review. ISPRS J. Photogramm. Remote Sens. 66, 247–259. https://doi.org/10.1016/j.isprsjprs.2010.11.001

Mueller, T.G., Pierce, F.J., 2003. Soil Carbon Maps. Soil Sci. Soc. Am. J. 67, 258–267. https://doi.org/10.2136/sssaj2003.2580

Muthusamy, M., Godiksen, P.N., Madsen, H., 2016. Comparison of Different Configurations of Quantile Regression in Estimating Predictive Hydrological Uncertainty. Procedia Eng., 12th International Conference on Hydroinformatics (HIC 2016) - Smart Water for the Future 154, 513–520. https://doi.org/10.1016/j.proeng.2016.07.546

Naghdi, R., Bagheri, I., Lotfalian, M., Setodeh, B., 2009. Rutting and soil displacement caused by 450C Timber Jack wheeled skidder (Asalem forest northern Iran). J. For. Sci. 55, 177–183. https://doi.org/10.17221/102/2008-JFS

Natural Resources Canada, 2013. Remote sensing in forestry [WWW Document]. URL https://www.nrcan.gc.ca/our-natural-resources/forests-forestry/sustainable-forest-management/measuring-reporting/remote-sensing-forestry/13429 (accessed 12.13.19).

Niemi, M.T., Vastaranta, M., Vauhkonen, J., Melkas, T., Holopainen, M., 2017. Airborne LiDAR-derived elevation data in terrain trafficability mapping. Scand. J. For. Res. 32, 762–773. https://doi.org/10.1080/02827581.2017.1296181

Nyland, R.D., 1992. Exploitation and greed in Eastern Hardwood Forests. J. For. 90, 33–37.

Odeha, I.O.A., McBratney, A.B., Chittleborough, D.J., 1994. Spatial prediction of soil properties from landform attributes derived from a digital elevation model. Geoderma 63, 197–214. https://doi.org/10.1016/0016-7061(94)90063-9

Padarian, J., Minasny, B., McBratney, A.B., 2017. Chile and the Chilean soil grid: A contribution to GlobalSoilMap. Geoderma Reg., Digital soil mapping across the globe 9, 17–28. https://doi.org/10.1016/j.geodrs.2016.12.001

Peng, C., 2000. Understanding the role of forest simulation models in sustainable forest management. Environ. Impact Assess. Rev. 20, 481–501. https://doi.org/10.1016/S0195-9255(99)00044-X

Pike, R.J., 1988. The geometric signature: Quantifying landslide-terrain types from digital elevation models. Math. Geol. 20, 491–511. https://doi.org/10.1007/BF00890333

Pouladi, N., Møller, A.B., Tabatabai, S., Greve, M.H., 2019. Mapping soil organic matter contents at field level with Cubist, Random Forest and kriging. Geoderma 342, 85–92. https://doi.org/10.1016/j.geoderma.2019.02.019

Quinlan, J.R., 2014. C4.5: Programs for Machine Learning. Elsevier.

Rahmati, O., Choubin, B., Fathabadi, A., Coulon, F., Soltani, E., Shahabi, H., Mollaefar, E., Tiefenbacher, J., Cipullo, S., Ahmad, B.B., Tien Bui, D., 2019. Predicting uncertainty of machine learning models for modelling nitrate pollution of groundwater using quantile regression and UNEEC methods. Sci. Total Environ. 688, 855–866. https://doi.org/10.1016/j.scitotenv.2019.06.320

Ratta, R., Lal, R., 1998. Soil Quality and Soil Erosion. CRC Press.

Roberta, P., 1948. Tree book: Learning to recognize trees of British Columbia (Technical Report) | ETDEWEB [WWW Document]. URL https://www.osti.gov/etdeweb/biblio/104035 (accessed 4.14.20).

Rossiter, D.G., 2018. Past, present & future of information technology in pedometrics. Geoderma 324, 131–137. https://doi.org/10.1016/j.geoderma.2018.03.009

Rousk, J., Bååth, E., Brookes, P.C., Lauber, C.L., Lozupone, C., Caporaso, J.G., Knight, R., Fierer, N., 2010. Soil bacterial and fungal communities across a pH gradient in an arable soil. ISME J. 4, 1340–1351. https://doi.org/10.1038/ismej.2010.58

Rudiyanto, Minasny, B., Setiawan, B.I., Saptomo, S.K., McBratney, A.B., 2018. Open digital mapping as a cost-effective method for mapping peat thickness and assessing the carbon stock of tropical peatlands. Geoderma 313, 25–40. https://doi.org/10.1016/j.geoderma.2017.10.018

Schneider, A., Hommel, G., Blettner, M., 2010. Linear Regression Analysis. Dtsch. Ärztebl. Int. 107, 776–782. https://doi.org/10.3238/arztebl.2010.0776

Schoenholtz, S.H., Miegroet, H.V., Burger, J.A., 2000. A review of chemical and physical properties as indicators of forest soil quality: challenges and opportunities. For. Ecol. Manag. 138, 335–356. https://doi.org/10.1016/S0378-1127(00)00423-0

Schreier, H., LOUGHEED, J., TUCKER, C., LECKIE, D., 1985. Automated measurements of terrain reflection and height variations using an airborne infrared laser system. Int. J. Remote Sens. 6, 101–113. https://doi.org/10.1080/01431168508948427

Schultz, T.P., Mwabu, G., 1998. Labor Unions and the Distribution of Wages and Employment in South Africa. ILR Rev. 51, 680–703. https://doi.org/10.1177/001979399805100407

Shi, X., Girod, L., Long, R., DeKett, R., Philippe, J., Burke, T., 2012a. A comparison of LiDAR-based DEMs and USGS-sourced DEMs in terrain analysis for knowledge-based digital soil mapping. Geoderma 170, 217–226. https://doi.org/10.1016/j.geoderma.2011.11.020

Shi, X., Girod, L., Long, R., DeKett, R., Philippe, J., Burke, T., 2012b. A comparison of LiDAR-based DEMs and USGS-sourced DEMs in terrain analysis for knowledge-based digital soil mapping. Geoderma 170, 217–226. https://doi.org/10.1016/j.geoderma.2011.11.020

Shrestha, D.L., Solomatine, D.P., 2006. Machine learning approaches for estimation of prediction interval for the model output. Neural Netw., Earth Sciences and Environmental Applications of Computational Intelligence 19, 225–235. https://doi.org/10.1016/j.neunet.2006.01.012

Solomatine, D.P., Shrestha, D.L., 2009. A novel method to estimate model uncertainty using machine learning techniques. Water Resour. Res. 45. https://doi.org/10.1029/2008WR006839

Sondheim, M., Suttie, K., 1983. User Manual for the British Columbia Soil Information System (BCSIS – Volume 1).

Sparks, D.L., 2012. Advances in Agronomy. Academic Press.

Stumpf, F., Schmidt, K., Goebes, P., Behrens, T., Schönbrodt-Stitt, S., Wadoux, A., Xiang, W., Scholten, T., 2017. Uncertainty-guided sampling to improve digital soil maps. CATENA 153, 30–38. https://doi.org/10.1016/j.catena.2017.01.033

Subburayalu, S.K., Slater, B.K., 2013. Soil Series Mapping By Knowledge Discovery from an Ohio County Soil Map. Soil Sci. Soc. Am. J. 77, 1254–1268. https://doi.org/10.2136/sssaj2012.0321

Szaro, R.C., Langor, D., Yapi, A.M., 2000. Sustainable forest management in the developing world: Science challenges and contributions. Landsc. Urban Plan. 47, 135–142. https://doi.org/10.1016/S0169-2046(99)00082-1

Szatmári, G., Pásztor, L., 2019. Comparison of various uncertainty modelling approaches based on geostatistics and machine learning algorithms. Geoderma 337, 1329–1340. https://doi.org/10.1016/j.geoderma.2018.09.008

Taghizadeh-Mehrjardi, R., Nabiollahi, K., Kerry, R., 2016. Digital mapping of soil organic carbon at multiple depths using different data mining techniques in Baneh region, Iran. Geoderma 266, 98–110. https://doi.org/10.1016/j.geoderma.2015.12.003

Tan, X., Chang, S.X., Kabzems, R., 2005. Effects of soil compaction and forest floor removal on soil microbial properties and N transformations in a boreal forest long-term soil productivity study. For. Ecol. Manag. 217, 158–170. https://doi.org/10.1016/j.foreco.2005.05.061

Thomas, G.W., 1996. Soil pH and soil acidity. https://doi.org/10.2136/sssabookser5.3.c16

Thomas, M., Clifford, D., Bartley, R., Philip, S., Brough, D., Gregory, L., Willis, R., Glover, M., 2015. Putting regional digital soil mapping into practice in Tropical Northern Australia. Geoderma 241–242, 145–157. https://doi.org/10.1016/j.geoderma.2014.11.016

Urbanová, M., Šnajdr, J., Baldrian, P., 2015. Composition of fungal and bacterial communities in forest litter and soil is largely determined by dominant trees. Soil Biol. Biochem. 84, 53–64. https://doi.org/10.1016/j.soilbio.2015.02.011

van den Driessche, R., 1984. Soil Fertility in Forest Nurseries, in: Duryea, M.L., Landis, T.D., Perry, C.R. (Eds.), Forestry Nursery Manual: Production of Bareroot Seedlings, Forestry Sciences. Springer Netherlands, Dordrecht, pp. 63–74. https://doi.org/10.1007/978-94-009-6110-4_7

Van Steenbergen, N., Ronsyn, J., Willems, P., 2012. A non-parametric data-based approach for probabilistic flood forecasting in support of uncertainty communication. Environ. Model. Softw. 33, 92–105. https://doi.org/10.1016/j.envsoft.2012.01.013

Vanclay, J.K., 1994. Modelling forest growth and yield: applications to mixed tropical forests. CAB International, Wallingford, U.K.

Vanclay, J.K., 1988. Precision in Modelling Native Forests 4.

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

Veronesi, F., Schillaci, C., 2019. Comparison between geostatistical and machine learning models as predictors of topsoil organic carbon with a focus on local uncertainty estimation. Ecol. Indic. 101, 1032–1044. https://doi.org/10.1016/j.ecolind.2019.02.026

Wainer, J., Cawley, G., 2018. Nested cross-validation when selecting classifiers is overzealous for most practical applications. ArXiv180909446 Cs Stat.

Wallis, A., Stokes, D., Wescott, G., McGEE, T., 1997. Certification and Labelling as a New Tool for Sustainable Forest Management. Aust. J. Environ. Manag. 4, 224–238. https://doi.org/10.1080/14486563.1997.10648386

Wang, Z., Bovik, A.C., 2009. Mean squared error: Love it or leave it? A new look at Signal Fidelity Measures. IEEE Signal Process. Mag. 26, 98–117. https://doi.org/10.1109/MSP.2008.930649

Wehr, A., Lohr, U., 1999. Airborne laser scanning—an introduction and overview. ISPRS J. Photogramm. Remote Sens. 54, 68–82. https://doi.org/10.1016/S0924-2716(99)00011-8

Weil, R.R., Brady, N.C., 2017. The Nature and Properties of Soils, Fourteen. ed. Pearson.

Wiesmeier, M., Barthold, F., Blank, B., Kögel-Knabner, I., 2011. Digital mapping of soil organic matter stocks using Random Forest modeling in a semi-arid steppe ecosystem. Plant Soil 340, 7–24. https://doi.org/10.1007/s11104-010-0425-z

Witten, I.H., Frank, E., Hall, M.A., Pal, C.J., 2016. Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann.

Wu, C., Wu, J., Luo, Y., Zhang, L., DeGloria, S.D., 2009. Spatial Prediction of Soil Organic Matter Content Using Cokriging with Remotely Sensed Data. Soil Sci. Soc. Am. J. 73, 1202–1208. https://doi.org/10.2136/sssaj2008.0045

Wulder, M.A., Hall, R.J., Coops, N.C., Franklin, S.E., 2004. High Spatial Resolution Remotely Sensed Data for Ecosystem Characterization. BioScience 54, 511. https://doi.org/10.1641/0006-3568(2004)054[0511:HSRRSD]2.0.CO;2

Wulder, M.A., White, J.C., Nelson, R.F., Næsset, E., Ørka, H.O., Coops, N.C., Hilker, T., Bater, C.W., Gobakken, T., 2012. Lidar sampling for large-area forest characterization: A review. Remote Sens. Environ. 121, 196–209. https://doi.org/10.1016/j.rse.2012.02.001

Yaalon, D.H., 1989. The Earliest Soil Maps and Their Logic. Bulletin of the International Society of Soil Science. Int. Soc. Soil Sci. Wagening. 24.

Yadav, S., Shukla, S., 2016. Analysis of k-Fold Cross-Validation over Hold-Out Validation on Colossal Datasets for Quality Classification, in: 2016 IEEE 6th International Conference on Advanced Computing (IACC). Presented at the 2016 IEEE 6th International Conference on Advanced Computing (IACC), pp. 78–83. https://doi.org/10.1109/IACC.2016.25

Yang, L., Jiao, Y., Fahmy, S., Zhu, A.-X., Hann, S., Burt, J.E., Qi, F., 2011. Updating Conventional Soil Maps through Digital Soil Mapping. Soil Sci. Soc. Am. J. 75, 1044–1053. https://doi.org/10.2136/sssaj2010.0002

Young, G., Fenger, M.A., Luttmerding, H.A., 1992. SOILS OF THE ASHCROFT MAP AREA 240.

YuSheng, Y., GuangShui, C., BaoLong, H., 2000. Variation in soil water and nutrients between different rotation stands of Chinese fir. J. Nanjing For. Univ. 24, 25–28.

Zamanian, K., Pustovoytov, K., Kuzyakov, Y., 2016. Pedogenic carbonates: Forms and formation processes. Earth-Sci. Rev. 157, 1–17. https://doi.org/10.1016/j.earscirev.2016.03.003

Zhao, Y.-C., Shi, X.-Z., 2010. Spatial Prediction and Uncertainty Assessment of Soil Organic Carbon in Hebei Province, China, in: Boettinger, J.L., Howell, D.W., Moore, A.C., Hartemink, A.E., Kienast-Brown, S. (Eds.), Digital Soil Mapping: Bridging Research, Environmental Application, and Operation, Progress in Soil Science. Springer Netherlands, Dordrecht, pp. 227–239. https://doi.org/10.1007/978-90-481-8863-5_19

Zhu, A.X., Hudson, B., Burt, J., Lubich, K., Simonson, D., 2001. Soil Mapping Using GIS, Expert Knowledge, and Fuzzy Logic. Soil Sci. Soc. Am. J. 65, 1463–1472. https://doi.org/10.2136/sssaj2001.6551463x

Zhu, G., Blumberg, D.G., 2002. Classification using ASTER data and SVM algorithms;: The case study of Beer Sheva, Israel. Remote Sens. Environ. 80, 233–240. https://doi.org/10.1016/S0034-4257(01)00305-4