Modelling of Exposed Bedrock and Soil Depth in the Critical Zone of Southern British Columbia

by

Christopher Frank Scarpone

MSA, Ryerson University 2012

B.A., Ryerson University, 2011

Thesis Submitted in Partial Fulfillment of the

Requirements for the Degree of

Master of Science

in the

Department of Geography

Faculty of Environment

SIMON FRASER UNIVERSITY

Fall 2015

Approval

Name:	**Christopher Frank Scarpone**
Degree:	**Master of Science (Geography)**
Title:	***Modelling of Exposed Bedrock and Soil Depth in the Critical Zone of Southern British Columbia***
Examining Committee:	**Chair:** Nick Hedley Associate Professor

Margaret Schmidt
Senior Supervisor
Associate Professor _____

Anders Knudby
Supervisor
Assistant Professor _____

Chuck Bulmer
Supervisor
Soil Scientist (PhD)
Provincial Government of British Columbia
Ministry of Forest Lands and Natural Resource Operations _____

Maja Krzic
External Examiner
Associate Professor
Department of Applied Biology
University of British Columbia _____

Date Defended/Approved: _____

Abstract

The Critical Zone (CZ) is the complex interaction of the hydrosphere, atmosphere, lithosphere, biosphere and pedosphere. It is in the CZ where most biological activity on earth can be found. At the centre of the CZ, the pedosphere is the medium in which all other regions of the CZ interact. The main objective of this study was to model two aspects of the CZ: the presence of exposed bedrock (EB) areas and the depth of the pedosphere (soil depth) in the Tulameen region of Southern British Columbia. Random Forest (RF) a classification tree method was used to predict the presence of EB. Prediction accuracy was found to be 88% with an independent validation dataset. The top three predictors of EB presence, which are a Landsat 7 PCA, Topographic Ruggedness Index (TRI), and a Normalized Difference Vegetation Index (NDVI) were further explored with modified partial dependence plots (PDPs) to determine the probability of EB presence. The depth of the pedosphere was predicted with a Generalized Linear Model (GLM), Random Forest (RF) and Residual Kriging (RK). Depth measurements came from the predicted EB layer which acted as a proxy for 0 m depth. In addition well water and soil pit information were used to define deeper depths for the region. GLM with RK was determined to produce the best model to measure depth, with an RMSE of 0.9 m in the 0 to 2 m range for depth measurements. EB proved to be a reliable and efficient proxy in addition to conventional soil depth measurements which are time consuming and costly to generate. The obtained results indicate that GLM with RK and the use of EB layers can aid in further studies of the CZ.

Keywords: Generalized Linear Models, Random Forest, Residual Kriging, Critical Zone, Exposed Bedrock, Soil Depth

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. Anders Knudby for allowing me to drop in on him with many half-hazard ideas, but he always managed to turn me in the right direction. To Dr. Chuck Bulmer for the many long encouraging conversations to continue pursuing my degree with confidence.

Also to the Soil Lab group at SFU. Brandon Heung for his guidance in the early stages of my studies. To Maciej Jamrozik for teaching me the fundamentals of soils as my T.A and to Jin Zhang through all her help as my field assistant.

I would also like to thank my parents, Patrick and Suzie Scarpone. Even though they are 4300km away they were also there for me, no matter what I needed. And lastly, my partner in crime, Kelly Baldwin. She always gave me the support and courage I needed to keep going when things got tough and allowed me to enjoy them when they weren't.

Table of Contents

List of Tables

List of Figures

List of Acronyms

ASTER	Advanced Spaceborne Thermal Emission and Reflection Radiometer
cLHS	Condition Latin Hypercube
CZ	Critical Zone
DEM	Digital Elevation Model
DSM	Digital Soil Mapping
EB	Exposed Bedrock
GAM	General Additive Model
GLM	Generalized Linear Model
GLMRK	Generalized Linear Model Residual Kriging
MDA	Mean Decrease in Accuracy
MRVBF	Multi-Resolution Valley Bottom Flatness
NDVI	Normalized Difference Vegetation Index
OLT	Other Land Types
OOB	Out of Bag (error)
PCA	Principle Component Analysis
PDP	Partial Dependence Plot
PEM	Predictive Ecosystem Mapping
RF	Random Forest
RFRK	Random Forest Residual Kriging
RK	Residual Kriging
RMSE	Root Mean Square Error
TRI	Topographic Ruggedness Index
VRI	Vegetation Resource Index
WW	Well Water

Chapter 1.

Introduction

1.1 Context of Study

The Critical Zone (CZ) is the complex environment which is made up of the lithosphere, hydrosphere, atmosphere, biosphere, and pedosphere (NRC, 2001). As demand for natural resources increases, a better understanding of the influence of humans on the CZ is needed (Kriebel et al., 2001). The pedosphere is the region that resides at the centre of the four other spheres (Lin, 2010) and it is within the upper layers of the pedosphere that most of terrestrial life exists. Water storage, plant growth, and the production and sequestration of soluble nitrogen and carbon are all maintained in this zone (Lin, 2005; Richter & Mobley, 2009). Quantifying the depth of the pedosphere will aid in the understanding of large scale processes for nutrient and chemical exchange (Brantley et al., 2007; Chorover et al., 2007), water storage, water yields (Field et al., 2015; Yu et al., 2015; Yu et al., 2014) weather, erosion, and soil production (Heimsath et al., 2012; Jin et al., 2010; Ma et al., 2010; Moraetis et al., 2014). However, few studies have attempted to map the depth of the pedosphere on a landscape scale as the quantity and quality of data required is time consuming to obtain. In this study I have used both deterministic and combined geostatistical methods of Generalized Linear Model (GLM), Random Forest (RF), GLM with Residual Kriging (GLMRK) and RF with Residual Kriging (RFRK) to map both exposed bedrock and the depth of the pedosphere.

1.2 Theoretical Background and Methods

Mapping of soil properties has evolved from using simple linear relationships of soil properties and topography, to complex multivariate relationships between climate, organisms, topography, parent material and time. This evolution was driven by the increased access to sophisticated digital data and modeling methods, which allows for more soil forming factors at multiple scales to be utilized (McBratney et al., 2003). Digital soil mapping (DSM) has increasingly replaced conventional methods of soil survey. As a whole, soil mapping is becoming reliant on remotely acquired data sources (Mulder et al., 2011).

DSM is the process of representing soil properties and classes in a quantitative format (Schmidt et al., 2008). Over the last fifty years, mapping of soil properties and classes has seen drastic changes and innovations. These changes are due to improvements in computational power, aggregation of multiple data sources (remotely sensed, both raster and vector based), and the integration of more robust statistical packages (Dobos et al., 2006). DSM is a departure from the qualitative aspect of soil mapping, which was described by Hans Jenny in 1941 with his CLORPT model (McBratney et al., 2003). CLORPT states that soil is a function of climate, organisms, topography, parent material and time. The SCORPAN framework (McBratney et al., 2003) (shown below) incorporates the fundamentals of Jenny's CLORPT model, but states that soil is a function of: soil properties of an area (s); climate (c); organism (o); relief (r); parent material (p); age (a), and; space (n).

$$S = f\,(s, c, o, r, p, a, n)$$

McBratney's additions to the CLORPT model have allowed for the integration of multiple data sources and new methods of DSM. In this study, SCORPAN has

been used as a framework to map locations of exposed bedrock, as well as the depth of the pedosphere in the CZ. Specific modeling methods used include GLM, RF, GLMRK and RFRK.

1.2.1 Residual Kriging

Kriging is a stochastic spatial interpolation method of observed Z values from the surrounding area weighted by spatial variance from other values (Burgess & Webster, 1980) There are multiple forms of kriging, but all have the common strength of predicting variables without bias, minimizing variance and the ability to account for prediction error (Burgess & Webster, 1980).

A major component of kriging is the user's ability to visually identify spatial covariance through the variogram. Spatial covariance is the similarity amongst values in a landscape, influenced by their proximity in space (Webster & Oliver, 2007). The general direction (or trend) of the covariance can be displayed through the variogram and different semi-variogram models (spherical, gaussian and exponential) can be applied to represent the distribution. Each semi variogram model attempts to reduce the local influence of the landscape on each value in the distribution, allowing the trend to be spatially independent (Johnston et al., 2001). It is through the variogram that uncertainty can be accounted for as it relates to the spatial covariance (Odeh et al., 1995; Odeh et al., 1994).

Kriging has had many uses in soil science where early scientists attempted to map soil attributes in areas lacking robust sampling schemes. These scientists typically used a type of kriging called ordinary kriging which, had been used in mining studies to assess heavy metal concentrations in an area (Atteia et al., 1994; Von Steiger et al., 1996). Studies then began to assess the spatial uncertainty that was associated with mapping soil properties and how different kriging methods could be explored to adjust for this uncertainty (Goovaerts 1999, 2001). The ability to

assess uncertainty was seen as a necessity when soil mapping began to introduce more ancillary information that could account for local variation. New combination methods of kriging with deterministic models were also being proposed to find new ways to account for uncertainty.

Residual kriging (RK) is a combination method that allows for deterministic models to account for the relations amongst multiple independent variables and the dependent variable. With RK, deterministic models such as regression models are used to predict a layer and the residuals from that model are kriged. The residual layer and the regression layer are then combined to create a new layer (Hengl et al., 2004). It is through this process that spatial variance is accounted for and removed from the predicted layer (Goovaerts, 2001).

There were initially 3 models for residual kriging that were proposed by Odeh et al. (1994):

> Model A: A regression model was produced for the target variable, followed by an OK of the regressed values that were combined. Variance was replaced by the diagonal error terms to represent uncertainty. This method was called “Kriging with Uncertain Data”.
>
> Model B: A regression model was produced for the target variable and residuals were extracted from that model. Regressed values were then kriged, as were the residual values. Both layers were then combined to account for target variance.
>
> Model C: A regression model was produced of the target variable and a layer of the target variables was produced from the regression model. The residual values were then kriged to account for the model uncertainty.

Model C is considered the best method for residual kriging as it utilizes more of the local information from the independent variables, allowing for the residuals to remain as the main source of uncertainty in the model (Hengl et al., 2007; Odeh et al., 1994; Odeh et al., 1995).

Residual kriging has allowed for the introduction of more complex modelling approaches that can account for more variability from a wider source of data inputs. Such models include GLM and RF.

1.2.2 Generalized Linear Model (GLM)

Generalized linear model (GLM) is an extension of the common linear model, where data do not need to be normally distributed (Nelder & Wedderburn, 1972). To accommodate for the assumptions of classical models where functions are non-normal, a link function is added to establish this connection (Equation 2).

$$y = g(x_i B) + \varepsilon_i$$

(Equation 2)

The GLM follows the same equations as a linear model; where x_i is all the measured independent variables with estimated regression weights incorporated, B is the slope intercept and ε_i is the error term. However the g variable is the inclusion of the link function (Lane, 2002)

The link function establishes a connection between the predictor and the mean of the dependant distribution (Venables & Dichmont, 2004). The link is considered non-linear and allows the mean of the response to be a scale on which the effects of the model can be combined additively.

The GLM is often used for modelling species and habitat distribution (Guisan et al., 1999; Guisan & Zimmermann, 2000; Miller & Franklin, 2006; Venables & Dichmont, 2004) because these studies typically have data that do not fall under the normal distribution. The GLM offers a way through the link function to adequately model data that does not follow the normal distribution. GLM has also been used in soil mapping especially in conjunction with kriging methods. Hengl et al. (2007) used a GLM to describe a standard residual kriging approach. Steinnes et al. (2014) built on Hengl's approach and mapped the spatio-temporal change of heavy medal contamination in Norway. Due to the wide range of contaminants, multiple distributions were discovered in their data and GLM provided one method to map them all. It also allowed for the use of discrete variables, such as land use types to be included as a logit, which aided in further understanding the correlations of heavy metal contamination.

1.2.3 Random Forest

Random Forest is a machine learning technique that uses calibration data to predict categorical or continuous data through classification trees (Breiman, 2001). Trees are grown from a bootstrap selection of calibration data, and each split (node) is determined by the total amount of variables in the calibration data. Trees are grown to their maximum extent to reduce bias and increase the maximum amount of variance in the model. To determine accuracy of a prediction, an internal validation method called the "Out of Bag" (OOB) error is calculated. A subsample of 33% (the out of bag data) of the calibration data is removed at the beginning of the model, leaving 66% of the calibration data to be used for further model predictions. The out of bag data is then compared to the predictions of the Random Forest (Diaz-Uriarte & Alvarez de Andres, 2005). The OOB sample is then compared to the generated classification trees, where a variable importance measure, similar to CART, is created. Variable importance measures the influence that each variable in the calibration data had on the final prediction. Variable

importance is primarily measured through one methods which is the mean decrease in accuracy (MDA). The MDA compares the OOB data to the final output by iteratively removing the variable's information content removing variables. The variables information that contribute the most error from their removal are considered the most important (Liaw & Wiener, 2003).

Random Forest was first noted as a powerful classification tool in an ecological setting to predict the presence of rare lichen species (Cutler et al., 2007). RF was powerful, due to its ability to utilize noisy data, its ability to rank variable importance and its ability to internally validate its results (Cutler et al., 2007). RF has been widely used in DSM. Heung et al. (2014) used the Random Forest R package to map parent material in a regional setting. The authors found that Random Forest did not need optimization compared to the default values. This is a beneficial finding as it reduces the need for user optimization which can be time consuming and often difficult. RF also demonstrated its ability to handle landscape scale trends as parent material was successfully mapped for the lower mainland in British Columbia.

Random Forest has also recently been explored as a mixed method for residual kriging, which can lead to greater improvements in current mapping standards. Guo et al. (2015) explored mapping soil organic matter using RFRK and found that it significantly out-performed a stepwise linear regression and a RF in regression mode. Hengl et al. (2015) used RFRK to map soil nutrient contents for the continent of Africa. Mapping was done at a 250 m resolution, with minimal data.

1.2.4 Partial Dependence Plots (PDPs)

Random Forest is often considered a "black box" where the interactions of the model are hidden from the user (Breiman, 2001). With every decision that the RF makes, the user is unable to assess the decisions that the classifier has made. In

soil science, black box approaches are becoming more common (through neural networks and other machine learning techniques) and often hide valuable information that is found in the feature selection of these models (Schmidt et al., 2008). This can be especially problematic with new covariates being produced (Tesfa, et al., 2009) accompanied by higher resolution imagery which have yet to be fully understood (Behrens et al., 2010; Samuel-rosa et al., 2015; Wiesmeier et al., 2010). Variable importance plots are used to determine specifically how each variable influences the final prediction (Hengl et al., 2015; Pahlavan Rad et al., 2014; Tesfa et al., 2009). This is integral for comparing relative importance of the variables in the model, but it does not specify which aspect of a particular variable is important such as elevation ranges or NDVI measurements. To further investigate how decisions are made in RF, it is possible to further analyze the individual classification trees that are created (Guo et al., 2015; Veronesi & Hurni, 2014; Wiesmeier et al., 2010). The classification trees can allow the user to determine values of a variable where the breaks were made. However, this method is often cumbersome to visually assess as multiple segments or branches are derived from one variable. It also does not specify which tree segment influenced the final decision and by how much.

Partial dependence plots (PDPs) are features that are incorporated into RF but are underutilized despite the fact they can aid in further interpretations of the model. PDPs measure the relative dependence (influence) that a set of independent variables has on a dependent variable (Hastie et al., 2009). PDPs are measured plots of a specified variable and the relative probability that a given predictor value will generate a presence response in the (classification) model. This presence response has been influenced by that variable and all other variables in the model (Hastie et al., 2009). Cutler et al., (2007) first described the use of PDFs with RF while predicting species distribution of rare lichens. His usage of PDPs was primarily qualitative as the scales for PDPs were measured in log-odds, which

range from -∞ to +∞, and still left some ambiguity in their interpretations. However this scale can be converted to a probability scale (Hastie et al., 2009) which can give quantitative results for a probability of classification with variables and values. This has yet to be fully explored, but it could allow for a more critical understanding of which variables are influencing the model (the MDA plots) and how each variable is actually influencing prediction.

1.2.5 Conditioned Latin Hypercube Sampling

There are two major sampling methods, which are commonly employed in soil science. The first sampling scheme is a random sampling scheme, where points are randomly selected from a landscape without any external interference. This method is considered lacking in bias, but can often oversample areas and features leaving large amounts of variability unsampled (Webster & Oliver, 2007). The other sampling method is stratified sampling, where a specified pattern is constructed from where sampling points are derived. This method often allows for easier acquisition of designated samples, but it introduces sampling bias according to the specific design of the sampling (Webster & Oliver, 2007). With study areas that are easy to access or with areas that may not be too spatially diverse, either method is suitable. In highly variable landscapes that are difficult to sample, the majority of the variance can be severely undersampled with random sampling. When there is significant undersampling, a large margin of error is introduced in the sampling points compared to the entire landscapes. In order to address these issues for remote access and spatial diversity, Minasny & McBratney (2006) have developed a new sampling method called a conditioned latin hyper cube (cLHS).

The cLHS is a modification of the latin hypercube sampling technique (LHS) that was initially proposed by McKay et al. (1979). The LHS is a stratified random sampling technique that can sample across multivariate distributions of both discrete and continuous data. Equally probable strata are created for each

covariate, where the number of samples defined by the user determines the number of strata. One sample is randomly chosen from within each strata. Minasny & McBratney, (2006) discovered that the LHS, when applied to geographical space, would often find samples that were not represented in the real world. The conditioned modification forced sample points to be located in the data feature space.

The cLHS has already been explored as a beneficial sampling technique. Brungard & Boettinger, (2010) found that with a cLHS they were able to accurately represent a 40 km^2 landscape, using 6 covariates, with only 100 points. cLHS was also found to better represent the distribution of heavy metal contamination in soil when compared to a random sampling scheme (Chu et al., 2010). cLHS as a sampling method holds great promise for landscape scale studies because it reduces the total amount of necessary points to be acquired for a large area. It also ensures that the largest diversity in the distribution will be sampled.

1.3 Research Rationale and Objectives

Further research into the CZ has been noted as one of the most important areas of study to allow people to continue to live on Earth (NRC, 2001). A model and methodology for landscape scale evaluation of the depth of the pedosphere is greatly needed as a component of CZ research. To date, the majority of soil depth mapping for the CZ has been conducted on a local watershed scale (Kuriakose, Devkota, Rossiter, & Jetten, 2009; Pelletier & Rasmussen, 2009) or hillslope scale (Heimsath et al., 2001; Heimsath et al., 1999). These studies have focused on steady state prediction of soil depth through a process-based model; however, an assumed starting depth is applied to all of these models, which could contain large margins of error (Roering et al., 2001). Also, local scale processes of slope and convexity have been increasingly explored to predict soil depth (DiBiase et al., 2012; Heimsath et al., 1997); however, landscape scale processes of erosion and

deposition are not always reflected in local scale models (Brardinoni et al., 2009; Brardinoni & Hassan, 2006).

Data sources such as information on exposed bedrock need to be assessed for landscape scale soil depth predictions. When bedrock is exposed, there can be a significant decrease in hydrological and biological weathering of bedrock, reducing total sediment created (Anderson et al., 2002). These areas greatly affect the denudation rates of the landscape, which affects total available sediment in a landscape. Areas with prominent bedrock exposures indicate that local erosion exceeds soil production and supply, reducing the total amount of steady state soil (DiBiase et al., 2012). The location, lithology, and characteristics of these exposed rocks can give better estimates of soil production caused by physical weathering (Moore et al., 2009).

The location of exposed bedrock (EB) can be used in deterministic/stochastic modeling as a relative depth measurement. Karlsson et al., (2014) used exposed bedrock and well water depth to bedrock measurements to create a local landscape scale depth map. The wells were used as a depth measurement and each well's proximity from the bottom of the wells to the nearest bedrock outcrop were used to quantify depth between the well and the bedrock exposure. Kuriakose et al., (2009) also used EB points as a proxy in their model to determine soil depth. They did not include EB as its own data source, but only incorporated it if it was one of their random sample points. It was found that bedrock increased the overall sampling density required for more robust modelling techniques.

RF, as a classification method, was used in my study to better assess exposed bedrock in the landscape along with legacy land data, topographic data, and vegetation inventories. RF has proven to be a valuable tool in predicting soil properties and assessing variable importance (Heung et al., 2014).

The objectives of phase one of this study were to:

1. Accurately predict the occurrence of bedrock exposure in the landscape,
2. Determine which variables had the greatest importance in predicting the, occurrence of EB
3. Investigate how these variables influenced predictions.

Once EB was mapped, the location of the EB was used to assist in mapping the depth of the pedosphere in the CZ. To map the depth of the pedosphere, I used RF as a regression model and compared the results to those using GLM, GLMRK and RFRK. EB, well water data and in-situ field measurements were used to represent the wide range of soil depths found in the landscape. The objectives of phase two were to:

1. Map the depth of the pedosphere in the CZ on a landscape scale,
2. Compare and determine the best modelling approach to map the pedosphere depth in Southern British Columbia,
3. To demonstrate the predictive power that EB can add to soil depth models.

1.4 Study Area

The study area is approximately 3435 km^2 and is located in Southern British Columbia. It is centered on the town of Tulameen, British Columbia (N 49°32'45" W 120°45'36"; see Figure 1-1). Tulameen has four biogeoclimatic zones that fall within its boundaries: The Coastal Western Hemlock (CWH); Mountain Hemlock (MH); Englemann Spruce (ES); and; Subalpine Fir and Alpine Tundra. The extensive precipitation (510 mm annually) is due to the orographic effect of westerly winds flowing from the Fraser Valley. Soils in this region are primarily Humo-Ferric Podzols, which are of coarse texture and well drained. The geology is a mix of intrusive igneous rock and folded and faulted volcanic sedimentary rocks. Elevations range from sea level to, 2042 m in height. Geomorphic processes

in the region are glacial in origin with a wide range of moraines, fluvial deposits, colluvial materials and mass wasting (BC Ministry of Parks, 1999).

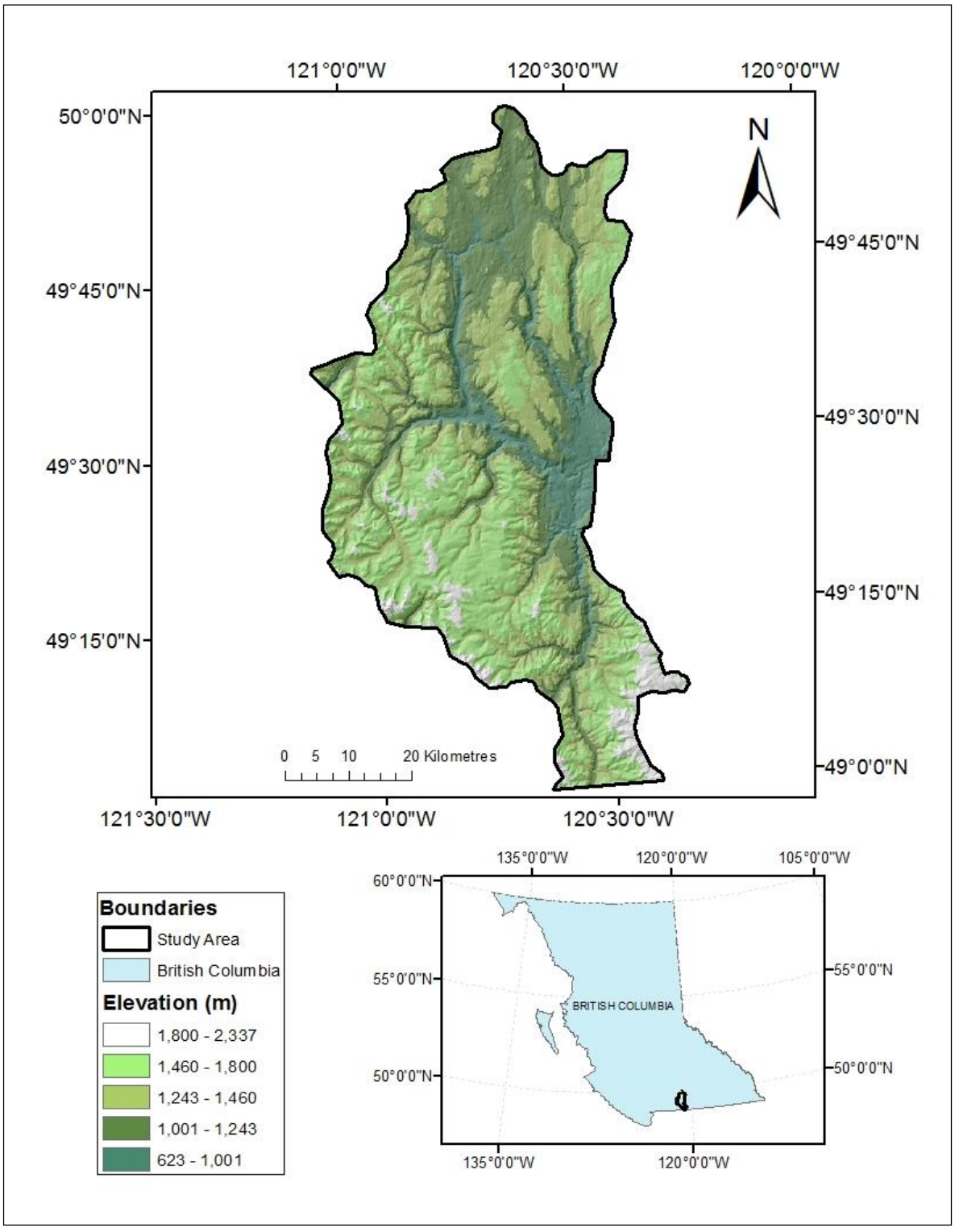


Figure 1-1 Tulameen study area.

1.5 Thesis Structure

The thesis is divided into four chapters, starting with this introduction, which has described context, current literature on the CZ, modelling methods, and the study area.

Chapter Two is a paper focused on accurately mapping EB at a 100 m resolution. RF was used to predict exposed bedrock occurrence with a combination of legacy land cover maps, topography, geological and vegetation indices. A variable reduction method was used to reduce the total amount of inputs needed, without affecting prediction accuracy. Partial dependence plots were then generated to assess the influence that each variable had on prediction rates for exposed bedrock. A comparison was made between the calibration data and the RF prediction rates (as seen in the PDPs) to determine the individual effect calibration data and each variable had on the final predictions of bedrock.

Chapter Three is a paper concerning the modelling of the pedosphere of the CZ (soil depth) using primarily the generated exposed bedrock layer that was created in chapter two. EB was treated as a 0 m depth measurement and was sampled at 100 and 5000 points and used with RF in regression mode, GLMRK, and RFRK. Well water and manually collected soil depth measurements, sampled from a cLHS method were also included with the calibration data. These calibration data were used to predict a continuous soil depth layer on the landscape scale for the Tulameen study area. Model effectiveness was assessed through the semi-variograms derived from the kriging procedures to access the removal of local trends. A validation dataset derived from a random subsample of exposed bedrock, well water and cLHS points were used to validate the predicted soil maps. These validations were presented with RMSE values.

Chapter Four concludes the thesis with a brief summary of results from chapters two and three; followed by a discussion of potential future research, and thesis contributions.

References

Anderson, S. P., Dietrich, W. E., & Brimhall, G. H. (2002). Weathering profiles, mass-balance analysis, and rates of solute loss: Linkages between weathering and erosion in a small, steep catchment. *Bulletin of the Geological Society of America*, *114*(9), 1143–1158.

Atteia, O., Dubois, J. P., & Webster, R. (1994). Geostatistical analysis of soil contamination in the Swiss Jura. *Environmental Pollution*, *86*(3), 315–327. http://doi.org/10.1016/0269-7491(94)90172-4

Behrens, T., Zhu, A.-X., Schmidt, K., & Scholten, T. (2010). Multi-scale digital terrain analysis and feature selection for digital soil mapping. *Geoderma*, *155*(3-4), 175–185. http://doi.org/10.1016/j.geoderma.2009.07.010

Brardinoni, F., & Hassan, M. A. (2006). Glacial erosion, evolution of river long profiles, and the organization of process domains in mountain drainage basins of coastal British Columbia. *Journal of Geophysical Research: Earth Surface*, *111*(1), 1–12. http://doi.org/10.1029/2005JF000358

Brantley, S. L., Goldhaber, M. B., & Vala Ragnarsdottir, K. (2007). Crossing disciplines and scales to understand the critical zone. *Elements*, *3*(5), 307–314. http://doi.org/10.2113/gselements.3.5.307

Brardinoni, F., Hassan, M. A., Rollerson, T., & Maynard, D. (2009). Colluvial sediment dynamics in mountain drainage basins. *Earth and Planetary Science Letters*, *284*(3-4), 310–319. http://doi.org/10.1016/j.epsl.2009.05.002

Breiman, L. (2001). Random forests. *Machine Learning*, 5–32.

Brungard, C. W., & Boettinger, J. L. (2010). Conditioned Latin Hypercube Sampling: Optimale Sample Size for Digitial Soil Mapping of Arid Rangelands in Utah, USA. In J. L. Boettinger, D. W. Howell, A. C. Moore, A. E. Hartemink, & S. Kienast-Brown (Eds.), *Progress in Soils Science 2.* Dordrecht: Springer Netherlands. http://doi.org/10.1007/978-90-481-8863-5

Burgess, T. M., & Webster, R. (1980). Optimal interpolation and isarithmic mapping of soil properties. I. The semi-variogram and punctual kriging. *Journal of Soil and Science*, *31*(2), 315–331, refs. http://doi.org/10.1111/j.1365-2389.1980.tb02084.

Chorover, J., Kretzschmar, R., Garica-Pichel, F., & Sparks, D. L. (2007). Soil biogeochemicial processes within the critical zone. *Elements*, *3*(5), 321–326. http://doi.org/10.2113/gselements.3.5.321

Chu, H.-J., Lin, Y.-P., Jang, C.-S., & Chang, T.-K. (2010). Delineating the hazard zone of multiple soil pollutants by multivariate indicator kriging and conditioned Latin hypercube sampling. *Geoderma*, *158*(3-4), 242–251. http://doi.org/10.1016/j.geoderma.2010.05.003

Cutler, D. R., Edwards, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., & Lawler, J. J. (2007). Random forests for classification in ecology. *Ecology*, *88*(11), 2783–92.

Diaz-Uriarte, R., & Alvarez de Andres, S. (2005). Variable selection from random forests : application to gene expression data, 1–11. Retrieved from http://arxiv.org/abs/q-bio.QM/0503025

DiBiase, R. A., Heimsath, A. M., & Whipple, K. X. (2012). Hillslope response to tectonic forcing in threshold landscapes. *Earth Surface Processes and Landforms*, *37*(8), 855–865.

Dobos, E., Carré, F., Hengl, T., Reuter, H.I., Tóth, G., 2006. Digital Soil Mapping as a support to production of functional maps. EUR 22123 EN, 68 pp. Office for Official Publications of the European Communities, Luxemburg.

Field, J. P., Breshears, D. D., Law, D. J., Villegas, J. C., López-hoffman, L., Brooks, P. D., Troch, P. A. (2015). Critical Zone Services : Expanding Context , Constraints , and Currency beyond Ecosystem Services. *Vadose Zone Journal*. http://doi.org/10.2136/vzj2014.10.0142

Goovaerts, P. (1999). Geostatistics in soil science: State-of-the-art and perspectives. *Geoderma*, *89*(1-2), 1–45. http://doi.org/10.1016/S0016-7061(98)00078-0

Goovaerts, P. (2001). Geostatistical modelling of uncertainty in soil science. *Geoderma*, *103*(1-2), 3–26. http://doi.org/10.1016/S0016-7061(01)00067-2

Guisan, A., Weiss, S. B., & Weiss, A. D. (1999). GLM versus CCA spatial modeling of plant species distribution. *Plant Ecology*, (143), 107–122.

Guisan, A., & Zimmermann, N. E. (2000). Predictive habitat distribution models in ecology. *Ecological Modelling*, *135*(2-3), 147–186. http://doi.org/10.1016/S0304-3800(00)00354-9

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

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. *Springer* (Vol. 27). http://doi.org/10.1007/b94608

Heimsath, A M., Dietrich, W. E., Nishiizumi, K., & Finkel, R. C. (2001). A: Stochastic processes of soil production and transport: erosion rates, topographic variation and cosmogenic nuclides. *In the Oregon Coast Range. Earth Surface Processes and Landforms*, *26*, 531–52.

Heimsath, A. M., DiBiase, R. a., & Whipple, K. X. (2012). Soil production limits and the transition to bedrock-dominated landscapes. *Nature Geoscience*, *5*(3), 210–214. http://doi.org/10.1038/ngeo1380

Heimsath, A. M., Dietrich, W. E., Nishiizumi, K., & Finkel, R. C. (1999). Cosmogenic nuclides, topography, and the spatial variation of soil depth. *Geomorphology*, *27*, 151–172.

Heimsath, A. M., Dietrich, W. E., Nishiizumi, K., Finkel, R. C., Mass, A., & National, L. L. (1997). The soil production function and landscape equilibrium, *388*(July), 358–361.

Hengl, T., Heuvelink, G. B. M., Kempen, B., Leenaars, J. G. B., Walsh, M. G., Shepherd, K. D.,Tondoh, J. E. (2015). Mapping Soil Properties of Africa at 250 m Resolution: Random Forests Significantly Improve Current Predictions. *PloS One*, *10*(6), e0125814. http://doi.org/10.1371/journal.pone.0125814

Hengl, T., Heuvelink, G. B. M., & Rossiter, D. G. (2007). About regression-kriging: From equations to case studies. *Computers & Geosciences*, *33*(10), 1301–1315. http://doi.org/10.1016/j.cageo.2007.05.001

Hengl, T., Heuvelink, G. B. M., & Stein, A. (2004). A generic framework for spatial prediction of soil variables based on regression-kriging. *Geoderma*, *120*(1-2), 75–93. http://doi.org/10.1016/j.geoderma.2003.08.018

Heung, B., Bakker, L., Schmidt, M. G., & Dragićević, S. (2013). Modelling the dynamics of soil redistribution induced by sheet erosion using the Universal Soil Loss Equation and cellular automata. *Geoderma*, *202-203*, 112–125. http://doi.org/10.1016/j.geoderma.2013.03.019

Heung, B., Bulmer, C. E., & Schmidt, M. G. (2014). Predictive soil parent material mapping at a regional-scale: A Random Forest approach. *Geoderma*, *214-215*, 141–154. http://doi.org/10.1016/j.geoderma.2013.09.016

Jin, L., Ravella, R., Ketchum, B., Bierman, P. R., Heaney, P., White, T., & Brantley, S. L. (2010). Mineral weathering and elemental transport during hillslope evolution at the Susquehanna/Shale Hills Critical Zone Observatory. *Geochimica et Cosmochimica Acta*, *74*(13), 3669–3691. http://doi.org/10.1016/j.gca.2010.03.036

Johnston, K., Hoef, J. Ver, Krivoruchko, K., & Lucas, N. (2001). *Using ArcGIS geostatistical analyst*. Retrieved from http://direitosminerarios.com/pdf/ESRI - Using ArcGIS Geostatistical Analyst.pdf

Karlsson, C. S. J., Jamali, I. A., Earon, R., Olofsson, B., & Mörtberg, U. (2014). Comparison of methods for predicting regolith thickness in previously glaciated terrain, Stockholm, Sweden. *Geoderma*, *226-227*, 116–129. http://doi.org/10.1016/j.geoderma.2014.03.003

Kriebel, D., Tickner, J., Epstein, P., Lemons, J., Levins, R., Loechler, E. L., Stoto, M. (2001). The precautionary principle in environmental science. *Environmental Health Perspectives*, *109*(9), 871–876. http://doi.org/10.2307/3454986

Kuriakose, S. L., Devkota, S., Rossiter, D. G., & Jetten, V. G. (2009). Prediction of soil depth using environmental variables in an anthropogenic landscape, a case study in the Western Ghats of Kerala, India. *Catena*, *79*(1), 27–38. http://doi.org/10.1016/j.catena.2009.05.005

Lane, P. (2002). Generalized linear models in soil science. *European Journal of Soil Science*, *53*(June), 241–251. http://doi.org/10.1046/j.1365-2389.2002.00440.x

Liaw, A., & Wiener, M. (2003). Package “randomForest.” *Retrieved December*. Retrieved from http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Package+“+randomForest+”#5

Lin, H. (2005) letter to the Editor on "From the Earth's Critical Zone to Mars Exploration: Can soil Science Enter Its Golden Age?". *Soil Science Society of America, 69*, 1351-1358

Lin, H. (2010) Earth's Cricial Zone and hydropedology: Concepts, Characteristics and advances. *Hydrology and Earth System Sciences,* 14(1), 25-45.

Ma, L., Chabaux, F., Pelt, E., Blaes, E., Jin, L., & Brantley, S. (2010). Regolith production rates calculated with uranium-series isotopes at Susquehanna/Shale Hills Critical Zone Observatory. *Earth and Planetary Science Letters*, *297*(1-2), 211–225. http://doi.org/10.1016/j.epsl.2010.06.022

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

Miller, J., & Franklin, J. (2006). Explicitly incorporating spatial dependence in predictive vegetation models in the form of explanatory variables: a Mojave Desert case study. *Journal of Geographical Systems*, *8*(4), 411–435. http://doi.org/10.1007/s10109-006-0035-8

Minasny, B., & McBratney, A. B. (2006). A conditioned Latin hypercube method for sampling in the presence of ancillary information. *Computers & Geosciences*, *32*(9), 1378–1388. http://doi.org/10.1016/j.cageo.2005.12.009

Ministry of Parks. (1999). *Coquihalla Recreation Summit Master Plan*.

Moore, J. R., Sanders, J. W., Dietrich, W. E., & Glaser, S. D. (2009). Influence of rock mass strength on the erosion rate of alpine cliffs. *Earth Surface Processes and Landforms*. http://doi.org/10.1002/esp

Moraetis, D., Paranychianakis, N. V., Nikolaidis, N. P., Banwart, S. A., Rousseva, S., Kercheva, M., Verheul, M. (2014). Sediment provenance, soil development, and carbon content in fluvial and manmade terraces at Koiliaris River Critical Zone Observatory. *Journal of Soils and Sediments*, *15*(2), 347–364. http://doi.org/10.1007/s11368-014-1030-1

Mulder, V. L., de Bruin, S., Schaepman, M. E., & Mayr, T. R. (2011). The use of remote sensing in soil and terrain mapping — A review. *Geoderma*, *162*(1-2), 1–19. http://doi.org/10.1016/j.geoderma.2010.12.018

National Research Council. (2001). *Basic Research Opportunities in Earth Science*. Retrieved from http://www.nap.edu/catalog/9981.html

Nelder, A. J. A, & Wedderburn, R. W. M. (1972). Generalized Linear Models. *J. R. Statist. Soc. A.*, *135*(3), 370–384.

Odeh, I., McBratney, A., & Chittleborough, D. (1994). Spatial prediction of soil properties from landform attributes derived from a digital elevation model. *Geoderma*, *63*(1994), 197–214. Retrieved from http://www.sciencedirect.com/science/article/pii/0016706194900639

Odeh, I. O. A., McBratney, A. B., & Chittleborough, D. J. (1995). Further results on prediction of soil properties from terrain attributes: heterotopic cokriging and regression-kriging. *Geoderma*. http://doi.org/10.1016/0016-7061(95)00007-B

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*. http://doi.org/10.1016/0016-7061(94)90063-9

Pahlavan Rad, M. R., Toomanian, N., Khormali, F., Brungard, C. W., Komaki, C. B., & Bogaert, P. (2014). Updating soil survey maps using random forest and conditioned Latin hypercube sampling in the loess derived soils of northern Iran. *Geoderma*, *232-234*, 97–106. http://doi.org/10.1016/j.geoderma.2014.04.036

Pan, B. T., Geng, H. P., Hu, X. F., Sun, R. H., & Wang, C. (2010). The topographic controls on the decadal-scale erosion rates in Qilian Shan Mountains, N.W. China. *Earth and Planetary Science Letters*, *292*(1-2), 148–157. http://doi.org/10.1016/j.epsl.2010.01.030

Pelletier, J. D., & Rasmussen, C. (2009). Quantifying the climatic and tectonic controls on hillslope steepness and erosion rate. *Lithosphere*, *1*(2), 73–80. http://doi.org/10.1029/2005JF000405

Prasad, A. M., Iverson, L. R., & Liaw, A. (2006). Newer Classification and Regression Tree Techniques: Bagging and Random Forests for Ecological Prediction. *Ecosystems*, *9*(2), 181–199. http://doi.org/10.1007/s10021-005-0054-1

Roering, J. J., Kirchner, J. W., & Dietrich, W. E. (2001). Correction to “Hillslope evolution by nonlinear, slope-dependent transport: Steady state morphology and equilibrium adjustment timescales” by Joshua J. Roering, James W. Kirchner, and William E. Dietrich. *Journal of Geophysical Research*, *106*(B11), 26787. http://doi.org/10.1029/2001JB900018

Samuel-rosa, A, Heuvelink, G. B. M., Vasques, G. M., & Anjos, L. H. C. (2015). Geoderma Do more detailed environmental covariates deliver more accurate soil maps ? *Geoderma*, *243-244*, 214–227. http://doi.org/10.1016/j.geoderma.2014.12.017

Schmidt, K., Behrens, T., & Scholten, T. (2008). Instance selection and classification tree analysis for large spatial datasets in digital soil mapping. *Geoderma*, *146*(1-2), 138–146. http://doi.org/10.1016/j.geoderma.2008.05.010

Steinnes, E., Nickel, S., Hertel, A., Pesch, R., Schr, W., & Uggerud, H. T. (2014). Modelling and mapping spatio-temporal trends of heavy metal accumulation in moss and natural surface soil monitored 1990 to 2010 throughout Norway by multivariate generalized linear models and geostatistics, *99*. http://doi.org/10.1016/j.atmosenv.2014.09.059

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., & Zeileis, A. (2008). Conditional variable importance for random forests. *BMC Bioinformatics*, *9*, 307. http://doi.org/10.1186/1471-2105-9-307

Tesfa, T. K., Tarboton, D. G., Chandler, D. G., & McNamara, J. P. (2009). Modeling soil depth from topographic and land cover attributes. *Water Resources Research*, *45*(10), n/a–n/a. http://doi.org/10.1029/2008WR007474

Venables, W. N., & Dichmont, C. M. (2004). GLMs, GAMs and GLMMs: An overview of theory for applications in fisheries research. *Fisheries Research*, *70*(2-3 SPEC. ISS.), 319–337. http://doi.org/10.1016/j.fishres.2004.08.011

Veronesi, F., & Hurni, L. (2014). Random Forest with semantic tie points for classifying landforms and creating rigorous shaded relief representations. *Geomorphology*, *224*, 152–160. http://doi.org/10.1016/j.geomorph.2014.07.020

Von Steiger, B., Webster, R., Schulin, R., & Lehmann, R. (1996). Mapping heavy metals in polluted soil by disjunctive kriging. *Environmental Pollution*, *94*(2), 205–215. http://doi.org/10.1016/S0269-7491(96)00060-7

Webster, R., & Oliver, M. (2007). *Geostatistics for environmental scientists*. Retrieved from http://books.google.com/books?hl=en&lr=&id=WBwSyvIvNY8C&oi=fnd&pg=PR11&dq=Geostatistics+for+Environmental+Scientists&ots=CAMqPMqJ_a&sig=mHzcLzcbVHkdWGQwuzt2JFOEhPU

Wiesmeier, M., Barthold, F., Blank, B., & Kögel-Knabner, I. (2010). Digital mapping of soil organic matter stocks using Random Forest modeling in a semi-arid steppe ecosystem. *Plant and Soil*, *340*(1-2), 7–24. http://doi.org/10.1007/s11104-010-0425-z

Chapter 2.
Investigating Variable Importance for Mapping Exposed Bedrock Cover in British Columbia's Southern Mountains Using a Random Forest Approach

2.1. Abstract

Accurate mapping of exposed bedrock (EB) is an essential component of landscape evolution modelling. This paper presents the use of Random Forest (RF) as a classifier to predict the location of EB in the landscape by using modified legacy land cover maps. The effect of variable reduction on prediction accuracy was evaluated using the VarSelRF package in R data analysis software. Variable influence was also analyzed through transformed partial dependence plots to understand the probability of classifying EB. Map accuracy of EB predictions was increased from 48% to 88% with the use of RF in comparison to legacy land cover maps. Reducing total variables from 43 to 17 had no effect on the prediction accuracy. These findings emphasize that more data in the form of predictors is not always necessary and confirms that future studies should focus on improving quality rather than quantity of input variables for spatial predictive modeling.

Keywords: Random Forest, Land Cover Mapping, Bedrock, Digital Soil Mapping, Partial Dependence Plots

A version of the following chapter will be submitted to a peer reviewed journal for publication under the co-authorship of Margaret Schmidt, Chuck Bulmer, and Anders Knudby.

2.2. Introduction

Understanding the factors that determine the location of exposed bedrock (EB) is a critical component in studies of long term landscape evolution, vegetation inventories, hydrological monitoring and lithological studies. When bedrock is exposed, there can be a significant decrease in hydrological and biological weathering that is affecting the exposed rock (Anderson et al., 2002). Areas with prominent bedrock exposures indicate that local erosion exceeds soil production, reducing the total amount of steady state soil (DiBiase et al., 2012). The location, lithology, and characteristics of these exposed rocks can give better estimates of soil production caused by physical weathering (Moore et al., 2009). Areas in the proximity of bedrock outcrops can be characterized by abrupt changes, often in the form of mass wasting events (Heimsath et al., 2012). These factors are important when attempting to model soil erosion and landscape evolution on larger scales. Landscape evolution models are becoming more reliant on higher accuracy erosion metrics, which is why bedrock mapping is integral to future modelling approaches (Heung et al., 2013; Krautblatter & Moore, 2014; Montgomery, 2003; Pan et al., 2010).

Existing bedrock mapping in British Columbia provides information on lithology and age of the subterranean rocks (British Columbia Ministry of Employment and Investment, 1996). Mapping efforts that include exposed bedrock cover are found in legacy land cover maps where accuracies are usually low (often less than 55%) and can be considered quite noisy (Saadat et al., 2008). However, even with lower accuracies, legacy maps can aid in modern approaches to mapping EB (Malone et al., 2014).

Legacy land cover maps are being used with modern statistical models to produce more accurate maps as is seen in approaches that are used in Digital Soil Mapping (DSM). These approaches try to enhance the information displayed in legacy maps and incorporate additional information such as terrain derivatives and satellite imagery (McBratney et al., 2003). DSM approaches have allowed for the improvement of spatial accuracy in legacy data (Grimm & Behrens, 2010), and has allowed for better sampling techniques to best represent legacy maps (Carré et al., 2007; Minasny & McBratney, 2006). New information can also be extracted through these approaches, for example relating to parent material, that has not previously been adequately mapped (Heung et al., 2014; Lemercier et al, 2012).

A machine learning technique often used for DSM is the decision tree classifier known as Random Forest (RF) (Breiman, 2001). RF is an ensemble of classification trees used to predict either discrete or continuous dependent variables. RF has been widely used in many fields including bioinformatics (Bureau et al., 2003; Diaz-Uriarte & Andrés, 2005; Mitra et al., 2014; Strobl et al., 2008; Svetnik et al., 2004), ecology and species distribution (Cutler et al., 2007; Knudby et al., 2011; Van Beijma et al., 2014), and DSM (Ghimire et al., 2010; Grimm et al., 2008; Pahlavan Rad et al., 2014; Wiesmeier et al., 2010). It has proven to be a very accurate modelling technique due to its ability to handle noisy data, generate variable importance measures and internally validate results (Breiman, 2001). Extensive testing has been done with the RF classifier for discrete and continuous datasets and it is found to have the highest accuracy rates with external validation, when tested against 179 other classifiers (Fernández-Delgad et al., 2014).

The objectives of this study were to (a) accurately predict the occurrence of bedrock exposures in the landscape, (b) determine which variables had the greatest importance in predicting the occurrence of EB and (c) investigate how these variables influenced predictions.

2.3. Methods

This study utilizes a RF classifier to integrate a wide variety of data sources, including provincial land cover maps, DEMs, satellite-derived spectral indices from both Landsat 7 and Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER) as well as vector data for geology, vegetation and climate, in order to produce spatially explicit predictions of the presence and absence of EB (Figure 2-1). Training data were derived from a combination of existing datasets that specify the location of EB in relation to all other land cover classes, which we refer to collectively as Other Land Types (OLT). An equal area sampling scheme was used to select training data points from both EB and OLT areas. RF and a variable reduction method called VarSelRF were used to identify the most important variables to predict EB. These variables were used in a second RF model and results were validated using a set of 200 randomly located data points with the land cover class (EB or OLT) assigned through photo interpretation of imagery available from Google Earth and Google Maps.

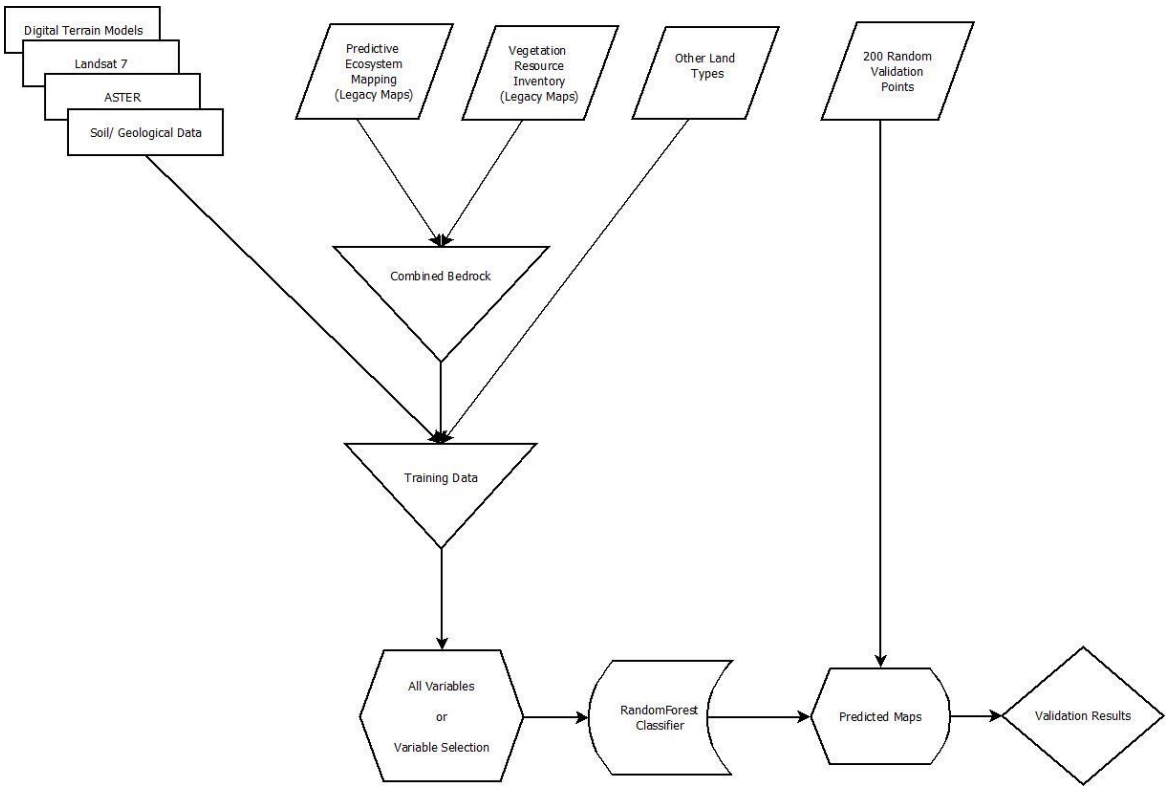


Figure 2-1. Workflow diagram of predictive land cover mapping using topographic indices, remotely sensed imagery, legacy land cover data and a RF classifier. Training data are derived from legacy land cover maps. Photo interpreted validation points were used to assess model accuracy.

2.3.1. Study Area

The study area covers approximately 3435 km^2 and is centred on the town of Tulameen, in the southern interior of British Columbia (N 49°32'45" W 120°45'36") (Figure 2-1). Two physiographic regions are represented in the Tulameen study area; the Coast-Cascade dry belt and the Thompson Plateau (Holland, 1964). Serrated peaks and ridges dominate the western portions of the landscape and often consist of EB. There is no present-day glacial activity within the study area. In addition, EB is often found on rounded summits and localized topographic highs below the tree line (1800 m) and in the eastern portion of the

region. The Tulameen study area has six biogeoclimatic zones that fall within its boundaries: The Coastal Western Hemlock; Ponderosa Pine; Interior Douglas Fir; Montane Spruce; Englemann Spruce Subalpine Fir; and Alpine Tundra (Lloyd et al. 1990). The region receives 510 mm of precipitation annually and the mean annual temperature varies from -7 to 18˚C. The extensive precipitation (including snowfall) in the western portion results from an orographic effect of westerly winds flowing from the Fraser Valley.

At higher elevations in the western part of the region, the dominant soils are Humo-Ferric Podzols, which have developed in coarse textured parent materials on well drained sites (Ministry of Parks, 1999). Brunisolic and Luvisolic soils predominate in the eastern portions of the region, where elevations are lower and less precipitation is received. The geology is a mixture of intrusive igneous rock and folded and faulted volcanic and sedimentary rocks with high mica content. Elevations range from 623 to 2337 m. The dominant geomorphic processes in the region are glacial in origin leading to a wide range of moraines, fluvial deposits, and colluvial materials. A significant amount of mass wasting has also occurred since deglaciation (Ministry of Parks, 1999).

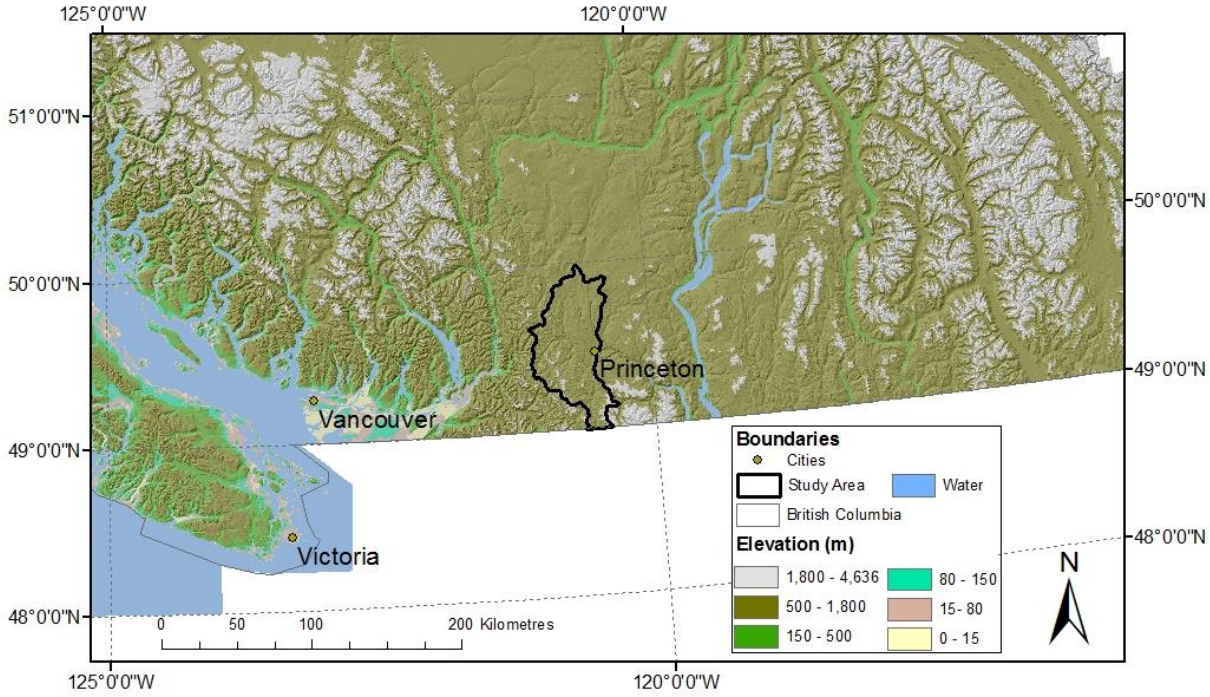


Figure 2-2. South Central British Columbia, depicting the Tulameen study area location.

2.3.2. Land Unit Classification

In this study, I consider EB to be a continuous or discontinuous naturally occurring land cover type where bedrock is the dominant feature visible with covering than 50% of the grid cell at the land surface. Areas of EB generally lack vegetation. Thin deposits of loose material, up to 10 cm thick, may overlay EB outcrops and such deposits typically are relatively unmodified by soil development processes. Other weathered or unweathered materials with discontinuous or continuous vegetation cover can be associated with EB in the landscape.

Generally, large continuous areas of EB are found above the tree line, which is located at approximately 1800 masl in the study area. However, EB can occur

at lower elevations in the presence of minor topographic highs, exposed ridges, or the upper portions of slopes with convex curvature. Unconsolidated material such as colluvium greater than 10 cm thick is excluded from the EB definition. Thicker deposits of colluvium are typically found in areas with localized weathering, erosion and deposition of materials, where material depth and vegetation types vary greatly.

In this study, EB is the only specific land type being modelled. All land types that do not meet the description of EB are labeled OLT. In the study area, OLT consist of, but are not limited to, agricultural land, urban land, densely vegetated land, and water. These features are considered together as a single land cover type, distinct from EB. Vegetated areas in close proximity or contained within bedrock areas are usually treeless, and may support shrubs, herbs, graminoids, byroids and lichens. Areas dominated by EB are often characterized by a rugged or topographically heterogeneous character compared to OLT. I consider a 1 hectare grid cell as my evaluation unit, and each grid cell is classified based on which land type (i.e. EB vs. OLT) occupies more than half of the area.

2.3.3. Calibration Data

Training data for this study was obtained from two legacy land cover datasets in vector format: The Vegetation Resource Inventory (VRI) and the Predictive Ecosystem Mapping (PEM) projects (Ministry of Forest, Lands and Natural Resource Operations, 2013).

The VRI is a province-wide database that maps British Columbia's vegetation at a 1:50,000 scale. The primary use of the VRI is to classify vegetation types and non-vegetated cover types through crown cover estimates and non-

vegetated coverage percentages. The database was created through photo interpretation and in-situ ground truthing (Ministry of Forest, Lands and Natural Resource Operations, 2013). VRI has a land cover type that consists of non-vegetated land units with a total cover of trees, shrubs, herbs and byroids less than 5%. These areas are usually upland and alpine regions that are primarily covered by rock, ice, and snow. Land cover units found in the VRI dataset classified as “exposed land” and “exposed soil” were considered as EB for the purposes of this study.

Predictive Ecosystem Mapping (PEM) is an automated inventory system, designed to create ecosystem maps from spatial data and ecology–landscape relationships. Mapped at a 1:20,000 scale, it is a province-wide initiative to improve on previous mapping techniques with regards to both scale and accuracy (The Resources Information Standards Committee, 2006). Many PEM mapping projects, including the project covering the study area, were accompanied by semi-detailed terrain surveys, and surficial material is an attribute in the PEM database that accompanies the maps. Polygons characterized as exposed bedrock in the PEM surficial materials attribute field (Ministry of Environment, 1988) were considered EB for the purposes of this study.

EB polygons from VRI and PEM were joined to create a combined EB layer that represented all occurrences of exposed bedrock for the study area. The resulting combined data layer was compared to aerial imagery from BING Maps (Figure 2-3 A) and in-situ photography where available (Figure 2-3 B). Land cover units that did not fit the definition of EB were either removed or modified to conform to what was visible on the underlying imagery. A 100 m grid was overlaid on the study area, and a point was generated at the centre of each grid cell. 8691 of these points fell within the boundaries of the combined EB polygons. An equal sampling method (Heung et al., 2014) was then employed to generate an additional 8691

OLT data points representing the rest of the study area. These two sets of points were used to calibrate the RF models.

A

B

Figure 2-3. A) BING imagery with a combined bedrock polygon, outlined in red, displaying the boundary of EB from both VRI and PEM. The orange dot represents the approximate location where image B was taken. B) A photo taken from the roadside showing a part of the combined bedrock polygon from A. Here EB is very prominent, with a vertical face of exposed rock. Overlying soil is a thin veneer with minimal vegetation.

2.3.4. Validation Data

Map validation was accomplished using 1,000 randomly generated points within the study area. Points were classified as either EB or OLT through visual inspection of Google Maps imagery. Each of the random points was assessed in turn, until 100 points of EB and 100 points of OLT were found.

OLT areas were generally easy to identify because they contained large expanses of trees, roads, cut blocks, lakes, rivers, and agricultural land. Generally bedrock outcrops were found in the upper slope of a catena. Grid cells for validation were compared to DEMs and derived slope values, the Google Maps satellite layer (30 cm resolution), Google Maps street view (where applicable), and Google Earth, using the vertical exaggeration. The Google maps satellite layer provided high resolution aerial imagery that allowed for identification of land types and local attributes for visual interpretation of vegetation cover. Google street view allowed for the highest resolution possible for interpretation of a grid cell when available.

For example, Figure 2-4 A. shows a Google Maps image with a randomly placed validation point. The immediate area around the point has a homogenous texture with a dark greyish colour. Figure 2-4 B. shows the same area in Google Earth, where vertical exaggeration is used to emphasize the vertical relief around the validation point, and Figure 2-4 C. shows the Google street view image of the validation point. Here EB is very prominent, and areas that are not exposed are covered with a thin veneer of soil and have a sparse vegetation cover with discontinuous coniferous tree cover.

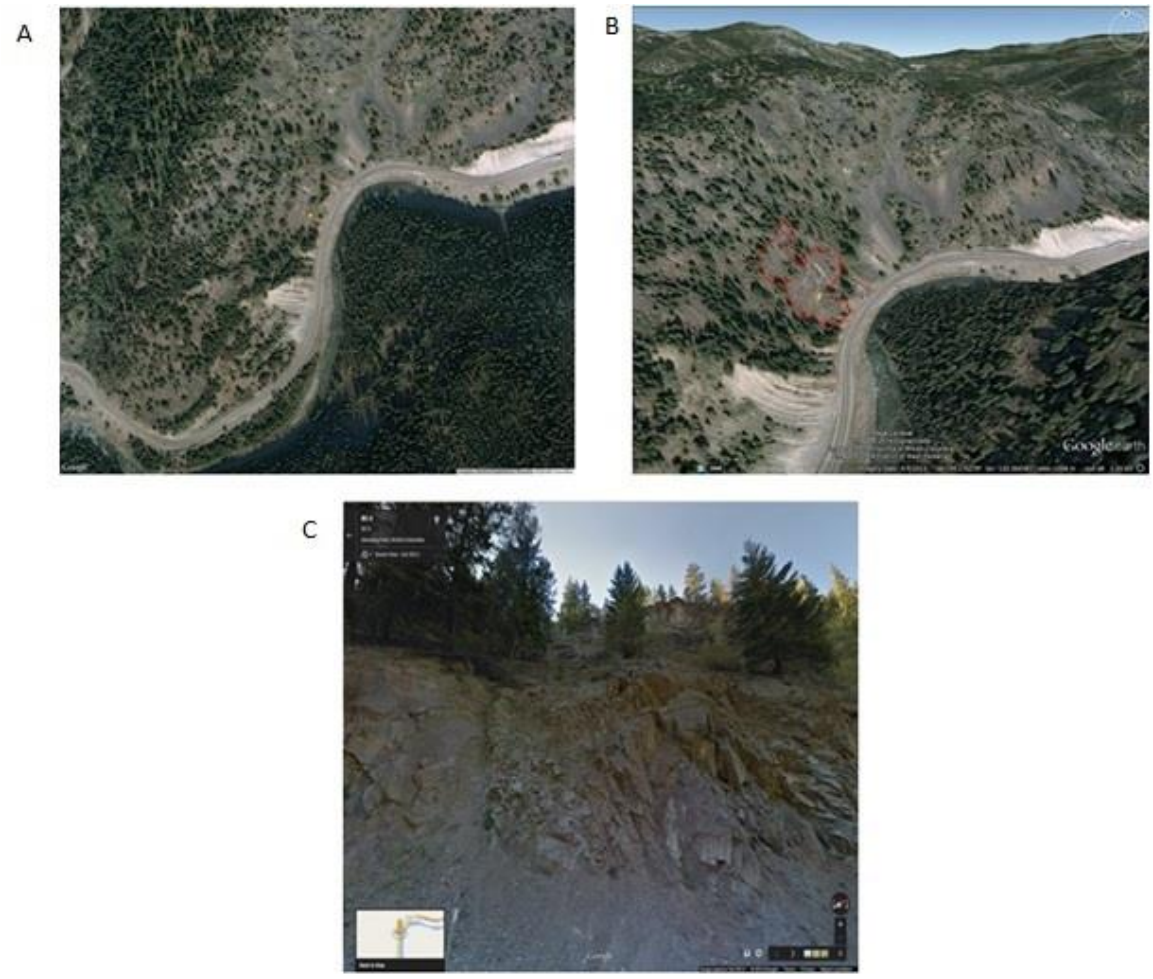


Figure 2-4. A) Image of a validation point that was classified as EB using Google Maps imagery. B) Google Earth image of the validation point from image with vertical exaggeration. The red boundary outlines all areas that would be considered as EB for this case study. C) Google Street View image of the validation point. EB is very prominent, and soil is a thin veneer. Vegetation consists of sparse tree cover and discontinuous bushes and shrubs.

2.3.5. Predictor Variables

Predictor variables used in this study were derived from a 100 m resolution DEM (Hectares BC, 2012), Landsat 7 and ASTER data as well as soil geological

and physiographic vector data (Table 2-1). 43 variables were used in total, all resampled to a 100 m resolution.

Table 2-1. List of 43 topographic, remotely sensed, and vector based land indices used as predictors in the RF classifier.

Landscape Representation	Terrain Derivatives	Code	Reference
DEM	Elevation	Elevation	
	Topographic Ruggedness Index	TRI	(Riley et al. 1999)
	Valley Depth	ValleyD	SAGA Development Team (2011)
	Topographic Profile Index	TPI	(Guisan et al. 1999)
	Slope	Slope	(Zevenbergen & Thorne, 1987)
	Multi-Resolution Valley Bottom Flatness	MRVBF	(Gallant & Dowling, 2003)
	Multi-Resolution Valley Bottom Flatness-Kilometre	MRVBF_KM	(Gallant & Dowling, 2003)
	Multi-Resolution Ridge Top Flatness	MRRTF	(Gallant & Dowling, 2003)
	Slope	Slope	(Zevenbergen & Thorne, 1987)
	Slope Height	SlopeH	(Boehner & Conrad 2008)
	Standardized Height	StandH	SAGA Development Team (2011)
	Profile Curvature	ProfileC	(Zevenbergen & Thorne, 1987)
	Plan Curvature	PlanC	(Zevenbergen & Thorne, 1987)
	Normalized Height	NormHeight	SAGA Development Team (2011)
	Mid Slope Position	MSP	SAGA Development Team (2011)
	Direct Insolation	DirectInso	(Böhner & Antonić, 2009)

	Diffuse Insolation	DiffuseInso	(Böhner & Antonić, 2009)
	Convergence Index	ConIndex	(Koethe and Lehmeier, 1996)
	Channel Network	ChanNet	SAGA Development Team (2011)
	Catchment Area	CatchArea	SAGA Development Team (2011)
	Altitude above Channel Network	AltAC	SAGA Development Team (2011)
	Slope Length Factor	LSFactor	(Moore et al. 1993)
	Topographic Wetness Index	TWI	(Beven & Kirkby, 1979)
	Ridge Hill Slope Position-Hectare	RHSP_HA	(MacMillan, R.A, 2005)
	Ridge Hill Slope Position-Kilometre	RHSP_KM	(MacMillan, R.A, 2005)
	Mass Balance Index	MB_IND	(Friedrich, K. 1996)
	Topographic Profile Index Classification	Landforms	(Guisan et al., 1999)
Landsat 7	Normalized Difference Vegetation Index	NDVI	(Tucker, 1979)
	Principal Component Analysis (7 Bands, First PC)	LandsatPCA	
ASTER	Visible and Near Infrared Principal Component Analysis (Bands 1-3, First PC)	VNIRPCA	
	Normalized Difference Vegetation Index	AsterNDVI	
	Iron Oxides	ironoxides	(Kilby & Kilby, 2006)
	Sericite and Illites	Sericite	(Kilby & Kilby, 2006)
	Siliceous Rocks	Siliceous	(Kilby & Kilby, 2006)
	Short Wave Infrared Principal Component Analysis (Bands 5-10, First PC)	SWIRPCA	

	Thermal Infrared Principal Component Analysis (Bands 11-15, First PC)	TIRPCA
Soil, Geological and	British Columbia Watersheds	Watershed
Physiographic	British Columbia Geology	bcgeology
Vector Data	Biogeoclimatic Zones-Codes	Beccodes
	Biogeoclimatic Zones-Zones	Beczones
	British Columbia Geological/ Ecological Classes	GeoEcoClas
	Glacial landforms Canada	Glaciallf
	Mathew's Physiographic Areas	PhysioArea
	Soil Landscapes of Canada	SLC

2.3.6. Random Forest Classification

The randomForest package (Liaw and Wiener, 2002) was used in the R statistical program (R Development Core Team, 2012) to model the relationships between EB and the predictor variables. RF is an assemblage of classification trees that are fit to a data set, where the final predicted output is the combination of all trees. Trees are grown from a random selection of training data variables, and are allowed to reach the largest extent without pruning (Chen et al, 2004). As a supervised classifier, RF uses a bootstrapped portion (default 63.2%) of the training data in order to generate each tree. The remaining data points, called the Out-Of-Bag (OOB) portion, is used to generate an error estimate by comparing predictions and observations for these data (Breiman, 2001). Two forms of variable

importance measures called the mean decrease in accuracy (MDA) and the mean decrease in Gini coefficient (MDG) are also generated from the RF model,. The MDA uses the OOB measure to determine the change in prediction error from permutation of a single predictor in the validation (OOB) data. A large increase in error from such permutation indicates that the variable is important for making good predictions. The second variable importance measure, MDG, relies in the Gini coefficient that quantifies the impurity of each node in a tree. At each split in each tree, the Gini coefficients of the resulting two nodes are lower than that of the initial node, and the decrease in Gini coefficient is a measure of the importance of the variable used in the split (Liaw and Wiener, 2002). Default parameters were used for the RF model, as further optimization proved to have negligible improvements on prediction accuracy.

2.3.7. Variable Selection

Interpretation of RF can become complicated when attempting to assess variable importance and model predictions. Correlation between predictors and its influence on predictions is often hidden in the “Black Box” of the RF model (Strobl et al., 2008). Furthermore, prediction accuracy is often not improved by ‘over-complicating’ a model by adding predictors, especially when those added variables are redundant (Svetnik et al., 2003). Accounting for data acquisition and processing, increased dimensionality can also be time consuming and costly and lowers or inhibits replicability in future studies. To reduce complexity a data reduction technique from Diaz-Uriarte and Andrés (2005) was employed to systematically remove redundant variables in the model.

Variable reduction was completed using the ‘varSelRF’ package in R (Genuer, Poggi, & Tuleau-Malot, 2010), starting with the original RF model and using a backwards variable elimination method. First the OOB error of the original

model is calculated. Then the predictor with the lowest MDA importance is removed, a new RF model generated, and the OOB error of the new model calculated. Each new model's OOB error is compared to the original model OOB error. This process is repeated, each time removing the predictor with the lowest MDA importance until there is only one variable left in the model. Each successively reduced model that was produced was compared to the original 43 variable model. The model with the least amount of variables, but still within one standard error of the original model OOB error, was chosen by comparing the final binomial error count from all forests. It is important to note that variable importance is not recalculated with removal of each variable, but instead follows the order that was determined from the original MDA because this method reduces overfitting of the model (Svetnik et al., 2004). Default values were used for all parameters, except for the "Vars.frac.drop" function, which was set to 0.1, as any higher value return only a single predictor.

2.3.8. Partial Dependence Plots (PDPs)

PDPs generated from the RF model illustrate the effect of x (a predictor) on $f(x)$ (the predicted probability of EB presence) when all other predictors are fixed at their mean value (Equation 1).

$$f(x) = \frac{1}{n}\sum_{i=1}^{n} f(x, x_{iC})$$

(Equation 1)

PDPs can be interpreted to identify values of x that are associated with particularly high or low probabilities of EB presence. It is important to note that the probabilities that are generated have also been influenced by all other variables in

the model (x_{iC}) (Hastie and Tibshirani, 2005), and may not represent environmental conditions actually found in the study area.

PDPs have a y axis that ranges from -∞ to ∞ and quantifies the log-odds of a positive classification for the total range of values in x. Log-odds are logarithmic transformations of the probabilities for values in x (Hastie and Tibshirani, 2005). Using Equation 2, the y-axis of the PDPs were converted from a log-odds to a probability scale that determines the probability of EB presence. The transformation of the y-axis was applied to the three variables (LandsatPCA, Landsat NDVI and Topographic Ruggedness Index (TRI) with the highest MDA importance after variable reduction.

$$p = EXP(y)/(1 + EXP(y))$$

(Equation 2)

2.4. Results and Discussion

2.4.1. Partial Dependence

RF is referred to as a “Black Box” due to the use of multiple classification trees, where detailed interpretation of the relations between the predictors and the response variable becomes very difficult (Prasad et al., 2006). Through the use of PDPs, the relationships between predictors and predictions can be at least partially understood. Cutler et al. (2007) used PDPs in their analysis of species distribution; however they only made a qualitative assessment of values. Here I use PDPs to quantify the probability of generating an EB presence prediction given the full range of values for a predictor. It is important to note that probabilities illustrated in these PDPs quantify the predicted probability of EB presence given a specific value of the predictor in question and the mean value of all other predictors present

in the model. The LandsatPCA, Landsat NDVI and TRI variables will be further analyzed due to their rank generated from the RF MDA plot. While all other 14 variables also had a non-negligible influence on the prediction of bedrock, only LandsatPCA, NDVI and TRI were explored further in this study.

2.4.1.1 Landsat PCA

LandsatPCA obtained the highest MDA variable importance. Similar studies that have used RF to map landforms, with the inclusion of EB, have found that elevation was ranked the most important. However these studies did not include a LandsatPCA layer in the analysis (Veronesi & Hurni, 2014). To further understand the influence that LandsatPCA had on the prediction of EB, a PDP was generated (Figure 2-5) to show the probabilities of generating a presence response for EB. Histograms from the training data were plotted with the PDPs to better understand the relationship between the data and the final model predictions.

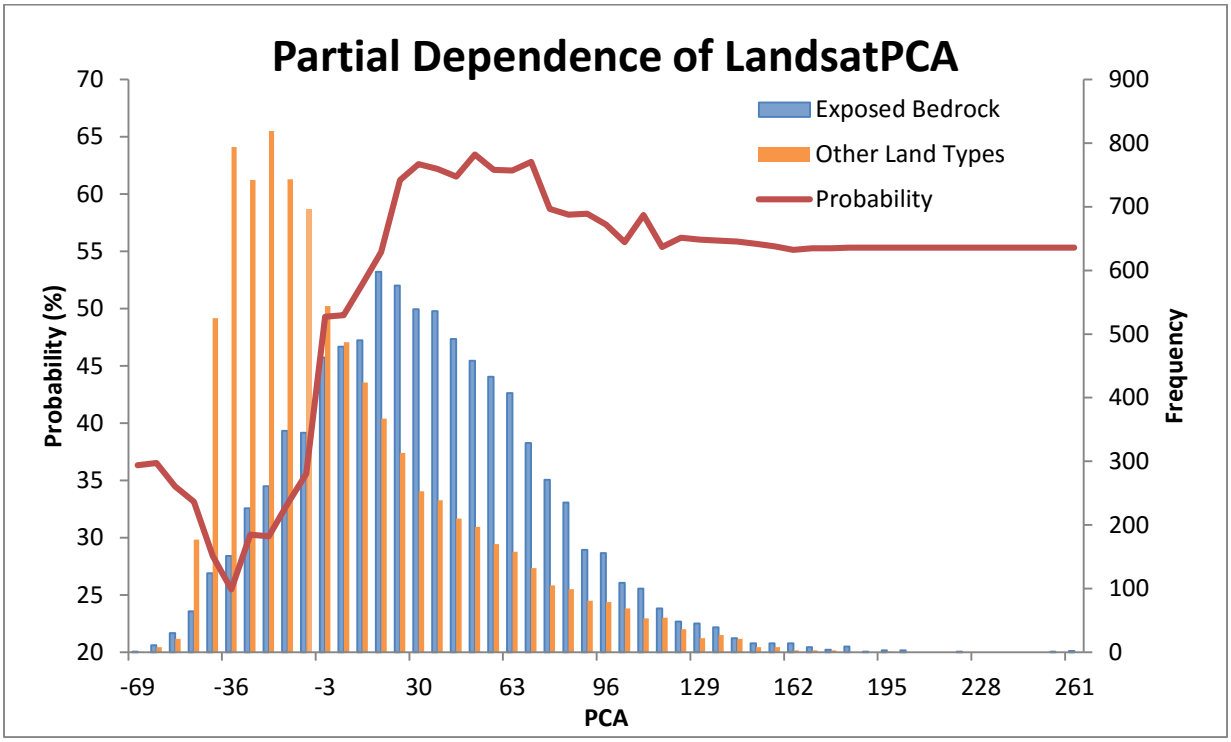


Figure 2-5. A Partial Dependence Plot for the first principal component of a Landsat 7 image. Partial dependence is the partial effect that a predictor will have for predicting the dependent variable, when all other predictors in the model are set to their mean value. The x-axis shows the full range of values of the predictor variable. The primary y-axis (left) shows the backwards log transformation of a log-odd to produce a probability metric for a presence prediction for EB. The secondary y-axis (right) is the cell count for LandsatPCA values in the data used to train the RF model.

The PDP shows that an increase in LandsatPCA values from -36 to approximately 30 lead to an increase in the predicted probability of EB presence from 25% to around 60%. This probability remains relatively stable at higher LandsatPCA values (Figure 2-5). The histograms also indicate a transition from a high frequency of OLT to EB observations over this range of LandsatPCA. The highest frequency values of Landsat PCA for the EB training data is 17, but the LandsatPCA values with the highest rate on the classification occur between

LandsatPCA 30 and 70. This indicates two processes. One, that RF favours values where one class has a significantly higher frequency than the other to increase prediction probability. Two, that there is a divergence between frequency of values in the training dataset and prediction rate and shows that higher frequencies do not directly correlate with increased prediction rates. EB presence is least likely in the -40 to -3 range of LandsatPCA, in which LandsatPCA values are predominantly associated with the OLT class.

Factor loadings from the PCA report attributed the majority of the variance to bands 4 and 6 from the Landsat 7 images. Band 4 is the near infrared band often used for biomass and shoreline interpretations (Bou Kheir et al., 2010;Gowardet al., 2001; Velmurugan & Carlos, 2009) and band 6 is for thermal infrared and is used for thermal and soil moisture mapping (Goward et al., 2001; Hashemimanesh et al., 2012). The greater the PCA value is, the more likely that RF will classify an area as bedrock. From the high factor loadings these areas are assumed to be warmer and lacking biomass.

2.4.1.2 Normalized Difference Vegetation Index

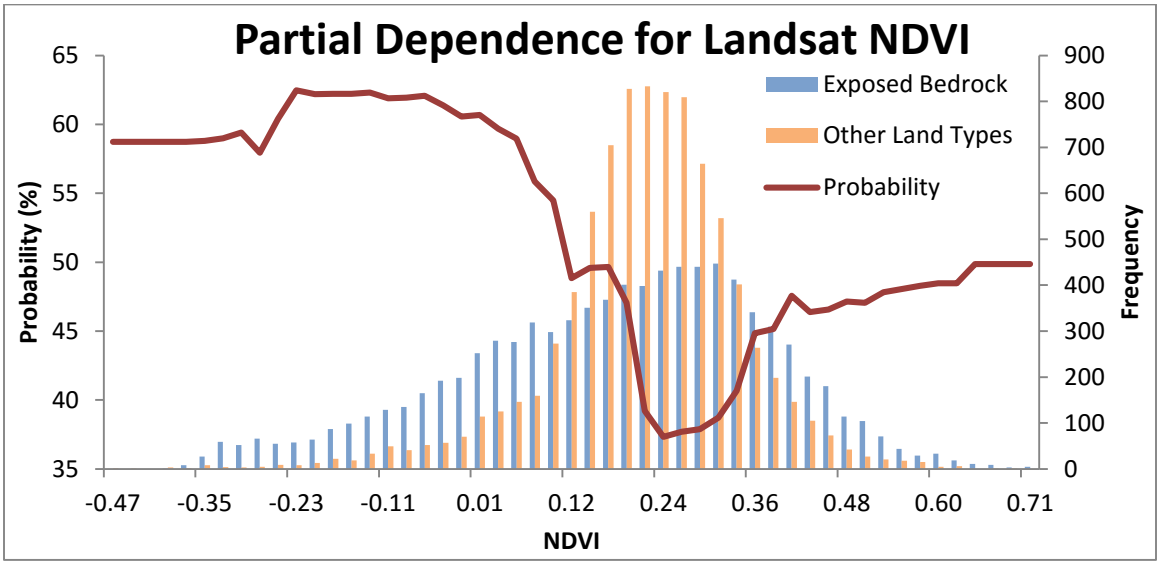


Figure 2-6. A Partial Dependence Plot for Landsat NDVI with count data of the Landsat NDVI training data values used to generate the PDP.

NDVI was the second most important variable in the MDA in the reduced RF model. The PDP for Landsat NDVI (Figure 2-6) shows an irregular u-shaped relationship with the prediction for EB, with three distinct regions. Lower prediction effects for EB can be seen in the 0.19 to 0.38 range of NDVI. The prediction of EB is as low as 37% in this range, which also coincides with significant overlap in the training data between the two classes (EB and OLT). NDVI values less than 0.19 represent a second region of the PDP, where much higher prediction effects (as high as 62%) were observed. Valor & Caselles (1996) used field testing in the Hérault region of southern France and found NDVI values for EB to be below 0.17, which is consistent with the results presented in this PDP. The last region of the PDP plot for NDVI occurs where values are greater than 0.38. Here another increase in prediction effect is observed in EB, where classification is seen as high as 49%. These values in the training data were found to be on fringe areas between EB and densely vegetated areas and are being represented as pure

bedrock. Values in this range tend to be closely related to vegetated environments, which is not typical for EB, having led to decreases in the prediction accuracy of the bedrock maps. The misrepresentation in the training data can be attributed to the changing scale for the combined bedrock polygons (1:20,000 scale) to a 100 m resolution.

2.4.1.3 Topographic Ruggedness Index

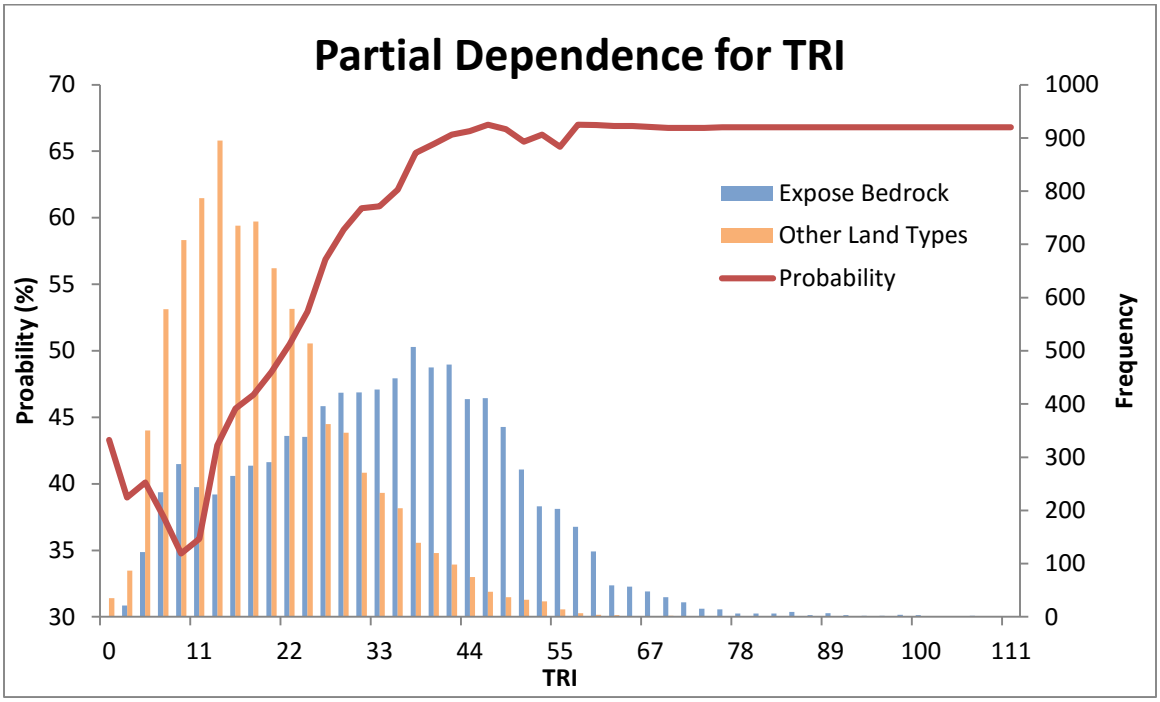


Figure 2-7 A Partial Dependence Plot for topographic ruggedness index with count data of the TRI training data values used to generate the PDP.

The third most important variable in the RF model was topographic ruggedness index (TRI) (Figure 2-7). Here a positive linear relationship with the prediction effect is seen; where the higher the value for TRI, the higher its prediction effect for classifying EB. TRI values between 0 and 24 have a great amount of overlap between EB and the OLT classes, and also the highest frequency of OLT training points. The effect that TRI has on the presence response

in the lower portion of this range can be as low as 34%, due to the higher frequencies of OLT training points found in this range. Classification reaches 67% as the frequency of EB exceeds OLT with values exceeding 24. TRI is closely related to slope (Sappington et al., 2007) and may be why it supersedes slope as an important predictor (DiBiase et al., 2012; A. M. Heimsath et al., 2012). However, it is important to note that TRI accounts for variance in a landscape and does not give the degree of change as does slope. The PDP suggests that highly rugged or highly variable terrain has more EB. The highest occurrence can be seen with a TRI value of 67. In this landscape, values do not tend to be any more variable than a TRI of 67 and this is why there is less frequency for higher TRI values.

It is important to note that the probabilities shown in the PDPs are not true probabilities of finding EB in an area with the given environmental conditions, but rather the probability of EB presence predicted by the RF model. These values are influenced by study area (location and scale), the training data, and the predictors used in the model.

2.4.2. Accuracy of Legacy Land Cover Maps

The 200 validation points were compared with the legacy land cover data that depicted EB and OLT for the study area (Figure 2-8). Validation results show the percentage of cells correctly classified in each map for EB and OLT out of 100. Both accuracy ratings for EB and OLT are then summed and averaged for the total map accuracy (Figure 2-9). Validation accuracy of EB for VRI and PEM was low, reported at 46% and 28%, respectively. The combined EB layer only marginally increased its accuracy at 48%.These results reflect in part the coarser map scale used to produce the legacy land cover maps, and also they possibly reflect that EB was not the focus of the VRI maps, where the vegetated and productive areas of the landscape received much more attention than those with EB. Total accuracy

for VRI and PEM were between 63% and 75%, however these values are skewed by the high prediction probability for the OLT class. Values for EB (the minority class) were much lower, and were seen to be as low as 28% with the PEM data. Not only was EB misclassified as OLT classes in areas with bedrock, there were significant portions of bedrock that were not included in the PEM and VRI datasets.

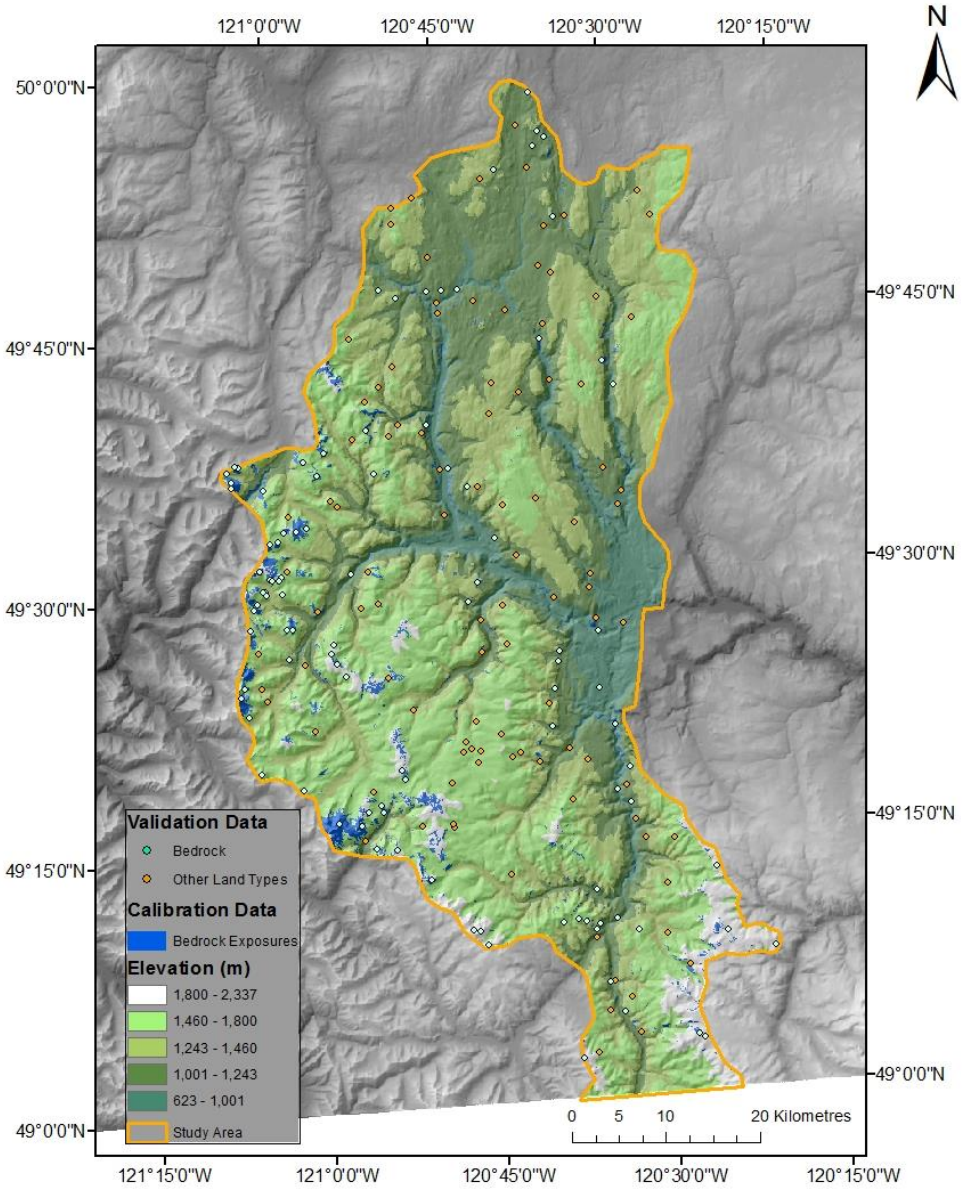


Figure 2-8. EB map from legacy land cover data (blue polygons), in relation to 200 validation points assessed for areas of EB and OLT.

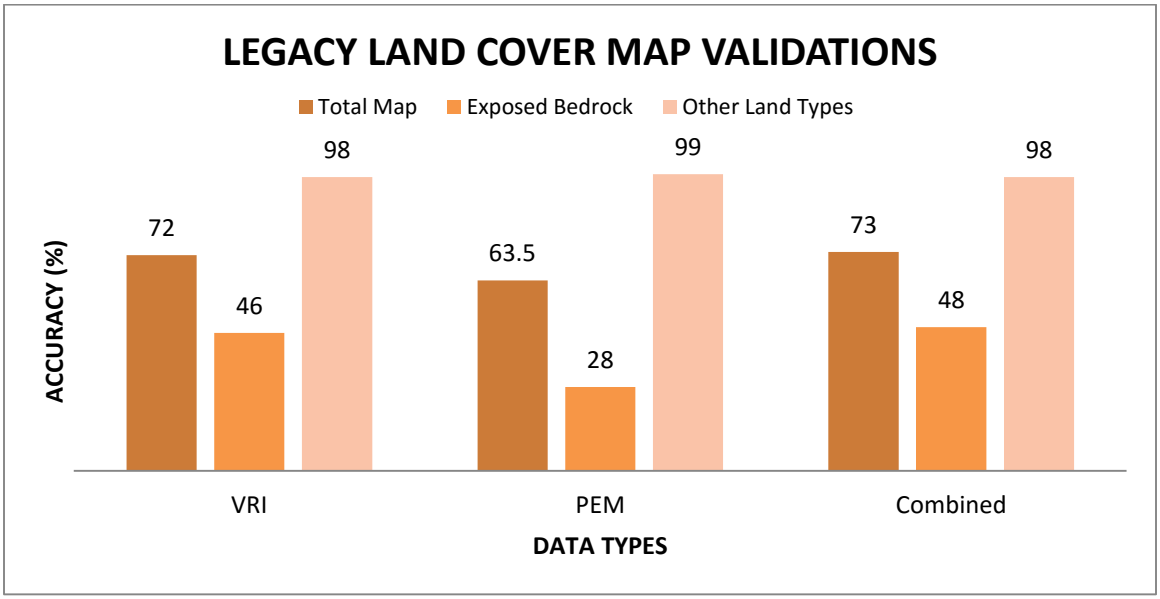


Figure 2-9 Validation accuracy (%) for the 200 validation points (100 for EB and 100 for OLT). Each cover type for validation was compared to the map and resulting accuracies (out of 100) for each class are presented. The total map class represents the average of both class accuracies.

When the validation points were compared to the RF predictions, overall accuracy for RF predictions using all variables and the reduced variable model were 88.5% and 89%, respectively. EB prediction accuracy was also 85% and 86% respectively and OLT prediction at 92% for both models. There was an average of 30% higher accuracy for the predicted RF maps compared to the legacy land cover maps. Emphasizing that legacy maps can greatly be improved through RF.

2.4.3. Predicted Maps

Both RF models for the Tulameen study area (Figure 2-10) produced very similar maps. The model with all 43 variables covered 425.95 km^2 (12.4%) of the study area and the reduced model covered 428.23 km^2 (12.5%) of the study area. Showing that even though there was a reduction in total predictors, seemingly identical maps were created. The total area mapped for both RF models was significantly greater than the area mapped by PEM and VRI. PEM, VRI and the combined bedrock layers had 46.06 km^2 (1.3%), 105.26 km^2 (3%), and 108.01 km^2 (3.1%) area, respectively. Figure 2-11 shows a comparison of the combined EB, and the predicted RF maps. The RF model captures more of the EB in the landscape that was not mapped from the PEM and VRI mapping procedures. However, it is important to note that VRI and PEM did not have exclusive EB layers. The lack of exclusive EB layers may have contributed to the lower accuracies, which emphasizes the need for this study's methods. It is also important to emphasize that even though more EB has been mapped in the landscape, this study focused on a 100 m resolution, where the VRI and PEM datasets were mapped at a 1:20,000 scale. VRI and PEM have higher precision with their boundaries, while the RF model includes more of the fringe landscape that is not EB but dense vegetation.

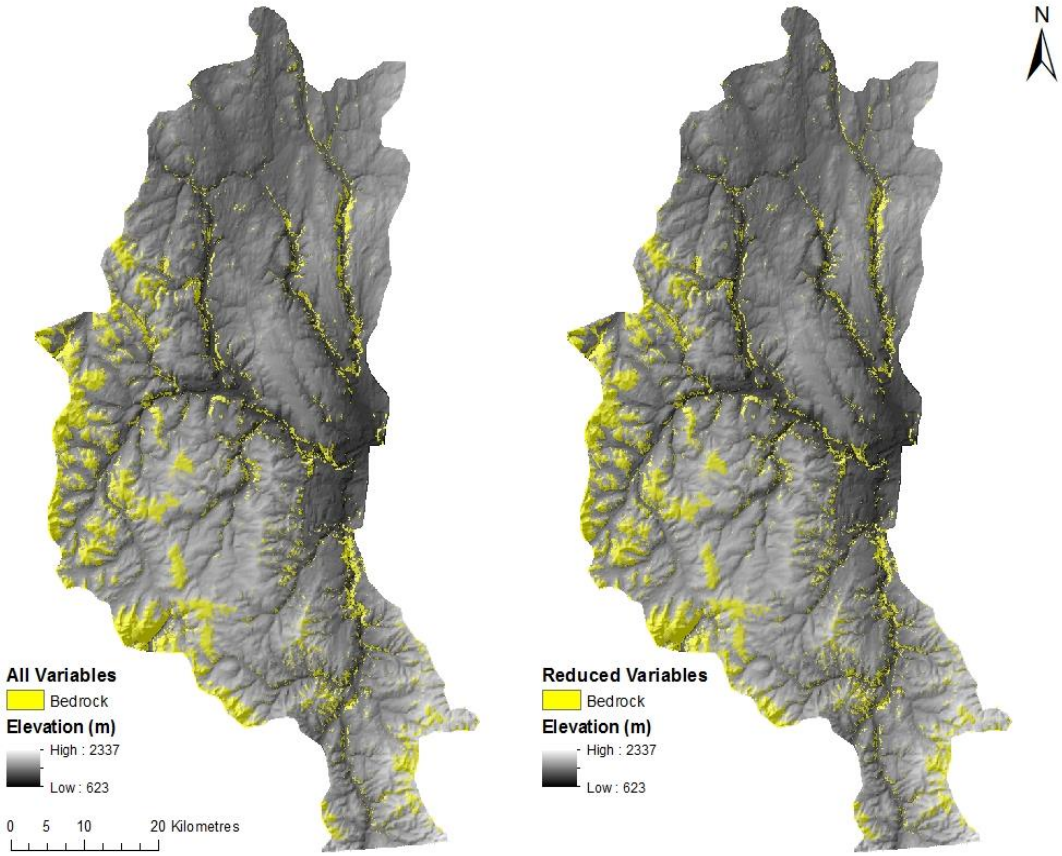


Figure 2-10. Predictive EB cover type maps using RF at a 100 m spatial resolution for the Tulameen study area.

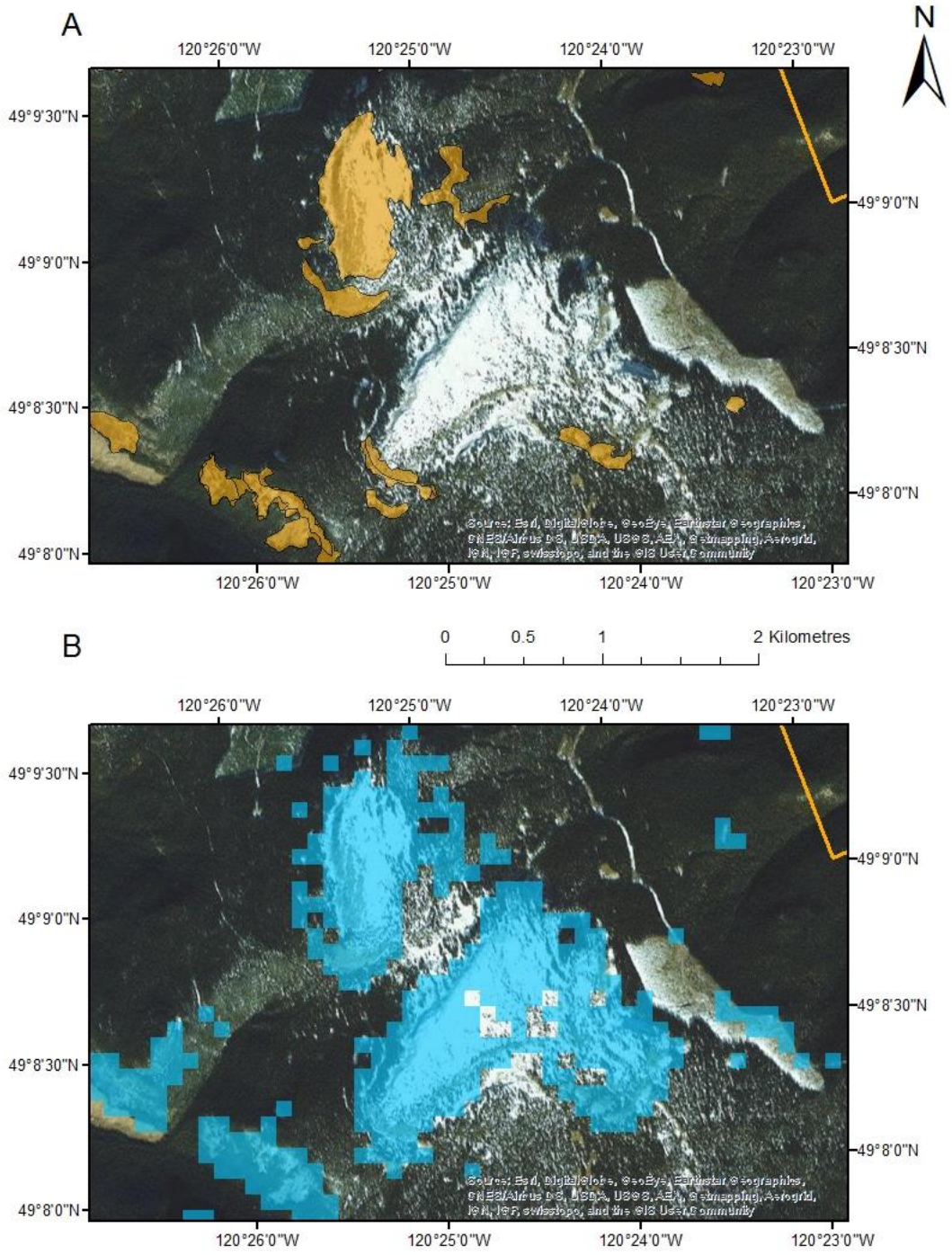


Figure 2-11. A comparison of the (A) combined EB and the (B) RF predicted EB maps.

2.5. Conclusion

This paper presented the use of a RF classifier to map exposed bedrock. The RF classifier was applied to a study area in Southern BC to demonstrate how modified legacy land cover maps can be used, in combination with a range of predictors derived from DEMs, satellite imagery and soil, geological and physiographic vector data to produce accurate bedrock maps.

Using the VarSelRF package, backward elimination of predictor variables showed that the 43 predictors used for the original model could be reduced to 17 variables, with no substantial effect on prediction accuracy. This reduction allowed for a simpler model and could save time in dataset acquisition and preparation if bedrock predictions are desired for different areas with the same model.

To further understand how variables influenced the prediction of bedrock, transformed PDPs from the reduced RF model were generated for the three most important predictors (Landsat PCA, NDVI and TRI). Key value ranges were described for values that were associated with accurate classification of EB for each variable. Anomalies in the data such as seen with NDVI were also easier to visualize and interpret with PDPs.

Finally, this study demonstrates that RF can greatly improve on legacy land cover maps that could be considered noisy and lacking in accuracy. These findings highlight the importance of machine learning techniques in land cover mapping, as field data collection is expensive and time-consuming and thus rarely carried out and there is an increasing reliance on remotely acquired information.

2.6. References

Anderson, S. P., Dietrich, W. E., & Brimhall, G. H. (2002). Weathering profiles, mass-balance analysis, and rates of solute loss: Linkages between weathering and erosion in a small, steep catchment. Bulletin of the Geological Society of America, 114(9), 1143–1158.

Bou Kheir, R., Greve, M. H., Bøcher, P. K., Greve, M. B., Larsen, R., & McCloy, K. (2010). Predictive mapping of soil organic carbon in wet cultivated lands using classification-tree based models: the case study of Denmark. Journal of Environmental Management, 91(5), 1150–60.

Breiman, L. (2001). Random forests. Machine Learning, 5–32.

British Columbia Ministry of Employment and Investment. (1996). Specifications and Guidelines for Bedrock Mapping in British Columbia. Retrieved from Minstry of Energy and Mines.

Bureau, A., Dupuis, J., Hayward, B., Falls, K., & Van Eerdewegh, P. (2003). Mapping complex traits using Random Forests. BMC Genetics, 4, S64.

Carré, F., McBratney, A. B., & Minasny, B. (2007). Estimation and potential improvement of the quality of legacy soil samples for digital soil mapping. Geoderma, 141(1-2), 1–14.

Chen, C., Liaw, A., & Breiman, L. (2004). Using random forest to learn imbalanced data. University of California, Berkeley, (1999), 1–12.

Cutler, D. R., Edwards, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., & Lawler, J. J. (2007). Random forests for classification in ecology. Ecology, 88(11), 2783–92.

Diaz-Uriarte, R., & Alvarez de Andres, S. (2005). Variable selection from random forests : application to gene expression data, 1–11.

DiBiase, R. A., Heimsath, A. M., & Whipple, K. X. (2012). Hillslope response to tectonic forcing in threshold landscapes. Earth Surface Processes and Landforms, 37(8), 855–865.

Exelis Visual Information Solutions. (2010). Boulder, Colorado: Exelis Visual Information Solutions.

Fernández-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we Need Hundreds of Classifiers to Solve Real World Classification Problems? Journal of Machine Learning Research, 15, 3133–3181.

Robin Genuer, Jean-Michel Poggi, Christine Tuleau-Malot. Variable selection using Random Forests. Pattern Recognition Letters, Elsevier, 2010, 31 (14), pp.2225-2236. <hal-00755489>

Ghimire, B., Rogan, J., & Miller, J. (2010). Contextual land-cover classification: incorporating spatial dependence in land-cover classification models using random forests and the Getis statistic. Remote Sensing Letters, 1(1), 45–54.

Goward, S. N., Masek, J. G., Williams, D. L., Irons, J. R., & Thompson, R. J. (2001). The Landsat 7 mission: Terrestial Research and Applications for the 21st Century. Remote Sensing of Environment, 78(1-2), 3–12.

Grimm, R., & Behrens, T. (2010). Uncertainty analysis of sample locations within digital soil mapping approaches. Geoderma, 155(3-4), 154–163.

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(1-2), 102–113.

Hashemimanesh, M. M., Matinfar, H. R., Alavipanah, S. K., & Zehtabian, G. (2012). Landsat thermal band efficiency on characterizing mulched soil surface. International Agrophysics, 26(3), 249–257.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer (Vol. 27).

Heimsath, A. M., DiBiase, R. A., & Whipple, K. X. (2012). Soil production limits and the transition to bedrock-dominated landscapes. Nature Geoscience, 5(3), 210–214.

Hectares, B.C., (2015.) Hectares BC. Available at http://hectaresbc.org/app/habc/HaBC.html (verified 18 January 2015).

Heung, B., Bakker, L., Schmidt, M. G., & Dragićević, S. (2013). Modelling the dynamics of soil redistribution induced by sheet erosion using the Universal Soil Loss Equation and cellular automata. Geoderma, 202-203, 112–125.

Heung, B., Bulmer, C. E., & Schmidt, M. G. (2014). Predictive soil parent material mapping at a regional-scale: A Random Forest approach. Geoderma, 214-215, 141–154.

Holland, S. S. (1964). Landforms of British Columbia: a physiographic outline. British Columbia Department of Mines and Petroleum Resources Bulletin.

Knudby, A., Roelfsema, C., Lyons, M., Phinn, S., & Jupiter, S. (2011). Mapping Fish Community Variables by Integrating Field and Satellite Data, Object-Based Image Analysis and Modeling in a Traditional Fijian Fisheries Management Area. Remote Sensing, 3(3), 460–483.

Krautblatter, M., & Moore, J. R. (2014). Rock slope instability and erosion: toward improved process understanding. Earth Surface Processes and Landforms, 39(9), 1273–1278.

Lemercier, B., Lacoste, M., Loum, M., & Walter, C. (2012). Extrapolation at regional scale of local soil knowledge using boosted classification trees: A two-step approach. Geoderma, 171-172, 75–84.

Liaw, A., & Wiener, M., (2002). Classification and Regression by randomForest. R News 2, 18–22.

MacMillan, R.A., (2005). A new approach to automated extraction and classification of repeating landform types. . Naples Florida Frontiers in Pedometrics, p. 54

Malone, B. P., Minasny, B., Odgers, N. P., & McBratney, A. B. (2014). Using model averaging to combine soil property rasters from legacy soil maps and from point data. Geoderma, 232-234, 34–44.

McBratney, A., Mendonça Santos, M., & Minasny, B. (2003). On digital soil mapping. Geoderma (Vol. 117).

Minasny, B., & McBratney, A. B. (2006). A conditioned Latin hypercube method for sampling in the presence of ancillary information. Computers & Geosciences, 32(9), 1378–1388.

Ministry of Environment, B. C. (1988). Terrain classification system for British Columbia.

Ministry of Parks, B. C. (1999). Coquihalla Recreation Summit Master Plan.

Mitra, J., Bourgeat, P., Fripp, J., Ghose, S., Rose, S., Salvado, O., & Carey, L. (2014). Lesion segmentation from multimodal MRI using random forest following ischemic stroke. NeuroImage, 98, 324–35.

Montgomery, D. R. (2003). Predicting landscape-scale erosion rates using digital elevation models. Surface Geosciences, 335(16), 1121–1130.

Moore, J. R., Sanders, J. W., Dietrich, W. E., & Glaser, S. D. (2009). Influence of rock mass strength on the erosion rate of alpine cliffs. Earth Surface Processes and Landforms.

Moore, I.D., Turner, A.K., Wilson, J.P., Jenson, S.K., Band, L.E., 1993. GIS and land- surface–subsurface process modeling. Environmental Modeling with GIS. Oxford University Press, pp. 196–230.

Pahlavan R, M. R., Toomanian, N., Khormali, F., Brungard, C. W., Komaki, C. B., & Bogaert, P. (2014). Updating soil survey maps using random forest and conditioned Latin hypercube sampling in the loess derived soils of northern Iran. Geoderma, 232-234, 97–106.

Pan, B. T., Geng, H. P., Hu, X. F., Sun, R. H., & Wang, C. (2010). The topographic controls on the decadal-scale erosion rates in Qilian Shan Mountains, N.W. China. Earth and Planetary Science Letters, 292(1-2), 148–157. http://doi.org/10.1016/j.epsl.2010.01.030

Prasad, A. M., Iverson, L. R., & Liaw, A. (2006). Newer Classification and Regression Tree Techniques: Bagging and Random Forests for Ecological Prediction. Ecosystems, 9(2), 181–199.

R Development Core Team, (2012). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria (URL http://www.R-project. org.).

Saadat, H., Bonnell, R., Sharifi, F., Mehuys, G., Namdar, M., & Ale-Ebrahim, S. (2008). Landform classification from a digital elevation model and satellite imagery. Geomorphology, 100(3-4), 453–464.

Saga Development Team, (2011). System for Automated Geoscientific Analyses (SAGA). Available at http://www.saga-gis.org/en/index.html

Sappington, J. M., Longshore, K. M., & Thompson, D. B. (2007). Quantifying Landscape Ruggedness for Animal Habitat Analysis: A Case Study Using Bighorn Sheep in the Mojave Desert. Journal of Wildlife Management, 71(5), 1419–1426.

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., & Zeileis, A. (2008). Conditional variable importance for random forests. BMC Bioinformatics, 9, 307. http://doi.org/10.1186/1471-2105-9-307

Svetnik, V., Liaw, A., Tong, C., Culberson, J. C., Sheridan, R. P., & Feuston, B. P. (2003). Random forest: a classification and regression tool for compound classification and QSAR modeling. Journal of Chemical Information and Computer Sciences, 43(6), 1947–58.

Svetnik, V., Liaw, A., Tong, C &.,Wang, T., 2004. Application of Breiman's Random Forest to modeling structure–activity relationships of pharmaceutical molecules. In: Roli, F., Kittler, J., Windeatt, T. (Eds.), Multiple Classifier Systems, Fifth International

Workshop, MCS 2004, Proceedings, 9–11 June 2004, Cagliari, Italy. Lecture Notes in Computer Science, vol. 3077. Springer, Berlin, pp. 334–343.

The Resources Information Standards Committee. (2006). Standard for mapping ecosystems at risk in British Columbia: an approach to mapping ecosystems at risk and other sensitive ecosystems. Terrestrial Ecosystems Task Force.

Hengl, T., (2006). Finding the right pixel size. Computers and Geosciences, 32(9), 1283-1298

Valor, E., & Caselles, V. (1996). Mapping Land Surface Emissivity from NDVI : Application to European, African, and South American Areas. Remote Sensing of Environment, 57, 167–184.

Van Beijma, S., Comber, A., & Lamb, A. (2014). Random forest classification of salt marsh vegetation habitats using quad-polarimetric airborne SAR, elevation and optical RS data. Remote Sensing of Environment, 149, 118–129.

Velmurugan, A., & Carlos, G. (2009). Soil Resource Assessment and Mapping using Remote Sensing and GIS, 1(September), 511–525.

Veronesi, F., & Hurni, L. (2014). Random Forest with semantic tie points for classifying landforms and creating rigorous shaded relief representations. Geomorphology, 224, 152–160.

Wiesmeier, M., Barthold, F., Blank, B., & Kögel-Knabner, I. (2010). Digital mapping of soil organic matter stocks using Random Forest modeling in a semi-arid steppe ecosystem. Plant and Soil, 340(1-2), 7–24.

Zevenbergen, L.W., & Thorne, C.R., (1987). Quantitative analysis of land surface topography. Earth Surface Processes and Landforms. 12, 47–56.

Chapter 3.
Modelling Soil Depth in the Critical Zone for Southern British Columbia

3.1. Abstract

The Critical Zone (CZ) is defined as the outer layer of the solid Earth, extending from the vegetation canopy to the pedosphere and down to the bottom of the weathered bedrock zone. Most biological, chemical and physical interactions take place in the CZ, and it is here that most terrestrial life is found. Being able to predict the depth of the pedosphere (i.e. soil depth) in the CZ can lead to a better understanding of rates of physical/chemical change, such as carbon sequestration, soil erosion, and water storage. The objective of this study was to accurately map the depth of the pedosphere on a landscape scale for Southern British Columbia. The data inputs used were exposed bedrock (EB) points, well water (WW) data, and manually sampled soil depth measurements, for which conditioned latin hypercube sampling (cLHS) was used to define a set of locations where soil depth was measured from soil pits. Four methods were then used to model soil depth as a function of environmental data layers derived from a digital elevation model and satellite imagery: Generalized Linear Model (GLM), Random Forest (RF), GLM Residual Kriging (GLMRK) and RF Residual Kriging (RFRK). An equal weighted random sampling scheme of 100 EB, WW, and soil pit points was used with each model. A second sampling scheme was used with the same WW and soil pit points and an additional 5000 randomly sampled EB points, used to improve prediction accuracy with limited data sources. Of the modelling methods used, GLMRK proved to be the best method for shallow depths (0 to 2 m) as assessed by Root Mean Square Error (RMSE) values (1.87 m) with the equal weighted sample scheme.

The addition of the 5000 EB points substantially improved predictions for shallow depths (RMSE 0.9 m) as well as soil depths in the 2-5 m range, while having negligible impact on predictions at deeper depths. This demonstrates that an exposed bedrock layer can help constrain shallow soil depth predictions when used in conjunction with geostatistical approaches such as GLMRK and RFRK for mapping soil depth.

Keywords: Critical Zone, Soil Depth, Random Forest, GLM, Kriging, Landscape Modelling

A version of the following chapter will be submitted to a peer reviewed journal for publication under the co-authorship of Margaret Schmidt, Chuck Bulmer, and Anders Knudby.

3.2. Introduction

The critical zone (CZ) is the environment in which complex interactions between soil, rock, air, water, and living organisms occur to maintain and regulate natural processes. The CZ extends from the top of the vegetation to the solid fresh bedrock below the soil surface (NRC, 2001). The active regions within which these processes take place are the atmosphere, biosphere, hydrosphere, lithosphere, and the pedosphere. The pedosphere, here considered as encompassing all unconsolidated material from the top of the mineral soil to the top of the fresh bedrock, acts as a transition zone between the atmosphere above and the hard rock below (Brady & Weil, 2007) and as the interface for the other four spheres (Lin, 2010).

The pedosphere influences nutrient and chemical exchange rates (Brantley et al., 2007; Chorover et al., 2007), the effects of water storage, water yield and boundary transfer (Field et al., 2015; Yu et al., 2015; Yu et al., 2014), and weathering, erosion and soil production rates (Heimsath et al., 2012; Jin et al., 2010; Ma et al., 2010; Moraetis et al., 2014). These processes are all strongly

influenced by the vertical extent of the pedosphere (henceforth: soil depth). Empirical and quantitative models are needed to predict and describe this aspect of the CZ due to the complexity of mapping the depth of the pedosphere directly (Lin, 2010). Four categories of such methods have been used to model soil depth:

1. Process based
2. Deterministic
3. Stochastic/ geostatistical
4. Combined

(Li et al., 2011; Pelletier & Rasmussen, 2009)

Each of these methods attempts to predict the steady state depth of soil. The first models to predict soil depth were process based models that incorporated the relationships of slope angle and slope convexity with soil production rates (Heimsath et al, 1999). These models were very susceptible to stochastic elements in the landscape that could alter erosion rates. An assumed starting soil depth was required for each model, which introduced errors as an approximate depth was given. Later, local depth measurements were introduced to calibrate the models, and higher resolution elevation data were introduced to aid in predictions. This generally improved the model predictions of soil depth (Pelletier & Rasmussen, 2009). However, these models are typically very complicated to reproduce and are only applicable to small hillslopes which require a large amount of in-situ measurements.

Deterministic methods are also commonly used to predict soil depth, for example regression models have been used to predict soil depth from topographic variables (Ziadat, 2010). Quantitative models incorporating the relationship between the distance to bedrock and well water depth have proven to be beneficial

in data limited areas of Sweden (Karlsson et al., 2014). However, one issue with these models is the inability to incorporate local variance related to heterogeneous landscapes. These models can therefore only be applied to small areas, where the effect of local trends in the landscapes can be minimized.

Many studies have used geostatistical methods such as kriging and combined methods of residual kriging to predict soil depth (Kuriakose et al., 2009; Odeh et al., 1994; Sarkar et al., 2013). The combination of both geostatistical and deterministic methods has allowed for the integration of topographic variables to account for spatial uncertainty in the landscape, and thus has led to improvements in soil depth mapping. Generalized Linear Model (GLM) residual kriging was one of the first combined methods proposed (Odeh et al., 1994). GLM is a weighted linear model that uses a link function to allow distributions other than a normal distribution to be used for predictions (Lane, 2002). Residual values were generated from the GLM model and kriged, then observed soil depth measurements were independently kriged and were added to the kriged residuals. It was found that only residual values (which incorporated the uncertainty in the landscape) should be kriged and combined with the layer produced by the regression model, not a kriged layer of observed soil depths (Hengl et al., 2004; Odeh et al., 1995). Other studies have explored different combinations of regression models and kriging procedures to improve upon prediction accuracies. Kuriakose et al., (2009) compared the use of a linear model, block kriging and residual block kriging and found that the residual block kriging performed the best for predicting soil depth. Residual block kriging was able to account for more of the variability in the landscape when comparing error rates; although validation results were similar to ordinary kriging with residual kriging due to the controlled environment. That study had 259 augured soil points; however the depth range was limited from 0 to 4 m, and the study site comprised a single watershed.

Random Forest (RF) is a non-parametric decision tree classifier that can predict both discrete and continuous data (Breiman, 2001). The ability of RF to handle large and noisy datasets and to generate variable importance plots has made it an important tool to map soil properties (Heung et al., 2014; Rad et al., 2014; Wiesmeier et al., 2010). Tesfa et al., (2009) used RF in regression mode and compared it to a General Additive Model (GAM) to predict soil depth at watershed-scale. Their study focused on the creation of new predictors that could help to predict soil depth. It was found that RF outperformed the GAM model and was able to express more of the spatial variability in the watershed than the GAM model; however they did not further explore the abilities of RF and residual kriging (RFRK).

RFRK has been explored in mapping other soil properties but has not been used to map soil depth. Guo et al., (2015) explored mapping soil organic matter using RFRK and found that it significantly out-performed a stepwise linear regression and a RF in regression mode. Hengl et al., (2015) used RFRK to map soil nutrient contents for the continent of Africa. Mapping was done at a 250 m resolution, with minimal data and these authors also found that it outperformed RF. Their results also demonstrate that RFRK can handle a large amount of spatial variation in the landscape in comparison to other methods, even at a continental scale. Even though RFRK has not been explored to map soil depth, it has proven its ability to map highly variable soil properties, indicating that it could be used for mapping soil depth at the landscape scale.

Most soil depth studies are conducted on a local scale, typically on a single watershed or a hillslope, with very extensive sampling methods. Combination methods such as Residual Kriging can prove to be an important tool for mapping soil depth at a larger landscape scale, especially in data limited areas.

The objectives of this study were to first map the depth of the pedosphere (soil depth) on a landscape scale; then to compare two deterministic models with residual kriging and assess which model had optimal performance; and finally to assess the predictive power that data on the location of exposed bedrock (EB) adds for mapping soil depth. The modelling methods GLM, RF, GLMRK and RFRK, were used to predict soil depth and the methods were compared through variogram variances and prediction errors to determine which method performs the best for mapping soil depth. The methods were applied to the Tulameen region, Southern British Columbia.

3.3. Methods

This study compares GLM, RF, GLMRK and RFRK for mapping the depth of a steady-state soil layer in Southern British Columbia (Figure 3-1). Explanatory environmental variables were derived from a digital elevation model (DEM) and Landsat 7 and ASTER satellite imagery. Calibration/validation soil depth data points were derived from well water (WW) records (n = 239), soil pits (n = 174), and exposed bedrock (EB) locations (n = 300). A separate set of models was created by expanding the set of EB locations from 300 to 5000. All models were calibrated with a subset of the available data (n = 300, n = 5200 with the expanded data set), applied to the environmental data layers to produce mapped predictions, and validated with a random subset of the soil depth data (n = 222) not used for model development. Map accuracy was quantified as the Root-Mean-Squared Error (RMSE) calculated separately for soil depth ranges 0-2 m, 2-5 m, 5-10 m, and >10m.

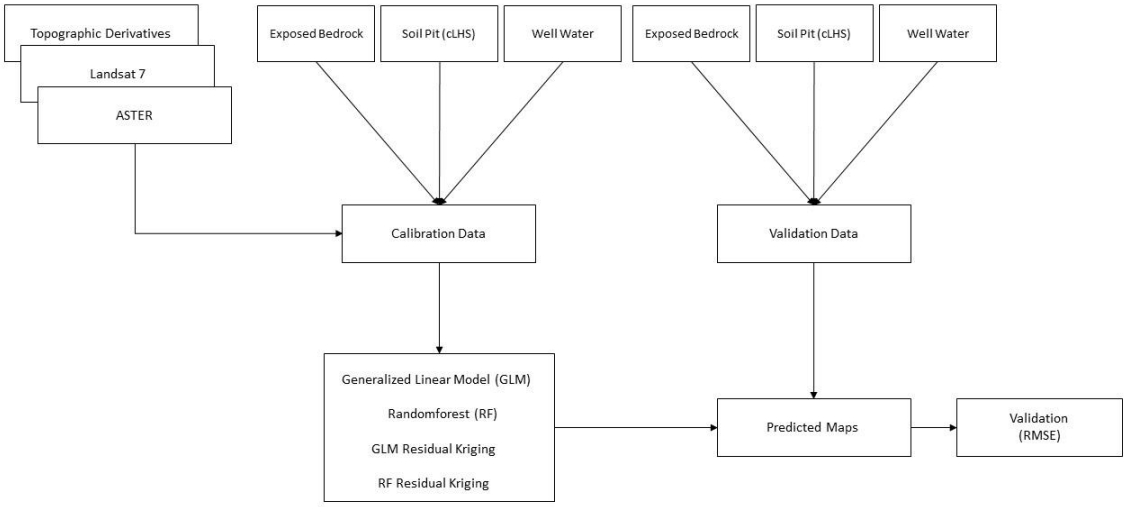


Figure 3-1 Workflow diagram for predictive mapping of soil depth using topographic indices, remotely sensed imagery and GLM, RF, GLMRK, and RFRK modelling methods. Calibration data consisted of well water data, in-situ soil depth measurements and exposed bedrock areas. Validation data were a subset of the calibration data and assessed the prediction accuracy of each model using RMSE values.

3.3.1. Study Area

The Tulameen study area is located in the south central interior of British Columbia, Canada (N 49°32' W 120°45') (Figure 3-2). The area occupies 3435 km^2 of primarily coniferous forest in the Cascade Dry Belt and Thompson Plateau. The Cascade Mountains here are considered to be part of the Coast Mountain range (Holland, 1976) and elevations range from 623 to 2337 masl. The biogeoclimatic zones that are found in this region are the Coastal Western Hemlock, Mountain Hemlock, Alpine Tundra, Interior Douglas-Fir, and the Engleman Spruce zones. Monthly average temperatures range from -12 to 27º C with an average of 550 mm annual precipitation (B.C. Ministry of Parks, 1999).

The majority of the soils in this region developed after the recession of the Wisconsin glaciation, approximately 12,000 years ago. The dominant parent material in the region is glacial till, with some areas of glacial fluvial and glacial lacustrine deposits. Typical soil types found in this region are Dystric Brunisols in the higher elevations and Humic Podzols in the lower elevations; soil pH ranges from 3.6 to 5.2 (Fraser et al., 1989). The depth of the soil in this region is highly variable, in part because glacial erosion and deposition during the Pleistocene Epoch likely resulted in six times more movement and deposition of material than subaerial erosion (Church & Ryder, 2010).

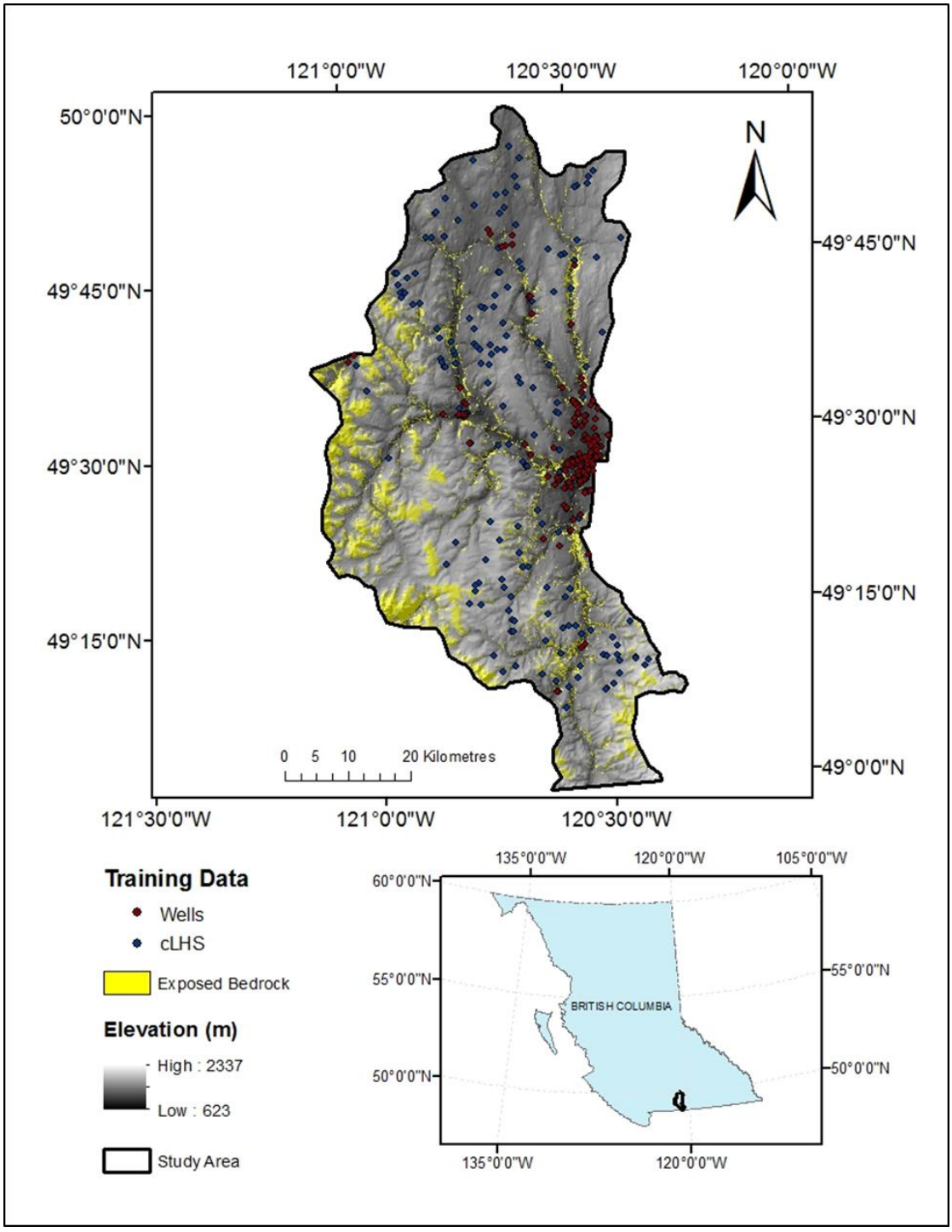


Figure 3-2 Map of the Tulameen study area in southern interior of British Columbia, Canada.

3.3.2. Soil Depth Data

Soil depth data were obtained from three independent sources: provincially regulated well water records, in-situ depth measurements from soil pits, and a predictively modeled EB layer. Due to the methods with which these datasets were created, they are inherently clustered, primarily near roads or populated areas. The creation of the EB data layer is briefly described below, full details were outlined in Chapter 2.

3.3.2.1 Exposed Bedrock Layer

A 100 m resolution map of EB for the study area was created using predictive modeling in a separate study (Chapter 2), with an overall accuracy of 88%. Calibration data for the EB map included land cover maps, a DEM, satellite-derived spectral indices from both Landsat 7 and ASTER as well as vector data for geology, vegetation and climatic zones. Modelling was facilitated by user interpretation of imagery with comparison to legacy cover data, and was validated with high-resolution imagery. Sample points for the present study were generated from the centroid of each EB cell (42,823 total cells were available) and were assigned a soil depth value of 0 m.

3.3.2.2 Well Water

The Province of British Columbia has made it mandatory through the Water Act that all private and public wells have a well report produced. Each report details artesian flow rates, limnology and depth to bedrock, along with GPS coordinates (http://www.env.gov.bc.ca/wsd/data_searches/wells/). All well data (568 points) in the Tulameen area were accessed and the depth to bedrock values were extracted; however, only 239 of the wells had complete reported depth measurements and were used in further analysis.

3.3.2.3 In-situ Soil Depth Data

Field measurements of soil depth were made at locations chosen using conditioned latin hypercube sampling (cLHS) (Minasny & McBratney, 2006). cLHS is a modified sampling technique based on the latin hypercube (LHS) initially discussed by McKay et al. (1979). The cLHS is a stratified random sampling technique that can sample across multivariate distributions of both discrete and continuous data. Equally probable strata are created for each covariate, where the number of strata is determined by the number of samples defined by the user. One sample value is randomly chosen from within each strata. cLHS ensures that samples are found in data feature space (Minasny & McBratney, 2006).

Eight predictor variables were chosen to represent the feature space for the Tulameen study area which were elevation, slope, total insolation, NDVI (Landsat 7), topographic profile index (TPI), plan-curvature, profile-curvature, and multiresolution index of valley bottom flatness (MRVBF).

A 150 m buffer was applied to all roads, logging routes and fire access roads in the study area. The buffer was applied to reduce the total potential area to be sampled, enabling more samples to be collected with shorter distances to travel, while still capturing the variability of the landscape. 200 sample points were chosen due to time constraints associated with field collection. Other studies have used a similar sample size, with suggestions that 200 to 300 sample points are representative for soil field studies using the cLHS method of sampling (Brungard & Boettinger, 2010).

3.3.2.3.1 Soil Depth Measurements

174 of the 200 points were sampled in the field, while the remaining 26 points were either inaccessible or located in dangerous areas. A one metre soil pit was excavated at each of the 174 points if possible. When soil depth (depth of

unconsolidated material to solid bedrock) was less than 1 m, the depth was recorded. When solid bedrock was not encountered in the 1 m profile, topographic details of the sampled area were evaluated and a generalized depth was visually estimated for the 1 hectare area that the plot represented. All pits regardless of depth had relevant site conditions recorded. Field information from local soil surveys were also used to assist in making soil depth estimates (Lord & Green, 1974).

An example of the method used in the field collection process for soil areas with a depth greater than 1 m is shown in Figure 3-3. Here the sample point was located on a 20° slope adjacent to the road. Bedrock was not found within the 1 m soil pit (Figure 3-3A). This point (Figure 3-3B) was found to be at a mid-slope position with a straight surface shape (slanting hill from right to left) with minor mounds. The soils here were derived from a morainal blanket deposit. Due to the presence of visible exposed bedrock in the area surrounding the pit, and the pit being located on a relatively steep slope, close to the peak of a local mountain, soil depth was estimated as 3 m for the 1 hectare grid cell that it represented.

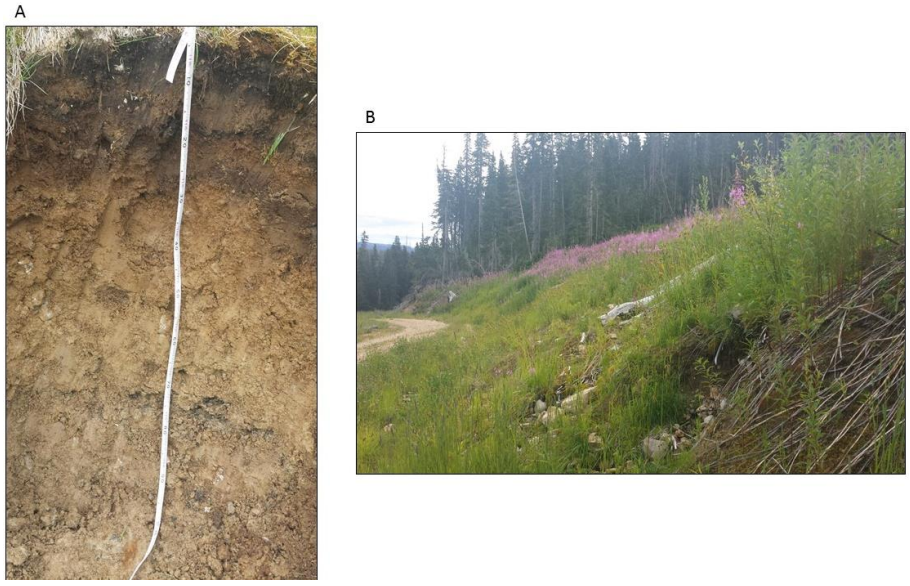


Figure 3-3 Images displaying the estimation of soil depth in the field. A) A 100 cm deep soil pit. B) The slope position of the sample point and the surrounding area.

3.3.3. Environmental Variables

Environmental predictor variables included both topographic variables derived from a DEM and remotely sensed variables (Table 3-1). A 100 m DEM (Hectares BC, 2012) was used to generate all topographic variables. Remotely sensed variables were used to quantify visible surficial mineral presence, vegetation cover and other land cover. Images were acquired from Landsat 7 and the ASTER satellites.

Table 3-1 List of 36 topographic and remotely sensed indices used in the GLM, RF, GLMRK, and RFRK models.

Landscape Representation	Terrain Derivatives	Code	Reference
DEM	Elevation	Elevation	
	Topographic Ruggedness Index	TRI	(Riley et al. 1999)

	Valley Depth	ValleyD	SAGA Development Team (2011)
	Topographic Profile Index	TPI	(Guisan et al. 1999)
	Slope	Slope	(Zevenbergen & Thorne, 1987)
	Multi-Resolution Valley Bottom Flatness	MRVBF	(Gallant & Dowling, 2003)
	Multi-Resolution Valley Bottom Flatness-Kilometre	MRVBF_KM	(Gallant & Dowling, 2003)
	Mulit-Resolution Ridge Top Flatness	MRRTF	(Gallant & Dowling, 2003)
	Slope (in radians)	Slope	(Zevenbergen & Thorne, 1987)
	Slope Height	SlopeH	(Boehner & Conrad 2008)
	Standardized Height	StandH	SAGA Development Team (2011)
	Profile Curvature	ProfileC	(Zevenbergen & Thorne, 1987)
	Plan Curvature	PlanC	(Zevenbergen & Thorne, 1987)
	Normalized Height	NormHeight	SAGA Development Team (2011)
	Mid Slope Position	MSP	SAGA Development Team (2011)
	Direct Insolation	DirectInso	(Böhner & Antonić, 2009)
	Diffuse Insolation	DiffuseInso	(Böhner & Antonić, 2009)
	Convergence Index	ConIndex	(Koethe and Lehmeier, 1996)
	Channel Network Base Level	ChanNet	SAGA Development Team (2011)
	Catchment Area	CatchArea	SAGA Development Team (2011)
	Altitude above Channel Network	AltAC	SAGA Development Team (2011)
	Slope Length Factor	LSFactor	(Moore et al. 1993)
	Topographic Wetness Index	TWI	(Beven & Kirkby, 1979)
	Ridge Hill Slope Position-Hectare	RHSP_HA	(MacMillan, R.A, 2005)
	Ridge Hill Slope Position-Kilometre	RHSP_KM	(MacMillan, R.A, 2005)
	Mass Balance Index	MB_IND	(Friedrich, K. 1996)
	Topographic Profile Index Classification	Landforms	(Guisan et al., 1999)
Landsat 7	Normalized Difference Vegetation Index	NDVI	(Tucker, 1979)
	Principal Component Analysis (7 Bands, First PC)	LandsatPCA	

ASTER	**Visible and Near Infrared Principal Component Analysis (Bands 1-3, First PC)**	**VNIRPCA**	
	Normalized Difference Vegetation Index (Bands 3 and 2)	**AsterNDVI**	
	Iron Oxides	**ironoxides**	**(Kilby & Kilby, 2006)**
	Sericite and Illites	**Sericite**	**(Kilby & Kilby, 2006)**
	Siliceous Rocks	**Siliceous**	**(Kilby & Kilby, 2006)**
	Short Wave Infrared Prinicpal Component Analysis (Band 5-10, First PC)	**SWIRPCA**	
	Thermal Infrared Principal Component Analysis (Bands 11-15, First PC)	**TIRPCA**	

3.3.4. Calibration and Validation Data Samples

3.3.4.1 Calibration Data

Of the 42,823 EB, 239 WW and 174 soil pit data points, a randomly selected, equal weighted subset of 100 EB, 100 WW and 100 soil pit points were used for model calibration. The EB points represent the shallowest depths, whereas, the WW point represent the deepest depths and the soil pit points are mid-range depth points. An alternative calibration data set, increasing the number of EB data points from 100 to 5000, was also created. The increased EB points were paired with the 100 WW and 100 soil pit points to test the effect of using this abundant but specific (depth = 0 m) data source on model behaviour. Greater sample sizes have been shown to increase estimation precision for deterministic models (Mendez & Lohr, 2011) and for variogram predictions when applied to residual kriging, however it is often unknown how large a sample needs to be in order to achieve optimal precision (Webster & Oliver, 2007). Values of all

environmental predictors (Table 3-1) were derived for all data points to form the calibration data set.

3.3.4.2 Validation Data

In order to assess model performance, a random subset of equal weighted samples was created from the WW, soil pit, and EB points that were not used in the generation of the models. Validation samples consisted of 74 well water points, 74 soil pit points, and 74 randomly sampled EB points for a total of 222 validation points.

3.3.5. Modelling Approaches

Four modelling approaches were used to estimate soil depth from the calibration data sets: GLM, RF in regression mode, GLM with residual kriging, and RF with residual kriging. Calibration data for the RF models were not normalized because RF is a non-parametric model, while predictors were log-transformed as necessary prior to use in the GLM models. All residuals used for kriging were checked for normality (Hengl et al., 2004) using a Shapiro-Wilk and Anderson Darling test and log-transformed as necessary prior to kriging. When residuals were below zero, the absolute value for the lowest negative residual was added to all residuals in order to allow log transformation prior to normalization. Residuals were normalized in order to meet all assumptions for kriging.

3.3.5.1 Generalized Linear Model

The generalized linear model (GLM) is an iterative weighted regression model that does not require the data be normally distributed (Nelder & Wedderburn, 1972). To ensure that the classical assumptions of linear models are still met, a link function is incorporated into the GLM. The link function establishes a connection between the mean of the dependant variables and the array of

independent variables. The dependant variable is scaled to the independent variables through the link function to allow for the effects of the model to be additively combined (Venables & Dichmont, 2004). For this study, the Gaussian distribution with the identity link function was used. This implied that my data followed a normal distribution (Lane, 2002). The GLM models were created in the R statistical program (R Development Core Team, 2012) with the glm package.

3.3.5.2 Random Forest

RF is an ensemble of decision trees capable of being used in either regression mode or classification mode, and can accept both categorical and continuous data (Breiman, 2001). Individual trees are grown from a bootstrapped sample (default 63.2%) of calibration data, and are allowed to grow to the largest extent without pruning to ensure that the largest amount of variance can be expressed. The remaining 36.8% (the 'out of bag', OOB, data) are used in an internal validation test that, in regression mode, estimates the mean square error (MSE) of predictions. Variable importance is also generated from the OOB data, ranking variables according to their influence on model predictions. Variable importance is measured in two forms, of which the mean decrease in accuracy (MDA) was used here. The MDA measures the increase in prediction error when permuting the values of a variable in the OOB data, thus effectively removing that variable's information content. The greater the increase in prediction error, the greater the importance of the variable (Liaw & Wiener, 2003). The RF models were created in the R statistical program (R Development Core Team, 2012) with the randomForest package (Liaw and Wiener, 2002).

3.3.5.3 Residual Kriging

Residual kriging was used to correct for any spatial trend that may exist in residuals from the GLM and RF models. Residuals are created and are used to generate a variogram. Residuals are interpolated with ordinary kriging as these

observations are then considered independent (Bivand et al., 2008). The predicted layer from the GLM or RF model is then added to the kriged residual layer to create the final predictions (Odeh et al., 1995; Odeh et al., 1994). Uncertainty is incorporated into the model through the addition of residuals from the GLM or RF that represents drift for the data (Hengl et al., 2004). The structure of the variogram (Gaussian, Spherical, and or Exponential) was chosen by comparing mean square value (MSE) from the cross validation results and choosing the structure with the smallest MSE (Oliver & Webster, 2014).

3.4. Results and Discussion

3.4.1. Descriptive Statistics

Calibration data for the models (Figures 3-4A and 3-4B) are positively skewed due to the relative abundance of EB points, all with depth = 0 m. EB only covers 438 km2, which is 7.8% of the landscape (Chapter 2) but currently represents approximately 33% of the calibration data (Figure 3-4A), or approximately 96 % if 5000 EB points are used (Figure 3-4B). The effect of the skew on mean depth in the calibration data increases significantly from the data set with 300 points (mean depth = 7.6 m) to the data set with 5200 calibration points (mean depth = 0.44 m). The average soil depth for the Tulameen area is 5.5 m (Lord & Green, 1974). Soil depths for the calibration data range from 0 m to 93.3 m. This is a relatively large range compared to other studies, which typically only include depths ranging from 0 m to 5 m (Kuriakose et al., 2009; Tesfa et al., 2009).

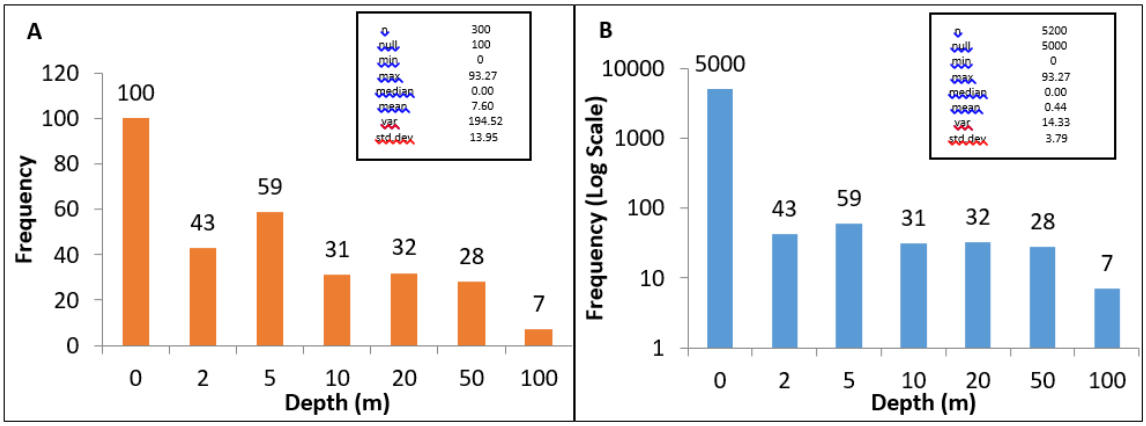


Figure 3-4 Frequency of the calibration data and descriptive statistics A) The equal weighted calibration points with 300 total points. B) The experimental calibration dataset with 5200 total points.

3.4.2. Variograms

The variogram was used to represent the reduction in spatial autocorrelation from the residual values of RF and the GLM in comparison to the target (observed depth) calibration data (Figure 3-5). The 300-point calibration data set (Figure 3-5A) had a fairly large spatial variance with an extent of 1.0 to 2.5. There is a general trend of increased variance with the target calibration data the farther away the points are. When comparing the residuals from both RF and GLM to the target variance, there is a substantial drop in variance from the target that implies a large portion of the overall variance was removed by the models (Odeh et al., 1995; Odeh et al., 1994). The noise in both the RF and GLM models is seen with the increase in spatial variance and then a sudden decrease at distances greater than 10,000 m. The noise being expressed indicates that there is not enough variation explained from the small sample sizes, which is influenced by the very heterogeneous landscape being mapped.

The 5200 target calibration data (Figure 3-5B) had a fairly small variance with an extent of 0.04 to 0.4. This smaller range is largely due to the large number of points in this data set and the significant clustering that is inherent with these selected calibration data. There is a small difference in the target calibration data variance until 23,000 m, where a large increase in variance can be seen. Overall the trend linearly increases in variance at greater distances for the target calibration data and GLM and RF residuals. Both the GLM and RF residuals see a substantial drop in variance from the target calibration data, implying that the trend has been removed (Odeh et al., 1995; Odeh et al., 1994). The variance is almost 0 for both the GLM and RF residuals, implying that the point density is useful for kriging and that the residuals have had most of their spatial autocorrelation removed.

When comparing both variograms (Figures 3-5A and 3-5B) there is a substantial drop of variance from the 300 to the 5200 target calibration data as the extent for variance drops from 0.3 to 2.5, to 0.04 to 0.4. The GLM and RF residuals using 5200 calibration points has substantially lower variance compared to the target of the 300 calibration point models. It is important to note that these are only the residuals of the models. The layers of depth have already been predicted and the models themselves hold most of the prediction power. The kriging residuals are designed to remove uncertainty trend that the landscape may have. This is important for landscape scale procedures as there could be multiple localized trends found within a larger landscape. With the addition of the RK it is seen that most of the spatial autocorrelation has been removed from the GLM and RF models, implying that the predicted maps would have higher accuracies due to the removed uncertainty from the kriged residual layer.

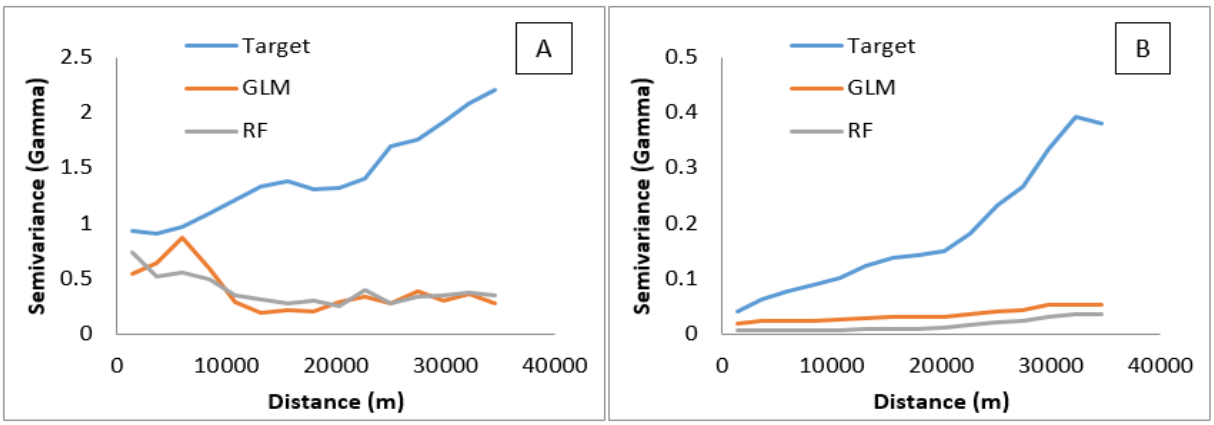


Figure 3-5 **Experimental variogram of observed depth calibration data (Target) and their associated residuals with GLM and RF A) 300 point calibration data, and B) 5200 point calibration data.**

3.4.3. Soil Depth Maps, Model Accuracy, and Variable Importance

3.4.3.1 Soil Maps

The resulting maps of soil depth (Figure 3-6) show the expected trend for depositional landscapes: valley bottoms and lower areas in the catena have deeper soil depths where more deposition of material can take place (Heimsath et al, 2001; Heimsath et al., 1997), and areas higher in the catena and closer to the peaks of mountains have little to no material, implying that rates of erosion have exceeded soil production and have left predominantly EB (DiBiase et al., 2012). Maximum ranges vary for each map according to prediction method and the residual kriging that was applied to each map (Tesfa et al., 2009).

Due to the effect that different models have, maximum soil depths range from 19.1 m (GLM5200) to 78.3 m (GLMRK300). Generally, the GLM maps (Figure 3-6A and 3-6B) have low maximum depths. This is rectified with the addition of RK (Figure 3-6C and 3-6D) as the errors contained in the residuals correct this shorter range. GLMRK300 (Figure 3-6C) is seen to have the most representative range of

all the maps. The range for calibration data is 0 m to 93.3 m and GLMRK300 has a range of 0 m to 78.4 m. All other maps have a range from 0 m to a maximum of 50-56 m. These maps do not fully capture the maximum range of depths for the Tulameen region.

Values at the ends of the depth range are susceptible to large over- and under-prediction errors (Hengl et al., 2007) indicating that deeper depths and shallower depths could have more error than mid-range depths. Under-prediction was easily identified with all kriging models, which included values below zero. To account for the prediction error, such negative soil depths were converted to a depth of 0 m (Sarkar et al., 2013). Generally the GLM models suffered less from this problem (never predicting below -1 m) than the RF models (predicting down to -17 m).

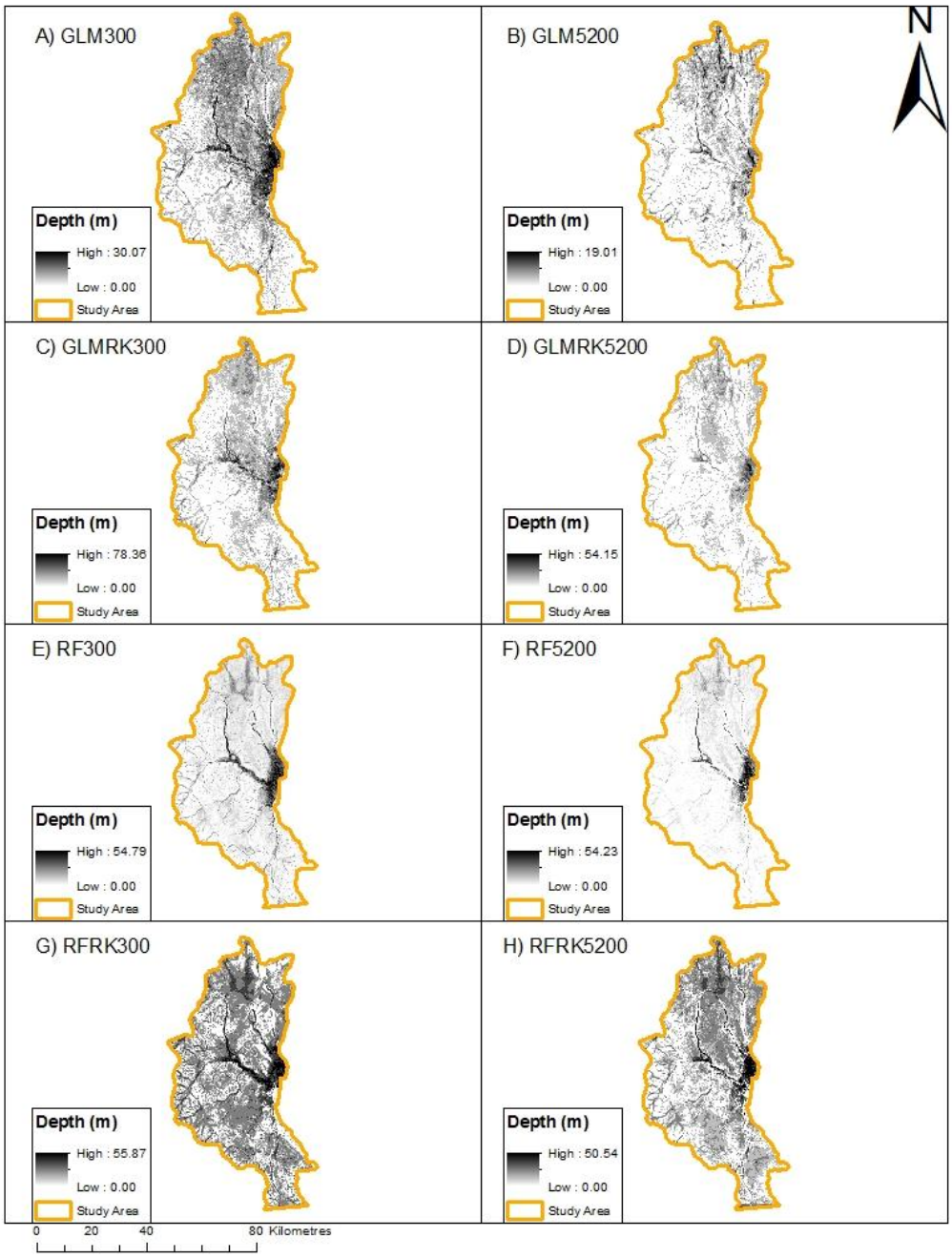


Figure 3-6 The 8 predicted soil depth (m) maps for the Tulameen study area. Depth maps are named after the model used to predict them and the number of EB points used A) GLM300 is the GLM model using 300 calibration points B) GLM5200 is the GLM model using 5200 calibration points C) GLMRK300 is the GLM model with residual kriging using.

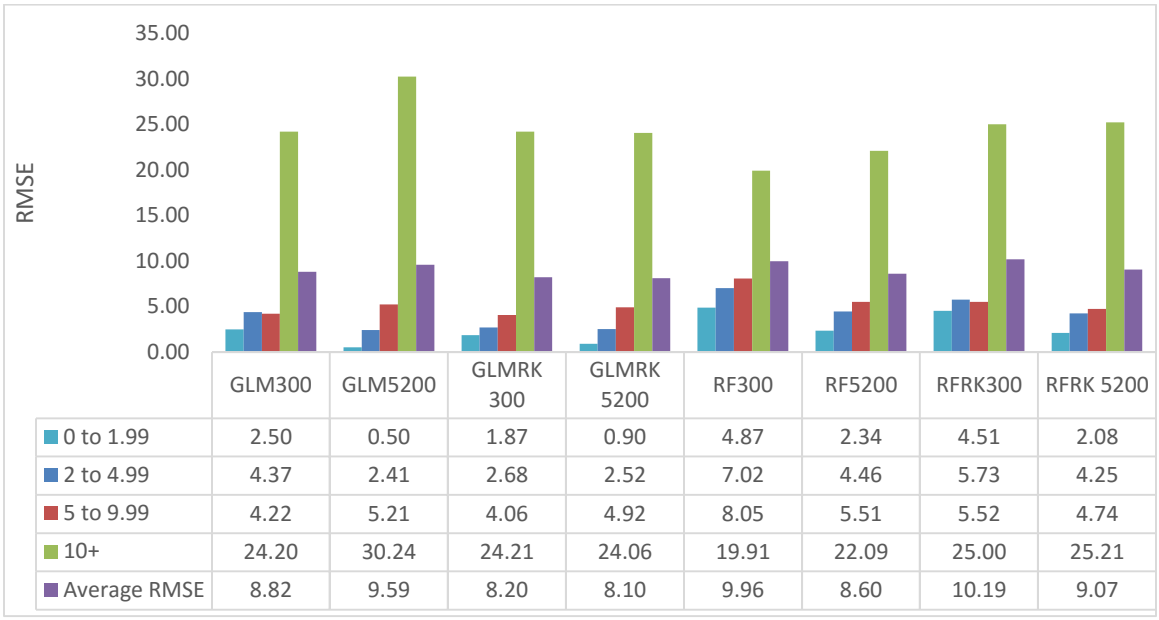
3.4.3.2 Model Fit and Prediction Accuracy

The model GLMRK5200 had the lowest average RMSE value (Figure 3-7). The GLMRK300 model performed second best in terms of lowest average RMSE especially considering the smaller calibration dataset used. Increasing the total amount of EB points increased prediction accuracy for all models and depths up to 5 m (Figure 3-7).

The use of EB as a depth proxy allows for an increased sample size without additional sampling through ancillary data or in-situ sampling. This indicates the strength of EB as an effective proxy for depth modelling in the future. The RF models with 5000 EB points had lower average RMSE than did corresponding models with 100 EB points, while the GLM models did not (Figure 3-7). RF models performed very well with regards to explained variance for model prediction in comparison to the GLM models (Table 3-2). However, the GLM models with RK outperformed all RF models in terms of RMSE values (Figure 3-7) showing that higher r2 values, implying greater model fit, do not always indicate a better prediction.

Table 3-2 Model coefficients of determinations (r2) for calibration data and predicted results for both GLM and RF.

Model Type	r^2
GLM 300	0.28
GLM 5200	0.20
RF 300	0.86
RF 5200	0.88



	GLM300	GLM5200	GLMRK 300	GLMRK 5200	RF300	RF5200	RFRK300	RFRK 5200
0 to 1.99	2.50	0.50	1.87	0.90	4.87	2.34	4.51	2.08
2 to 4.99	4.37	2.41	2.68	2.52	7.02	4.46	5.73	4.25
5 to 9.99	4.22	5.21	4.06	4.92	8.05	5.51	5.52	4.74
10+	24.20	30.24	24.21	24.06	19.91	22.09	25.00	25.21
Average RMSE	8.82	9.59	8.20	8.10	9.96	8.60	10.19	9.07

Figure 3-7 **Validation results using RMSE values (m) for the eight depth maps created for this study. Four models were used: GLM, RF, GLMRK and RFRK. Each model used two sets of calibration data, a 300 point calibration dataset and a 5200 point calibration dataset. Results are named after the model type and the calibration dataset used. Each layer was validated for 4 depth intervals and an average RMSE value for all depths. The number of points to validate each group are presented in the top right portion of the chart.**

The 0 m to 4.99 m soil depth range is very important because this is where the majority of biological activity takes place in the pedosphere (Al-Agely & Reeves, 1995; Canadell et al., 1996). A large number of EB points likely over-fitted the models with an imbalance of 0 m data allowing for the shallower depths to be favoured in the predictions (Chen et al., 2004; Van Hulse & Khoshgoftaar, 2009). As a consequence, deeper depths (5 m and greater) are underrepresented in the predictions, resulting in larger error rates for these depths. However, the addition of EB points greatly reduces each model's error for the 0 m to 5 m range when comparing the 300 models to the 5200 models of the same model type. Addition of a large number of EB points may prove similarly useful in other areas where soil

depth mapping is focused on this shallow depth range (Kuriakose et al., 2009; Tesfa et al., 2009).

Overall, RK had a mixed effect for reducing reported errors. The greatest reductions with RK are seen in the 0 m to 4.99 m range, with the exception of the GLM5200 models. These models had lower errors in the 0 m to 4.99 m depths, but significantly higher errors for depths >5 m. Part of the reduction in error can be associated with the removal of the spatial variances that is seen with all models from the reported variograms (Figure 3-5). Variograms do not look at reported values (depth ranges), but rather are designed to visualize and quantify spatial autocorrelation in the landscape. This is why variograms cannot directly relate the removal of spatial uncertainty to having better predicted maps. Variograms express that spatial uncertainty has been removed, however these are associated with modeled residuals, which can introduce further modelling error if the models used are inaccurate. This is reported in Figure 3-7, with lower RMSE values for a majority of the RK but the reduced accuracy is especially apparent with the RFRK predictions in comparison to the GLMRK maps as the modelled residuals for RF had much larger ranges than the GLM.

3.4.3.3 Variable Importance and Environmental Influences on Soil Depth

Deeper soil depths are usually associated with more transport-limited landscapes, and thus less variability can be explained by the environmental predictors that are used. At the landscape scale, glacial and periglacial activity override the normal debris flow equilibrium. Slope and convexity relations typical of un-glaciated environments do not always apply here (Brardinoni & Hassan, 2006). Soil production rates are higher where soil depth is shallower. Slopes that are too steep (greater than 30°) could contribute to a net loss of soil as the erosion rate is greater than the rate of soil production (DiBiase et al., 2012; Heimsath et al., 2012). As soil becomes deep, soil production decreases, and soil depth is no

longer a function of slope and convexity as explored through process-based models (Dietrich et al., 1995; Heimsath et al., 1999; Minasny & McBratney, 1999; Pelletier & Rasmussen, 2009). However, at a landscape scale these deeper soils differ from previous findings in heavily studied convexity and slope studies, as local processes of soil production become harder to quantify with DEMs and topographic derivatives. This is due to the fact that determinants that predict soil erosion and deposition on a local scale do not respond to determinants on the landscape scale (Brardinoni & Hassan, 2006). Depths in the Tulameen region are closely related to the deposition of sediment from glaciers that differ from typical expectation of local lithological sediment exchange (Brardinoni et al., 2009).

Figure 3-8 is the MDA chart generated from the RF model (5200 calibration points). Variables are ranked based on the MSE increase caused by their permutation. In this landscape-scale model, slope is ranked 30th and planform- and profile-curvature are also ranked very low. It is clear that on a landscape scale, variables such as channel network base level (prediction for water table heights in the landscape), elevation and MRVBF (indicator for areas of high erosion and deposition) which indicate regions of deposition are better predictors for depth than previously described slope and convexity. This indicates that other site specific topographic variables affect the distribution of material on a smaller scale (Behrens et al., 2010). Further research is needed on the interactions of glacial deposits and soil distribution as it relates to soil depth.

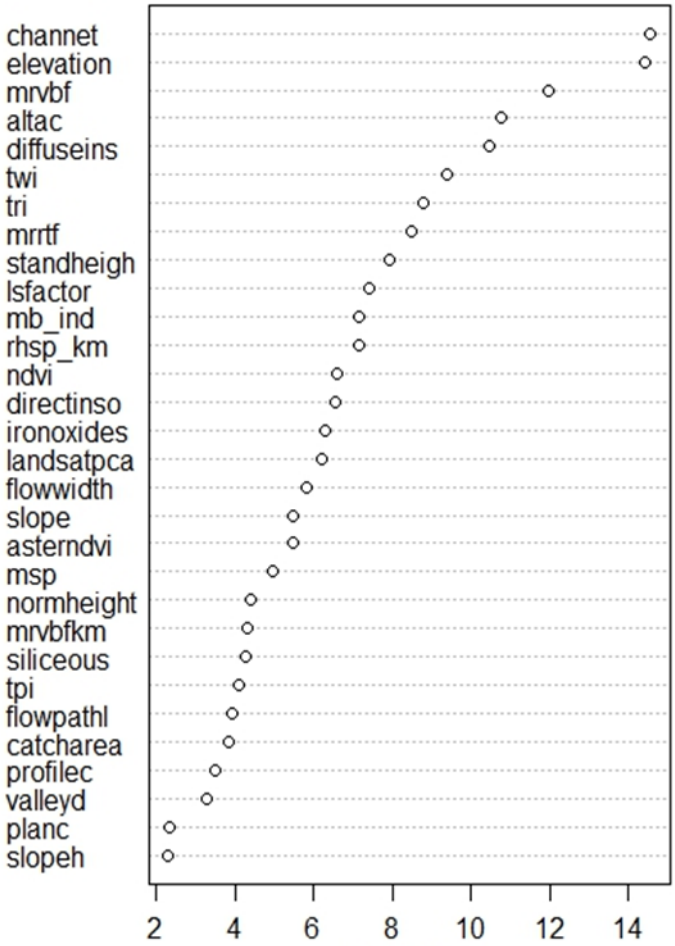


Figure 3-8 Mean Decrease Accuracy (MDA) plot from the RF model with 5200 calibration points. This plot shows relative importance for predictor variables in the RF model. Higher values imply a greater reduction in model accuracy if that variable is not included.

3.5. Conclusion

This paper presented the use of GLM, RF, GLMRK and RFRK as methods to predict soil depth. The modelling of soil depth is important for a better understanding of the CZ. Not many studies have been conducted on mapping soil depth on a landscape scale, especially to depths greater than 5 meters. Data that are easy to access and can be acquired remotely is noted as a critical need for

mapping soil depth (Riebe & Chorover, 2013). In this study, novel data inputs were used in order to supplement the lack of available data for mapping soil depth. WW data provided a free, accurate and relatively abundant source of data at deeper depths. The soil pit information collection through a cLHS method proved to be useful to include calibration data at moderate soil depths that are rarely found in WW data (deeper soils) or EB data (by definition depth = 0 m). EB points also allowed for far more accurate models to be generated for the 0 to 4.99 m depth range as an abundance of points could be generated. The inclusion of EB to constrain model predictions at the shallowest depths created more sample points for kriging, which produced maps that were more accurate.

When analyzing soil depth on a landscape scale, it is also important to understand different predictors affecting the deposition and erosion of soil material. My study found that the top 3 predictors for soil depth were the relationship to a channel network base level, elevation and MRVBF (indicator for valley bottoms and deposition). Due to the multiple scales at which the depth of the pedosphere can be mapped, it is important to further understand which predictors are relevant at each scale.

When comparing the three models used, GLMRK was the superior method. RF is often cited as a better prediction method for soil attributes, however, with the introduction of kriging, my results showed that GLMRK models can produce better maps. RMSE values remained relatively consistent until greater than 4.99 m depths. Here it is hypothesized that the models were unable to accurately predict these depths because deeper depths do not correlate with topographic variables as much as shallower depths. The addition of RK was found to reduce RMSE for most models, at most depths. The RMSE values that were obtained in this study suggest that soil depth has been effectively modelled and mapped with limited data. The maps of soil depth can be beneficial for further study of soils within the

CZ. This study has highlighted other determinants can help to predict soil deposition, erosion and production on a landscape scale. These findings convey that the CZ should be analyzed as a multiscale phenomenon to better understand how each predictor influences the depth of the pedosphere.

3.6. References

Al-Agely, A., & Reeves, F. (1995). Inland Sand Dune Mycorrhizae: Effects of Soil Depth, Moisture and pH on Colonization of Oryzopsis hyenoides. Mycologia, 87(1), 54–60.

Behrens, T., Zhu, A.-X., Schmidt, K., & Scholten, T. (2010). Multi-scale digital terrain analysis and feature selection for digital soil mapping. Geoderma, 155(3-4), 175–185. http://doi.org/10.1016/j.geoderma.2009.07.010

Bivand, R. S., Pebesma, E., & Gómez-Rubio, V. (2008). Applied spatial data analysis with R.

Brady, N. C., & Weil, R. R. (2007). The Nature and Properties of Soils (14th ed.). New York, New York, USA: Macmillian.

Brantley, S. L., Goldhaber, M. B., & Vala Ragnarsdottir, K. (2007). Crossing disciplines and scales to understand the critical zone. Elements, 3(5), 307–314. http://doi.org/10.2113/gselements.3.5.307

Brardinoni, F., & Hassan, M. A. (2006). Glacial erosion, evolution of river long profiles, and the organization of process domains in mountain drainage basins of coastal British Columbia. Journal of Geophysical Research: Earth Surface, 111(1), 1–12. http://doi.org/10.1029/2005JF000358

Brardinoni, F., Hassan, M. A., Rollerson, T., & Maynard, D. (2009). Colluvial sediment dynamics in mountain drainage basins. Earth and Planetary Science Letters, 284(3-4), 310–319. http://doi.org/10.1016/j.epsl.2009.05.002

Breiman, L. (2001). Random forests. Machine Learning, 5–32.

Brungard, C. W., & Boettinger, J. L. (2010). Conditioned Latin Hypercube Sampling: Optimale Sample Size for Digitial Soil Mapping of Arid Rangelands in Utah, USA. In J. L. Boettinger, D. W. Howell, A. C. Moore, A. E. Hartemink, & S. Kienast-Brown (Eds.), Progress in Soils Science 2. Dordrecht: Springer Netherlands. http://doi.org/10.1007/978-90-481-8863-5

Canadell, J., Jackson, R. B., Ehleringer, J. B., Mooney, H. A., Sala, O. E., & Schulze, E.-D. (1996). Maximum rooting depth of vegetation types at the global scale. Oecologia, 108(4), 583–595. http://doi.org/10.1007/BF00329030

Chen, C., Liaw, A., & Breiman, L. (2004). Using random forest to learn imbalanced data. University of California, Berkeley, (1999), 1–12.

Chorover, J., Kretzschmar, R., Garica-Pichel, F., & Sparks, D. L. (2007). Soil biogeochemicial processes within the critical zone. Elements, 3(5), 321–326. http://doi.org/10.2113/gselements.3.5.321

Church, M., & Ryder, J. (2010). Physiography of British Columbia. In R. Pike, T. Redding, D. Moore, R. Winkler, & K. Bladon (Eds.), Compendium of Forest Hydrology and Geomorphology in British Columbia (pp. 17–46). Victoria.

DiBiase, R. A., Heimsath, A. M., & Whipple, K. X. (2012). Hillslope response to tectonic forcing in threshold landscapes. Earth Surface Processes and Landforms, 37(8), 855–865.

Dietrich, W. E., Reiss, R., Hsu, M. L., & Montgomery, D. R. (1995). A process-based model for colluvial soil depth and shallow landsliding using digital elevation data. Hydrological Processes. http://doi.org/10.1002/hyp.3360090311

Field, J. P., Breshears, D. D., Law, D. J., Villegas, J. C., López-hoffman, L., Brooks, P. D., … Troch, P. A. (2015). Critical Zone Services : Expanding Context , Constraints , and Currency beyond Ecosystem Services. Vadose Zone Journal. http://doi.org/10.2136/vzj2014.10.0142

Fraser, D., Farr, D., Ramsay, L., & Turner, N. (1989). Natural and Human Histroy Interpretive Theme Document for Manning Provincial Park. Victoria.

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

Heimsath, A M., Dietrich, W. E., Nishiizumi, K., & Finkel, R. C. (2001). A: Stochastic processes of soil production and transport: erosion rates, topographic variation and cosmogenic nuclides. In the Oregon Coast Range. Earth Surface Processes and Landforms, 26, 531–52.

Heimsath, A. M., DiBiase, R. A., & Whipple, K. X. (2012). Soil production limits and the transition to bedrock-dominated landscapes. Nature Geoscience, 5(3), 210–214. http://doi.org/10.1038/ngeo1380

Heimsath, A. M., Dietrich, W. E., Nishiizumi, K., & Finkel, R. C. (1999). Cosmogenic nuclides, topography, and the spatial variation of soil depth. Geomorphology, 27, 151–172.

Heimsath, A. M., Dietrich, W. E., Nishiizumi, K., Finkel, R. C., Mass, A., & National, L. L. (1997). The soil production function and landscape equilibrium,Letters to Nature, 388(July), 358–361.

Hengl, T., Heuvelink, G. B. M., Kempen, B., Leenaars, J. G. B., Walsh, M. G., Shepherd, K. D., Sila, A., MacMillan, R. A., Jesus, J.M., Tamene, K., Tondoh, J. E. (2015). Mapping Soil Properties of Africa at 250 m Resolution: Random Forests Significantly Improve Current Predictions. PloS One, 10(6), e0125814. http://doi.org/10.1371/journal.pone.0125814

Hengl, T., Heuvelink, G. B. M., & Rossiter, D. G. (2007). About regression-kriging: From equations to case studies. Computers & Geosciences, 33(10), 1301–1315. http://doi.org/10.1016/j.cageo.2007.05.001

Hengl, T., Heuvelink, G. B. M., & Stein, A. (2004). A generic framework for spatial prediction of soil variables based on regression-kriging. Geoderma, 120(1-2), 75–93. http://doi.org/10.1016/j.geoderma.2003.08.018

Heung, B., Bulmer, C. E., & Schmidt, M. G. (2014). Predictive soil parent material mapping at a regional-scale: A Random Forest approach. Geoderma, 214-215, 141–154. http://doi.org/10.1016/j.geoderma.2013.09.016

Jin, L., Ravella, R., Ketchum, B., Bierman, P. R., Heaney, P., White, T., & Brantley, S. L. (2010). Mineral weathering and elemental transport during hillslope evolution at the Susquehanna/Shale Hills Critical Zone Observatory. Geochimica et Cosmochimica Acta, 74(13), 3669–3691. http://doi.org/10.1016/j.gca.2010.03.036

Karlsson, C. S. J., Jamali, I. A., Earon, R., Olofsson, B., & Mörtberg, U. (2014). Comparison of methods for predicting regolith thickness in previously glaciated terrain, Stockholm, Sweden. Geoderma, 226-227, 116–129. http://doi.org/10.1016/j.geoderma.2014.03.003

Kuriakose, S. L., Devkota, S., Rossiter, D. G., & Jetten, V. G. (2009). Prediction of soil depth using environmental variables in an anthropogenic landscape, a case study in the Western Ghats of Kerala, India. Catena, 79(1), 27–38. http://doi.org/10.1016/j.catena.2009.05.005

Lane, P. (2002). Generalized linear models in soil science. European Journal of Soil Science, 53(June), 241–251. http://doi.org/10.1046/j.1365-2389.2002.00440.x

Li, J., Heap, A. D., Potter, A., & Daniell, J. J. (2011). Application of machine learning methods to spatial interpolation of environmental variables. http://doi.org/10.1016/j.envsoft.2011.07.004

Liaw, A., & Wiener, M. (2003). Package “randomForest.” Retrieved December. Retrieved from http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Package+“+randomForest+”#5

Lin, H. (2010). Earth’s Critical Zone and hydropedology: Concepts, characteristics, and advances. Hydrology and Earth System Sciences, 14(1), 25–45. http://doi.org/10.5194/hess-14-25-2010

Lord, T., & Green, G. (1974). Soils of the Tulameen Area of British Columbia. Ottawa.

Ma, L., Chabaux, F., Pelt, E., Blaes, E., Jin, L., & Brantley, S. (2010). Regolith production rates calculated with uranium-series isotopes at Susquehanna/Shale Hills Critical Zone Observatory. Earth and Planetary Science Letters, 297(1-2), 211–225. http://doi.org/10.1016/j.epsl.2010.06.022

McKay, M. D., Beckman, R. J., & Conover, W. J. (1979). Comparison of three methods for selecting values of input variables in the analysis of output from a computer code. Technometrics, 21(2), 239–245. http://doi.org/10.2307/1271432

Mendez, G., & Lohr, S. (2011). Estimating residual variance in random forest regression. Computational Statistics and Data Analysis. http://doi.org/10.1016/j.csda.2011.04.022

Minasny, B., & McBratney, A. B. (1999). A rudimentary mechanistic model for soil production and landscape development. Geoderma, 90(1-2), 3–21. http://doi.org/10.1016/S0016-7061(98)00115-3

Minasny, B., & McBratney, A. B. (2006). A conditioned Latin hypercube method for sampling in the presence of ancillary information. Computers & Geosciences, 32(9), 1378–1388. http://doi.org/10.1016/j.cageo.2005.12.009

Ministry of Parks. (1999). Coquihalla Recreation Summit Master Plan.

Moraetis, D., Paranychianakis, N. V., Nikolaidis, N. P., Banwart, S. A., Rousseva, S., Kercheva, M., … Verheul, M. (2014). Sediment provenance, soil development, and carbon content in fluvial and manmade terraces at Koiliaris River Critical Zone Observatory. Journal of Soils and Sediments, 15(2), 347–364. http://doi.org/10.1007/s11368-014-1030-1

National Research Council. (2001). Basic Research Opportunities in Earth Science. Retrieved from http://www.nap.edu/catalog/9981.html

Nelder, A. J. A, & Wedderburn, R. W. M. (1972). Generalized Linear Models. J. R. Statist. Soc. A., 135(3), 370–384.

Odeh, I., McBratney, A., & Chittleborough, D. (1994). Spatial prediction of soil properties from landform attributes derived from a digital elevation model. Geoderma, 63(1994), 197–214. Retrieved from http://www.sciencedirect.com/science/article/pii/0016706194900639

Odeh, I. O. A., McBratney, A. B., & Chittleborough, D. J. (1995). Further results on prediction of soil properties from terrain attributes: heterotopic cokriging and regression-kriging. Geoderma. http://doi.org/10.1016/0016-7061(95)00007-B

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. http://doi.org/10.1016/0016-7061(94)90063-9

Oliver, M. A., & Webster, R. (2014). A tutorial guide to geostatistics: Computing and modelling variograms and kriging. Catena, 113, 56–69. http://doi.org/10.1016/j.catena.2013.09.006

Pahlavan Rad, M. R., Toomanian, N., Khormali, F., Brungard, C. W., Komaki, C. B., & Bogaert, P. (2014). Updating soil survey maps using random forest and conditioned Latin hypercube sampling in the loess derived soils of northern Iran. Geoderma, 232-234, 97–106. http://doi.org/10.1016/j.geoderma.2014.04.036

Pelletier, J. D., & Rasmussen, C. (2009). Geomorphically based predictive mapping of soil thickness in upland watersheds. Water Resources Research, 45(9), 1–15. http://doi.org/10.1029/2008WR007319

Riebe, C. S., & Chorover, J. (2013). Report on Drilling, Sampling, and Imaging the Depths of the Critical Zone, an NSF Workshop. Retrieved from http://criticalzone.org/sierra/publications/pub/riebe-and-chorover-2014-report-on-drilling-sampling-and-imaging-the-depths-/

Sarkar, S., Roy, A. K., & Martha, T. R. (2013). Soil depth estimation through soil-landscape modelling using regression kriging in a Himalayan terrain. International Journal of Geographical Information Science, 27(12), 2436–2454. http://doi.org/10.1080/13658816.2013.814780

Tesfa, T. K., Tarboton, D. G., Chandler, D. G., & McNamara, J. P. (2009). Modeling soil depth from topographic and land cover attributes. Water Resources Research, 45(10). http://doi.org/10.1029/2008WR007474

Van Hulse, J., & Khoshgoftaar, T. (2009). Knowledge discovery from imbalanced and noisy data. Data & Knowledge Engineering, 68(12), 1513–1542. http://doi.org/10.1016/j.datak.2009.08.005

Venables, W. N., & Dichmont, C. M. (2004). GLMs, GAMs and GLMMs: An overview of theory for applications in fisheries research. Fisheries Research, 70(2-3 SPEC. ISS.), 319–337. http://doi.org/10.1016/j.fishres.2004.08.011

Webster, R., & Oliver, M. (2007). Geostatistics for environmental scientists. Retrieved from http://books.google.com/books?hl=en&lr=&id=WBwSyvIvNY8C&oi=fnd&pg=PR11&dq=Geostatistics+for+Environmental+Scientists&ots=CAMqPMqJ_a&sig=mHzcLzcbVHkdWGQwuzt2JFOEhPU

Wiesmeier, M., Barthold, F., Blank, B., & Kögel-Knabner, I. (2010). Digital mapping of soil organic matter stocks using Random Forest modeling in a semi-arid steppe ecosystem. Plant and Soil, 340(1-2), 7–24. http://doi.org/10.1007/s11104-010-0425-z

Yu, X., Duffy, C., Baldwin, D. C., & Lin, H. (2014). The role of macropores and multi-resolution soil survey datasets for distributed surface-subsurface flow modeling. Journal of Hydrology, 516, 97–106. http://doi.org/10.1016/j.jhydrol.2014.02.055

Yu, X., Lamačová, A., Duffy, C. J., Krám, P., Hruška, J., White, T., & Bhatt, G. (2015). Modeling the long term water yield effects of forest management in a Norway spruces forest. Hydrological Sciences Journal, (July). http://doi.org/10.1080/02626667.2014.897406

Ziadat, F. M. (2010). Prediction of Soil Depth from Digital Terrain Data by Integrating Statistical and Visual Approaches. Pedosphere, 20(3), 361–367. http://doi.org/10.1016/S1002-0160(10)60025-2

Chapter 4. Conclusions

4.1. Thesis Conclusions

The focus of this research was to develop a methodology to model and map the location of EB and the depth of the pedosphere in the CZ. A landscape scale EB map was created with the use of a RF classifier. Modified PDPs were used to further analyze the probability of prediction for EB. The resulting EB map was a critical component for mapping the depth of the pedosphere in the CZ. GLM, RF, GLMRK and RFRK were each used to map the depth of the pedosphere and the results were compared. The results showed that with minimal data, accurate predictions for the CZ with RF and GLM could be made. In addition, EB was found to be an effective proxy for soil depths in remote regions with minimal data to facilitate depth mapping of the pedosphere.

In the first component (Chapter 2) of this thesis, RF was used to map the occurrence of exposed bedrock from legacy land cover data and topographical, satellite and land cover data. RF was used because of its ability to handle highly variable data, rank variable importance and generate partial dependence plots. Calibration data were created from legacy land cover data from vegetation inventories and land cover maps. Attributes from the vegetation inventory and land cover maps were selected to best represent exposed bedrock in the landscape.

Once the model had been run, variable importance was used to reduce the total amount of variables in the model without having an effect on prediction accuracy. The original model had 48 variables used to predict exposed bedrock and the final prediction only had 17. Both maps had a prediction accuracy of 88%. These results show that having more information for calibration does not always produce better maps.

Landsat PCA, Topographic Ruggedness Index (TRI), and a Normalized Difference Vegetation Index (NDVI) were explored to determine their influence on exposed bedrock prediction through the use of modified PDPs. It was found that higher values in the Landsat PCA (with high factor loadings of the near infrared band and thermal band) could produce prediction probabilities as high as 63%. This indicated that warmer areas lacking biomass are good indicators for EB. Through analysis of TRI, it was found that TRI values greater than 34 (implying highly rugged landscapes) could have prediction probabilities as high as 67%. Lastly NDVI values less than 0.17 showed prediction probabilities as high as 62%. PDPs outlined not only the influence that a given variable had, but also which values are the most important for predicting the occurrence of EB. This research allowed for a closer look inside the black box of RF to better understand the influence of individual variables on the prediction of EB occurrence.

In the second component of this thesis (Chapter 3), the depth of the pedosphere was modelled for the Tulameen region of Southern British Columbia. Calibration data included point data obtained from the EB map with soil depth set to 0 metres for randomly selected exposed bedrock pixels; well water data; and soil depth data collected in the field. Environmental variables used in the models included topographic and satellite indices to account for the local variability of the landscape. There was high spatial co-variance for the region when assessing the co-variance of the calibration data with semi-variograms. Both GLM and RF models had similar abilities to remove the co-variance for the landscape, implying that both models perform well when predicting soil depth.

It was found that the GLMRK model performed the best out of all the models for predicting soil depth. RMSE values for the 0 to 2 m range were 1.87 m for the equal weighted random sampling, and 0.9 m for the boosted sampling scheme. Deeper depth ranges had similar results for prediction accuracy, where GLMRK

was the superior model and RFRK was the second best. The RF model without kriging had the worst results, implying that deterministic models alone were unable to fully model the spatial structure in soil depth across the study area. These results showed that GLMRK and RFRK are both powerful tools to map landscape scale soil attributes such as soil depth. It also shows that data collected remotely, such as EB and WW data, are powerful supplements for soil depth modelling. This implies that more studies could use EB as a data input for mapping soil depth.

4.2. Future Research

This research has presented a new generic approach to quantify the CZ on a landscape scale, however this study's specific results are limited to the environmental conditions in the study area. The study has been conducted in a primarily mountainous region that has recently been glaciated in the last 15,000 years. Due to the topographical heterogeneity of the landscape (especially presented with the TRI), there was an abundance of EB. Not all landscapes have an abundance of EB (approximately 10% of the landscape was EB) and therefore it may not be applicable as a data source. Possible solutions and areas of research would be to use landform classifications in order to supplement regions that have little to no EB. Mapping of landform units has already been heavily studied (Jasiewicz et al., 2014; Vannametee et al., 2014) with automated processes designed to map these units. Relative depths can be assigned to land units and used as proxies for depth (BC Ministry of Environment, 2010; Vannametee et al., 2014) and landforms and spatial interpolation methods can be further explored for regions lacking abundant EB.

Slope and convexity in the glacial landscape have less of an influence on the deposition of material than in other landscapes that have not been recently glaciated (Brardinoni et al., 2009). Slope and convexity are cited as the main

drivers to quantify depth in studies that quantify soil production and regolith depth on a hill slope scale (Heimsath et al., 2001; Heimsath et al., 1999). This research has shown that slope and convexity are not major drivers for soil depth, where elevation, depth of channel network and MRVBF can have more influence. This disparity between local scales and global (landscape scale) areas has created another gap in understanding the drivers and relationships for soil production and deposition. Testing the models from this study in a mountainous region that has not been recently glaciated, could provide further insights into these processes.

Lastly due to the scale at which the soil depth and EB are being mapped, local processes of soil production, erosion, and deposition become harder to quantify because the scale of these events are not as responsive on the landscape scale (Brardinoni & Hassan, 2006). Soil forming processes have been shown to respond at different spatial resolutions (Behrens et al., 2010). Multi-scale studies would provide a better understanding for the soil production processes and could produce higher quality maps.

4.3. Thesis Contributions

The research in this thesis was the first to map exposed bedrock and depth of the pedosphere on a landscape scale using RF, GLM and kriging in Southern British Columbia, Canada. This study is unique in its use of ancillary data sources. Modified legacy land cover data of EB was used as primary data inputs to map the landscape scale variability for EB. The resulting EB maps were used to produce high quality soil depth maps. The use of these data sources allowed for a reduction in extensive fieldwork that typically would have been required for such an area, and therefore I suggest that future studies use similar data inputs.

With the use of modified PDPs that were presented in Chapter 2, I allowed for a critical view inside the black box of RF that is typically unavailable. Where other studies focus primarily on relative influence or cumbersome regression trees, the PDPs allow for an intuitive look at how decisions were made in RF with relation to the data in the model. This allows for a better understanding of the data available and the relationships between them.

GLMRK and RFRK proved to be very powerful tools for mapping the depth of the pedosphere. Typical studies use intensive process-based models to map steady- state soil depth, where my methodology proposes a simpler approach that can yield just as accurate results. This method can also extend past typical scales of hill slope or watershed scale studies into the landscape scale as I have done.

These contributions can be used in many fields, such as GIScience, soil science, hydrology and geomorphology (landscape evolution modelling) since soil depth is required for their modelling. My approaches to map EB and soil depth, draw from all of these fields of study and extend current standards and practices that are available to improve on current mapping methodologies.

4.4. References

BC Ministry of Forests and Range and BC Ministry of Environment. (2010). *Field manual for describing terrestiral ecosystems - 2nd edition*. Victoria, British Columbia.

Behrens, T., Zhu, A.-X., Schmidt, K., & Scholten, T. (2010). Multi-scale digital terrain analysis and feature selection for digital soil mapping. *Geoderma*, *155*(3-4), 175–185. http://doi.org/10.1016/j.geoderma.2009.07.010

Brardinoni, F., & Hassan, M. A. (2006). Glacial erosion, evolution of river long profiles, and the organization of process domains in mountain drainage basins of coastal British Columbia. *Journal of Geophysical Research: Earth Surface*, *111*(1), 1–12. http://doi.org/10.1029/2005JF000358

Brardinoni, F., Hassan, M. a., Rollerson, T., & Maynard, D. (2009). Colluvial sediment dynamics in mountain drainage basins. *Earth and Planetary Science Letters*, *284*(3-4), 310–319. http://doi.org/10.1016/j.epsl.2009.05.002

Heimsath, a M., Dietrich, W. E., Nishiizumi, K., & Finkel, R. C. (2001). a: Stochastic processes of soil production and transport: erosion rates, topographic variation and cosmogenic nuclides. *In the Oregon Coast Range. Earth Surface Processes and Landforms*, *26*, 531–52.

Heimsath, A. M., Dietrich, W. E., Nishiizumi, K., & Finkel, R. C. (1999). Cosmogenic nuclides, topography, and the spatial variation of soil depth. *Geomorphology*, *27*, 151–172.

Jasiewicz, J., Netzel, P., & Stepinski, T. F. (2014). Landscape similarity, retrieval, and machine mapping of physiographic units. *Geomorphology*, *221*, 104–112. http://doi.org/10.1016/j.geomorph.2014.06.011

Vannametee, E., Babel, L. V., Hendriks, M. R., Schuur, J., de Jong, S. M., Bierkens, M. F. P., & Karssenberg, D. (2014). Semi-automated mapping of landforms using multiple point geostatistics. *Geomorphology*, *221*, 298–319. http://doi.org/10.1016/j.geomorph.2014.05.032