Regional-Scale Digital Soil Mapping in British Columbia using Legacy Soil Survey Data and Machine-Learning Techniques

by

Brandon Heung

B.Sc. (Hons.), Simon Fraser University, 2011

Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of
Doctor of Philosophy

Department of Geography
Faculty of Environment

SIMON FRASER UNIVERSITY

Summer 2017

Approval

Name: **Brandon Heung**

Degree: **Doctor of Philosophy (Geography)**

Title: ***Regional-Scale Digital Soil Mapping in British Columbia using Legacy Soil Survey Data and Machine-Learning Techniques***

Examining Committee: **Chair:** Dr. Nadine Schuurman
Professor

Dr. Margaret G. Schmidt
Senior Supervisor
Associate Professor

Dr. Chuck E. Bulmer
Supervisor
Soil Scientist
BC Ministry of Forest Lands and
Natural Resource Operations

Dr. Anders Knudby
Supervisor
Assistant Professor
Department of Geography,
Environment and Geomatics
University of Ottawa

Dr. Suzana Dragićević
Internal Examiner
Professor

Dr. Brian Klinkenberg
External Examiner
Professor
Department of Geography
University of British Columbia

Date Defended/Approved: April 19, 2017

Abstract

Digital soil mapping (DSM) is the intersection of geographical information systems (GIS), and (spatial) statistics and is a sub-discipline of soil science that has been increasingly relevant in helping to address emerging issues such as food production, climate change, land resource management, and the management of earth systems. Even with the need for digital soil information in the raster format, such information is limited for British Columbia (BC) where much of it is digitized from legacy soil survey maps with inherent spatial problems related to polygon boundaries; attribute specificity due to multi-component map units; and map scale where small-scale surveys have limited use in addressing local and regional needs. In spite of these issues, legacy soil survey data are still useful as sources of training data where machine-learning techniques may be used to extract soil-environmental relationships from a survey and a suite of digital environmental covariates.

This dissertation describes a framework for developing training data from conventional soil survey maps and compares various machine-learning techniques for predicting the spatial patterns of qualitative soil data such as soil parent material and soil classes. Results of this research included maps of soil parent material, Great Groups, and Orders for the Lower Fraser Valley and a soil Great Group map for the Okanagan-Kamloops region at a 100 m spatial resolution. Key findings included (1) the recognition of Random Forest being the most effective machine-learner based on two model comparison studies; (2) the conclusion that model choice greatly impacted the accuracy of predictions; (3) the method for developing training data greatly impacted the accuracy through a comparison of four methods; and (4) that training data derived from soil survey maps were more effective in representing the feature space of various classes in comparison to using training data derived from soil pits. This study advances the understanding of model selection and training data development in DSM and may facilitate the future development of methodologies for provincial maps of BC.

Keywords: Digital Soil Mapping; Machine-Learning; Soil Classification; Pedometrics; Model Comparison; Ensemble-Learning

Dedication

This dissertation is dedicated to my parents, Raymond and Terry Heung, for their endless support and for teaching their children the value of humility and hard-work.

I also dedicate this dissertation to the memory of Dr. Owen Hertzman who inspired me, and many of my undergraduate classmates, to pursue a career in geography. During my undergraduate years, he treated all his students with the greatest respect and fairness. Early on in my graduate career, I was privileged to be one of Owen’s teaching assistants, where he always stressed the importance of providing the best possible education to students. I share a similar dedication and attitude towards teaching. Owen was a teacher, mentor, colleague, and friend to many of us.

Acknowledgements

First and foremost, I would like to thank my senior supervisor, Dr. Margaret Schmidt, for teaching me about soil science; for providing me with the opportunity to work in her lab as an undergraduate research assistant; for her willingness to accept me as a graduate student; for allowing me to figure my way through the research and giving me some freedom; for putting up with me; and also for her unwavering support. There is a lot of other stuff that I thank Margaret for too.

Secondly, I would like to acknowledge Dr. Chuck Bulmer for introducing me to the field of digital soil mapping; for your willingness to have long conversations about our work; and for providing me with a ton of field experience. When looking back to 2010, it is amazing how much we learned together and how much progress we have made over these years.

I also acknowledge Dr. Anders Knudby for taking an early interest in my work. Anders' willingness to have long discussions about technical details; his willingness to engage in debate; and his overall support in my career makes him one of the finest colleagues I've had the opportunity to work with.

This research was made possible with the help of funding from NSERC for providing me with an Alexander Graham Bell Award (2014-2017); Graduate Fellowship Awards from SFU; BC Ministry of Forests, Lands and Natural Resource Operations; BC Future Forest Ecosystem Scientific Council; and the Financial Institute of Mom & Pop.

I also thank my fellow lab mates, which include Khaled Hamdan, Maciej Jamrozik, Chris Scarpone, Debby Reeves, and Jin Zhang. Jin was an awesome field assistant and fellow graduate student that really put up with me and all my antics. Recognition should also go to the many research assistants that helped me with my work: Darren Murray, Sarah Robertson, Lillian Fan, and especially Darrell Hoffman. In addition, I would also like to thank my other research collaborators and all their contribution: Mr. Derrick Ho and Matus Hodúl. Special thanks also go to Ravinder Multani and especially Joyce Chen for their friendship and encouragement along the way.

Many, many, many thanks to my partner, Rena Li, for all the emotional support that you have provided me over all these years. We got through a lot of the tough stuff together and I don’t think I would have been able to manage all of it without you. I would also like to recognize the rest of the Li family (Jeffery, Cherry, and Ricky) for all their kindness, generosity, and support.

Finally, I would also like to thank my siblings for their support and specifically: Alicia, for taking care of a lot of the complicated stuff so I can focus on my school work; Elsie, for being my sounding board; and Ernie, for being an all-around great brother. And of course, I would like to thank my parents, Raymond & Terry Heung, for their support and for not kicking me out of the house.

Table of Contents

List of Tables

List of Figures

List of Acronyms

ANN	Artificial Neural Network
BCSIS	British Columbia Soil Information System
CART	Classification and Regression Tree
CART+	Classification and Regression Tree with Bagging
CV	Cross-Validation
DEM	Digital Elevation Model
DSM	Digital Soil Mapping
EVI	Enhanced Vegetation Index
*k*NN	*k*-Nearest Neighbour
*k*NN+	*k*-Nearest Neighbour with Bagging
LMT	Logistic Model Tree
LMT+	Logistic Model Tree with Bagging
LST	Land Surface Temperature
MART	Multiple Adaptive Regression Spline
MLR	Multinomial Logistic Regression
MLR+	Multinomial Logistic Regression with Bagging
MDA	Mean Decrease in Accuracy
MDG	Mean Decrease in Gini
MSAVI	Modified Soil Adjusted Vegetation Index
NDVI	Normalized Difference Vegetation Index
NDWI	Normalized Difference Water Index
NSC	Nearest Shrunken Centroid
OOB	Out-of-Bag
PCA	Principal Component Analysis
RF	Random Forest
SAVI	Soil Adjusted Vegetation Index
SLC	Soil Landscapes of Canada
SVM-Lin	Linear Support Vector Machine
SVM-RBF	Support Vector Machine with Radial Basis Function

Chapter 1.

Introduction

The demand for up-to-date soil information has been increasing in order to address emerging environmental issues such as sustainable food production; climate change regulation, adaptation, and mitigation; soil degradation; land resource management; and the provision of earth system services across all geographical extents (Sanchez et al., 2009; FAO and Global Soil Partnership, 2016). Additionally, better soil information is necessary for performing soil assessments; and the reduction and informing of risks for decision-making (Carré et al., 2007; Finke, 2012; Arrouays et al., 2014).

Within the soil science discipline, pedometrics, a branch of pedology that aims to quantitatively characterize soil variation over space, has been used to provide soil information through the development of digital soil maps (Burrough et al., 1994; Sanchez et al., 2009). Although digital soil mapping - the intersection of soil science, geographical information science/systems (GIS), and (spatial) statistics – has existed as early as the 1970s (e.g. Webster and Burrough, 1972a, 1972b), the advancements in computing technology, remote-sensing technology, GIS, data-mining and machine-learning techniques, and the increasing availability of spatial datasets have greatly facilitated the production of DSM products since the 2000s (McBratney et al., 2003; Scull et al., 2003; Minasny and McBratney, 2016). Furthermore, technological advancements have also allowed for digital soil maps to be produced at progressively larger spatial extents and higher resolutions (Minasny and McBratney, 2016).

Digital soil maps (DSM) have been developed at a large range of scales for a wide array of applications. At the global scale, organizations such as the International Soil Reference and Information Centre (ISRIC) have recently produced predictions of soil attributes (e.g. soil organic carbon, pH, particle size fractions, bulk density, cation exchange capacity, soil depth, and coarse fragments) for six standard depth intervals and the prediction of soil taxonomic units (Hengl et al., 2014) at a 1 km spatial resolution as part of the *SoilGrids1km* project. Continentally, similar developments have been made for the prediction of soil attributes and classes by way of an African project through the African Soil Information Service (AfSIS) and ISRIC at a 250 m spatial resolution (Hengl et al., 2015); or in the case of Australia, as part of the Soil and Landscape Grid of Australia project at a 90 m spatial resolution. In Europe, an extensive number of projects have applied DSM through the Joint Research Centre of the European Commission in order to predict soil erosion due to wind (Borelli et al., 2014) and water (Panagos et al., 2015) as well as the prediction of total organic carbon stocks in order to test potential climate and land cover change scenarios (Yigini and Panagos, 2016). At regional and local scales, DSM methods have often been used as a tool to address specific environmental issues; for instance, to predict the spatial distribution of biological soil crusts in order to assess soil stability in (semi-) arid environments (Brungard and Boettinger, 2012); to identify distinct wine-producing soils (Hughes et al., 2012); to monitor seasonal changes in soil salinity in order to better mitigate the impacts of salinization (Berkal et al., 2012); or to produce crop-specific suitability maps and maps of gross margins for those crops (Harms et al., 2015; Kidd et al., 2015). This dissertation provides a framework to test various DSM methodologies, and presents three case-studies for using existing soil data to train a variety of machine-learning techniques that will be extended for mapping the province of British Columbia where high-resolution digital soil data is currently limited.

1.1. Background Information

The objective of this section is (1) to provide a brief overview of conventional soil surveys and the limitations pertaining to them; (2) to summarize key concepts in digital soil mapping and provide a description of various soil-environmental covariates that facilitate the soil predictions; and (3) to provide an overview of various machine-learning techniques that are applicable in digital soil mapping.

1.1.1. Conventional Soil Surveys

Two major achievements contributed to the development of conventional soil survey methods in North America. The first achievement was the formalization of Jenny's Factors of Soil Formation (1941) and the second was the formalization of national soil taxonomic systems in the US (Soil Taxonomy; Soil Survey Staff, 1975) and Canada (The Canadian System of Soil Classification, CSSC; Canada Soil Survey Committee, 1978). The classification systems described how soils were classified based on soil morphology using morphological properties that could easily be measured and quantified in the field. Jenny's *clorpt* model (Eq. 1.1) characterizes the environmental conditions for which soils are found as a function of climate (*cl*), organisms (*o*), relief (*r*), parent material (*p*), time (*t*), and other local factors that influence soils (*…*):

Eq. (1.1) $S = f(cl, o, r, p, t, \ldots)$.

The *clorpt* model was originally proposed as a method for studying how soils varied, quantitatively, as a function of various state factors. As such, soil properties were examined across a gradient of a single factor while the other factors were held constant in order to develop quantitative functions that related the change in one factor to the change in soils. It is necessary to note that the model does not describe *how* each factor influences soil formation nor does it treat the variables as *formers* of soil formation (a common misconception); and thus, system dynamics are not represented in the *clorpt* formulation. Rather, the *clorpt* factors are just variables that represent the environment from which soils and their properties are found. In other words, the factors define the *state* of the soil system. In principle, each of the soil-environmental variables were thought to be independent; however in practice, it was eventually realized that all the variables were interrelated with each other and, furthermore, they were also not truly independent from each other – with the exception of *time* (Phillips, 1998). Despite some of the limitations of the model, Jenny's soil-environmental variables were still extremely useful in assisting with soil surveys because the variables were generally observable at the field level or through the use of aerial photography, where changes in soil-environmental conditions resulted in differences in soil characteristics – thus facilitating the delineation of map units.

Soils may be represented and viewed as *profiles*, *pedons*, *polypedons*, or map units. The soil *profile* is a 2-dimensional representation and the pedon is a 3-dimensional representation with an areal extent of approximately 1 m^2. In principle, the properties of the pedon should not vary horizontally (only vertically) and when several similar pedons are connected, the resulting body of soil is represented as a polypedon. The polypedon concept differs from a *map unit* because the map unit – a distinct polygon that is mapped by soil surveyors – is an areal representation of a polypedon or a polypedon-complex. There are two types of mapping units: *consociations* and *associations*. Consociations are map units delineated based on a single taxonomic unit or soil class and may be referred to as a *simple* mapping unit, whereas associations consist of two or more dissimilar soil taxa (multiple components) that occur in a pattern that is too complex to be resolved at the selected mapping scale (Hole and Campbell, 1987; Schaetzl and Anderson, 2005). In the case of complex map units, the proportion of each soil class within the map unit is specified in the map legend or the map symbol (Bie and Beckett, 1971)

In a conventional soil survey, the mapper first uses preconceived hypotheses of what types of soils to expect at a location based on existing knowledge of soil-environmental relationships. The mapper then uses aerial photographs to identify patterns where the soil-environmental variables exhibit an external expression on the landscape in order to attempt the correlation of landscape characteristics to soil boundaries (soil-landscape relationships). The underlying concept behind Jenny (1941) was that areas with similar soil-environmental characteristics should share similar soil characteristics based on a type of rule-based reasoning (Abraham, 2005). Once a preliminary reconnaissance map has been developed, the map is then tested in the field where morphological classification is performed on the preliminary map units and linked to a soil taxonomic unit. When the soil class has been identified, the mapper attempts to further delineate or adjust the boundaries of map units based on where the rate of change in soil properties is the greatest. Supplemented with field data and profile descriptions, map units with similar morphological characteristics are grouped into the same taxonomic unit (or series) and the soil properties and the range of environmental conditions from which the soils are found are then described in the soil legend.

Limitations with Conventional Soil Surveys

Several concerns arise with conventional soil surveys. Firstly, the data is represented as discrete classes where the conditions are assumed to be homogenous within the polygons (Hole, 1978; Zhu and Band 1994). It is recognized that a significant amount of spatial generalization within the map unit occurs due to inclusions of subdominant soils that are too small to be resolved at the spatial scale of the map (Hole and Campbell, 1987). As a result, the purity of the mapping units are dependent on the complexity of the terrain, external expression of boundaries, survey effort, and mapping scale (Beckett, 1971). Although it would be ideal to have soil maps that consisted of only simple mapping units, increasing the proportion of 'pure' map units has been shown to result in an exponential cost increase for developing the soil map (Bie et al., 1973).

Other limitations may be related to the imprecision in map unit boundaries where the variability (or lack thereof) of the topographic surface does not necessarily coincide with the variability that may be occurring belowground (Hole, 1978). In addition the changes in soil are not necessarily discrete (as suggested by the use of boundaries), but rather, they are fuzzy where the soil attributes between two neighbouring map units may be an intergrade of the soil properties of the two units (Zhu and Band, 1994; Schaetzl and Anderson, 2005).

The final set of challenges for conventional surveys stems from the soil surveyors, themselves, where the delineation of map units are based on the mental-models of soil-environmental relationships, which are rarely ever recorded or reported. In addition, the mental-models may be based on a false hierarchy where some of the five soil-environmental variables are given preference over others when delineating boundaries – as a result, this may also lead to inconsistencies in mapping amongst different surveyors and also inconsistencies throughout time and landscapes. Consequently, these inconsistencies manifest themselves within and between soil maps as mismatching map unit boundaries between different map sheets, counties, states/provinces, and countries (Thompson et al., 2012; Dewitte et al., 2013). Such inconsistencies may also lead to problems where multiple soil series share the same soil properties, which results in redundancy, or even worse, where two soil series with the same name have completely different soil properties (Thompson et al., 2012).

Despite these issues, conventional soil maps have great value as sources of training data for machine-learning tools, where soil map units may be intersected with multiple soil-environmental variables and used to predict soils elsewhere (Bui, 2004). Such approaches have been demonstrated numerous times using decision trees (Bui and Moran, 2001, 2003; Moran and Bui, 2001; Grinand et al., 2008) and the Random Forest (RF) algorithms (Häring et al., 2012); and more recently, decision trees have also been used to disaggregate complex map units (Odgers et al., 2014; Subbarayalu, et al., 2014).

1.1.2. Digital Soil Mapping

Despite the wide use of Jenny's (1941) *clorpt* model, it is still largely a conceptual model. With the increasing capabilities of GIS and computers, coupled with the availability of geospatial data in digital format, the widely used *clorpt* model becomes inadequate for the purposes of modelling soils as a spatial phenomenon. McBratney et al. (2003) recognized the significance of a spatial component in soil formation theory and proposed the *scorpan* model:

Eq. (1.2) $S_{c,a} = f\,(s_{x,y,\sim t},\ c_{x,y,\sim t},\ o_{x,y,\sim t},\ r_{x,y,\sim t},\ p_{x,y,\sim t},\ a_{x,y},\ n)$

The *scorpan* model includes the five factors from Jenny's *clorpt* model, which are climate (*cl*), organisms (*o*), relief (*r*), and parent material (*p*) at spatial position, (*x*,*y*), and the time of which an environmental covariate represents, *t,* and the age of the soil (*a*). In addition, the *scorpan* model includes existing soil knowledge or soil properties at a point (*s*), the spatial position (*n*) of a soil observation and a quantitative function, *f*(), that empirically links the *scorpan* variables to a soil class, S_c, or to a soil attribute, S_a. The *n* factor was included with the intention that it would capture the spatial trends that were not captured by the other environmental covariates. As a result, *scorpan* allows for the digital mapping and modelling of soils as it takes into account where in geographical-space a particular soil attribute or class occurs. The following sections provide a brief overview of soil-environmental covariates used in the DSM literature.

Relief (r)

Based on McBratney et al. (2003), which reviewed 132 papers on DSM, it was observed that of all seven *scorpan* factors, the most widely used factor in DSM studies was ‘relief’ (*r*), from which nearly 80% of the studies used digital elevation models (DEMs) and other terrain derivatives calculated from it. DEMs are particularly useful because they are readily available and are consistent in coverage. In addition, many terrain derivatives, such as slope, curvature, aspect, and drainage may be calculated from a DEM. The terrain derivatives may be used to inform the hydrologic processes that influence soil formation. Furthermore, it has previously been demonstrated that landscape classification could easily be produced using only a DEM (MacMillan et al., 2000, 2003), and where soil properties such as organic matter, in particular, are often linked to landform elements (Pennock et al. 1987); in addition, landform classes have also been used as environmental covariates in soil mapping (Smith et al., 2012) and in predictive ecosystem mapping (MacMillan et al., 2007).

Soils (s)

The second most used *scorpan* factor was soil information, *s*, where it was used by nearly 40% of the reviewed studies in McBratney et al. (2003). Conventional soil survey data is commonly used to train models or build knowledge bases. Hewitt (1993) notes that the soil mapping rules, mental models, and the soil-landscape relationships originally used by soil surveyors were generally not recorded; however, soil maps can still provide valuable knowledge on the soil-landscape model. In Qi and Zhu (2003), it was recognized that the soil-landscape relationships could be extracted and used for soil mapping and classification. The extraction of knowledge from legacy data sources can be useful in cases where there are no experts or when the soil-landscape relationships are not recorded. In examples such as Bui et al. (1999), Qi and Zhu (2003), Moran and Bui (2002), and Grinand et al. (2008), soil-landscape relationships were extracted using classification trees and various other machine-learning algorithms. Furthermore, Qi and Zhu (2003) provided a method for extracting point data from soil survey polygons and similarly in Lagacherie et al. (1995), a reference area (or training area) with a small extent was used to identify and formulate the soil pattern rules from which the soil survey was developed and where those rule-sets could then be used for extrapolation purposes.

In addition to obtaining soil information from conventional soil maps, the *s* factor includes the use of remote sensing data. In these cases, soil samples are collected from the field and taken to a laboratory in order to determine the relationships between a soil's attributes and its spectral characteristics where airborne or space-borne data may then be used to map a soil attribute based on the soil-spectral relationships. In a review of the applications of remote sensing on soil mapping, Mulder et al. (2011) noted that remotely sensed data has been particularly useful for mapping soil mineralogy, soil texture, soil moisture, organic carbon, iron and carbonate contents, and salinity on bare soil. Coupled with the wide coverage and availability of satellite imagery and the expanding size of spectral libraries, mapping soils using only remotely sensed data will increasingly be utilized; however, issues related to atmospheric influences and spectral and spatial resolution still remain (Mulder et al., 2011). Furthermore, other sources of data may be obtained through the use of multi-spectral, hyper-spectral, and radar sensors; electoral conductivity; or gamma radiometric data (McBratney et al., 2003).

Organisms (o)

Compared to relief and soils, soil-environmental layers that represent organisms (*o*) have been used to a lesser degree (30% of the reviewed studies) (McBratney et al., 2003). In DSM a major source of vegetation data may be derived from satellite imagery where numerous vegetative indices have been developed based on satellite band-ratios (Mulder et al., 2011). The Normalized Difference Vegetation Index (NDVI) is one such example and has been shown to be fairly effective as a covariate in mapping soil organic carbon (Boettinger, 2010; Marchetti et al., 2010; Zhao and Shi, 2010). Other similar indices that are adapted from the NDVI include the Soil Adjusted Vegetation Index (SAVI), Transformed SAVI (TSAVI), Modified SAVI (MSAVI) and the Global Environment Monitoring Index (GEMI) (Mulder et al., 2011). In addition to NDVI and its various renditions, remotely sensed data may also be used to determine other vegetative characteristics such as the Leaf Area Index (LAI), fractional canopy cover, plant water content, aboveground biomass, evapotranspiration, and vegetation height (Dorigo et al., 2007). Applications of some of these vegetative indices have yet to be tested in DSM.

Crop data has also been used as a covariate for spatial prediction; for instance, crop yields are the result of the interaction between soils and the atmosphere. Therefore, crop yield data may be used as an indicator of soil properties since plant growth is influenced by properties such as clay content, moisture content, and nutrient content (e.g. Shatar and McBratney, 1999; McBratney et al., 2000). In a forested setting, a possible opportunity may lie in the use of forest inventory data where forest variables such as basal area, gross total volume, stand density, stand height, and aboveground biomass could provide some insight into the soil properties. Especially with the developments in LiDAR imagery, forest inventory data might become more available in the future (Woods et al., 2011; Treitz et al., 2012). Land class and vegetation class data may also be a useful source of data to represent *o*. In Smith et al. (2012), *o* was represented using Biogeoclimatic Ecosystem Classification data and the CIRCA land classification data for mapping soil classes in the Okanagan.

Parent Material (p)

With respect to parent materials (*p*), only 25% of the reviewed studies in McBratney et al. (2003) included a parent material layer; in addition, 75% of the cases that used a parent material map used geological maps rather than surficial material maps – an approach that may be appropriate for soils derived from residual parent material. Consequently, transported parent materials are poorly represented and the parent material maps used for DSM become biased in favor of residual parent materials (Lawley and Smith, 2008). Geological maps also suffer from the same problems as conventional soil maps, which sometimes use complex map units. As a result, parent materials have occasionally been mapped as a property of the soil rather than used as an environmental covariate for predicting soils (e.g. Bui and Moran, 2001; Lacoste et al., 2011; Lemercier et al., 2012).

Climate (c)

Climate, *c*, is the least used of the environmental covariates of *scorpan* - several possible explanations are apparent. Firstly, the local climate is largely influenced by topography and as a result, topographic indices such as elevation and aspect may be used as a proxy for climate variables due to the relationship between elevation and the environmental lapse-rate and the relationship between slope-face direction and temperature (Schaetzl and Anderson, 2005). In terms of climate layers, common covariates include mean annual temperature, mean annual precipitation, and evapotranspiration – all of which may be derived from satellite imagery (McBratney et al., 2003). The usefulness of these covariates are largely dependent on the extent of the study area where constant climatic conditions may be assumed as the study area decreases in size.

1.1.3. Machine-Learning Techniques for Classification[1]

A brief overview of various machine-learning techniques is presented here. The objective is not to provide a detailed explanation of each approach but rather to provide a summary of several approaches, and their relevance in DSM. In addition to the learners used in DSM, approaches that have been used in other disciplines but have yet to be explored in DSM are also summarized. The objective here is to examine machine-learners for mapping soil taxonomic units, and therefore, the context of this overview is focused mainly on the use of machine-learners as classifiers for mapping qualitative soil properties rather than for the numerical mapping of soil attributes.

Tree-Based Learners

Tree-based algorithms are perhaps the most commonly used learners in the DSM literature. Tree-based learners consist of nodes and leaves where each node is a partition of the training dataset that aims to maximize the within-node homogeneity and the between-node heterogeneity, based on node splitting rules that are generated from a set of predictor variables - a type of *if-then* statement (Breiman et al., 1984). The leaves are the terminal nodes where a decision is made with regards to the response variable of interest. As a result of their hierarchical structure, tree-based learners are able to represent non-linear and non-smooth relationships between predictor and response variables as well as interaction effects where the relationship between a predictor and the response depends on one or more other predictors. In addition, tree-based learners

[1] A version of this section has been published in "Heung, B., Ho, H.C., Zhang, J., Knudby, A., Bulmer, C.E., Schmidt, M.G., 2016. An overview and comparison of machine-learning techniques for classification purposes in digital soil mapping. Geoderma 265, 62-77."

are also flexible as they are able to handle numerical, ordinal, or discrete predictors, and do not require assumptions on normality (Hastie et al., 2009).

Tree-based learners have commonly been used for classification to map soil taxonomic units (e.g. Behrens et al., 2010; Bui and Moran, 2001, 2003; Bui et al., 1999; Grinand et al., 2008; Jafari et al., 2014; Moran and Bui, 2002; Nelson and Odeh, 2009; Schmidt et al., 2008; Scull et al., 2005; and Taghizadeh-Mehrjardi et al., 2014) or soil parent material classes (e.g. Bui and Moran, 2001; Lacoste et al., 2011; and Lemercier et al., 2012), and more recently, for the disaggregation of complex map units from conventional soil maps (e.g. Nauman and Thompson, 2014; Odgers et al., 2014; and Subburayalu et al., 2014). In addition, they have also been used to map soil attributes such as pH, soil depth, organic C, clay content, and total N and P using regression modeling (e.g. Bui et al., 2006, 2009; Henderson et al., 2005; and McKenzie and Ryan, 1999).

The RF learner is conceptually similar to tree-based learners and shares the same advantages; however, multiple decision trees are trained and the results are based on the predictions from an ensemble of the individual trees (Breiman, 2001). For the RF learner, each tree is trained from a randomized bootstrap sample of the entire training set and a subset of predictors used for the node-splitting rules is also randomly selected. Although the RF learner was adopted early on to analyze large datasets in the bioinformatics literature (e.g. Díaz-Uriarte and Alvarez de Andrés, 2006; Qi, 2012; and Svetnik et al., 2003), its usage in DSM appears to become increasingly more prominent. DSM applications of the RF learner, similar to those of the decision trees, have included the mapping of soil organic C (e.g. Grimm et al., 2008; Guo et al., 2015; and Wiesmeier et al., 2011;), soil texture (Ließ et al., 2012) as well as for classification purposes such as the mapping of soil parent material classes (Heung et al., 2014) or the updating and disaggregation of conventional soil survey maps (Häring et al., 2012; and Rad et al., 2014). Despite the similarities between single tree-based learners and RF, few studies in DSM have compared the two, with the exception of Ließ et al. (2012) who compared them for the prediction of particle size fractions using regression and found that RF performed better.

Logistic Regression

A review of DSM approaches by McBratney et al. (2003) identified that linear models (e.g. multiple linear regression and generalized linear models) have commonly been used for mapping soil attributes and have regularly been hybridized with kriging in regression kriging (e.g. Odeh et al., 1995; Hengl et al., 2007). For classification purposes, however, the most frequently used linear approach is through the use of multinomial logistic regression models (e.g. Kempen et al., 2009; Debella-Gilo and Etzelmüller, 2009; Collard et al., 2014; Jafari et al., 2012).

Logistic regression models are a type of generalized linear model that is well suited for datasets where the dependent variable is categorical. These models are able to describe the relationships between a set of predictor variables and a dichotomous dependent variable that has values of 0 or 1. In the binomial case, outputs of logistic regression are expressed in probabilistic terms where values close to 0 indicate a low probability of occurrence, and values close to 1 represent a high probability of occurrence (Kleinbaum et al. 2008).

In order to extend the logistic regression model approach to predict multinomial categorical response variables, both Kempen et al. (2009) and Debella-Gilo and Etzelmüller (2009) propose a multinomial logistical regression (MLR) approach. In both cases, logistic regression models were developed for each soil class that was found in the study area. The relationships between topography and soil taxonomic units were determined from legacy soil data. In order to convert a set of binomial logistic regression models into a generalized multinomial model, the following equation is used:

Eq. (1.3) $$p_i = \frac{exp\ (p_i)}{exp(p_1) + exp(p_2) + \ldots + exp(p_n)}\ ,$$

where p_i represents the probability of occurrence for class i, and the denominator of the equation represents the sum of the probabilities of occurrence for n classes. In the final classification, each data point is then assigned to the class with the highest probability of occurrence.

Distance-Based Learners

Although the *k*-nearest neighbour (*k*NN) learner has been used for a wide variety of different applications in the natural sciences, such as the classification of agricultural land cover (Samaniego and Schultz, 2009), and is increasingly used in forest inventory studies (e.g. Meng et al., 2007; Bernier et al., 2010; Beaudoin et al., 2014), its usage is rare in DSM. One recent exception is in Mansuy et al. (2014), where the *k*-nearest neighbour method was used to predict forest soil properties such as forest floor thickness, total organic carbon and nitrogen concentrations, soil particle size fractions, and bulk density; however, the use of *k*NN as a classifier is still relatively limited. One example of using *k*NN for classification in the DSM literature may be found in Subburayalu and Slater (2013) and the disaggregation of Soil Survey Geographic database (SSURGO) polygons for mapping soil series.

The main concept of the *k*NN learner is related to Tobler's First Law of Geography, where near things are more related than distant things, with the nuance that the *k*NN is concerned with distance in feature space rather than physical distance. Therefore, within the feature space, predictions are made based on a neighbourhood that is defined by the *k* number of training points that are located nearest to the predicted point (Hastie et al., 2009). When $k = 1$, a predicted location is assigned the value of the closest training point and when $k > 1$, classification is determined through majority vote. For a detailed explanation of the *k*NN, refer to Hastie et al. (2009).

Nearest centroid learners are another type of distance-based learners; however, unlike in *k*NN where a pixel is assigned a class based on its distance to the nearest sample point(s), nearest centroid learners calculate the distance based on each feature's inverse coefficient of variation for each class. The most common nearest centroid learner is the nearest shrunken centroid learner (NSC). For the NSC learner, the class centroid is 'shrunken' to the overall centroid based on a shrinkage threshold parameter. Predictions are based on the distance of a new sample to the closest shrunken centroid in feature space. If a predictor is shrunken to zero for all classes, the influence of that predictor on the classification rules become negligible. As a result, an advantage of the NSC approach is its inherent ability to determine variable importance based on the distance of the shrunken class centroids to the overall centroids (Tibshirani et al., 2002).

Logistic Model Trees

Model trees are a relatively new approach that hybridizes linear models with a nonlinear tree model. Typically, linear models such as MLR have been shown to produce a stable model with a low variance and a potentially high bias and thus run the risk of under-fitting the model, whereas tree-based models may exhibit a low bias with a high variance as they capture non-linear relationships and thus risk over-fitting. As a result, the complementary nature of tree-based and linear models would seem appropriate for classification purposes, where the structure of a logistic model tree (LMT), proposed in Landwehr et al. (2005), consists of a decision tree where the leaves consist of individual logistic regression models.

To summarize the construction of the LMT (Landwehr et al., 2005), at the stump of the tree, a logistic regression model is initially fitted to the entire training dataset and iteratively refined using the *LogitBoost* algorithm (Friedman et al., 2000) that optimizes the number of predictor variables and coefficient values. Once the initial regression model may no longer be refined, a node-splitting rule is applied and local regression models are fitted to the subset data points within the child nodes using *LogitBoost*. The fitting of partial logistic regression models on smaller subsets of data increase the overall fit of the model. The partial logistic regression models are incrementally refined to increasingly smaller subsets of the data and thus the decision tree is grown until a stopping criterion is met based on the size of the terminal nodes. To reduce model complexity and the computational demand of predictions, the size of the tree is reduced based on the CART pruning scheme. Because of this hybridized structure, the LMT has the advantage of being able to capture the nonlinearities and interaction effects in the dataset while minimizing the risk of over-fitting (Landwehr et al., 2005). A further advantage of the LMT is its flexibility in adapting to the complexity and size of a dataset where the structure of the tree becomes increasingly elaborate.

Artificial Neural Networks

The origin of the artificial neural network (ANN) learner may be traced back to the 1940’s, where McCulloch and Pitts (1943) initially planned to develop a virtual “central nervous system” for computer modelling, which had similar data processes to a biological nervous system. The structure of an ANN consists of a set of interconnected units, or ‘neurons’ that estimate the non-linear correlations between each variable. The input neurons, which represent predictor variables, are connected to a single or multiple layer(s) of hidden neurons, which are then linked to the output neurons that represent the target soil variable. In an ANN, the user parameterizes the number of hidden layers and neurons within each hidden layer. During the ANN training process, the connections between the neurons are established by assigning weights based on an intrinsic learning process where the weights are iteratively adjusted to match the outputs of the training dataset (Behrens et al., 2005).

In DSM, ANN has typically been used to predict continuous soil variables such as particle size fractions (Chang and Islam, 2000; McBratney et al., 2000; Priori et al., 2014), and soil erosion rates (Licznar and Nearing, 2003); the use of ANN for predicting discrete soil data still remains limited with some exceptions including Behrens et al. (2005) and Silveira et al. (2013).

Support Vector Machines

The support vector machine (SVM) classifier is a learner that is designed to construct an optimal separating hyperplane, in the feature space, between the various classes (Hastie et al., 2009). As such, the SVM classifier predicts the maximum margin of possibility between each class (Pal and Mather, 2005; Ocak and Seker, 2013). In the case of binary classification, or ‘one-class-classification’, SVM detects the closest points between two classes in feature space and assigns a margin based on the distance between the hyperplane and the points. Following this, the margins are maximised by the ‘support vectors’ (the optimal points that should be lying on the boundary) in order to estimate an optimal separating hyperplane between the two classes (Witten and Frank, 2005). This hyperplane, from the maximized margins, is used as a criterion for subsequent classification.

In the case of models such as logistic regression and linear discriminant analysis, the decision boundaries between classes are separable using linear class boundaries. The SVM classifier is an extension of the linear approaches for cases where the classes are non-separable or overlap in feature space (Hastie et al., 2009). To account for overlap between classes, the SVM classifier reduces the weight of data points that fall into the wrong side of the hyperplane in order to reduce their influence on the classification. In order to implement nonlinear class boundaries for the classification of complex datasets, a SVM applies a linear model to the feature space of the training dataset, which has been transformed into a higher-dimensional space using a polynomial or radial basis expansion, in order to create a linear-like space that may be separated using a hyperplane. As a result, a linear hyperplane in the transformed space becomes a nonlinear hyperplane in the original non-transformed space (Witten and Frank, 2005).

SVM is a relatively common classification technique used for land-use and land cover mapping with remotely sensed data (e.g. Huang et al., 2002; Melgani and Bruzzone, 2004; Mountrakis et al., 2011; Pal and Mather, 2005; Ocak and Seker, 2013); however, the use of SVM for classification is less common in DSM. Several applications of SVM have included Kovačević et al. (2010) for predicting soil chemical and physical properties and taxonomic units; Ahmad et al., (2010) for estimating soil moisture with remote sensing data; and Priori et al. (2014) where soil texture and stoniness was mapped using γ-radiometric data.

1.2. Research Problem

Despite the many uses of DSM products, the availability of high-resolution digital soil data for British Columbia (BC) – especially at regional scales – remains limited. Currently, the digital soil dataset with the most comprehensive coverage of British Columbia is the Soil Landscapes of Canada (SLC) dataset, which is provided by the Canadian Soil Information Service (CanSIS). The SLC dataset was created through the digitization of a combination of provincial and regional scale soil survey maps and is provided in a polygon format at a 1:1,000,000 scale (Schut et al., 2011). The portion of the SLC that covers BC consists of 2,651 multi-component polygons with an average polygon size of 380 km^2 (Geng et al., 2010). In order to meet the requirements for global-scale DSM products (e.g. *GlobalSoilMap.net*) and to provide DSMs of soil properties in the raster format, the SLC data has previously been rasterized and the data was harmonized in accordance with global specifications (Hempel et al., 2012 and 2013). Although the SLC may be useful for addressing national- and global-scale DSM needs, its usefulness is diminished when addressing regional- or local-scale needs due to its small mapping scale. Another potential DSM product for BC may be extracted from the *SoilGrids1km* project, which captures more detail in comparison to the SLC; however, very few data points were obtained for BC to train the models. Consequently, any predictions made for the province may not be reliable in comparison to predictions made in the USA and Mexico, where the sample densities were drastically higher (Hengl et al., 2014).

Although detailed soil surveys have been developed for BC at scales ranging from 1:25,000 to 1:125,000, those detailed soil surveys cover less than 50% of BC (Bulmer et al., 2016). In addition, existing soil surveys for forested and northern regions were either mapped at smaller map scales than used when mapping agricultural regions or they were not mapped at all. However, existing legacy soil maps are still a useful resource in developing training data for predicting soil distributions for unmapped regions, as well as in the refinement of existing soil surveys using machine-learning techniques.

Examples of where legacy soil survey maps have been used as training data for DSMs have included studies such as Bui and Moran (2001, 2003), Moran and Bui (2002), Grinand et al. (2008), Odgers et al. (2008), and Kempen et al. (2008), where legacy soil survey maps for Australia, France, USA, and Netherlands were used to fill the gaps in coverage as well as the refinement of existing surveys. However, these studies, as well as most other studies in the DSM literature, provide a limited rationalization for their model choice where only one type of model was used. For instance, Bui and Moran (2001, 2003), Moran and Bui (2002), Grinand et al. (2008) and Odgers et al. (2008) only tested decision tree algorithms, while Kempen et al. (2008) only tested a multinomial logistic regression. In fact, there is a clear research gap in terms of how different models perform in DSM when using the same training data. Model comparison studies have generally been few in the DSM literature, with some notable exceptions such as Taghizadeh-Mehrjardi et al. (2015) and Brungard et al. (2015) that compared 6 and 11 models, respectively, using pit-derived training data. However, a model comparison study under the context of using soil surveys as training data has yet to be performed. Furthermore, an additional research gap exists regarding a direct comparison of the different types of training data where DSM studies typically only use pit-derived training data (e.g. Taghizadeh-Mehjardi et al., 2015; Brungard et al., 2015) or only soil survey-derived training data – a comparison of these training data using the same set of models and environmental covariates has yet to be performed.

In order to address the limited availability of digital soil data within the context of British Columbia; the limited number of model comparison studies; and the lack of a comparison between training data, the following research questions provide the impetus for this dissertation:

1. Given the availability of digitized legacy soil survey maps for British Columbia, how may machine-learning techniques be used to extract the soil-environmental data from these maps and be used to predict the spatial patterns of soils?

2. With the diversity of machine-learning techniques found within and beyond the DSM literature, how similar or dissimilar are soil predictions produced from different models using the same training data? What is the significance of performing model comparison studies in DSM?

3. Within the DSM literature, data used to train models often come in the form of soil pit data or through the use of map units from legacy soil surveys in the polygon format – what are the differences in their usage as training data and how do they compare in terms of prediction accuracy?

1.3. Research Objectives

The main objectives of this study are:

1. To develop a framework for extracting soil-environmental training data from detailed soil surveys and to test the framework in the prediction of soil types and parent materials at a regional-scale using machine-learning techniques.

2. To perform a comprehensive comparison of machine-learning techniques in DSM.

3. To compare the accuracies of soil predictions produced using soil pit-derived training data and soil survey-derived training data.

1.4. Overview of Dissertation

This dissertation consists of five chapters, where Chapters 2-4 are individual papers that were designed to address aspects of the objectives.

Chapter 1 provides an introduction to DSM, where the background information, research problems and objectives are presented. Section 1.2 “Background Information” was designed to explore three themes: (1) a background of conventional soil surveys and their challenges; (2) a summary of common environmental covariates used in the DSM literature; and (3) an overview of machine-learning techniques that have been used for DSM and some techniques that have yet to be used in DSM.

Chapter 2 directly addresses Objective 1 of the dissertation and explores three methods for extracting training data from conventional soil survey maps and the Random Forest (RF) machine-learner was tested as a classification algorithm. The framework and the RF algorithm were tested for mapping soil parent material for the Lower Fraser Valley of BC. Secondary objectives included (1) testing the necessity of optimizing the parameters of the RF model and (2) testing a variable reduction technique for improving predictions.

Chapter 3 was designed to demonstrate the importance of model comparison in DSM studies by comparing a suite of 10 machine-learners (CART, CART with bagging, MLR, LMT, RF, *k*NN, SVM-Lin, SVM-RBF, and ANN) and comparing their accuracies in order to address Objective 2. Here, the model comparison was performed on the Lower Fraser Valley for mapping soil Great Groups and Orders as a case study, where training data was extracted using the same framework presented in Chapter 2. Secondary objectives included (1) the testing of methods for addressing the issue of class imbalance in training data – an issue identified in Chapter 2 and (2) the use of allocation and quantitative disagreement as accuracy metrics.

Chapter 4 addresses a research gap that was identified in Chapter 3 and Brungard et al. (2015), where most DSM literature used only training data derived from either soil pit data or soil survey data and where comparisons between the two had yet to be made. To meet Objective 3, this study developed training data from soil surveys using the same framework in Chapters 2 and 3, and compared it to the predictions made from training data derived using legacy soil pit data obtained from the BC Soil Information System. Predictions were made for soil Great Groups for the Okanagan-Kamloops region of BC as a case study. Secondary objectives included (1) the comparison of 9 machine-learners; (2) the comparison between single-model learners (CART, MLR, LMT, and *k*NN) and ensemble-model learners (CART with bagging, MLR with bagging, LMT with bagging, *k*NN with bagging; and RF); and (3) the development of classification uncertainty maps.

Chapter 5 summarizes the key research findings of this dissertation and summarizes contributions to the DSM literature.

1.5. References

Abrahan, A. 2005.Rule-based expert systems. *In* Handbook of Measuring System Design, John Wiley & Sons, 909-919.

Ahmad, S., Kalra, A., Stephen, H., 2010. Estimating soil moisture using remote sensing data: A machine learning approach. Advances in Water Resources 33, 69-80.

Arrouays, D., Grundy, M.G., Hartemink, A.E., Hempel, J.W., Heuvelink, G.B.M., Hong, S.Y., Lagacherie, P., Lelyk, G., McBratney, A.B., McKenzie, N.J., Mendonça -Santos, M.L., Minasny, B., Montanarella, L., Odeh, I.O.A., Sanchez, P.A., Thompson, J.A., Zhang, G-L., 2014. GlobalSoilMap: Toward a fine-resolution global grid of soil properties. Advances in Agronomy 125, 93-134.

Beaudoin, A., Bernier, P.Y., Guindon, L., Villemaire, P., Guo, X.J., Stinson, G., Bergeron, T., Magnussen, S., Hall, R.J., 2014. Mapping attributes of Canada's forests at moderate resolution through *k*NN and MODIS imagery. Canadian Journal of Forest Research 44, 521-532.

Beckett, P.H.T. 1971. The cost-effectiveness of soil survey. Outlook Agriculture 6: 191-198.

Behrens, T., Förster, H., Scholten, T., Steinrücken, U., Spies, E-D, Goldschmitt, M., 2005. Digital soil mapping using artificial neural networks. Journal of Plant Nutrition and Soil Science 168, 21-33.

Behrens, T., Zhu, A-X., Schmidt, K., Scholten, T., 2010. Multi-scale digital terrain analysis and feature selection for digital soil mapping. Geoderma 155, 175-185.

Berkal, I., Walter, C., Michot, D., Djili, K., 2014. Seasonal monitoring of soil salinity by electromagnetic conductivity in irrigated sandy soils from a Saharan oasis. Soil Research 52, 769-780.

Bernier, P.Y., Daigle, G., Rivest, L-P., Ung, C-H., Labbé, F., Bergeron, C., Patry, A., 2010. From plots to landscape: A *k*-NN-based method for estimating stand-level merchantable volume in the Province of Québec, Canada. The Forestry Chronicle 86, 461-468. Bie, S.W., Beckett, P.H.T., 1971. Quality control in soil survey. Introduction: I. The choice of mapping unit. Journal of Soil Science 22, 32-49.

Bie, S.W., Ulph, A., Beckett, P.H.T., 1973. Calculating the economic benefits of soil survey. Journal of Soil Science 24, 429-435.

Boettinger, J.L., Environmental covariates for digital soil mapping in Western USA. *In* Digital Soil Mapping: Bridging Research, Environmental Application, and Operations. Progress in Soil Science 2. Springer.

Borrelli, P., Ballabio, C., Panagos, P., Montanarella, L., 2014. Wind erosion susceptibility of European soils. Geoderma 232-234, 471-478.

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

Breiman, L., Friedman, J., Stone, C.J., Olshen, R.A., 1984. Classification and Regression Trees. CRC Press LLC, Boca Raton, FL.

Brungard, C.B., Boettinger, J.L., 2012. Spatial prediction of biological soil crust classes: Value added DSM from soil survey. *In* Digital Soil Assessments and Beyond. CRC Press, Leiden, Netherlands, pp. 57-64.

Brungard, C.W., Boettinger, J.L, Duniway, M.C., Wills, S.A., Edwards, T.C., 2015. Machine learning for predicting soil classes in three semi-arid landscapes. Geoderma 239-240, 68-83.

Bui, E.N., 2004. Soil survey as a knowledge system. Geoderma 120, 17-26.

Bui, E.N., Henderson, B.L., Viergever, K., 2006. Knowledge discovery from models of soil properties developed through data mining. Ecological Modelling 191, 431-446.

Bui, E.N., Henderson, B.L., Viergever, K., 2009. Using knowledge discovery with data mining from the Australian Soil Resource Information System database to inform soil carbon mapping in Australia. Global Biogeochemical Cycle 23, GB4033. http://dx.doi.org/10.1029/2009GB003506.

Bui, E.N., Loughead, A., Corner, R., 1999. Extracting soil-landscape rules from previous soil surveys. Australian Journal of Soil Research 37, 495-508.

Bui, E.N., Moran, C.J., 2001. Disaggregation of polygons of surficial geology and soil maps using spatial modeling and legacy data. Geoderma 103, 79-94.

Bui, E.N., Moran, C.J., 2003. A strategy to fill gaps in soil survey over large spatial extents: an example from the Murray-Darling basin of Australia. Geoderma 111, 21-44.

Bulmer, C.E., Schmidt, M.G., Heung, B., Scarpone, C., Zhang, J., Filatow, D., Finvers, M., Smith, C.A.S., 2016. Improved soil mapping in British Columbia, Canada, with legacy soil data and Random Forest. *In* Digital Soil Mapping Across Paradigms, Scales and Boundaries. Springer Environmental Science and Engineering, pp. 291-303.

Burrough, P.A., Bouma, J., Yates, S.R., 1994. The state of the art in pedometrics. Geoderma 62, 311-326.

Canada Soil Survey Committee, 1978. The Canadian System of Soil Classification. 1st ed. Research Branch, Canadian Department of Agriculture.

Carré, F., McBratney, A.B., Mayr, T., Montanarell, L., 2007. Digital soil assessments: Beyond DSM. Geoderma 142, 69-79.

Chang, D-H., Islam, S., 2000. Estimation of soil physical properties using remote sensing and artificial neural networks. Remote Sensing of Environment 74, 534-544.

Collard, F., Kempen, B., Heuvelink, G.B.M, Saby, N.P.A., Richer de Forges, A.C., Lehmann, S., Nehlig, P., Arrouays, D., 2014. Refining a reconnaissance soil map by calibrating regression models with data from the same map (Normandy, France). Geoderma Regional 1, 21-30.

Debella-Gilo, M., Etzelmüller, B. 2009. Spatial prediction of soil classes using digital terrain analysis and multinomial logistic regression modeling integrated in GIS: examples from Vestfold County, Norway. Catena 77, 8-18.

Dewitte, O., Jones, A., Spaargaren, O., Breuning-Madsen, H., Brossard, M., Dampha, A., Deckers, J., Gallali, T., Hallett, S., Jones, R., Kilasara, M., le Roux, P., Michéli, E., Montanarella, L., Thiombiano, L., van Ranst, E., Yemefack, M., Zougmore, R., 2013. Harmonisation of the soil map of Africa at the continental scale. Geoderma 211-212, 138-153.

Díaz-Uriarte, R., Alvarez de Andrés, S., 2006. Gene selection and classification of microarray data using Random Forest. BMC Bioinformatics 7, 3-15.

Dorigo, W.A., Zurita-Milla, R., de Wit, A.J.W., Brazile, J., Singh, R., Schaepman, M.E., 2007. A review on reflective remote sensing and data assimilation techniques for enhanced agroecosystem modeling. International Journal of Applied Earth Observation 9, 165-193.

FAO and Global Soil Partnership, 2016. Global Soil Partnership Pillar 4 Implementation Plan: Towards a Global Soil Information System. Pillar 4 Working Group, 38 pp.

Friedman, J., Hastie, T., Tibshirani, R., 2000. Additive logistic regression: a statistical view of boosting. The Annals of Statistics 38, 337-374.

Finke, P.A., 2012. On digital soil assessment with models and the Pedometric agenda. Geoderma 171-172, 3-35.

Geng, X., Fraser, W., VandenBygaart, B., Smith, C.A.S., Wadell, A., Jiao, Y., Patterson, G., 2010. Towards digital soil mapping in Canada: Existing soil survey data and related expert knowledge. In Digital Soil Mapping: Bridging Research, Environmental Application, and Operation, Springer, 325-337.

Grimm, R., Behrens, T., Märker, M., Elsenbeer, H., 2008. Soil organic carbon concentrations and stocks on Barro Colorado Island – digital soil mapping using Random Forest analysis. Geoderma 146, 102-113.

Grinand, C., Arrouays, D., Laroche, B., Martin, M.P., 2008. Extrapolating regional soil landscapes from an existing soil map: sampling intensity, validation procedures, and integration of spatial context. Geoderma 143, 180-190.

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 residual kriging approach. Geoderma 237-238, 49-59.

Häring, T., Dietz, F., Osenstetter, S., Koschitzki, T., Schröder, B., 2012. Spatial disaggregation of complex soil map units: a decision-tree based approach in Bavarian forest soils. Geoderma 185-186, 37-47.

Harms, B., Brough, D., Philip, S., Bartley, R., Clifford, D., Thomas, M., Willis, R., Gregory, L. 2015. Digital soil assessment for regional agricultural land evaluation. Global Food Security 54, 25-36.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed. Springer, New York, NY, 734 pp.

Hempel, J.W., Libohova, Z., Odgers, N.P., Thompson, J.A., Smith, S.S., Lelyk, G.W., 2012. Versioning of GlobalSoilMap.net rasters property maps for the North American Node. *In* Digital Soil Assessments and Beyond. CRC Press, Leiden, Netherlands, pp. 429-433.

Hempel, J.W., Libohova, Z., Thompson, J.A., Odgers, N.P., Smith, C.A.S., Lelyk, G.W., Geraldo, G.E.E., 2014. GlobalSoilMap North American Node progress. *In* GlobalSoilMap – Basis of the Global Spatial Soil Information System. CRC Press, Leiden, Netherlands, pp. 41-45.

Hengl, T., de Jesus, J.M., MacMillan, R.A., Batjes, N.H., Heuvelink, G.B.M., Ribeiro, E., Samuel-Rosa, A., Kempen, B., Leenaars, J.G.B., Walsh, M.G., Gonzalez, M.R., 2014. SoilGrids1km – Global soil information based on automated mapping. PLOS ONE 9, 17 pp.

Hengl, T., Heuvelink, G.B.M., Kempen, B., Leenaars, J.G.B., Walsh, M.G., Shepherd, K.D., Sila, A., MacMillan, R.A., de Jesus, J.M., Tamene, L., Tondoh, J.E., 2015. Mapping soil properties of Africa at 250 m resolution: Random Forests significantly improve current predictions. PLOS ONE 10, 26pp.

Hengl, T., Heuvelink, G.B.M., Stein, A., 2007. About regression-kriging: From equations to case studies. Computers & Geosciences 33, 1301-1315.

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

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.

Hewitt, A.E., 1993. Predictive modelling in soil survey. Soils and Fertilizers 56, 305-314.

Hole, F.D. 1978. An approach to landscape analysis with emphasis on soils. Geoderma 21, 1-23.

Hole, F.D., Campbell, J.B., 1985. Soil Landscape Analysis. Rowman and Allanheld, Totowa, NJ.

Huang, C., Davis, L., Townshend, J., 2002. An assessment of support vector machines for land cover classification. International Journal of Remote Sensing 23, 725-749.

Hughes, P., McBratney, A.B., Malone, B.P., Minasny, B., 2012. Development of terrons for the Lower Hunter Valley wine-growing regions. *In* Digital Soil Assessments and Beyond. CRC Press, Leiden, Netherlands, pp. 31-36.

Jafari, A., Khademi, H., Finke, P., Wauw, J.V.D., Ayoubi, S., 2014. Spatial prediction of soil Great Groups by boosted regression trees using a limited point dataset in an arid region, southeastern Iran. Geoderma 232-234, 148-163.

Jenny, H., 1941. Factors of Soil Formation: A System of Quantitative Pedology. McGraw-Hill, NY.

Kempen, B., Brus, D.J., Heuvelink, G.B.M., Stoorvogel, J.J. 2009. Updating the 1:50,000 Dutch soil map using legacy soil data: a multinomial logistic regression approach. Geoderma 51, 311-326.

Kidd, D., Webb, M., Malone, B., Minasny, B., McBratney, A.B., 2015. Digital soil assessment of agricultural suitability, versatility and capital in Tasmania, Australia. Geoderma Regional 6, 7-21.

Kleinbaum, D.G., Kupper, L.L., Nizam, A., and Muller, K.E. 2008. Logistic regression analysis. *In* Applied Regression Analysis and Other Multivariate Methods 4th ed., Thomson Brooks/Cole, Belmont, CA: 604-634.

Kovačevic, M., Bajat, B., Gajić, B., 2010. Soil type classification and estimation of soil properties using support vector machines. Geoderma 154, 340-347.

Lacoste, M., Lemercier, B., Walter, C., 2011. Regional mapping of soil parent material by machine learning based on point data. Geomorphology 133, 90-99.

Lagacherie, P., Gomez, C., Bailly, J.S., Baret, F., Coulouma, G., 2010. The use of hyperspectral imagery for digital soil mapping in Mediterranean areas. *In* Digital Soil Mapping: Bridging Research, Environmental Application, and Operations. Progress in Soil Science 2. Springer.

Landwehr, N., Hall, M., Frank, E., 2005. Logistic model trees. Machine Learning 59, 161-205.

Lawley, R., Smith, B. 2008. Digital soil mapping at a national scale: a knowledge and GIS based approach to improving parent material and property information. *In* Digital Soil Mapping with Limited Data, Springer.

Lemercier, B., Lacoste, M., Loum, M., and Walter, C., 2012. Extrapolation at regional sale of local soil knowledge using boosted classification trees: a two-step approach. Geoderma 171-172, 75-84.

Licznar, P., Nearing, M.A., 2003. Artificial neural networks of soil erosion and runoff prediction at the plot scale. CATENA 51, 89-114.

Ließ, M., Glaser, B., Huwe, B,. 2012. Uncertainty in the spatial prediction of soil texture: comparison of regression tree and Random Forest models. Geoderma 170, 70-79.

MacMillan, R.A., Martin, T.C., Earle, T.J., McNabb, D.H., 2003. Automated analysis and classification of landforms using high-resolution digital elevation data: Applications and issues. Canadian Journal of Remote Sensing 29, 592-606.

MacMillan, R.A., Moon, D.E., Coupé, R.A., 2007. Automated predictive ecological mapping in a forest region of B.C., 2001-2005. Geoderma 140, 353-373.

MacMillan, R.A., Pettapiece, W.W., Nolan, S.C., Goddard, T.W., 2000. A generic procedure for automatically segmenting landforms into landform elements using DEMs, heuristic rules and fuzzy logic. Fuzzy Sets and Systems 113, 81-109.

Mansuy, N., Thiffault, E., Paré, D., Bernier, P., Guindon, L., Villemaire, P., Poirier, V., Beaudoin, A., 2014. Digital mapping of soil properties in Canadian managed forests at 250 m of resolution using the *k*-nearest neighbor method. Geoderma 235-236, 59-73.

Marchetti, A., Piccini, C., Francaviglia, R., Santucci, S., Chiuchiarelli, I., 2010. Estimating soil organic matter content by regression kriging. *In* Digital Soil Mapping: Bridging Research, Environmental Application, and Operations. Progress in Soil Science 2. Springer.

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

McBratney, A.B., Odeh, I.O.A., Bishop, T.F.A., Dunbar, M.S., Shatar, T.M., 2000. An overview of pedometrics techniques for use in soil survey. Geoderma 97, 293-327.

McCulloch, W.S., Pitts, W., 1943. A logical calculus of the ideas immanent in nervous activity. The Bulletin of Mathematical Biophysics 54, 115-133.

McKenzie, N., Ryan, P., 1999. Spatial prediction of soil properties using environmental correlation. Geoderma 89, 67-94.

Melgani, F., Bruzzone, L., 2004. Classification of hyperspectral remote sensing mages with support vector machines. IEEE Transactions on Geoscience and Remote Sensing 42, 1778-1790.

Meng, Q., Cieszewski, C.J., Madden, M., Borders, B.E., 2007. K Nearest neighbour method for forest inventory using remote sensing data. GIScience & Remote Sensing 44, 149-165.

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

Moran, C.J., Bui, E.N., 2002. Spatial data mining for enhanced soil map modeling. International Journal of Geographical Information Systems 16, 533-549.

Mountrakis, G., Im, J., Ogole, C., 2011. Support vector machines in remote sensing: A review. ISPRS Journal of Photogrammetry and Remote Sensing 66, 247-259.

Mulder, V.L., de Bruin, D., Schaepman, M.E., Mayr, T.R., 2011. The use of remote sensing in soil and terrain mapping – A review. Geoderma 162, 1-19.

Nauman, T.W., Thompson, J.A., 2014. Semi-automated disaggregation of conventional soil maps using knowledge driven data mining and classification trees. Geoderma 213, 385-399.

Nelson, M.A., Odeh, I.O.A., 2009. Digital soil class mapping using legacy soil profile data: a comparison of a genetic algorithm and classification tree approach. Australian Journal of Soil Research 47, 632-647.

Ocak, I., Seker, S.E., 2013. Calculation of surface settlements caused by EPBM tunnelling using artificial neural network, SVM, and Gaussian processes. Environmental Earth Sciences 70, 1263-1276.

Odeh, I.O.A., McBratney, A.B., and Chittleborough, D.J., 1995. Further results on prediction of soil properties from terrain attributes: heterotropic cokriging and regression-kriging. Geoderma 67, 215-226.

Odgers, N.P., Sun, W., McBratney, A.B., Minasny, B., and Clifford, D., 2014. Disaggregating and harmonising soil map units through resampled classification trees. Geoderma 214, 91-100.

Pal, M., Mather, P.M., 2005. Support vector machines for classification in remote sensing. International Journal of Remote Sensing 26, 1007-1011.

Panagos, P., Borrelli, P., Poesen, J., Ballabio, C., Lugato, E., Meusburger, K., Montanarella, L., Alewell, C., 2015. The new assessment of soil loss by water erosion in Europe. Environmental Science & Policy 54, 438-447.

Pennock, D.J., Zebarth, B.J., De Jong, E., 1987. Landform classification and soil distribution in hummocky terrain, Saskatchewan, Canada. Geoderma 40, 297-315.

Phillips, J.D., 1998. On the relations between complex systems and the factorial model of soil formation (with Discussion). Geoderma 86, 1-21.

Priori, S., Bianconi, N., Constantini, E.A.C., 2014. Can γ-radiometrics predict soil textural data and stoniness in different parent materials? A comparison of two machine-learning methods. Geoderma 226-227, 354-364.

Qi, Y., 2012. Random forest for bioinformatics. Ensemble Machine Learning. Springer, pp. 307-323.

Qi, F., Zhu, A.X. 2003. Knowledge discovery from soil maps using inductive learning. International Journal of Geographical Information Systems 17, 771-795.

Rad, M.R.P., 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.

Samaniego, L., Schultz, K., 2009. Supervised classification of agricultural land cover using a modified *k*-NN technique (MNN) and landsat remote sensing imagery. Remote Sensing 1, 875-895.

Sanchez, P.A., Ahamed, S., Carré, F., Hartemink, A.E., Hempel, J., Huising, J., Lagacherie, P., McBratney, A.B., McKenzie, N.J., Mendonça Santos, M.L., Minasny, B., Montanarella, L., Okoth, P., Palm, C.A., Sachs, J.D., Shepherd, K.D., Vågen, T-G., Vanlauwe, B., Walsh, M.G., Winowiecki, L.A., Zhang, G-L., 2009. Digital soil map of the world. Science 325, 680-681.

Schaetzl,RJ., Anderson, S., 2005. Soils: Genesis and Geomorphology. Cambridge University Press.

Schmidt, K., Behrens, T., Scholten, T., 2008.Instance selection and classification tree analysis for large spatial datasets in digital soil mapping. Geoderma 146, 138-146.

Scull, P., Franklin, J., Chadwick, O.A., 2005. The application of classification tree anlaysis to soil type prediction in a desert landscape. Ecological Modelling 181, 1-15.

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

Shatar, T.M., McBratney, A.B., 1999. Empirical modeling of relationships between sorghum yield and soil properties. Precision Agriculture 1: 249-276.

Silveira, C.T., Oka-Fiori, C., Santos, L.J.C., Sirtoli, A.E., Silva, C.R., Botelho, M.F., 2013. Soil prediction using artificial neural networks and topographic attributes. Geoderma 195-196, 165-172.

Smith, C.A.S., Daneshfar, B., Frank, G., Flager, E., Bulmer, C.E., 2012. Use of weights of evidence statistics to define inference rules to disaggregate soil survey maps. *In* Digital Soil Assessments and Beyond. CRC Press, Leiden, Netherlands, pp. 215-220.

Soil Survey Staff, 1975. Soil Taxonomy. A Basic System of Soil Classification for Making and Interpreting Soil Surveys. 2nd ed. U.S. Government Print Office, Washington, DC.

Subburayalu, S.K., Jenhani, I., Slater, B.K., 2014. Disaggregation of component soil series on an Ohio County soil survey map using possibilistic decision trees. Geoderma 213, 334-345.

Subburayalu, S.K., Slater, B.K., 2013. Soil series mapping by knowledge discovery from an Ohio County soil map. Soil Science Society of America Journal 77, 1254-1268.

Svetnik, V., Liaw, A., Tong, C., Culberson, 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, 1947-1958.

Taghizadeh-Mehrjardi, R., Sarmadian, F., Minasny, B., Triantafilis, J., 2014. Digital mapping of soil classes using decision tree and auxiliary data in the Ardkan region, Iran. Arid Land Research and Management 213, 15-28.

Thompson, J.A., Nauman, T.W., Odgers, N.P., Libohova, Z., Hempel, J.W., 2012. Harmonization of legacy soil maps in North America: Status, trends, and implications for digital soil mapping efforts. *In* Digital Soil Assessments and Beyond. CRC Press, Leiden, Netherlands, pp. 97-102.

Tibshirani, R., Hastie, T., Narasimhan, B., Chu, G., 2002. Diagnosis of multiple cancer types by shrunken centroids of gene expression. PNAS 99, 6567-6572.

Treitz, P., Lim, K., Woods, M., Pitt, D., Nesbitt, D., Etheridge, D., 2012. LiDAR sampling density for forest resource inventories in Ontario, Canada. Remote Sensing 4, 830-848.

Viscarra Rossel, R.A., Chen, C., Grundy, M.J., Searle, R., Clifford, D., Campbell, P.H., 2015. The Australian three-dimensional soil grid: Australia's contribution to the *GlobalSoilMap* project. Soil Research 53, 845-864.

Webster, R., Burrough, P.A., 1972a. Computer-based soil mapping of small areas from sample data - I: Multivariate classification and ordination. Journal of Soil Science 23, 210-221.

Webster, R., Burrough, P.A., 1972a. Computer-based soil mapping of small areas from sample data - II: Classification smoothing. Journal of Soil Science 23, 222-234.

Wiesmeier, M., Barthold, F., Blank, B., Kögel-Knabner, I., 2011. Digital mapping of soil organic matter stocks using Random Forest modeling in a semi-arid steppe ecosystem. Plant Soil 340, 7-24.

Witten, I.H., Frank, E., 2005. Data Mining: Practical Machine Learning Tools and Techniques, 2nd ed. Elsevier, San Francisco, CA, 525 pp.

Woods, M., Pitt, D., Penner, M., Lim, K., Nesbitt, D., Etheridge, D., Treitz, P., 2011. Operational implementation of LiDAR inventory in Boreal Ontario. The Forestry Chronicle 87: 512-528.

Yigini, Y., Panagos, P., 2016. Assessment of soil organic carbon stocks under future climate and land cover changes in Europe. Science of the Total Environment 557-558, 838-850.

Zhao, Y.V., Shi, X.Z., 2010. Spatial prediction and uncertainty assessment of soil organic carbon in Hebei Province, China. *In* Digital Soil Mapping: Bridging Research, Environmental Application, and Operations. Progress in Soil Science 2. Springer.

Zhu, A.X., Band, L.E., 1994. A knowledge-based approach to data integration for soil mapping. Canadian Journal of Remote Sensing 20, 408-418.

Chapter 2.

Predictive Soil Parent Material Mapping at a Regional Scale: A Random Forest Approach[2]

2.1. Abstract

This study evaluated the application of a Random Forest (RF) classifier as a tool for understanding and predicting the complex hierarchical relationships between soil parent material and topography using a digital elevation model (DEM) and conventional soil survey maps. Single-component soil polygons from conventional soil survey maps of the Langley-Vancouver Map Area, British Columbia (Canada), were used to generate randomized training points for 9 parent material classes. Each point was intersected with values from 27 topographic indices derived from a 100 m DEM. RF's m_{try} parameter was optimized using multiple replicates of 5-fold cross validation and parent material predictions were made for the region. Predictive parent material maps were validated through comparisons with legacy soil survey maps and 307 field points. Results show that predictions made by a non-optimized RF resulted in a kappa index of 89.6% when validated with legacy soil survey data from single-component polygons and a kappa index of 79.5% when validated with field data. Variable reduction and m_{try} optimization resulted in minimal improvements in RF predictions. Our results demonstrate the effectiveness of RF as a machine-learning and data mining approach; however, the need for reliable training data was highlighted by less reliable results for polygon disaggregation in portions of the map where fewer training data points could be established.

[2] A version of this chapter has been published in "Heung, B., Bulmer, C.E., Schmidt, M.G., 2016. Predictive soil parent material mapping at a regional-scale: A Random Forest approach. Geoderma 214-215, 141-154."

2.2. Introduction

Soil parent material is the initial state of the soil system and the material from which soils are derived (Jenny, 1941). Soil type, soil development and the physical and chemical properties of soils are influenced by parent material. Information on parent material and its texture is recognized as a useful factor in soil erosion (Weaver, 1991; le Roux et al., 2007; Heung et al., 2013) and would also be beneficial to the evaluation of forest and agriculture productivity potential; the hydrologic characteristics of watersheds; the suitability of materials for construction; and the assessment of terrain stability. Furthermore, information on soil parent material may also be used for predictive ecosystem mapping (MacMillan et al., 2007) and digital soil mapping studies (McBratney et al., 2003).

Soil parent material is the product of geomorphic processes interacting with bedrock over long periods of time. In British Columbia, glaciation during the Pleistocene Epoch was a dominant process in the evolution of the modern landscape, where the majority of parent materials in the region now consist of unconsolidated sediments deposited on the land surface by ice, gravity, water and wind (Luttmerding, 1981; Howes and Kenk 1988). The geomorphic processes of erosion and deposition that were active at a particular location during glacial, post glacial, and modern times have also created a mosaic of distinct landforms across the region where a close association exists between the topographic landscape form and the characteristics of the unconsolidated parent material. Parent materials are classified in this area, and throughout Canada, based on their mode of formation and transport (Howes and Kenk, 1997).

The majority of digital soil mapping studies reviewed by McBratney et al. (2003) used bedrock geology as a surrogate predictor for parent material - an approach that may be adequate for environments where the soils are predominantly derived from residual materials. However, for environments influenced by glaciation where geomorphic transport processes have significantly influenced the nature and distribution of parent material; bedrock geology likely provides an incomplete depiction of the influence of parent materials on soil properties when it is used alone. Consequently, transported parent materials may not be well represented and the resulting maps would potentially become biased in favor of residual materials (Lawley and Smith, 2008). For these reasons, improving the quality and accuracy of digital soil maps in glaciated areas require more detailed parent material maps that have been derived with the consideration of transport processes.

Conventional soil maps and other resource inventories are commonly developed by delineating map units based on climate, ecological features, topography, parent material, bedrock geology, soil, and vegetation (Resource Inventory Committee, 1998). The importance of parent material as a soil-environmental variable is illustrated by the use of this variable as both a fundamental and distinguishing characteristic between soil types at all mapping scales. There is an especially strong relationship between map unit boundaries and topography, since the topography reflects the dominant geomorphic process and parent material characteristics (Hole and Campbell, 1985), and also because topography has a significant influence on vegetation and other ecological attributes that are often of interest to map makers. In addition, soil surveys are commonly based on aerial photo interpretation and the boundaries of the mapping units are determined based from the external expression of soil-environmental variables on the landscape (Webster and Wong, 1969; Beckett, 1971). Therefore, the derived map units on a conventional soil map tend to contain soil types with a defined set of parent material attributes while also maintaining a close association with topographic features in the landscape.

In British Columbia, the only comprehensive soil parent material map is the national level Soil Landscapes of Canada geographic dataset (SLC; Schut et al., 2011). The SLC database consists of 12,728 multi-component map units, with multiple taxonomic soil classes, that are generalized from detailed soil surveys and are mapped at a 1:1,000,000 scale (Geng et al., 2010). Despite having a consistent map database and comprehensive geographic coverage, the use of such highly aggregated polygon data may not be appropriate for mapping the spatial patterns of parent materials at regional or local scales. At regional-scales, existing soil surveys are available for many, but not all, parts of British Columbia and may be used to obtain information on parent materials. Examples include the 1:25,000 and 1:50,000 scale maps for the Langley-Vancouver Area (Luttmerding, 1980); 1:126,720 scale maps for the Tulameen Area

(Lord and Green, 1974); and 1:20,000 scale maps for the Okanagan and Similkameen Valleys (Wittneben, 1986) with areal extents of approximately 5472 km^2, 4008 km^2, and 3895 km^2, respectively. In addition, other sources of parent material information may be taken from surficial and bedrock geology maps such as those for the New Westminster Area (Armstrong, 1957) and the Vancouver Area (Armstrong, 1956); however, such examples were mapped at smaller spatial extents in comparison to soil surveys.

Extracting the knowledge from existing soil maps is complicated because such maps typically include a large number of multi-component map units, and therefore lack a spatially explicit representation of the soil's class and attributes (Webster and Beckett, 1968; Hole and Campbell, 1985; Zhu and Band, 1994). Despite these spatial challenges, soil maps still have the potential to provide useful information about soil-landscape relationships (McBratney et al., 2003; Bui, 2004); for instance, Bui and Moran (2001) have previously used the map units from soil surveys to train the C5.0 decision tree algorithm and to validate the algorithm's outputs – an approach that was further extended in subsequent studies that mapped the soils of the Murray-Darling Basin, Australia (Moran and Bui, 2001; Bui and Moran, 2003).

Decision trees are data mining, machine-learning, and rule-induction algorithms that classify data by inferring the relationships between a dependent variable and a set of predictors (Bui and Moran, 2001). They consist of nodes and leaves where each node represents an if-then statement and the leaves are terminal nodes where a decision is made with respect to the class variable (Breiman et al., 1984). The aim of a tree-based model is to examine all predictors in order to identify optimal node splitting rules where the within-node homogeneity is maximized. However, the manner in which the splits are made is dependent on the tree-splitting algorithm that is used.

The decision tree modeling approach has many advantages. Firstly, it is a particularly useful modeling approach for handling non-parametric data where the predictors are not characterized as having a specific distribution (Breimann et al., 1984). Secondly, decision trees are not sensitive to missing data, to the inclusion of irrelevant predictors, or to the presence of outliers. Furthermore, decision trees operate effectively using numerical, ordinal, binary, and categorical datasets. Finally, decision trees are well suited for identifying complex hierarchical relationships between predictors and response variables, as well as the relationships between predictors (Díaz-Uriarte and Alvarez de Andrés, 2006; Hastie et al., 2009). Decision trees have been used extensively to map soil classes (Bui et al., 1999; Bui and Moran, 2001, 2003; Moran and Bui, 2002; Scull et al., 2005; Grinand et al., 2008); soil properties such as pH (Henderson et al., 2005), organic C, % clay, and total N and P (Bui et al., 2006, 2009), or natural drainage (Lemercier et al., 2012). In addition, decision trees have also been used for the purposes of predictive soil parent material mapping (e.g. Bui and Moran, 2001; Lacoste et al., 2011; and Lemercier et al., 2012); however, further evaluation of these methods would be valuable and incorporating detailed predictions of the distribution of parent materials, based on topographic characteristics, would likely help such efforts.

The Random Forest (RF) classifier is conceptually similar to a decision tree; except, an ensemble of decision trees are combined in order to improve the classification accuracy (Breiman, 2001; Cutler et al., 2007). For each decision tree in the RF, a random selection of predictors and training points are used to identify splits when building the tree. RF, a hierarchical non-parametric modeling approach, shares similar model advantages to decision trees (e.g. insensitive to missing data, to the inclusion of irrelevant predictors and outliers, and is flexible with various types of datasets); however, RF provides a stronger prediction as it is less susceptible to over-fitting and it provides a better error measurement in comparison to decision trees (Breiman, 2001). Furthermore, RF has the advantage of incorporating 'randomness' into its predictions through reiterative bootstrap sampling and randomized variable selection when generating each decision tree. Additional characteristics of RF include its ability to provide variable importance measures and its ability to provide good predictions when noisy training data is used (Hua et al., 2005).

RF has widely been used in the field of bioinformatics (e.g. Svetnik et al., 2003; Svetnik et al., 2004; Díaz-Uriarte and Alvarez de Andrés, 2006; Statnikov et al., 2008; Qi, 2012). In ecology, examples of studies that have used RF include the mapping of tree species distribution (e.g. Prasad et al., 2006); land cover classification (e.g. Gislason et al., 2006); ecological classification (e.g. Cutler et al., 2007); the mapping of soil organic matter (Grimm et al., 2008; Wiesmeier et al., 2011); and soil texture (Ließ et al., 2012). With the exception of Häring et al. (2012) in using RF to disaggregate multi-component soil polygons, RF has not been used extensively for mapping categorical soil properties such as soil taxonomic units or parent materials.

The objectives of this study were to first evaluate the methods for extracting training data from soil survey data and the optimization of RF parameters; then, to test the reliability of using the RF classifier within single-component polygons in learning the relationship between parent material and topography; and finally, to evaluate RF as a potential method for disaggregating multi-component parent material polygons. The approach was based on the assumption that changes in parent material were closely associated with changes in topography; and hence, all environmental covariates were derived from a digital elevation model (DEM) at a 100 m spatial resolution. The proposed approach may be extended to other resource inventory mapping studies such as ecosystem mapping (Resource Inventory Committee, 1998) and forest inventory mapping (Natural Resources Canada, 2004) where conventional mapping also uses a combination of single and multi-component map units.

2.3. Methodology

The workflow for this study is based on the integration of a DEM and conventional soil survey maps for the development of training data; RF for modeling the hierarchical relationships between parent material and topography; and the use of point data and a conventional soil survey map for assessing model outputs (Figure 2.1). In order to select suitable training areas, the map units from a conventional soil survey map were first separated into two categories: map units with a single parent material (single-component) used as training areas and map units with multiple parent materials (multi-component). To produce a topography-parent material matrix for submission into the RF classifier, random points were generated within training areas and intersected with a suite of topographic indices derived from a DEM of the study area. Using the inputted matrix, the RF parameters were optimized and a variable reduction procedure was tested. The output of the RF classifier was a parent material map of the study area, which was then assessed using the original soil survey map and also external point data. In addition, the ability of RF to disaggregate polygons with multiple parent material components was assessed using the multi-component map units that are not used in the development of the training dataset.

2.3.1. Study Area

The 5472 km^{2} study area ranges from 49^{o}00’ N to 49^{o}56’ N latitude and 121^{o}16’ W to 123^{o}11’ W longitude with an elevational range of 0 - 2555 m above mean sea level and located in the Coastal Western Hemlock biogeoclimatic zone (Figure 2.2) (Pojar et al., 1991). The zone receives mean annual precipitation of 2228 mm where snowfall constitutes less than 15% of the precipitation. The study area consists of the Lower Fraser Valley, which has predominantly an agricultural and urban land coverage, and includes portions of the predominantly forested Coastal Mountain Range located along the northern region of the area.

The pre-existing soil survey identifies 139 distinct soil series with 9 mineral parent material classes (Luttmerding, 1981). Organic parent materials are found in depressions and cover 6% of the landscape; however, they were not predicted for this study. Although the distribution of organic parent materials is affected by topography, these parent materials are also strongly dependent on climatic as well as vegetative factors, which were not included in this study. In the soil survey, the parent material classes were subdivided into 20 subclasses (Table 2.1). In this glaciated landscape there are very few residual materials except at high elevation, and parent materials are almost exclusively derived from the depositional and erosive processes of glaciation, gravity, wind and water. At low elevations, fluvial material is the dominant parent material; however, both marine and glaciomarine materials are also common. At higher elevations, morainal materials are the dominant parent material.

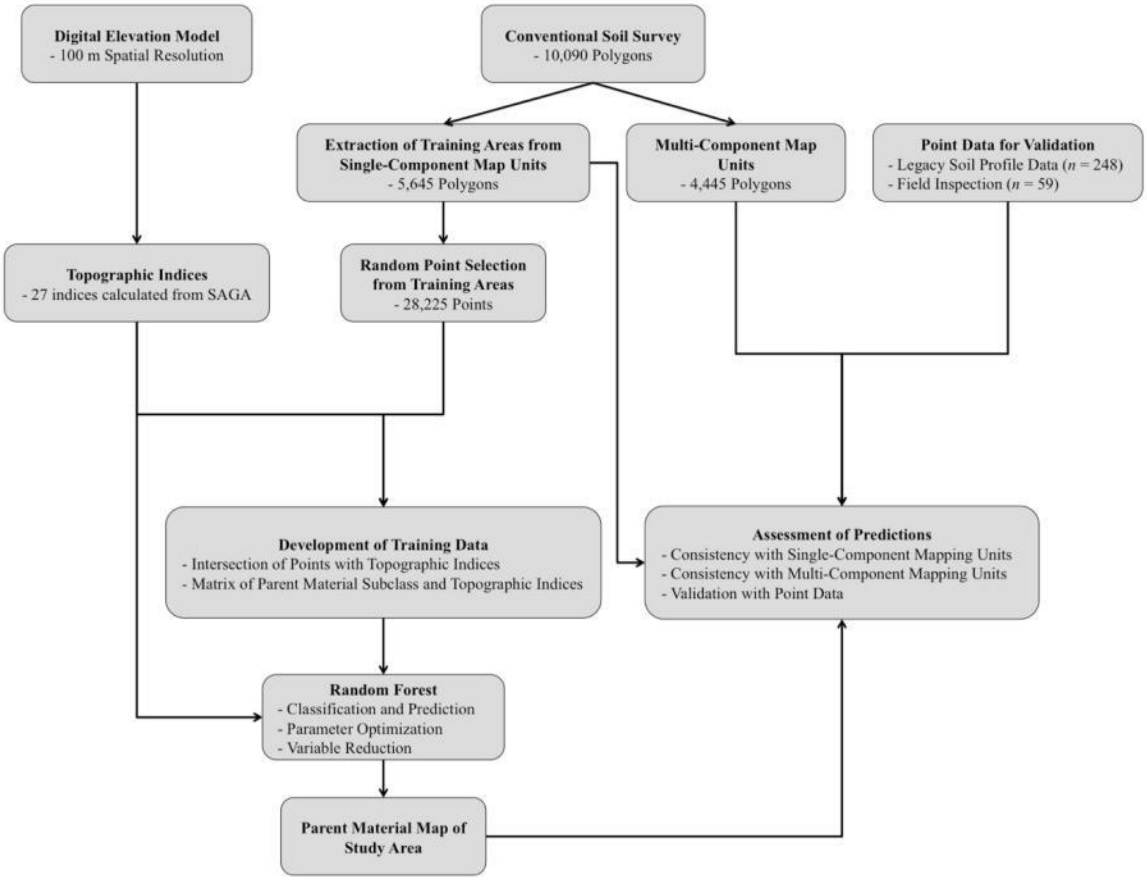

Digital Elevation Model
- 100 m Spatial Resolution
Conventional Soil Survey
- 10,090 Polygons
Extraction of Training Areas from Single-Component Map Units
- 5,645 Polygons
Multi-Component Map Units
- 4,445 Polygons
Point Data for Validation
- Legacy Soil Profile Data (n = 248)
- Field Inspection (n = 59)
Topographic Indices
- 27 indices calculated from SAGA
Random Point Selection from Training Areas
- 28,225 Points
Development of Training Data
- Intersection of Points with Topographic Indices
- Matrix of Parent Material Subclass and Topographic Indices
Assessment of Predictions
- Consistency with Single-Component Mapping Units
- Consistency with Multi-Component Mapping Units
- Validation with Point Data
Random Forest
- Classification and Prediction
- Parameter Optimization
- Variable Reduction
Parent Material Map of Study Area

Figure 2.1. Workflow diagram of predictive soil parent material mapping using a digital elevation model, conventional soil survey data, and Random Forest algorithm. Digital elevation model is used to generate topographic indices; the conventional soil survey is used to train the Random Forest model and to assess model outputs.

Most of the Lower Fraser Valley is underlain by sedimentary rocks from the Cretaceous period (and younger) with approximately 30 m to 150 m of unconsolidated deposits overlying the bedrock (Armstrong, 1957; Valentine et al., 1978). Due to glacial advance during the Pleistocene, ice accumulation with a thickness of 2500 m resulted in the submergence of land into the Pacific Ocean. Both glacial till and glaciofluvial materials were deposited over large areas during this time and during the subsequent ice retreat. As a result of the melting of glacial ice and isostatic rebound, marine and glaciomarine sediments from the Pacific are also common in the Lower Fraser Valley (Luttmerding, 1981). The mountainous area in the northern portion of the study area is part of the Pacific Ranges of the Coast Mountains, where the bedrock is derived from Late Mesozoic intrusive igneous rocks (Valentine et al., 1978). On steep slopes, the dominant parent material is colluvium while depositions of glacial till are most common in areas with gentle and moderate slopes (Luttmerding, 1981). Exposed bedrock is uncommon even in the upland portions of the study area.

2.3.2. Development of Training Data

The soil map for the study area was created at a 1:25,000 scale for the Lower Fraser Valley and at a 1:50,000 scale for the Southern Sunshine Coast and Southern Coast Mountains (Luttmerding, 1981). The soil surveys were subsequently digitized into a seamless coverage and made freely available through Agriculture and Agri-Food Canada and the British Columbia Ministry of Environment (Kenney and Frank, 2010).

Data layers for topographic predictors were calculated using British Columbia's Terrain Resource Information Management (TRIM) DEM (B.C. Ministry of Sustainable Resource Management). The 25 m DEM, originally derived from a triangulated irregular network (TIN) built from TRIM mass-points and break-lines, was then aggregated to a 100 m spatial resolution. The 100 m DEM is freely available from HectaresBC.org (Hectares BC, 2012).

Three successive mean filters with window sizes of 3 x 3, 3 x 3, and 5 x 5 cells were applied to the DEM in order to remove anomalous pits and peaks. Similar to MacMillan et al. (2003) and Li et al. (2011), it was found through preliminary work that the successive smoothing procedure reduces local-scale noise and improves landscape-scale signals. In Grinand et al. (2008), it was also demonstrated that the application of an adaptive mean filter was able to incorporate spatial context into their outputs and improve predictions using the Multiple Additive Regression Tree algorithm (MART).

Table 2.1. Mineral parent material classes and subclasses from The Soils of the Langley-Vancouver Map Area (Luttmerding, 1981).

Parent Material Class	Code	Parent Material Subclass	Code
Colluvial	C	Colluvial Deposits (>1m thick)	Cb
		Shallow Colluvial (<1m thick) over Bedrock	Cv
Eolian	E	Eolian	E
Fluvial	F	Fluvial Deposits - Deltaic (Sandy)	sF-D
		Fluvial Deposits - Deltaic (Silty or Clayey)	zcF-D
		Fluvial Deposits - Floodplain (Sandy)	sFp
		Fluvial Deposits - Floodplain (Silty or Clayey)	zcFp
		Fluvial Deposits - Local Streams (Sandy)	sF-S
		Fluvial Deposits - Local Streams (Silty or Clayey)	zcF-S
		Fluvial Deposits - Fans	Ff
Glaciofluvial	FG	Glaciofluvial Deposits	FG
		Eolian Veneer over Glaciofluvial Deposits	E/FG
Lacustrine	L	Lacustrine Deposits (Sandy)	sL
		Lacustrine Deposits (Silty or Clayey)	zcL
Glaciolacustrine	LG	Glaciolacustrine Deposits	LG
Morainal	M	Morainal (Glacial Till) Deposits	M
		Eolian Veneer over Morainal Deposits	E/M
Marine	W	Marine Deposits (Clayey)	cW
		Marine Deposits (Lag or Littoral)	W
Glaciomarine	WG	Glaciomarine Deposits	WG

Table 2.2. Topographic derivatives derived from a 100 m spatial-resolution DEM.

Landscape Representation	Terrain Derivative	Code	Reference
Local Landscape Characteristics	Transformed aspect	ASPECT	Zevenbergen and Thorne, 1987
	Curvature	CURVE	Zevenbergen and Thorne, 1987
	Elevation	ELEV	
	Slope length factor	LS	Moore et al., 1993
	Plan curvature	PLAN	Zevenbergen and Thorne, 1987
	Profile curvature	PROF	Zevenbergen and Thorne, 1987
	Slope	SLOPE	Zevenbergen and Thorne, 1987
	Tangential curve	TANCUR	Florinsky, 1998
	Terrain ruggedness index	TRI	Riley et al., 1999
	Total curvature	TCURVE	Wilson and Gallant, 2000
Hydrologic Characteristics	Convergence index	CONV	Koethe and Lehmeier, 1996
	Distance to nearest river	RiDIST	
	Distance to nearest stream	StDIST	
	Modified relative hydrologic slope position	mRHSP	MacMillan, 2005
	Relative hydrologic slope position	RHSP	MacMillan, 2005
	Stream power index	StPI	Moore et al., 1991
	SAGA wetness index	SWI	Böhner et al., 2002
	Topographic wetness index	TWI	Beven and Kirkby, 1979
Landscape Context	Multiresolution ridge top flatness index	MRRTF	Gallant and Dowling, 2003
	Multiresolution valley bottom flatness index	MRVBF	Gallant and Dowling, 2003
	Midslope position	MSLOPE	SAGA Development Team, 2011
	Normalized height	NHEIGHT	SAGA Development Team, 2011
	Slope height	SLOPEH	SAGA Development Team, 2011
	Valley depth	VDEPTH	SAGA Development Team, 2011
Landscape Exposure	Sky view factor	SKYVIEW	Häntzshel et al., 2005
	Terrain view	TERVIEW	Häntzshel et al., 2005
	Visible sky	VISSKY	SAGA Development Team, 2011

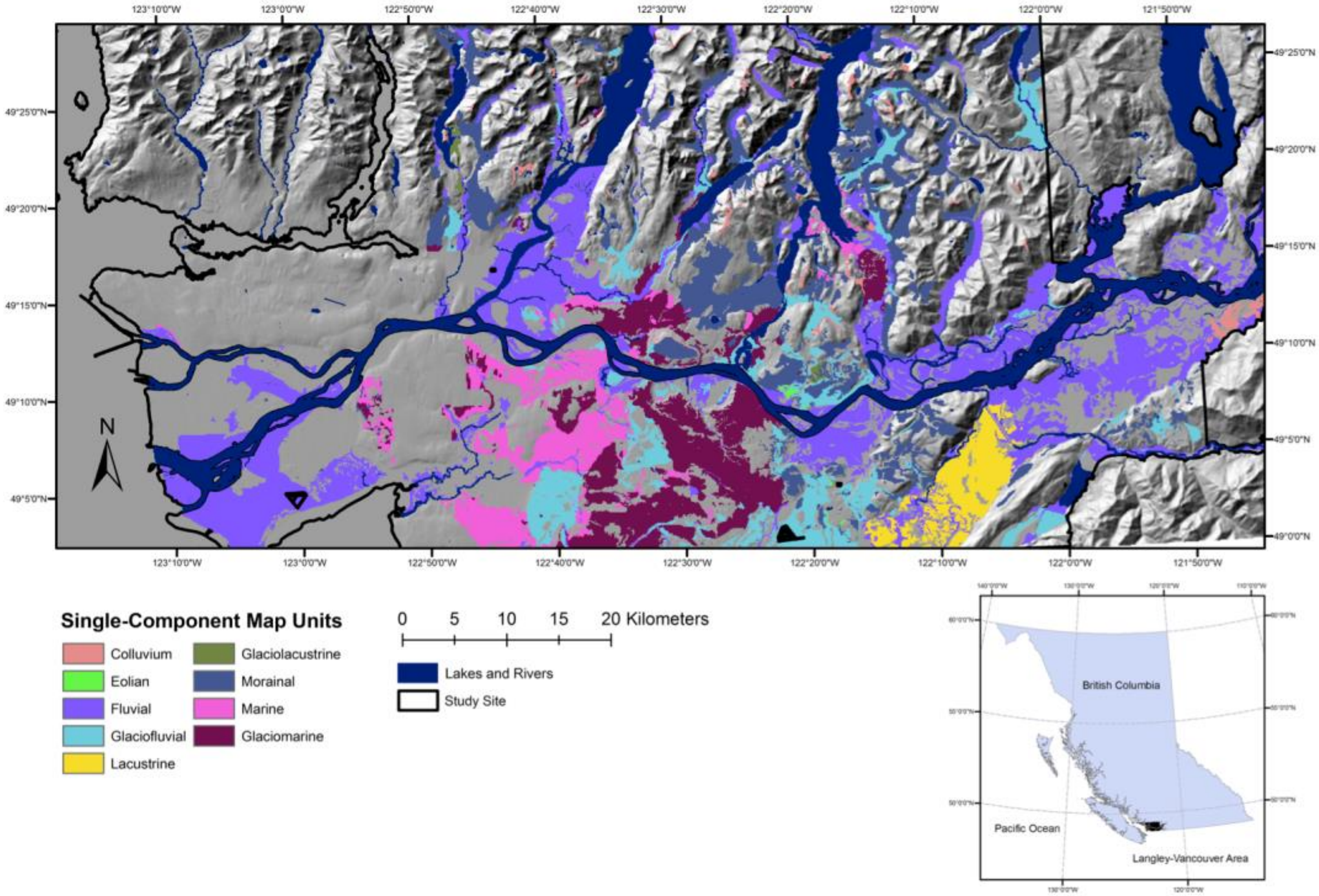


Figure 2.2. **Single-component parent material map units from the Langley-Vancouver Map Area (Luttmerding, 1981). Inset: study area in relation to British Columbia.**

Topographic and hydrologic attributes for 27 topographic indices (Table 2.2) were calculated from the successively filtered DEM using the System for Automated Geoscientific Analysis (SAGA) (SAGA Development Team, 2011). The indices were selected based on their ability to represent basic landscape characteristics of the local neighbourhood (e.g. elevation, slope, aspect, and curvature); hydrologic characteristics at the watershed scale (e.g. wetness index, convergence index, and relative hydrologic slope position); and landscape context (normalized height, slope height, sky view and terrain view). In addition, the distance to nearest stream and distance to nearest river was calculated in order to account for the presence of local streams as well as the Fraser River that runs through the study area.

2.3.3. Development of Training Data

Soil survey map units include attribute data for parent material subclass where 5645 polygons contained a single parent material subclass that covered 29.8% of the study area while 3025 multi-component polygons contained either 2 or 3 subclasses that covered 55.0% of the study area (Figure 2.2). The remaining 15.2% of the study area included miscellaneous land types such as anthropogenic land, bedrock, gravel pits, ice, recent alluvium, rock outcrops, talus, and tidal flats where bedrock only accounted for 0.4% of the study extent. To minimize the uncertainty in the training data, only polygons with a single-component of parent material subclass were used to develop the predictive model; however, it was also recognized that these polygons may have small inclusions of other components. Overall, the dominant parent material subclasses included silty or clayey fluvial floodplain material and glaciomarine sediments for the Fraser Valley; and morainal material (glacial till) along the Coastal Mountain Range (Figure 2.3). The most common parent material class is fluvial, which accounts for approximately 41% of the single-component polygon training data. In addition, many of the soils in the area have had varying amounts of eolian material added as a veneer (>1 m thick) to the surface layers. Where such additions were present only in the surface layers, or where they were considered to have a minor influence, the soil parent material was classified based on the dominant material below the eolian veneer. In other areas where eolian materials were dominant throughout the soil profile, the area was classified as having an eolian parent material.

Predictive models were developed using randomly generated training points within each single-component polygon, where the points were intersected with the values for each topographic attribute and its parent material subclass. Three different methods for developing the training dataset were used with different allocations of the training points according to the following approaches: (1) equal number of points per parent material subclass, (2) equal number of points per polygon, and (3) the number of points was determined as an area-weighted proportion of the subclass' extent over the entire study area. For each sampling strategy, n = 28225 training points, with an average sampling density of 9.4 samples/km2, were used as inputs for RF. The number of training points was selected based on the equal number per polygon sampling scheme where 5 points were randomly generated within each of the 5645 polygons with a single parent material subclass.

2.3.4. Random Forest

To establish the hierarchical relationships between parent materials and topography, the *randomForest* package in the statistical software, *R*, was used (Liaw and Wiener, 2002; R Development Core Team, 2012). The RF classifier uses numerous decision trees, n_{tree}, that are grown from bootstrap samples of the entire sample population, n (Breiman, 2001). The bootstrap sampling makes RF less sensitive to over-fitting in comparison to decision trees. Initially, the RF classifier uses a bootstrapped sample to grow a single RF tree. At each binary split, the predictor that produces the best split is chosen from a random subset, m_{try}, of the entire predictor set, p, where the number of predictors tried at each split, m_{try}, is defined by the user. As a result, m_{try}_is recognized as the main tuning parameter of RF and should therefore by optimized (Svetnick et al., 2003; 2004) The tree growing procedure is performed recursively until the size of the node reaches a minimum, k, which is parameterized by the user (Hastie et al., 2009). Secondly, the remaining samples from the training dataset that was not used in the growing of a decision tree, the out-of-bag (OOB) samples, X_i, are inputted through the decision tree and a predicted class is assigned to each OOB sample, $Y_{OOB}(X_i)$.

The resulting output of the RF is a single model that is accompanied with a single aggregated error estimate – the overall OOB error rate, ER_{OOB}, using the following:

Eq. (2.1) $$ER_{OOB}=n^{-1}\sum_{i=1}^{n} I[Y_{OOB}(X_i)\neq Y_i],$$

where the predicted class of a sample, $Y_{OOB}(X_i)$, is compared against its actual class, Y_i, using the indicator function, *I* (Breiman, 2001; Liaw and Wiener, 2002; Svetnick et al., 2003). In Eq. (2.1), *I* has a value of 1 when $Y_{OOB}(X_i)\neq Y_i$ - otherwise *I* is 0. The OOB error is similar to *k*-fold cross validation (CV) and provides comparable values (Hastie et al. 2009). As a result, RF and its OOB error rates may potentially be used when an independent validation dataset is not available.

In addition, the RF algorithm also provides two measures of variable importance: mean decrease in accuracy (MDA) and mean decrease in Gini (MDG). The MDA is a permutation-based measure of variable importance based on evaluating a variable's contribution to the prediction accuracy. The MDG also measures variable importance; however, it is based on the quality of each split (node) on a variable in a decision tree. A variable that produces high homogeneity in the descendent nodes results in a high MDG (Breiman, 2001).

In this study, the parent material training points, and their associated topographic attributes were used to train the RF classifier. The resulting non-spatial RF model was then applied to all unknown points in the study area using the set of topographic indices (Table 2.1). The output was a map of parent material subclasses in a raster format for the entire study area.

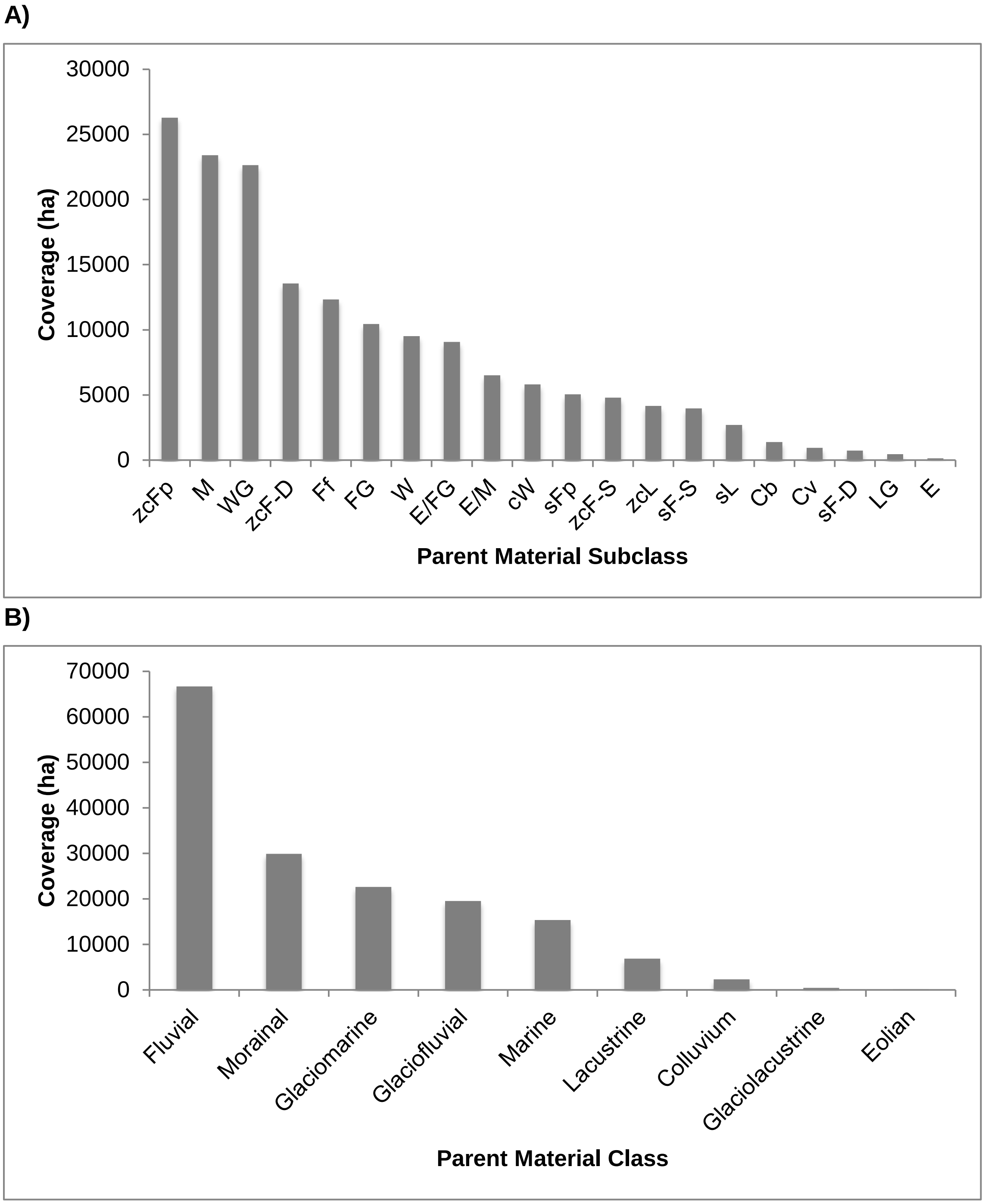


Figure 2.3. **Coverage of single-component parent material polygons by (A) subclass and by (B) class from Soils of the Langley-Vancouver Area (Luttmerding, 1981). See Table 2.1 for a description of the parent material classes.**

Optimization of m_{try}

Based on preliminary results, n_{tree} = 750 was use as it produced stable OOB error rates and was also small enough to maximize computational efficiency. In addition, a terminal node size of k = 1 was selected as an increasing k resulted in a monotonic increase in the OOB error rates.

To optimize the primary tuning parameter, m_{try} values ranging from 1 to 27 were tested and the OOB error rates from 50 replicates for each m_{try} value were assessed. In addition, m_{try} values were further assessed using error rates obtained from 20 replications of a 5-fold CV. Where m_{try} = 1, a random predictor variable is selected at each node; contrarily, $m_{try} = p$ has the same effect as bagging the predictors.

Variable Reduction

Variable reduction has previously been shown to result in slight error reductions (Svetnik et al., 2003; 2004), or to have minimal effect on the RF classifier (Xiong et al., 2012) through the removal of potentially irrelevant predictor variables. In this study, variable reduction was tested in order to examine whether or not a smaller set of predictors would lead to an improvement in RF predictions based on the following algorithm adopted from Svetnik et al. (2003):

1. The RF classifier was initially applied using the entire set of predictors. Variable importance, based on the mean decrease in accuracy, was used to rank the predictor variables.

2. Using the variable rankings, the three least important predictors were removed.

3. The training data was then partitioned into 5-folds for cross-validation and the error rates for each of the 5 cross-validation partitions were aggregated into a mean error rate. 20 replicates of 5-fold CV was performed.

4. Steps 2 and 3 were repeated until 3 predictors remained.

To test the effects of variable reduction, an initial variable importance plot was generated using the default settings of RF and the area-weighted sampling approach. Variable ranking was done using the MDA as it provides a more reliable measure of variable importance in comparison to the MDG (Bureau et al., 2003).

Since the choice of m_{try} depends on the total number of predictor variables (*p*), m_{try} was calculated as a function of *p*. Here, the m_{try} functions were defined as follows: $m_{try} = p$ (bagging), *p/2*, *p/4*, and $p^{1/2}$ (default setting). A resulting parent material prediction was generated using a reduced number of predictors with an optimal m_{try} function as a basis for comparison to the map produced using the entire variable set.

2.3.5. Assessment of Predictions

Three approaches for assessing the predictions made by RF were used. Firstly, RF predictions were compared to the single-component polygons used as training areas from the soil survey. Secondly, RF predictions were compared to the multi-component polygons from the soil survey in order to examine RF's ability to disaggregate complex mapping units. Finally, the RF predictions were validated using point data.

For assessment purposes, the raw parent material maps were reclassified by generalizing the 20 parent material subclasses to 9 classes in order to offset the limited number of field validation points for each parent material subclass. In addition, OOB error rates were recalculated in order to reflect this reclassification procedure. Furthermore, a preliminary study showed that RF performed better when the training data were derived from parent material subclasses and the results were later generalized to classes.

Consistency with Single-Component Polygons

Using *Map Comparison Kit 3* (Van Vliet, 2003), *overall agreement* was calculated as the percentage of pixels that were correctly classified by RF and the *disagreement with soil survey* was calculated as the percentage of pixels that were incorrectly classified by RF. In addition, the kappa index, a measure of map agreement that considers map agreement that occurs by 'chance', was also calculated for map comparison using the following (Visser and de Nijs, 2006):

Eq. (2.2) $$\kappa = \frac{\mathrm{P}(A)\text{-}\mathrm{P}(E)}{1\text{-}\mathrm{P}(E)},$$

where *P(A)* represents the actual agreement fraction and *P(E)* represents the expected agreement fraction between the soil survey and the RF predictions. Because the 'chance' factor is taken into account, the kappa index is consistently lower than the overall agreement.

Disaggregation of Single-Component Polygons

To evaluate the effectiveness of RF in disaggregating multi-component map units, the proportion of parent material classes specified for each unit in the soil survey was compared to the proportional extent of the predicted classes. Model residuals, $\varepsilon_{c,j}$, were calculated from the difference between RF's predicted extent, $\hat{n}_{c,j}$ [% of polygon], of a parent material class, *c*, for a polygon, *j*, and the parent material's estimated extent from the soil survey, $n_{c,j}$ [% of polygon] under the same polygon in the following:

$$\text{Eq. (2.3)} \qquad \varepsilon_{c,j} = \hat{n}_{c,j} - n_{c,j} \ .$$

Validation with Point Data

Legacy soil pit data from the British Columbia Soil Information System (BCSIS) (Sondheim and Suttie, 1983), which consists of *n* = 248 points, were supplemented with additional data collected from fieldwork (between April and August, 2009; *n* = 59) in order to form an external validation point dataset with *n* = 307 points. Because the BCSIS data points were primarily located in the agricultural landscapes of the Lower Fraser Valley, supplemental data points were established along the Coastal Mountain on forested landscapes. Due to the forested and mountainous terrain, the supplemental data points were located along areas with good access and in places that reflected the range of materials present. To account for uncertainty in the location of the original soil pits, two levels of validation were used. At the first level, a predicted cell was considered valid if the validation point matched the prediction at that exact location (*r* = 0). At the second level, if a validation point matched a predicted cell that was located within a radius of 1 cell (*r* = 1), or within 100 m, surrounding the validation point, the predicted cell was considered valid.

2.4. Results & Discussion

2.4.1. Development of Training Data

In general, the OOB error rates produced from RF's internal validation were similar to the disagreement with soil survey rates (Figure 2.4A). For the equal sampling by polygon approach, however, the OOB error rates were more than 10% lower than the disagreement with soil survey for the colluvium, glaciolacustrine, and morainal parent material classes. These discrepancies suggest that the OOB error rates may not be the most reliable measure of class error; hence, the overall agreement with soil survey and kappa indices were used to select the sampling approach used for the parameter optimization and variable reduction analyses.

Overall agreement and kappa were highest for the area-weighted sampling approach when compared to the single-component polygons from the soil survey (Table 3). Moran and Bui (2002) noted that the area-weighted sampling approach performed better because more training data points were used to represent geographically extensive classes in order to capture a greater amount of variability that occurs under these classes. In this study, it was observed that the area-weighted sampling approach resulted in a lower error in agreement with soil survey for the most common (majority) classes, which include fluvial, morainal, and glaciomarine parent materials. In comparison, the equal-class sampling approach was superior in predicting the minority classes (e.g. eolian, glaciolacustrine, colluvium, and lacustrine parent materials) (Figure 2.4B). The discrepancy in performance between majority and minority classes was expected as machine-learning algorithms are recognized for their poor performance for minority classes (e.g. Kubat and Matwin, 1997; Van Hulse et al., 2007).

Table 2.3. Overall agreement within single-component soil survey polygons based on sampling by equal number per polygon, equal-class, and area-weighted.

Sampling Method	**RF Internal Validation**	**Soil Survey**	
	Out-of-Bag Error (%)	Overall Agreement (%)	Kappa (%)
By Polygon	7.8	86.6	82.9
Equal-Class	7.0	90.3	87.1
Area-Weighted	8.3	92.2	89.6

It was recognize that a potential problem that may arise with the use of the area-weighted sampling approach was that such an approach would lead to an unbalanced training dataset – a common problem for various machine-learning approaches (Van Hulse et al., 2007; Van Hulse and Khoshgoftaar, 2009; Galar et al., 2011). Despite these differences, this study was primarily aimed at producing a parent material map with the lowest overall error and, hence, the area-weighted sample set was selected for all remaining analyses. A rigorous study in addressing the issue of an unbalanced dataset was beyond the scope of this study; however, such a study would be of use in cases where a study's objective is to predict the presence of rare soils or unique features in a landscape.

2.4.2. Optimization of m_{try}

The CV error rates reached a minimum when m_{try} ranged from 15 to 21; although, the increase of m_{try} from 11 to 15 only amounted to a minor decrease in CV error rate of 0.1% (Figure 2.5). In the optimization of RF's main tuning parameter, m_{try}, it was determined that the OOB error rate was a fairly adequate measure of the model error when compared to the 5-fold CV error rates. Generally, the OOB error rates were consistently lower than the CV error rates by a margin of roughly 1%. The lower OOB error rates were expected because fewer training points were used to build a RF using the partitioned 5-fold CV training dataset, which was further partitioned through bootstrap sampling. Based on 20 replicates of 5-fold CV, m_{try} = 11 was used; in addition, the smaller value of m_{try} was used in order to retain more of the 'randomness' in RF's randomized variable selection process.

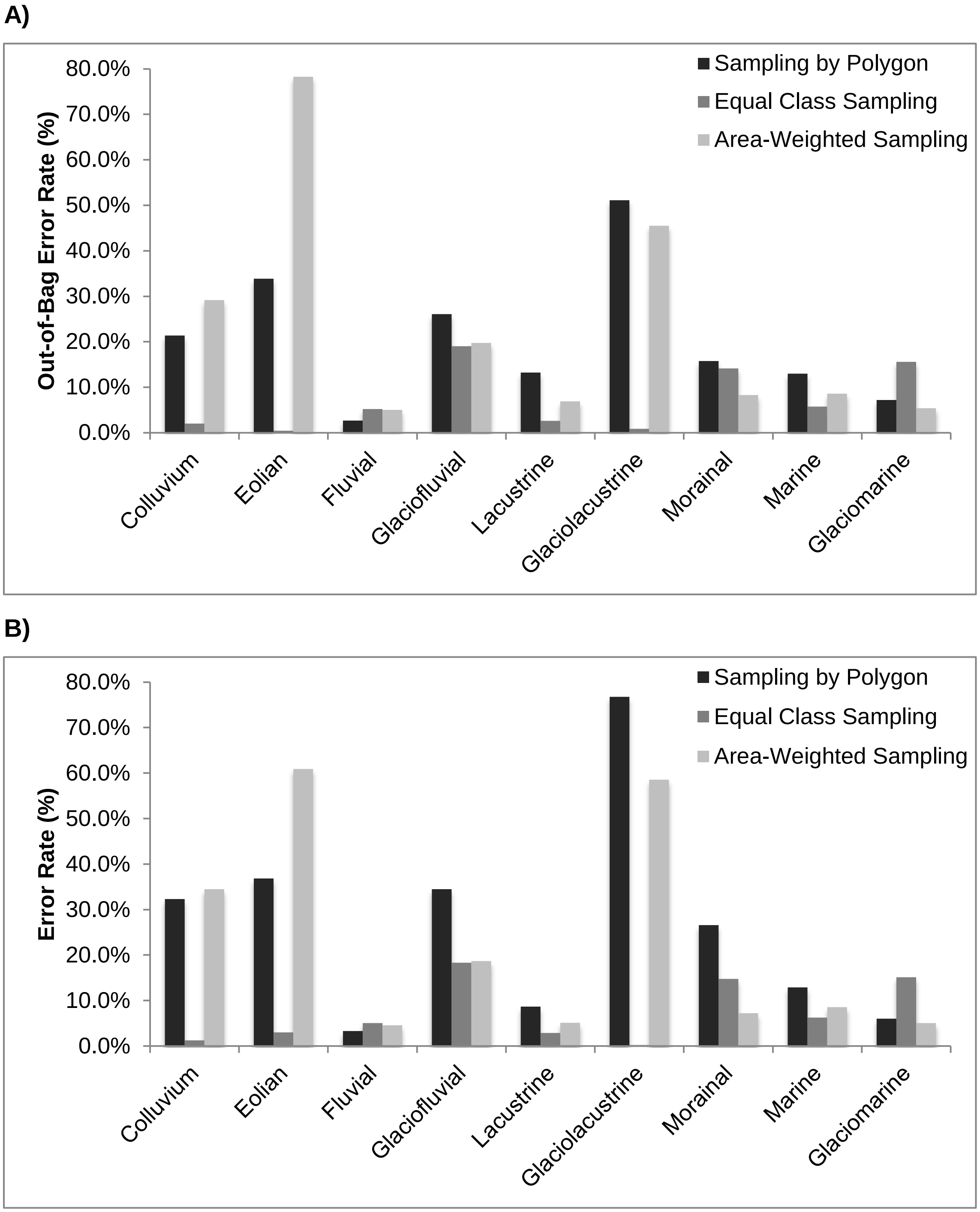


Figure 2.4. **(A) Out-of-bag error rates (%) and (B) disagreement with soil survey (%) by parent material class using sampling by polygon, equal-class sampling, and area-weighted sampling.**

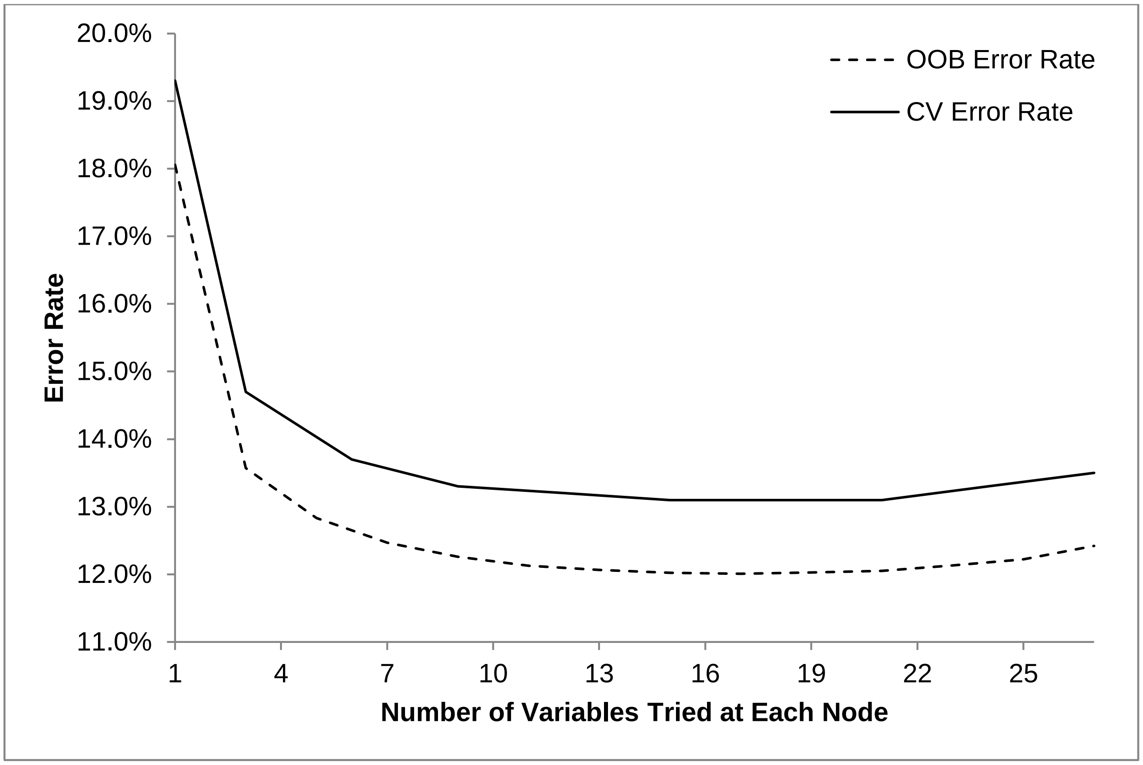


Figure 2.5. **Non-aggregated overall out-of-bag error rates and mean 5-fold CV error rates with respect to the number of predictor variables tried at each node (m_{try}).**

2.4.3. Variable Reduction

Based on the MDA values from the variable importance plot generated by RF, it was observed that the most important variables were aspect, distance to nearest stream, convergence index, and distance to nearest river whereas slope-length, slope, plan curvature, and profile curvature were the least important (Figure 2.6). A detailed further examination between environmental covariates and parent material classes was beyond the scope of this study since the topographic indices were all derived from the same DEM; and hence, an inherently high level of cross-correlation between the indices would make such a detailed analysis to be highly complex.

From this study it was determined that the CV error rates produced when m_{try} = p, $p/4$, $p/2$, and $p^{1/2}$ remained fairly consistent until the number of predictors were reduced to p = 9 (Figure 2.7). As the number of variables reduced to p = 3, the CV error rates increased to 64% for each m_{try} function (not shown in Figure 2.7). Overall, m_{try} = $p/2$ resulted in a slightly better overall performance; however, the difference in CV error rates in comparison to other m_{try} functions were less than 0.5%. Hence, it was found that variable reduction did not necessarily result in an improvement in RF performance with respect to the CV error rates. Furthermore, the minimal degradation in RF predictions when the predictors were reduced to p = 9 indicates that, for our study area, RF is insensitive to the presence of irrelevant predictors. These findings corroborate the results in Svetnick et al. (2003) and Xiong et al. (2012) where variable reduction algorithms were also tested. Although this study only examined a single approach for variable reduction, further studies may explore alternative dimension reduction approaches, such as the use of principal components as predictors for RF.

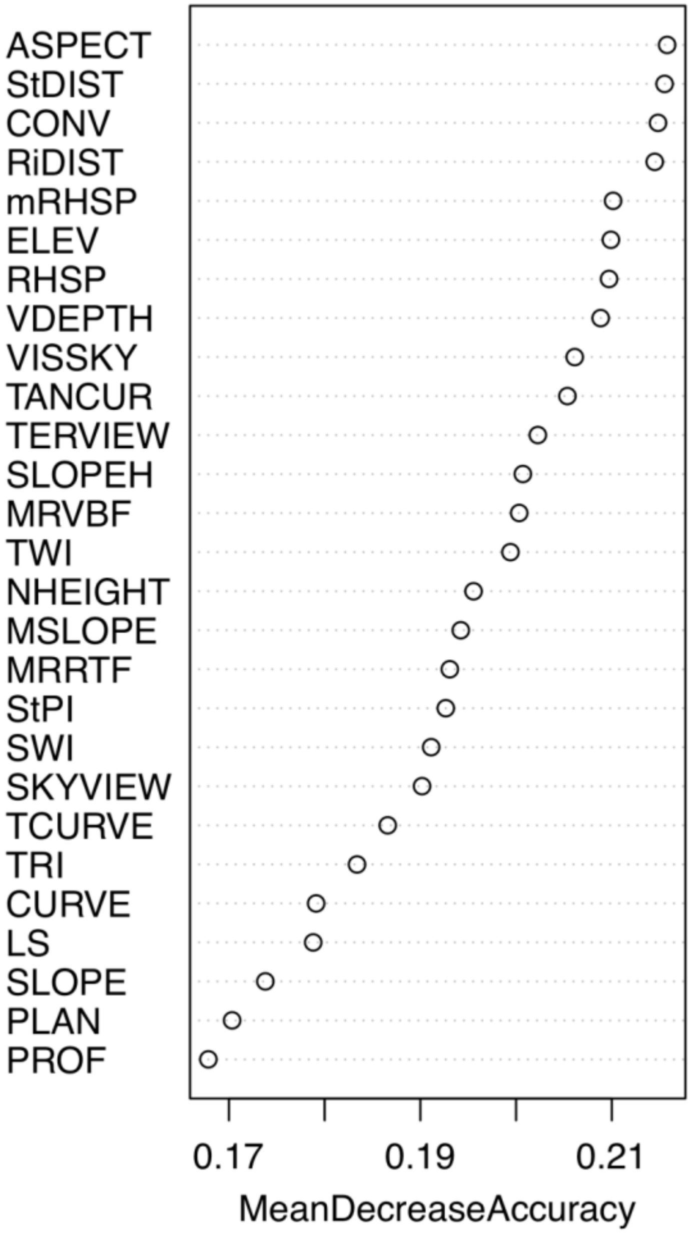


Figure 2.6. Variable importance plots based on mean decrease in accuracy using area-weighted sampling. See Table 2.2 for the description of predictor variables.

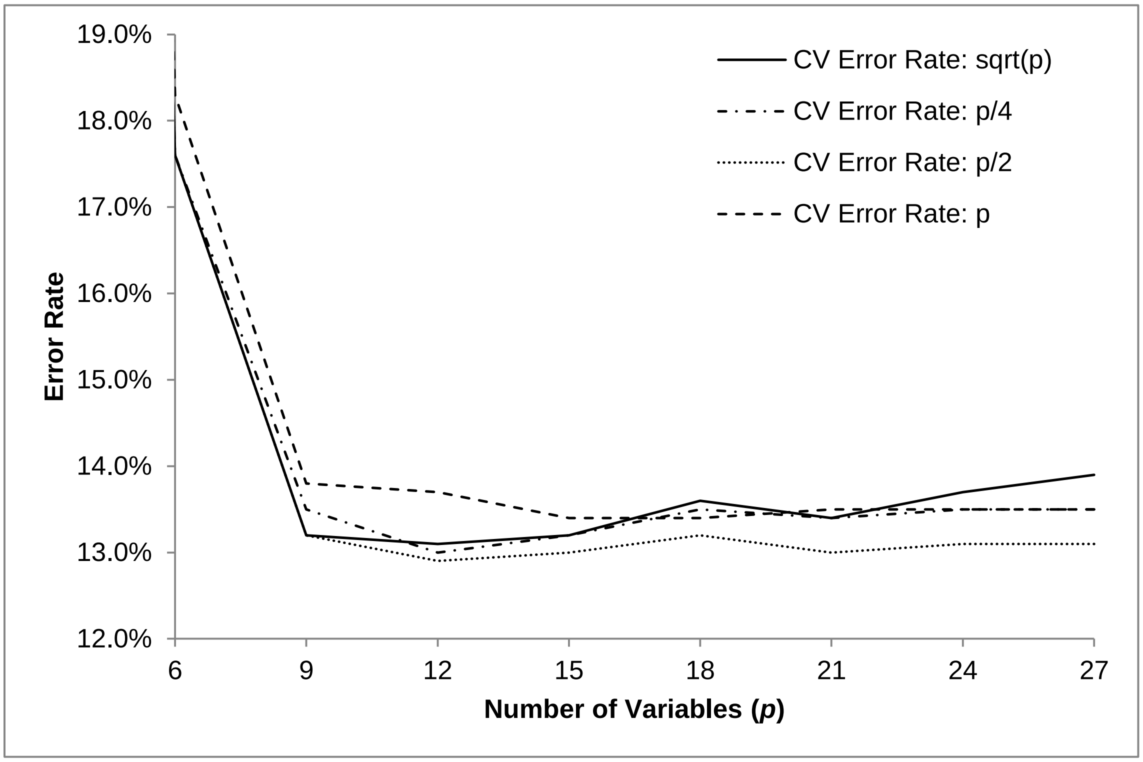


Figure 2.7. Non-aggregated mean cross-validation (CV) test error rates with 3 predictor variables removed at each step using various m_{try} functions: m_{try} = sqrt(*p*); *p*/4; *p*/2; and *p*.

2.4.4. Assessment of Predictions

Consistency with Single-Component Polygons

There was a high overall agreement between the RF predictions and single-component polygons from soil survey data (Table 2.4). Based on the overall agreement with soil survey data, m_{try} optimization resulted in a minimal effect when the entire variable set was used while variable reduction increased the overall agreement by 0.9% when compared to the optimized RF using 27 predictors. Kappa indices indicated a high overall agreement between predicted parent material maps and the soil survey data. The optimization of m_{try} and variable reduction, however, increased kappa minimally.

By examining the various error rates for each parent material class (Figure 2.8), it was observed that m_{try} optimization had a minimal effect on improving the agreement with soil survey data in the cases of fluvial, lacustrine, morainal, and marine parent materials. Improvements in agreement for these classes were less than 0.5%. Improvements in agreement with soil survey data occurred primarily for the minority classes such as eolian, colluvium, and glaciolacustrine, which had an increase in agreement of 3.0%, 4.3%, and 12.5%, respectively.

Table 2.4. Classification accuracy measurements using non-optimized RF, optimized RF with no variable reduction, and optimized RF with variable reduction.

Number of Predictors	**RF Internal Validation**	**Soil Survey**		**External Validation (*R* = 0 cell)**		**External Validation (*R* = 1 cell)**	
	Out-of-bag Error (%)	Overall Agreement (%)	Kappa (%)	Overall Accuracy (%)	Kappa (%)	Overall Accuracy (%)	Kappa (%)
Not Optimized							
27	8.3	92.2	89.6	77.5	69.1	85.0	79.5
Optimized							
12	7.3	93.0	90.7	77.9	69.7	85.7	80.3
27	7.7	92.8	90.4	77.5	69.2	85.7	80.4

By way of a visual comparison between the single-component parent material polygons and the continuous surface generated using RF (Figure 2.9), it was observed that RF was able to produce results that had patterns and boundaries that were qualitatively similar to the single-component polygons. Figure 2.9A shows a close-up of an area with low relief terrain, adjacent to the Fraser River, which is typical of the southern region of the study area and where fluvial, glaciofluvial, marine, and glaciomarine parent materials are most common. Based on the visual comparison and a low disagreement with soil survey for the listed parent materials, the RF results were fairly consistent with the single-component polygons for low relief terrain. This was to be expected since the majority of the training points were located in low relief terrains. In comparison, Figure 9b shows a close-up of an area with high relief terrain, which is typical of the northern region where the dominant parent material is moraine; however, the presence of colluvial, fluvial, and glaciofluvial deposits are also common. RF was able to produce parent material boundaries that were similar to single-component polygon boundaries; however, it was also noted that RF over-predicted the presence of morainal deposits and under-predicted the presence of colluvial deposits since the steep slopes were classified as morainal when colluvial materials were expected.

Disaggregation of Multi-Component Polygons

Histograms of the distribution of model residuals, $\varepsilon_{c,j,}$ were produced based on polygons where a parent material class, *c*, was a component of multi-component polygon, *j* (Eq. 2.3). Examples for glaciomarine, fluvial, colluvial, and morainal parent materials using an optimized RF and p = 27 predictors are presented in Figure 2.10. In Figure 2.10, $\varepsilon_{c,j}$ = 0 represents cases of multi-component polygons where the proportional extent of a parent material class estimated by the soil survey, $n_{c,j}$, matched the extent of the same parent material class predicted by RF, $\hat{n}_{c,j}$. Where $\varepsilon_{c,j} > 0$, the parent material class was over-predicted; conversely, where $\varepsilon_{c,j} < 0$, the parent material class was under-predicted for polygon, *j*.

These results confirm the initial visual assessment from Section 2.4.3 and suggests that the topographic distinctions between morainal and colluvial deposits may be difficult for RF to detect. The distinctions between these parent materials were not entirely clear in the soil survey map (Luttmerding, 1981) for several reasons. Firstly, colluvial and morainal components were frequently coupled together in 481 multi-component polygons; and therefore not included in the training dataset. Secondly, the soil survey was carried out at a 1:50,000 scale for the Coast Mountains whereas the Lower Fraser Valley was carried out at the 1:25,000 scale. Consequently, the smaller scale mapping of the Coast Mountains inherently resulted in greater generalization of the map units where localized colluvial deposits would not have been mapped as a single-component map unit, but rather, a multi-component unit. The generalization of morainal and colluvial deposits into multi-component polygons would also explain the small number of single-component colluvium polygons that were available to train the RF and hence, the poor disaggregation of morainal-colluvial complexes.

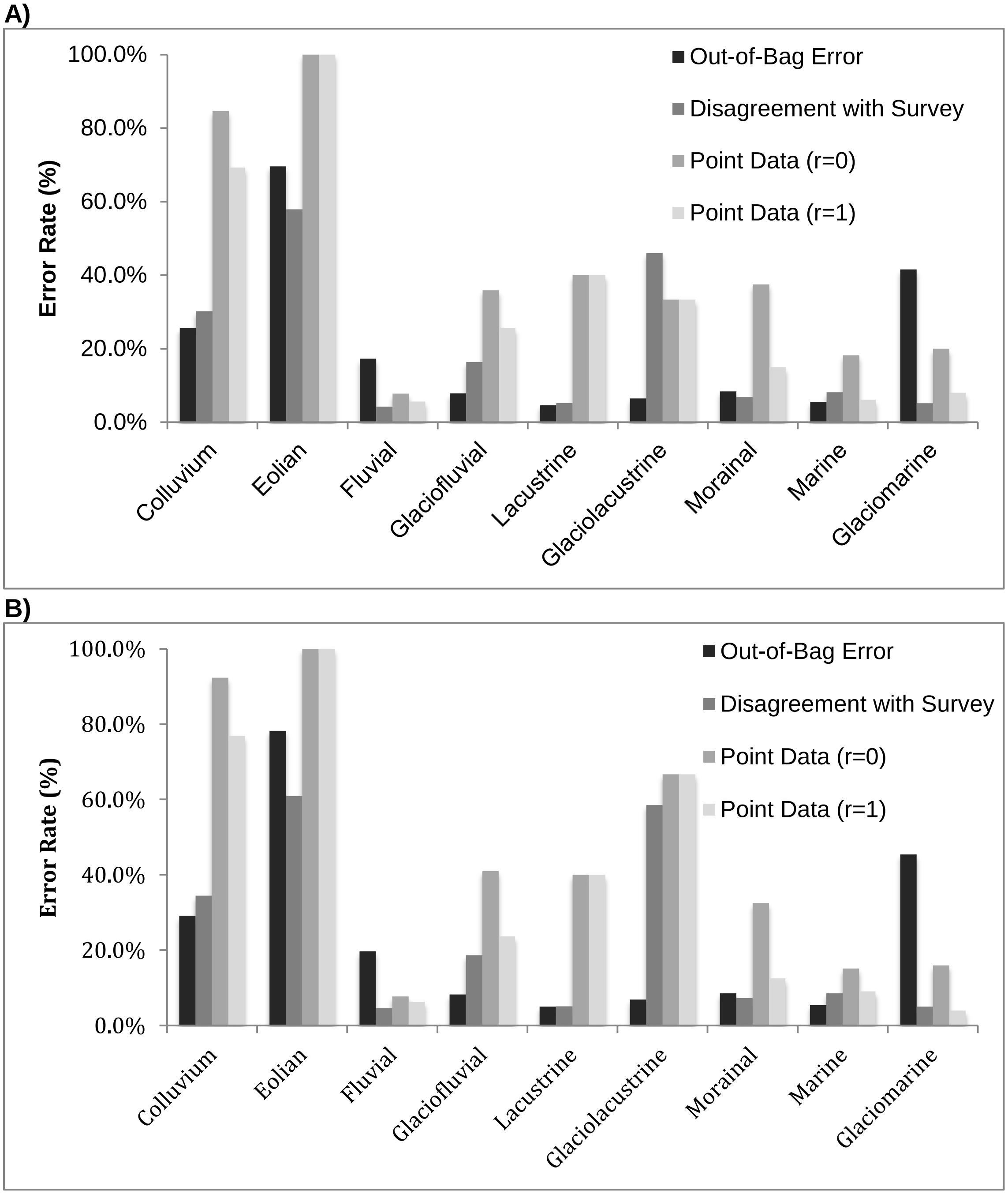


Figure 2.8. **Out-of-bag error rates (%), disagreement with soil survey (%), and error rates using validation points (%) by parent material class for a radius of *R* = 0 and *R* = 1 cells using (A) non-optimized Random Forest; (B) optimized Random Forest with *p* = 27 predictor variables.**

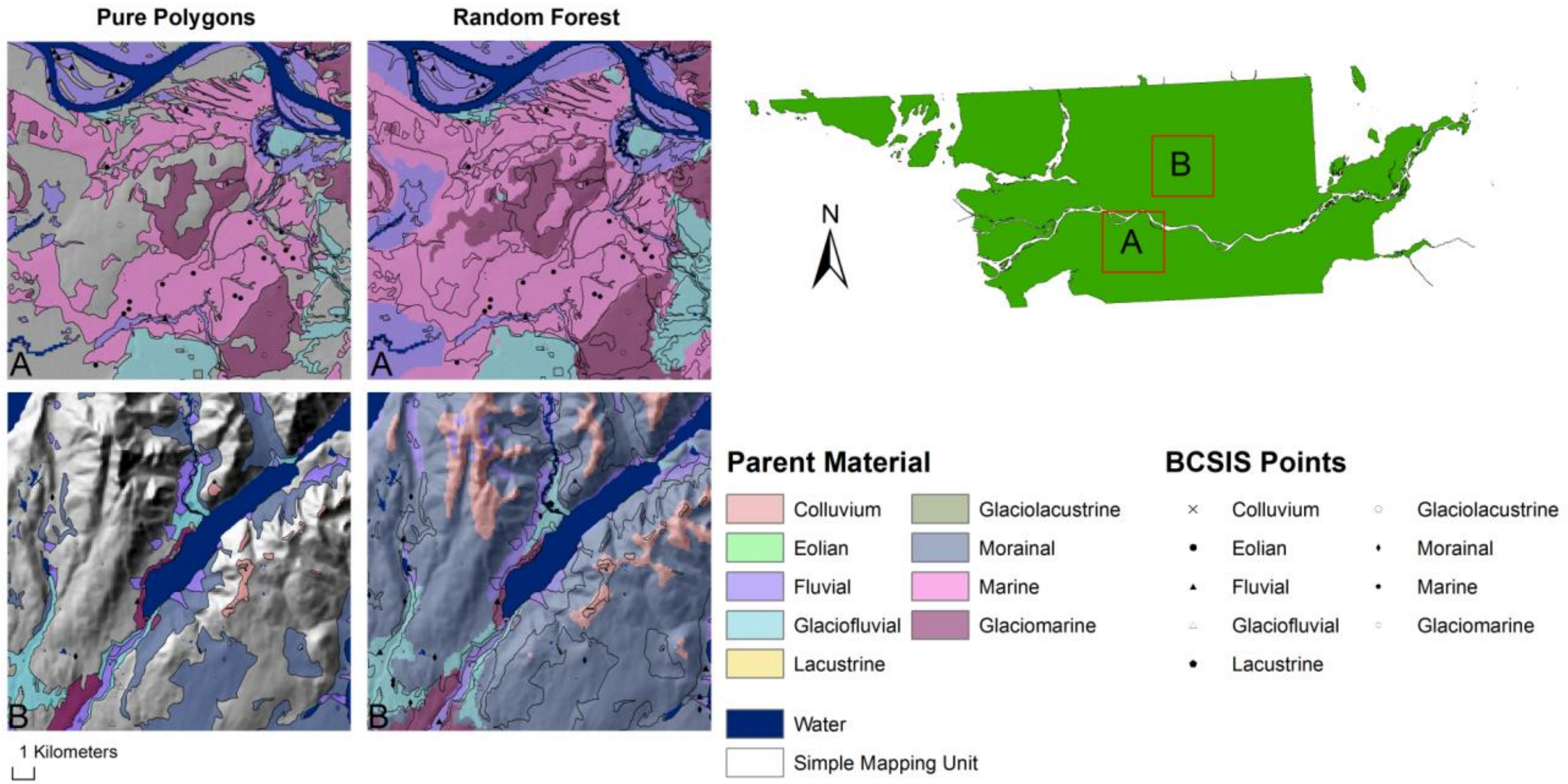


Figure 2.9. **Close-up map of single-component (pure) parent material polygons, RF results, and sample points overlaid on a hill-shade for (A) a low relief terrain and (B) a high relief terrain. The map uses an optimized m_{try} with p = 27 predictor variables.**

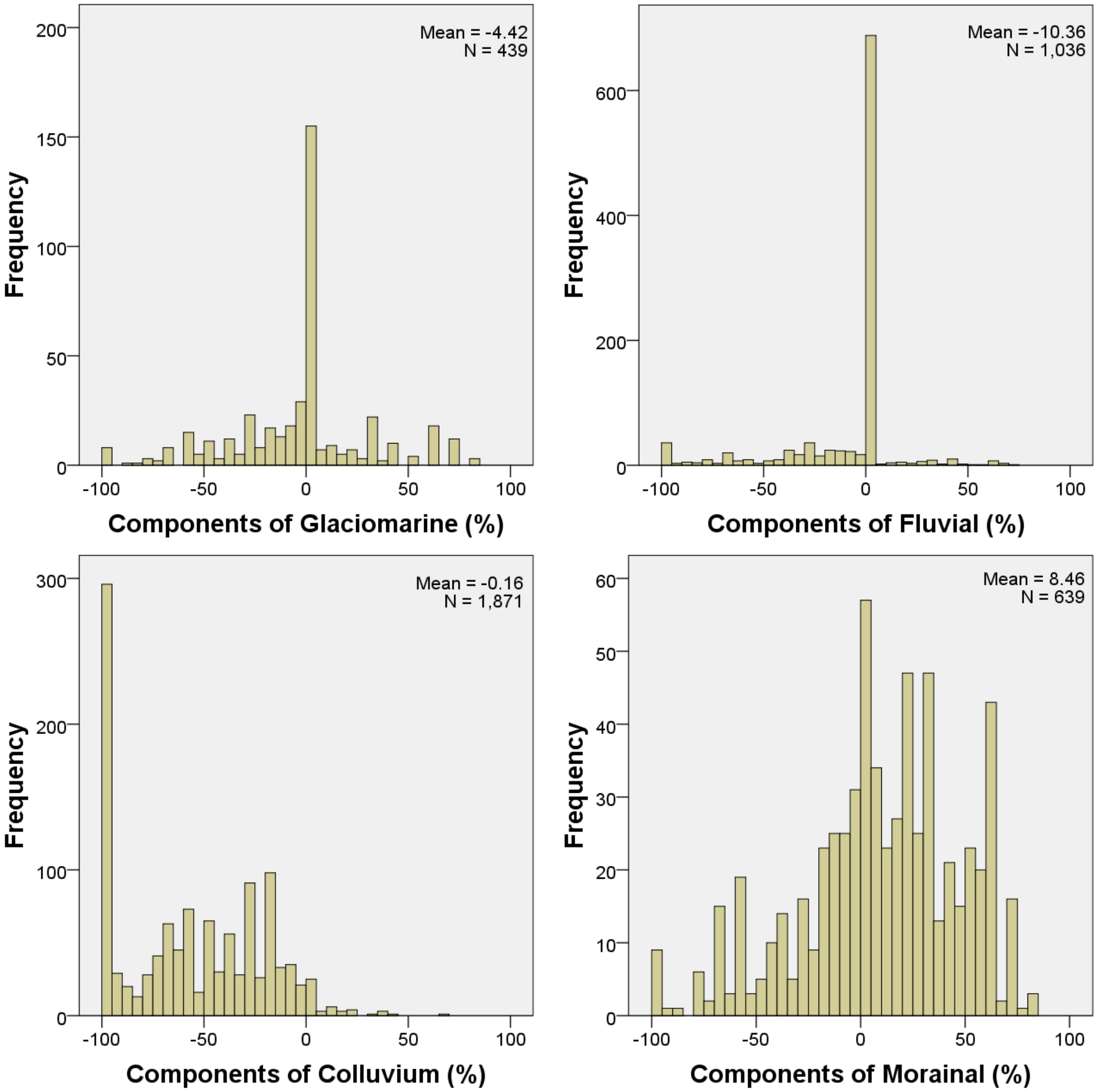


Figure 2.10. **Histograms of model residuals, $\varepsilon_{c,j}$, calculated as the difference between the predicted RF extent and the soil survey extent for each multi-component polygon, *j*. Histograms only consider polygons where parent material class, *c*, is a component of polygon, *j*. Histograms are based on an optimized m_{try} with p = 27 predictor variables.**

The findings of using RF for polygon disaggregation are further summarized in Table 2.5. Most parent materials were not predicted by RF when they were not a component of a soil survey polygon, with the exceptions of morainal and fluvial materials. Both the optimization of m_{try} and variable reduction had little influence on the disaggregation of multi-component polygons. Colluvial, eolian, fluvial, glaciofluvial, marine, and glaciomarine materials were under-predicted when they were included as components of a polygon; whereas, lacustrine and morainal materials were over-predicted. There were 815 instances where morainal materials were predicted in polygons where they were not identified in the soil survey; consequently, this inherently would have contributed to the under-prediction of the parent materials that were mapped as a component of a polygon. In contrast to the polygon disaggregation study in Häring et al. (2012), this study did not constrain the number of different parent material classes to the ones identified by the multi-component polygons in order to account for the inclusion of parent materials that were not recognized. Hence, the polygon components identified by the soil survey would have been under-predicted due to the presence of small inclusions of other parent materials in the polygons.

Validation with Point Data

When the predicted parent material maps were compared to the validation points (Figure 2.11), the overall accuracy and kappa indices were lower than the agreement with soil survey data (Table 2.4). It was observed that between 77% and 78% of the validation points matched the predicted parent material map exactly. Comparing the overall accuracy to the kappa index, there was a difference of 8%. Differences between the overall accuracy and kappa index suggest that there was a low to moderate probability that cells were correctly classified by chance. When examining the cells that were within a 1-cell (100 m) radius of a validation point, it was observed that the overall accuracy and kappa index increased by 8% and 10%, respectively on average. This suggests that the RF produced a fairly accurate map within 100 m with an average overall accuracy of 85% and an average kappa index of 80%.

Table 2.5. Descriptive statistics for model residuals, $\varepsilon_{c,j}$, calculated as the difference between the predicted RF extent and the soil survey extent of each parent material class, c, for each multi-component polygon, j.

	Parent Material Class	***n***	**27 Variables**		**27 Variables + Optimization**		**12 Variables + Optimization**	
			Mean (%)	St. Dev. (%)	Mean (%)	St. Dev. (%)	Mean (%)	St. Dev. (%)
Non-Component of Polygons[1]	Colluvium	1871	0.2	0.1	0.2	0.1	0.2	0.1
	Eolian	2969	0.0	0.0	0.0	0.0	0.0	0.0
	Fluvial	1989	7.4	0.5	7.8	0.5	7.8	0.5
	Glaciofluvial	2758	3.5	0.3	4.1	0.3	5.2	0.3
	Lacustrine	2924	0.2	0.1	0.2	0.1	0.2	0.1
	Glaciolacustrine	3015	0.0	0.0	0.0	0.0	0.0	0.0
	Morainal	2386	22.7	0.7	21.5	0.7	21.3	0.7
	Marine	2654	1.3	0.2	1.3	0.2	1.2	0.2
	Glaciomarine	2586	3.5	0.3	3.4	0.3	3.3	0.3
Component of Polygons[2]	Colluvium	1154	-57.4	1.0	-56.8	1.0	-57.7	1.0
	Eolian	56	-61.4	4.1	-60.4	4.2	-61.1	4.1
	Fluvial	1036	-10.7	0.9	-10.4	0.9	-10.6	0.9
	Glaciofluvial	267	-29.7	2.4	-27.0	2.4	-25.9	2.5
	Lacustrine	101	16.3	3.5	17.6	3.5	15.9	3.8
	Glaciolacustrine	10	23.5	10.1	-65.9	7.8	-65.9	7.7
	Morainal	639	10.0	1.5	8.5	1.5	7.9	1.5
	Marine	371	-25.7	2.1	-24.4	2.1	-23.3	2.2
	Glaciomarine	439	-2.8	1.6	-4.4	1.6	-6.2	1.7

[1] Polygons where parent material class, *c*, is not a component of multi-component polygon, *j*.

[2] Polygons where parent material class, *c*, is a component of multi-component polygon, *j*.

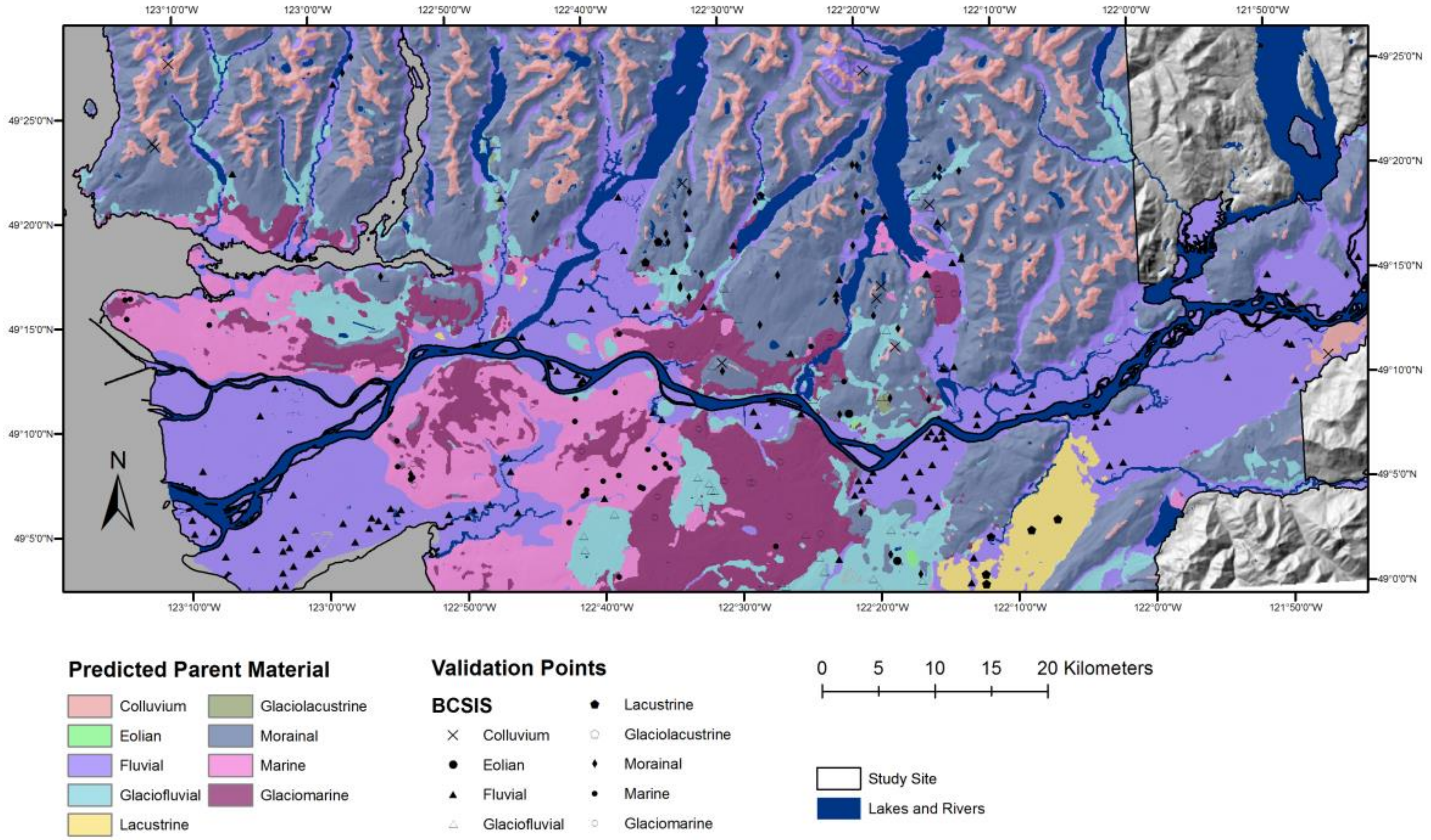


Figure 2.11. Predictive parent material map using Random Forest at a 100 m spatial resolution with underlying hill-shade and overlying sample points for the Langley-Vancouver Map Area, British Columbia. The map uses an optimized m_{try} with p = 27 predictor variables.

Based on the validation using point data, m_{try} optimization and variable reduction resulted in little improvement in predictions for each parent material class (Figure 2.8). A comparison between the results produced with optimized RF and the results with variable reduction were similar with minimal differences (<1 %) in agreement with soil survey as well as with the validation points. For glaciolacustrine and colluvium classes, m_{try} optimization resulted in a 33% and 7.7% increase for those respective classes; however, the seemingly large increase is the result of a small sample size for those classes.

Map comparison using the single-component polygons from the soil survey data resulted in higher prediction accuracy compared to the prediction accuracy using the point data. These differences in accuracy are likely caused in part by the initial use of the soil survey data to stratify the training data for the RF model. Secondly, the soil survey polygons represent an aggregation of the soil-environmental conditions for each map unit whereas the point data may not necessarily be representative of the average environmental conditions from which map units are derived and from which the RF model is based on.

2.5. Conclusions

The objective of this study was to first evaluate methods for the extraction of training data from legacy soil data and the optimization of RF parameters. It was determined that the imbalanced area-weighted sampling resulted in higher overall agreement with soil survey with lower error rates for majority parent material classes such as fluvial, morainal, marine and glaciomarine materials; however, the prediction of minority classes was less successful. Using a balanced dataset improved the prediction of the minority parent material classes such as eolian, glaciolacustrine, colluvium, and lacustrine materials; however, overall agreement with soil survey decreased as a consequence. This research suggests that the selection of a sampling approach for training data should reflect the objectives of the study and whether the goal is to maximize overall accuracy or to maximize the accuracy of the minority classes. Furthermore, this research also suggests that the relationship between imbalanced multi-class training data and machine-learning approaches should be investigated further.

In terms of the optimization of RF parameters, this study has found through extensive CV testing, that both the m_{try} optimization and variable reduction had little effect in improving RF outputs. As a result, it was concluded that RF performs well with minimal user intervention through the parameterization of the model. In addition, it was found that RF was able to identify important predictors, internally, as the reduction of predictors resulted in marginal improvements in overall agreement with soil survey and overall accuracy.

The second objective of this study was to assess the reliability of RF outputs within single-component polygons. It was determined that RF produced maps that had a high overall agreement with soil surveys and that RF was effective in extracting the relationships between parent material and topography. In comparison, however, it was also concluded that RF was not as effective in the disaggregation of multi-component parent material polygons. These results may illustrate the importance of the training dataset as much as the characteristics of RF, since our training data was concentrated in areas of the map with single-component polygons. This study has found that the RF classifier is an effective machine-learning and data mining approach. Our approach to developing a training dataset by extracting points from single-component polygons likely limited the performance of RF for disaggregation of multi-component polygons.

2.6. Acknowledgements

The authors are thankful for the support from the Forest Science Program of the Ministry of Forests, Lands, and Natural Resource Operations; the Future Forest Ecosystem Scientific Council of British Columbia and from a Natural Sciences and Engineering Research Council Discovery grant to M. Schmidt. The authors are also grateful for the field and lab assistance provided by Darren Murray, Sarah Robertson, Maciej Jamrozik, Jin Zhang, Lillian Fan, and in particular, Darrell Hoffman.

2.7. References

Armstrong, J.E., 1956. Surficial geology of Vancouver area, British Columbia. Paper 55-40, Geological Survey of Canada, Ottawa.

Armstrong, J.E., 1957. Surficial geology of New Westminster map-area, British Columbia. Paper 57-5 and Map 16, Department of Mines and Technical Surveys, Ottawa.

B.C. Ministry of Sustainable Resource Management, 2002. Gridded Digital Elevation Model Product Specifications 2nd edition. Base Mapping and Geomatics Services Branch Ministry of Sustainable Resource Management, Victoria.

Beckett, P.H.T., 1971. The cost-effectiveness of soil survey. Outlook Agriculture 6, 191-198.

Beven, K.J., Kirkby, M.J., 1979. A physically based, variable contributing area model of basin hydrology. Hydrological Sciences Bulletin 24, 43-69.

Böhner, J., R. Köthe, O. Conrad, J. Gross, A. Ringeler, Selige, T., 2002. Soil regionalization by means of terrain analysis and process parameterization. European Soil Bureau – Research Report 7, 213-222.

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

Breiman, L., Friedman, J., Stone, C.J., Olshen, R.A., 1984. Classification and Regression Trees. CRC Press LLC, Boca Raton, FL.

Bui, E.N., 2004. Soil survey as a knowledge system. Geoderma 120, 17-26.

Bui, E.N., A. Lougheed, Corner, R., 1999. Extracting soil-landscape rules from previous soil surveys. Academia Journal of Scientific Research 37, 495-508.

Bui, E.N., Moran, C.J., 2001. Disaggregation of polygons of surficial geology and soil maps using spatial modeling and legacy data. Geoderma 103, 79-94.

Bui, E.N., Moran, C.J., 2003. A strategy to fill gaps in soil survey over large spatial extents: an example from the Murray-Darling basin of Australia. Geoderma 111, 21-44.

Bui, E.N., B.L. Henderson, Viergever, K., 2006. Knowledge discovery from models of soil properties developed through data mining. Ecological Modelling 191, 431-446.

Bui, E.N., B.L. Henderson, Viergever, K., 2009. Using knowledge discovery with data mining from the Australian Soil Resource Information System database to inform soil carbon mapping in Australia. Global Biogeochemical Cycle 23: GB4033. http://dx.doi.org/10.1029/2009GB003506

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

Church, M. and J.N. Ryder. The physiography of British Columbia. P 17-44 in Pike, R.G., T.E. Redding, R.D. Moore, R.D. Winker and K.D. Bladon (editors). 2010.Compendium of Forest Hydrology and Geomorphology in British Columbia. B.C. Ministry of Forest and Range, Forest Science Program, Victoria, B.C. and FORREX Forum for Research and Extension in Natural Resources, Kamloops, B.C. Land Management Handbook. 66.

Cutler, R.D., T.C. Edwards, K.H. Beard, A. Cutler, K.T. Hess, J. Gibson, Lawler, J.J., 2007. Random forests for classification in ecology. Ecology 88: 2783-2792.

Díaz-Uriarte, R., Alvarez de Andrés, S., 2006. Gene selection and classification of microarray data using Random Forest. BMC Bioinformatics 7, 3-15.

Florinsky, I.V., 1998. Accuracy of local topographic variables derived from digital elevation models. International Journal of Geographical Information Science 12, 47-61.

Galar, M., Fernández, A., Barrenechea, E., Bustince, H., Herrera, F., 2011. A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. IEEE Transactions on Systems, Man, and Cybernetics – Part C: Applications and Reviews 42, 463-484.

Gallant, J.C., Dowling, T.I., 2003. A multiresolution index of valley bottom flatness for mapping depositional areas. Water Resource Research 39, 1347-1359.

Geng, X., Fraser, W., VandenBygaart, B., Smith, C.A.S., Wadell, A., Jiao, Y., Patterson, G., 2010. Towards digital soil mapping in Canada: Existing soil survey data and related expert knowledge. In Digital Soil Mapping: Bridging Research, Environmental Application, and Operation, Springer, 325-337.

Gislason, P.O., Benediktsson, J.A., Sveinsson, J.R., 2006. Random forests for land cover classification. Pattern Recognition in Remote Sensing 27, 294-300.

Grimm, R., Behrens, T., Märker, M., Elsenbeer, H., 2008. Soil organic carbon concentrations and stocks on Barro Colorado Island – digital soil mapping using Random Forests analysis. Geoderma 146, 102-113.

Grinand, C., Arrouays, D., Laroche, B., Martin, M.P., 2008. Extrapolating regional soil landscapes from an existing soil map: sampling intensity, validation procedures, and integration of spatial context. Geoderma 143, 180-190.

Häntzchel, J., Goldberg, V., Bernhofer, C., 2005. GIS-based regionalization of radiation, temperature and coupling measures in complex terrain for low mountain ranges. Meteorological Applications 12, 33-42.

Häring, T., Dietz, E., Osenstetter, S., Koschitzki, T., Schröder, B., 2012. Spatial disaggregation of complex soil map units: A decision-tree based approach in Bavarian forest soils. Geoderma 185-186, 37-47.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed. Springer, New York, NY, 734 pp.

Hectares BC, 2012. Hectares BC. Available at http://hectaresbc.org/app/habc/HaBC.html (verified 16 May 2012).

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

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.

Hole, F.D., Campbell, J.B., 1985. Soil Landscape Analysis. Rowman and Allanheld, Totowa, NJ.

Howes, D.E., Kenk, E., 1997. Terrain Classification System for British Columbia, Version 2.Resource Inventory Branch, Ministry of Environment, Lands and Parks, Victoria, BC.

Hua, J., Xiong, Z., Lowely, J., Suh, E., Dougherty, E.R., 2005. Optimal number of features as a function of sample size for various classification rules. Bioinformatics 21, 1509-1515.

Jenny, H., 1941. Factors of Soil Formation: A System of Quantitative Pedology. McGraw-Hill, NY.

Kenney, E., Frank, G., 2010. Creating a seamless soil dataset for the Okanagan Basin, British Columbia. *In* Proceedings of the Western Regional Cooperative Soil Survey Conference, Las Vegas, NV. USDA-NRCS. ftp://ftp-fc.sc.egov.usda.gov/NSSC/NCSS/Conferences/regional/2010/west/kenney.pdf (verified 8 Nov 2012).

Koethe, R., Lehmeier, F., 1996. SARA-Systeme Zur Automatischen Relief-Analyse, Benutzerhandbuch, 2, Goettingen University.

Kubat, M., Matwin, S., 1997. Addressing the curse of imbalanced training sets: one-sided selection. Proceedings of the 14th Annual International Conference on Machine Learning, Nashville, TN, 179-186.

Lacoste, M., Lemercier, B.,Walter, C., 2011. Regional mapping of soil parent material by machine learning based on point data. Geomorphology 133, 90-99.

Lawley, R., Smith, B., 2008. Digital soil mapping at a national scale: a knowledge and GIS based approach to improving parent material and property information. *In* Digital Soil Mapping with Limited Data, Springer, 173-182.

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.

Li, S., MacMillan, R.A., Lobb, D.A., McConkey, B.G., Moulin, A., Fraser, W.R., 2011. Lidar DEM error analyses and topographic depression identification in a hummocky landscape in the prairie region of Canada. Geomorphology 129, 263-275.

Liaw, A., Wiener, M., 2002. Classification and Regression by randomForest. R News 2, 18-22.

Ließ, M., Glaser, B., and Huwe, B., 2012. Uncertainty in the spatial prediction of soil texture: comparison of regression tree and Random Forest models. Geoderma 170, 70-79.

Lord, T.M., Green, A.J., 1974. Soils of the Tulameen Area of British Columbia. Report No. 13, British Columbia Soil Survey. Research Branch, Canada Department of Agriculture, Ottawa, ON, Canada.

Luttmerding, H.A., 1981. Soils of the Langley-Vancouver Map Area. Report No. 15, British Columbia Soil Survey. BC Ministry of Environment, Kelowna, BC, Canada.

MacMillan, R.A, 2005. A new approach to automated extraction and classification of repeating landform types. *In* Frontiers in Pedometrics. 2005. Naples Florida p. 54, http://www.conference.ifas.ufl.edu/pedometrics/Abstract%20Book.pdf (verified 8 Nov 2012).

MacMillan, R.A., Martin, T.C., Earle, T.J., McNabb, D.H., 2003. Automated analysis and classification of landforms using high-resolution digital elevation data: Applications and issues. Canadian Journal of Remote Sensing 29, 592-606.

MacMillan, R.A., Moon, D.E., Coupé, R.A., 2007. Automated predictive ecological mapping in a forest region of B.C., 2001-2005. Geoderma 140, 353-373.

McBratney, A.B., Mendoça Santos, M.L., Minasny, B., 2003. On digital soil mapping. Geoderma 117, 3-52

Moore, I.D., Grayson, R.B., Ladson, A.R., 1991. Digital terrain modelling: a review of hydrological, geomorphological, and biological applications. Hydrological Processes 5, 3-30.

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

Moran, C.J., Bui, E.N., 2002. Spatial data mining for enhanced soil map modeling. International Journal of Geographical Information Systems 16: 533-549.

Natural Resources Canada, 2004. Canada’s National Forest Inventory Photo Plot Guidelines, version. 1.1. Canadian Forest Service, Pacific Forestry Centre, Victoria, BC.

Pojar, J., Klinka, K., and Demarchi, D., 1991. Ecosystems of British Columbia. British Columbia Ministry of Forests and Ranges, 95–111

Prasad, A.M., Iverson, L.R., Liaw, A., 2006. Newer classification and regression tree techniques: bagging and Random Forests for ecological prediction. Ecosystems 9, 181-199.

Qi, Y., 2012. Random forest for bioinformatics. *In* Ensemble Machine Learning, Springer, 307-323.

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.

Resource Inventory Committee, 1998. Standards for Terrestrial Ecosystem Mapping in British Columbia. Ecosystem Working Group, Terrestrial Ecosystem Task Force, Resource Inventory Committee, Victoria, BC.

Riley, S.J., DeGloria, S.D., Elliot, R., 1999. A terrain ruggedness index that quantifies topographic heterogeneity. Intermountain Journal of Sciences 5, 23-27.

le Roux, J.J., Newby, T.S., and Sumner, P.D., 2007. Monitoring soil erosion in South Africa at a regional scale: review and recommendations. South African Journal of Science 103: 329-335.

Saga Development Team, 2011. System for Automated Geoscientific Analyses (SAGA). Available at http://www.saga-gis.org/en/index.html (verified 12 August, 2012).

Schut, P., Smith, S., Fraser, W., Geng, X., Kroetsch, D., 2011. Soil Landscapes of Canada: Building a national framework for environmental information. Geomatica 65, 293-309.

Scull, P., Franklin, J., Chadwick, O.A., 2005. The application of classification tree analysis to soil type prediction in a desert landscape. Ecological Modelling 181, 1-15.

Sondheim, M., Suttie, K., 1983. User manual for the British Columbia Soil Information System. Volume 1, BC Ministry of Forests Publication R28-82053, Victoria, BC.

Statnikov, A., Wang, L., Aliferis, C.F., 2008. A comprehensive comparison of Random Forests and support vector machines for microarray-based cancer classification. BMC Bioinformatics 9: 319-329.

Svetnik, V., Liaw, A., Tong, C., Culberson, 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, 1947-1958.

Svetnik, V., Liaw, A., Tong, C., Wang, T., 2004. Application of Breiman's Random Forest to modeling structure-activity relationships of pharmaceutical molecules. Multiple Classifier Systems, Fifth International Workshop, MCS 2004, Proceedings, 9-11 June 2004, Cagliari, Italy. Lecture Notes in Computer Science, vol. 3077. Roli, F., Kittler, J., and Windeatt, T. (eds). Berlin: Springer, 334-343.

Valentine, K.W.G., Sprout, P.N., Baker, T.E., Lavkulich, L.M., 1978. The Soil Landscapes of British Columbia. BC Ministry of Environment, Victoria, BC, Canada.

Van Hulse, J., Khoshgoftaar, T., 2009. Knowledge discovery from imbalanced and noisy data. Data & Knowledge Engineering 68, 1513-1542.

Van Hulse, J., Khoshgoftaar, T.M., Napolitano, A., 2007. Experimental perspectives on learning from imbalanced data. Proceedings of the 24th Annual International Conference on Machine Learning (ICML 2007), Corvalis, OR, 935-942.

Van Vliet J., 2003. Map Comparison Kit 3: User Manual. Research Institute for Knowledge Systems: Maastricht.

Visser, H., de Nijs, T., 2006. The map comparison kit. Environmental Modelling & Software 21, 346-358.

Weaver, A., 1991. The distribution of soil erosion as a function of slope aspect and parent material in Ciskei, Southern Africa. Geojournal 23, 29-34.

Webster, R., Beckett, P.H.T., 1968. Quality and usefulness of soil maps. Nature 219, 680-682.

Webster, R.,Wong, I.F.T., 1969. A numerical procedure for testing soil boundaries interpreted from air photographs. Photogrammetria 24, 59-72.

Wiesmeier, M., Barthold, F., Blank, B., Kögel-Knabner, I., 2011. Digital mapping of soil organic matter stocks using Random Forest modeling in a semi-arid steppe ecosystem. Plant Soil 340, 7-24.

Wilson, J.P., Gallant, J.C. (eds), 2000. Terrain Analysis: Principles and Applications. John Wiley & Sons, New York, NY.

Wittneben, U., 1986. Soils of the Okanagan and SimilkameenValleys. Report No. 52, British Columbia Soil Survey. Survey and Resource Mapping Branch, BC Ministry of Environment, Victoria, BC, Canada.

Xiong, X., Grunwald, S., Myers, D.B., Kim, J., Harris, W.G., Comerford, N.B., 2012. Which covariates are needed for soil carbon models in Florida. *In* Digital Soil Assessment and Beyond, CRC Press, 109-113.

Zevenbergen, L.W., Thorne, C.R., 2006. Quantitative analysis of land surface topography. Earth Surface Processes and Landforms 12, 47–56.

Zhu, A.X., Band, L.E., 1994. A knowledge-based approach to data integration for soil mapping. Canadian Journal of Remote Sensing 20, 408-418.

Chapter 3.

A Comparison of Machine-Learning Techniques for Classification Purposes in Digital Soil Mapping[3]

3.1. Abstract

Machine-learning is the automated process of uncovering patterns in large datasets using computer-based statistical models, where a fitted model may then be used for prediction purposes on new data. Despite the growing number of machine-learning algorithms that have been developed, relatively few studies have provided a comparison of an array of different learners – typically, model comparison studies have been restricted to a comparison of only a few models. This study evaluates and compares a suite of 10 machine-learners as classification algorithms for the prediction of soil taxonomic units in the Lower Fraser Valley.

A variety of machine-learners (CART, CART with bagging, Random Forest, *k*-nearest neighbour, nearest shrunken centroid, artificial neural network, multinomial logistic regression, logistic model trees, and support vector machine) were tested in the extraction of the complex relationships between soil taxonomic units (Great Groups and orders) from a conventional soil survey and a suite of 20 environmental covariates representing the topography, climate, and vegetation of the study area. Methods used to extract training data from a soil survey included by-polygon, equal-class, area-weighted, and area-weighted with random over sampling (ROS) approaches. The fitted models, which consist of the soil-environmental relationships, were then used to predict soil Great Groups and orders for the entire study area at a 100 m spatial resolution. The resulting maps were validated using 262 points from legacy soil data.

[3] A version of this chapter has been published in “Heung, B., Ho, H.C., Zhang, J., Knudby, A., Bulmer, C.E., Schmidt, M.G., 2016. An overview and comparison of machine-learning techniques for classification purposes in digital soil mapping. Geoderma 265, 62-77.”

On average, the area-weighted sampling approach for developing training data from a soil survey was most effective. Using a validation of *R* = 1 cell, the *k*-nearest neighbour and support vector machine with radial basis function resulted in the highest accuracy of 72% for Great Groups using ROS; however, models such as CART with bagging, logistic model trees, and Random Forest were preferred due to the speed of parameterization and the interpretability of the results while resulting in similar accuracies ranging from 65-70% when using the area-weighted sampling approach. Model choice and sample design greatly influenced outputs. This study provides a comprehensive comparison of machine-learning techniques for classification purposes in soil science and may assist in model selection for digital soil mapping and geomorphic modeling studies in the future.

3.2. Introduction

Data mining may be defined as the automated or semi-automated process of uncovering patterns from large electronic datasets using trained models, where the patterns may then be used on new data for the purposes of prediction (Witten and Frank, 2005). The process of 'training' a model is also synonymously described as a type of 'learning', where 'machine-learning' can be defined as the process of discovering the relationships between predictor and response variables using computer-based statistical approaches (Witten and Frank, 2005; Hastie et al., 2009).

In soil science, machine-learning techniques have most commonly been used in the subfield of pedometrics for the development of predictive or digital soil maps (DSM; Scull et al., 2003; McBratney et al., 2003) due to developments in geographical information systems, availability of digital spatial data, and constantly advancing computer technology (McBratney et al., 2003). In DSM, the workflow for the environmental-correlation approach (McKenzie and Austin, 1993; McKenzie and Ryan, 1999) entails the collection of soil point or polygon data that are co-located with a suite of *clorpt* soil-environmental variables (Jenny, 1941) in order to develop the training dataset (McBratney et al., 2003). The relationships between the soil and environmental covariates are fitted with a model, and the learned relationships are then applied to locations where soil data are not available. This generic procedure, a form of supervised learning, may be applied to the prediction of quantitative outputs (e.g. soil organic matter content, clay content, pH, or electrical conductivity) using regression, or the prediction of qualitative outputs (e.g. soil taxonomic units) using classification (McBratney et al., 2003; Hastie et al., 2009).

Numerous machine-learning algorithms are available, including the commonly used tree-based learners such as the classification and regression tree (CART) learner proposed in Breiman et al. (1984) and its extensions using bagging (Breiman, 1996) or boosting (Breiman, 1998) and, subsequently, the development of Random Forest (RF; Breiman, 2001). Other learners less commonly used in DSM include support vector machines (Kovačevic et al., 2010; and Priori et al., 2014), artificial neural networks (Aitkenhead et al., 2013; Priori et al., 2014; and Silveira et al., 2013), *k*-nearest neighbour (Mansuy et al., 2014), and linear approaches (Kempen et al., 2009; Vasques et al., 2014). With the notable exceptions of Brungard et al. (2015) and Taghizadeh-Mehrjardi et al. (2015), the number of models compared in DSM studies have generally been restricted to a few models for each study (e.g. Cavazzi et al., 2013; Ließ et al., 2012; Bourennane et al., 2014; Priori et al., 2014; Collard et al., 2014), rather than an expansive comparison where some learners, commonly used in other fields, have yet to be tested for DSM.

The objectives of this study are (1) to evaluate and compare a suite of 10 machine-learners as classifiers for the prediction of soil taxonomic units and (2) to evaluate different methods for generating training data from a conventional soil survey. The evaluation and comparison between the modeling approaches are based on a case study for the Lower Fraser Valley region of British Columbia, Canada, where the various classifiers are used to learn the relationships between soil taxonomic units and environmental covariates through the data mining of a conventional soil survey as described in Heung et al. (2014). In order to make a fair comparison between the learners, model parameters were all optimized to the training data.

3.3. Methodology

The methodology for this study follows the workflow provided in Heung et al. (2014) and is summarized here. The method entails the use of conventional soil survey maps where map units comprised of only a single taxonomic (single-component) unit with the same soil Great Group were used as training areas for the machine-learners. To produce the training dataset, random points were generated within the single-component mapping units and intersected with a suite of topographic, vegetative, and climatic indices produced from digital elevation models (DEM), satellite imagery, and climate model outputs. Similar methods of sampling from conventional soil survey data may be found in studies such as Collard et al. (2014), Odgers et al. (2014), and Subburayalu et al. (2014). The resulting soil-environmental covariate matrix was then used to train the various machine-learners, and predictions were made for unsampled locations. The resulting output was a map of soil taxonomic units for the study area where model predictions were assessed for consistency with the original training area and also validated using legacy soil point observations.

3.3.1. Study Area

The 5472 km^2 study area ranges from approximately 49^{o}00'N to 49^{o}56'N latitude and 121^{o}16'W to 123^{o}11'W longitude with an elevational range of 0-2555 m above mean sea level, and is located in the Coastal Western Hemlock biogeoclimatic zone (Figure 3.1; Pojar et al., 1991). This biogeoclimatic zone experiences a mean annual temperature range of 5.2-10.5 ^{o}C with a mean annual precipitation range of 1000-4400 mm. In the southern region of the zone, where the study area is located, 15% of the precipitation is in the form of snowfall.

The northern region of the study area encompasses portions of the Coastal Mountain Range and is predominantly covered by forests comprised mainly of a mixture of western hemlock (*Tsuga heterophylla*), Douglas-fir (*Pseudotsuga menziesii*), and western redcedar (*Thuja plicata*) tree species (Pojar et al., 1991). The soils of this area are classified as being mainly Ferro-Humic Podzols and Humo-Ferric Podzols derived dominantly from glacial deposits as well as colluvial deposits. In contrast, the southern region of the study area constitutes the Lower Fraser Valley where the land-use is primarily agriculture and urban. The soils of the Lower Fraser Valley are primarily derived from fluvial deposits at lower elevations while marine and glacio-marine deposits, originating from the Pacific Ocean, are found at higher elevations as a result of isostatic rebound due to glacial retreat. In general, Humic Gleysols and Rego Gleysols are common in this part of the study area (Luttmerding, 1981).

3.3.2. Environmental Covariates

A suite of 20 environmental covariates representing topographic, climatic, and vegetative indices were used as predictors for this study (Table 3.1). Information on parent material was not included in this study because it was previously shown in Heung et al. (2014) that the surficial materials were closely linked to topographic indices for this area. The use of bedrock geology was also considered; however, the soils in the region are primarily developed from transported sediments and hence the mineralogical characteristics of the underlying bedrock and transported sediments are most likely different. All environmental covariates were scaled because distance-based learners such as *k*NN and NSC require covariates to have a similar range in values.

Topographic Indices

26 topographic indices were calculated in the System for Automated Geoscientific Analysis (SAGA) (SAGA Development Team, 2011) using British Columbia's Terrain Resource Information Management (TRIM) DEM (B.C. Ministry of Sustainable Resource Management, 2002). The DEM was originally produced from a triangulated irregular network (TIN) developed from TRIM mass-points and break-lines at a 100 m spatial resolution. In order to reduce noise and anomalies in the DEM and its derived indices, three successive mean filters with window sizes of 3 x 3, 3 x 3, and 5 x 5 cells were applied based on Heung et al. (2014).

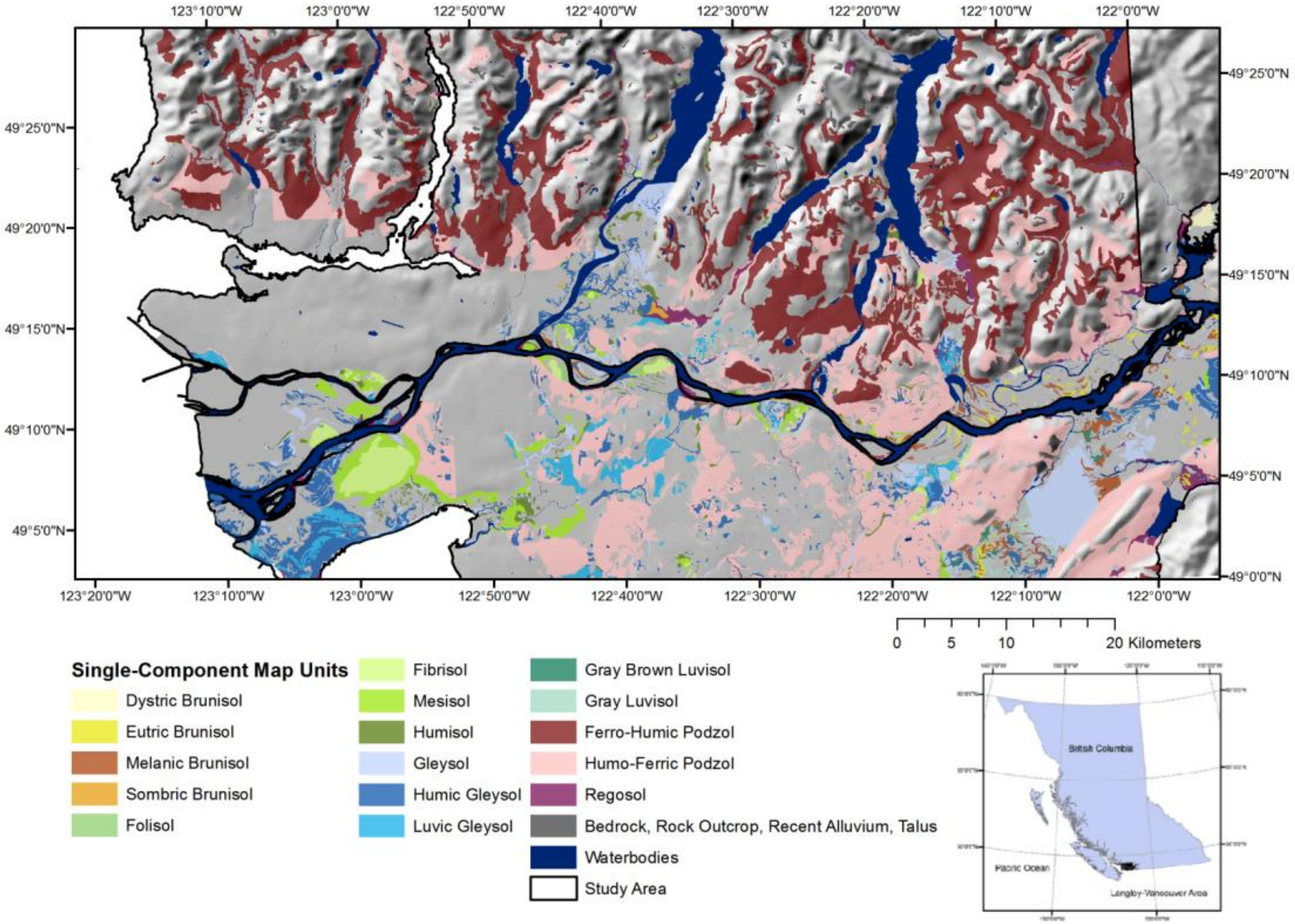


Figure 3.1. **Single-component soil Great Group map units from the Langley-Vancouver Map Area (Luttmerding, 1981). Inset: study area in relation to the province of British Columbia, Canada.**

In addition, two distance metrics were also included as predictors: distance to the nearest stream and distance to the Fraser River. These distance metrics were calculated from stream polyline mapping from HectaresBC.org (Hectares BC, 2012) – a data repository that provides gridded data layers for the province of BC. These covariates were included as they were previously found to be important for capturing the distribution of fluvial sediments in the region (Heung et al., 2014). The topographic indices were selected in order to represent local scale morphometry (e.g. elevation, slope, aspect, and curvature); landscape scale morphometry (e.g. slope height, multi-resolution ridge top flatness, and valley bottom flatness); hydrological characteristics (e.g. wetness index, hydrologic slope position, and distance to nearest stream and river); and landscape exposure (e.g. sky view factor and terrain view) (Table 3.1).

Since the topographic indices, with the exception of the two distance metrics, were calculated from the same DEM, a principal component analysis (PCA) was applied to the topographic covariates in order to reduce the number of variables used to train the learners as well as to remove predictor multi-collinearity. In the machine-learning literature, PCA is an appropriate step to take for the purposes of cleaning the covariate data (Witten and Frank, 2005) and has been shown to improve predictions for high-dimensional datasets (Howley et al., 2006). The original set of 26 topographic covariates was reduced to the first 13 principal components, which together accounted for 95% of the cumulative variance. Other methods of dimension reduction have included the use of techniques such as correlation-based feature selection (Taghizadeh-Mehrjardi et al., 2015); variable selection based on variable importance metrics from the RF algorithm (Brungard et al., 2015; Heung et al., 2014); and in Behrens et al. (2010), a PCA and an ANOVA filtering approach was compared as dimension reduction techniques.

Climatic Indices

Previous studies have applied temperature estimation from remotely sensed data to represent soil temperature for the prediction of soil taxonomic units (Chang and Islam, 2000; Mansuy et al., 2014); however, such estimations of temperature, at the sensor-level, often do not consider the effects from atmospheric transmission and absorption. Here, the land surface temperature (LST), corrected for atmospheric effects, was used as the climatic index where 14 cloud-free Landsat images retrieved from June to August and 5 images retrieved from November to February were used to calculate the average

LST for the years 2000-2014 during the winter (Nov – Feb) and the summer (Jun – Aug) seasons.

In order to estimate LST, blackbody radiance at LST was calculated using Eq. 3.1 (Coll et al., 2010):

Eq. (3.1)

$$B(LST) = \frac{L_{sen} - L^{\uparrow}}{\varepsilon \tau} - \frac{1 - \varepsilon}{\varepsilon} L^{\downarrow},$$

where B(LST) is the blackbody radiance at LST, L_{sen} is the at-sensor radiance, $L^{\uparrow}$ is the up-dwelling atmospheric radiance, $L^{\downarrow}$ is the down-welling atmospheric radiance, τ is the atmospheric transmittance, and ε is the emissivity value. $L^{\uparrow}$, $L^{\downarrow}$ and τ were obtained from NASA's Atmospheric Correction Parameter Calculator (Barsi et al. 2003) and ε was obtained for non-waterbody surfaces from Eq. 3.2 (Van de Griend and Owe, 1993):

Eq. (3.2)

$$e = \begin{cases} 0.923 & \mathrm{NDVI} < 0.157 \\ 1.0094 + 0.047 \cdot \ln \mathrm{NDVI} & 0.157 \le \mathrm{NDVI} \le 0.727 \\ 0.994 & \mathrm{NDVI} > 0.727 \end{cases} .$$

To calculate the LST, an inversion of Planck's Law was applied (Ho et al, 2014):

Eq. (3.3)

$$LST = \frac{K_2}{\ln\left(\frac{K_1}{B(LST)} + 1\right)},$$

where K1 and K2 are thermal band calibration constants. Landsat bands 5, 7, and 8 were obtained at a 120 m, 60 m, and 30 m spatial resolution, respectively, and resampled to a 100 m spatial resolution to match the grid size of the terrain attributes. In addition to LST, mean annual precipitation data was obtained from HectaresBC.org (Hectares BC, 2012).

Vegetation Indices

The Normalized Difference Vegetation Index (NDVI) and the Normalized Difference Water Index (NDWI) were used as vegetative indices. NDVI is commonly used to indicate the amount of healthy vegetation, and is defined as:

Eq. (3.4)

$$NDVI = (\rho_{NIR} - \rho_{VIS}) / (\rho_{NIR} + \rho_{VIS}),$$

where ρ_{NIR} and ρ_{VIS} are surface reflectances in the near-infrared and visible red wavelengths respectively.

NDWI can be used to quantify vegetation water content (Gao, 1996). Vegetation water content is known as one of the key variables that can strongly influence the land surface cooling by evapotranspiration. It may also reflect the size and health status of the vegetation, and be related to the amount of soil moisture. NDWI is defined as:

Eq. (3.5) $$\mathrm{NDWI} = (\rho_{NIR} - \rho_{MIR})/(\rho_{NIR} + \rho_{MIR}),$$

where ρ_{MIR} is the surface reflectance in the mid-infrared wavelengths.

As with the LST layers, 14 cloud-free Landsat images retrieved from June to August, and 5 images retrieved from November to February, were used to calculate the average NDVI and NDWI in the summer and winter. Landsat images were obtained at a 30 m spatial resolution and resampled to a 100 m spatial resolution.

3.3.3. Development of Training Data

The soil map used to determine the training areas was derived from a seamless digitized soil map for the Lower Fraser Valley, created at a 1:25,000 scale, and a map for the Southern Sunshine Coast and Southern Coast Mountains at a 1:50,000 scale (Luttmerding, 1981; Kenney and Frank, 2010). Using the soil survey, map units with a single-component at the Great Group taxonomic level of the Canadian System of Soil Classification (Soil Classification Working Group, 1998) were extracted and used as the training areas. The training area consisted of 3121 single-component map units (Figure 3.1). The soil survey included 16 soil Great Groups, 6 soil Orders, and 4 miscellaneous land classes (bedrock, recent alluvium, rock outcrops, and talus slopes). Of the training areas, Humo-Ferric Podzols and Ferro-Humic Podzols were the majority classes that occupied 35% and 29% of the training areas respectively; whereas, minority classes such as Gray Luvisols, Folisols, Dystric Brunisols, Grey Brown Luvisols, and Sombric Brunisols each occupied < 0.1% of the training area (Figure 3.2).

Table 3.1. Environmental covariates derived from a 100 m spatial resolution DEM and 30 m Landsat imagery.

Representation	Environmental Covariate
Local Scale Morphometry[1]	General Curvature
	Elevation
	Plan curvature
	Profile curvature
	Slope
	Slope length factor (Moore et al., 1993)
	Tangential curvature (Florinsky, 1998)
	Terrain ruggedness index (Riley et al., 1999)
	Total curvature (Wilson and Gallant, 2000)
	Transformed aspect
Landscape Scale Morphometry[1]	Multi-resolution ridge top flatness index (Gallant and Dowling, 2003)
	Multi-resolution valley bottom flatness index (Gallant and Dowling, 2003)
	Mid-slope position
	Normalized height
	Slope height
	Valley depth
Hydrologic Characteristics[1]	Distance to nearest river
	Distance to nearest stream
	Modified relative hydrologic slope position (MacMillan, 2005)
	Relative hydrologic slope position (MacMillan, 2005)
	SAGA wetness index (SAGA Development Team, 2011)
	Stream power index (Moore et al., 1991)
	Topographic wetness index (Beven and Kirkby, 1979)
Landscape Exposure[1]	Sky view factor (Häntzschel et al., 2005)
	Terrain view (Häntzschel et al., 2005)
	Visible sky
Climatic Indices	Land Surface Temperature (Summer) (Ho et al., 2014)
	Land Surface Temperature (Winter) (Ho et al., 2014)
	Mean Annual Precipitation
Vegetative Indices	Normalized Difference Vegetation Index (Summer)
	Normalized Difference Vegetation Index (Winter)
	Normalized Difference Water Index (Summer) (Gao, 1996)
	Normalized Difference Water Index (Winter) (Gao, 1996)

[1] Principal component analysis was applied to topographic indices where the 26 indices were reduced to the top 13 principal components representing 95% of the cumulative variance.

To create the soil-environmental training data matrix used to train the learners, 4 methods of sampling the training areas were used: (1) equal number of sample points per Great Group (equal-class sampling); (2) equal number of sample points per polygon (by-polygon sampling); (3) the number of sample points determined as an area-weighted proportion of a Great Group's extent (area-weighted sampling); and (4) random over sampling (ROS) applied to the area-weighted sampling approach. Approaches (1-3) have all been used in previous DSM studies; for instance, Moran and Bui (2002) compared the equal-class and the area-weighted sampling approaches, Odgers et al. (2014) used by-polygon sample, and Heung et al. (2014) used all three.

ROS is an approach that has been used with the intention of creating a 'balanced' dataset where the number of samples for each class is equal. Whereas equal-class sampling (method 1) is a type of 'random under sampling' (RUS) where training points are not duplicated, ROS duplicates the training points for the minority classes so that they equal the number of samples in the majority class (Van Hulse and Khoshgoftaar 2009; Van Hulse et al., 2008). The issue of class imbalance has been recognized as being influential on classification problems (e.g. Van Hulse and Khoshgoftaar 2009; Van Hulse et al., 2008), but has not been extensively studied in DSM, although recent studies such as Subburayalu and Slater (2013) and Heung et al. (2014) have examined the impact of different sampling methods on predictive mapping of discrete data. For each sampling method, with the exception of area-weighted with ROS, 15,605 training points were used with an average sampling density of 2.9 samples/km^2. The number of training points was chosen based on the by-polygon sample approach where 5 points were randomly generated within each map unit. In the case of ROS, where duplication of training points for minority classes are applied in order to match the number of training points for the majority class, 121,748 training points were generated.

3.3.4. Machine-learning Approaches

The learners used in this study included: CART, CART with bagging, *k*-nearest neighbour (*k*NN), logistic model tree (LMT), multinomial logistic regression (MLR), artificial neural network (ANN), nearest shrunken centroid (NSC), Random Forest (RF), linear support vector machine (SVM-Lin), and support vector machine with radial basis function (SVM-RBF) – all of which are available using the *R* statistical software (R Development Core Team, 2012) and the *caret* package (Kuhn, 2008). The *caret* package is particularly useful for model comparison studies as the package compiles a database of classification and regression algorithms from existing *R* packages and facilitates the optimization of model parameters and the selection of models through a repetitive cross validation (CV) procedure. As a result, users do not have to be familiar with the use of the original packages from which *caret* is compiled. Furthermore, the package is easily adaptable for spatial datasets.

Parameter Optimization

In order to make a fair comparison between the various learners, each learner was parameterized using a 5-fold CV procedure, where the training dataset was randomly partitioned into five subsets – four of the partitions, comprising 80% of the data, were then used to train the learner and the remaining 20% were used for validation. This process was repeated 5 times, using each fold for validation once. In order to account for the randomness from the partitioning, 10 replicates of 5-fold CV were used. For each learner, a range of parameter values were tested and the final predictions were made based on the combination of parameter values that produced the lowest averaged error rates from the CV procedure. Similar procedures were implemented in Brungard et al., (2015), Heung et al (2014), and Schmidt et al. (2008).

3.3.5. Assessment of Predictions

Two approaches were used for assessing the predictions made by the learners. First, the predictions were compared to the mapping of the original training areas in order to assess the consistency of the predictions with the original soil survey. Secondly, the predictions were further validated using legacy soil point data. Assessments were done at two taxonomic levels: Great Group and order. To assess the results at the ‘order’ level of taxonomy, the predictions made at the Great Group level were reclassified and aggregated to the higher level.

Consistency with Soil Survey

To assess the consistency of predictions with the soil survey, the overall agreement (or proportion correct), *C*, was calculated as the percentage of pixels that were correctly classified by the machine-learner when compared to the single-component map units.

Validation with Point Data

The external validation dataset used for calculating model accuracy consisted of *n* = 262 legacy soil pit data from the British Columbia Soil Information System (Sondheim and Suttie, 1983). In order to account for the uncertainties of the spatial location of the legacy soil pits, two levels of validation were done: (1) if the validation point matched the predicted pixel at the exact location, it was considered valid (r = 0) and (2) if the validation point matched a pixel within a radius of 1 pixel, it was considered valid (r = 1).

In addition to calculating the overall agreement, the use of quantity disagreement (*Q*) and allocation disagreement (*A*) were also introduced in this analysis – both of which may be derived from an error matrix (Pontius and Millones, 2011; Warrens, 2015). The quantity disagreement, *Q*, represents the amount of difference between the validation and prediction dataset that is the result of disagreement in the proportion of each category and is calculated as follows:

Eq. (3.6)
$$\mathrm{Q} = \tfrac{1}{2}\textstyle\sum_{i=1}^{j}\left|p_{i+} - p_{+i}\right| \, ,$$

where p_{i+} and p_{+i} represent the row and column totals of the error matrix for *i*th class for *j* number of classes. Values of *Q* range from 0 to 1 where values close to 0 represent conditions where the proportions of coverage for each class between the validation and prediction datasets are in agreement. The allocation disagreement, *A*, represents the amount of difference between the validation and prediction dataset that is the result of disagreement in the spatial allocation of classes, given the class totals of the two datasets, and is calculated as follows:

Eq. (3.7) $$A = \left[\sum_{i=1}^{j} \min(p_{i+}, p_{+i})\right] - C \text{ ,}$$

where C is the overall agreement. The values of A range from 0 to 1 where values close to 0 represent conditions where the spatial allocations for each class between the validation and prediction datasets are in agreement. Both Q and A are the result of decomposing the total disagreement, D, in the following relationship:

Eq. (3.8) $$D = 1 – C = Q + A.$$

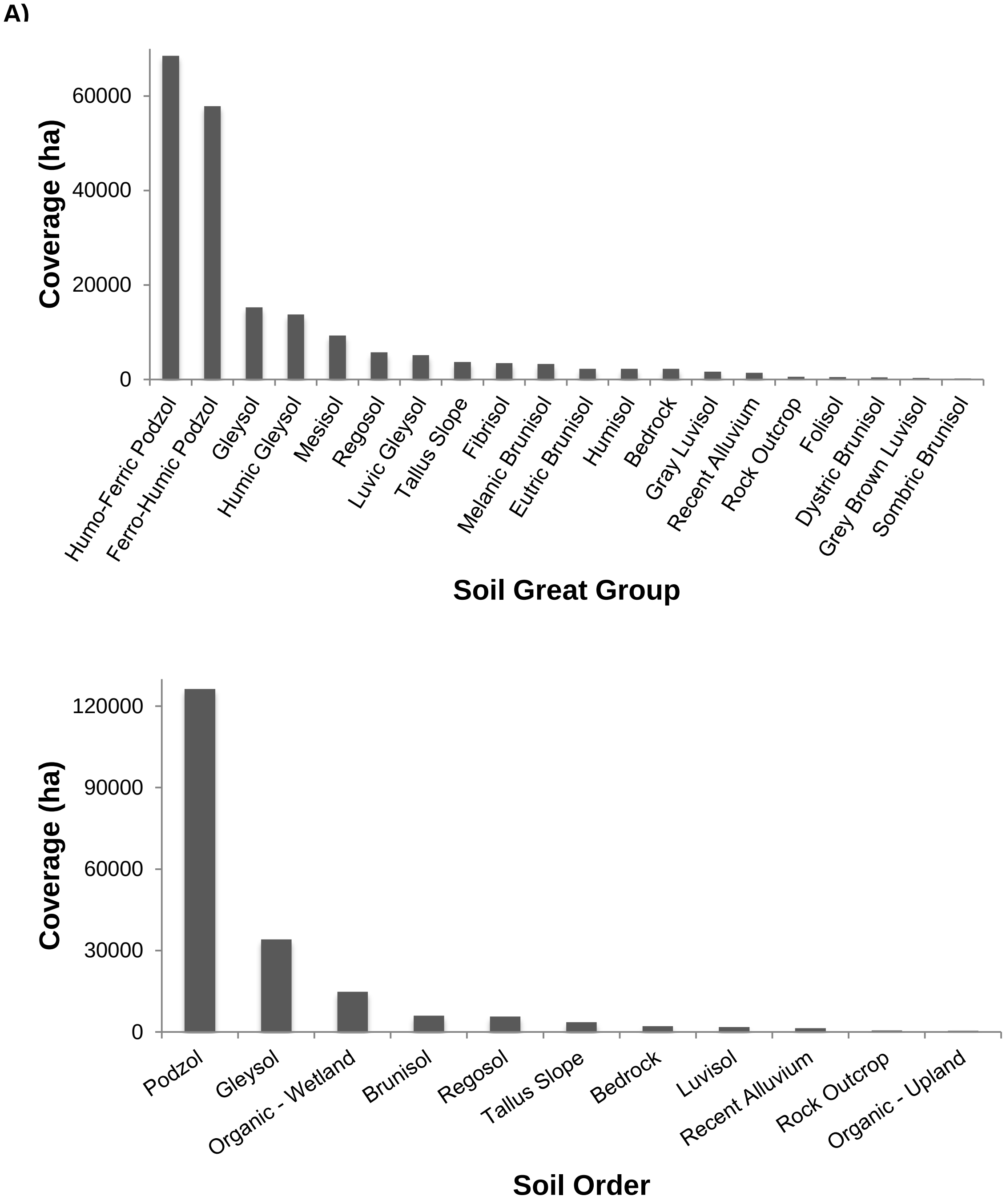


Figure 3.2. **Coverage of single-component map units by (a) soil Great Group and by (b) soil Order from the Soils of the Langley-Vancouver region (Luttmerding, 1981). Miscellaneous classes such as bedrock, recent alluvium, rock outcrops, and talus slopes are included.**

3.4. Results & Discussion

3.4.1. Parameter Optimization

The optimized parameter values and the averaged internal validation values with 5-fold CV for each model using the 4 training data sampling methods are summarized in Table 3.2. It should be noted that models such as CART with bagging and MLR did not require parameterization while LMT only required the number of iterations for the *LogitBoost* algorithm (and hence an arbitrarily large number would be sufficient). As a result, predictions were efficiently made for these three models. Models that required parameterization while retaining efficiency included CART, *k*NN, and NSC due to the simplicity of those models and the minimal amount of time required for generating models for each CV fold. In terms of complex models, as classified by Brungard et al. (2015), models such as ANN, SVM-Lin, and SVM-RBF were extremely time consuming to parameterize. In the case of these models, parameters such as the number of units within a hidden layer (*size*) and decay weights (*weights*) for the ANN model and the cost parameter (*c*) and Gaussian smoothing parameter (*sigma*) for SVM have an infinite number of combinations for their values; hence, the challenge is the identification of an 'optimal' combination of parameter values. Furthermore, SVM is inherently a computationally demanding model where an increase in the cost parameter results in an increase in processing time. This made the parameterization procedure extremely time consuming for large datasets in spite of the use of multi-core processing. Finally, parameters such as *size* for ANN and *c* for SVM have little intuitive meaning (Shawe-Taylor and Cristianini, 2004).

Table 3.2. Parameter optimization values and internal validation rates for machine-learners requiring parameterization using 10 replicates of 5-fold cross-validation.

MODEL[1,2]	EQUAL CLASS		BY POLYGON	
	Correctness	Optimized Parameters	Correctness	Optimized Parameters
CART	0.38	maxdepth = 16	0.33	maxdepth = 7
***k*-Nearest Neighbour**	0.63	k = 2	0.57	k = 2
Neural Networks	0.47	size = 20; decay = 0.3	0.40	size = 20; decay = 0.3
Nearest Shrunken Centroid	0.37	threshold = 0.2	0.35	threshold = 0.1
Random Forest[3]	0.78	mtry = 9; iter = 1000	0.72	mtry = 12; iter = 1000
Support Vector Machine - Linear	0.57	c = 1000	0.42	c = 100
Support Vector Machine - Radial Basis Function	0.75	c = 50; sigma = 0.1	0.69	c = 25; sigma = 0.25
MODEL	**AREA WEIGHTED**		**AREA WEIGHTED + ROS**	
	Correctness	**Optimized Parameters**	**Correctness**	**Optimized Parameters**
CART	0.61	maxdepth = 5	0.38	maxdepth = 14
***k*-Nearest Neighbour**	0.69	k = 4	0.98	k = 2
Neural Networks	0.64	size = 20; decay = 0.1	0.44	size = 20; decay = 0.1
Nearest Shrunken Centroid	0.55	threshold = 0.1	0.37	threshold = 0.1
Random Forest	0.77	mtry = 12; iter = 1000	0.99	mtry = 9; iter = 1000
Support Vector Machine - Linear	0.65	c = 100	0.63	c = 100
Support Vector Machine - Radial Basis Function	0.76	c = 10; sigma = 0.1	0.99	c = 25; sigma = 0.5

[1.] Multinomial logistic regression did not require parameter optimization.

[2.] CART with bagging and logistic model tree are tree ensemble methods where an arbitrarily large number of trees were grown (n_{trees} = 1000).

[3.] Random Forest model is a tree ensemble method where an arbitrarily large number of trees were grown (n_{trees} = 1000).

Table 3.3. Consistency of model prediction with single-component soil survey polygons.

	EQUAL CLASS		BY POLYGON		AREA WEIGHTED		AREA WEIGHTED + ROS	
	Great Group	**Order[1]**	**Great Group**	**Order**	**Great Group**	**Order**	**Great Group**	**Order**
CART with Bagging	0.67	0.77	0.54	0.70	0.79	0.88	0.78	0.88
CART	0.29	0.38	0.36	0.52	0.62	0.78	0.40	0.51
***k*-Nearest Neighbour**	0.59	0.69	0.45	0.63	0.76	0.86	0.78	0.87
Logistic Model Tree	0.61	0.73	0.43	0.63	0.73	0.87	0.73	0.84
Multinomial Logistic Regression	0.45	0.58	0.42	0.64	0.64	0.81	0.44	0.57
Neural Networks	0.49	0.61	0.44	0.64	0.63	0.78	0.22	0.33
Nearest Shrunken Centroid	0.37	0.48	0.42	0.72	0.56	0.70	0.38	0.50
Random Forest	0.69	0.79	0.55	0.72	0.79	0.88	0.80	0.89
Support Vector Machine - Linear	0.54	0.67	0.42	0.63	0.65	0.83	0.56	0.70
Support Vector Machine - Radial Basis Function	0.69	0.80	0.53	0.73	0.80	0.89	0.79	0.87

[1.] Results for soil Orders are aggregated from predictions made from the Great Group level of soil taxonomy.

Although RF may be considered to be a complex model (Brungard et al., 2015), the parameterization of the model was not a very time consuming task when compared to other complex models such as ANN, SVM-Lin, and SVM-RBF. The reason was that the main tuning parameter, m_{try}, which defined the number of random predictors tried at each node of a decision tree, had a finite number of potential values – the total number of predictors of the dataset. Furthermore, advances in parallel processing have led to the development of more efficient RF algorithms that greatly reduce the computational time required to parameterize RF through the use of *R* packages such as *caret* or *sprint*.

3.4.2. Consistency with Soil Survey

Comparison of Machine-Learners

Complex models such as CART with bagging, SVM-RBF, and RF generated models that were most consistent with the original soil survey – regardless of the sampling method used, with average overall agreements of 69%, 70%, and 71%, respectively (Table 3.3). In comparison, simple models such as CART and NSC resulted in very poor consistencies with soil surveys. The low consistency for NSC was likely the result of overlap in feature space between taxonomic units. The ambiguity in feature space was also observed by comparing the consistency between SVM-Lin and SVM-RBF where the radial basis expansion allowed the SVM to produce a nonlinear separation plane between classes and improves model results when compared to SVM-Lin.

Several interesting connections between model complexity and consistency may be made from these results. Firstly, CART, CART with bagging, and RF are all tree-based models that are successively more advanced than each other where ensemble learning and randomized variable selection each add a layer of complexity to the original CART model. With CART, only a single classification tree is expected to learn the complex relationships between a large number of categories (e.g. soil Great Groups and orders) and a large number of variables - the poorer predictions for this model might reflect the model’s inability to handle such complex relationships. When ensemble-learning methods were introduced to the model for CART with bagging, consistency is drastically increased for all sampling methods. In the bagging procedure, the instability of a single-tree model is minimized through the aggregation of multiple models, which results in improved consistency (Breiman, 1996). In the case of RF, randomized variable selection is also applied and minimizes the potential model bias that occurs when a few predictors are used more often than others to generate the node-splitting rules (Breiman, 2001). Therefore, it may be suggested that ensemble-learning and randomized variable selection are two techniques that improve the consistency of results with soil surveys where the relationships between variables are complex.

The second interesting connection occurs between the MLR and LMT classifiers. Both these models are related; however, LMT is an advancement of MLR because it incorporates a tree-based structure on top of a linear classifier and thus increases model complexity. When the consistencies between MLR and LMT are compared, LMT shows an improved consistency regardless of sampling method. The improvement would therefore seem to suggest that not only are the soil-environmental relationships complex, they also have a hierarchical structure. Although the reason for a hierarchical variable structure is unclear, it may be partially due to the effects of soil forming processes operating at multiple scales.

Influence of Training Data Development

Even when using the same training dataset, the consistency of model predictions with soil survey differed markedly between individual learners and the sampling approach for developing the training dataset (Table 3.3). Area-weighted sampling resulted in the highest consistency when using overall agreement for assessment where the single-component map units were used for comparison. When the area-weighted sampling approach was compared to a balanced training dataset, such as equal-class sampling (similar to RUS), there was a large decrease in overall agreement with an average decrease of 11% in agreement across all models. The implementation of ROS on the area-weighted sampling approach generally resulted in a decrease in consistency or a negligible increase – as was the case with *k*NN, LMT, and RF. The sampling by-polygon method, an intermediate between equal-class and area-weighted sampling methods, in terms of class balance, resulted in a lower consistency when compared to area-weighted sampling; however, comparisons between equal-class and by-polygon sampling were mixed.

The overall model consistency metrics from Table 3 did not describe how well each model predicted individual soil taxonomic units for different sampling methods because the overall metrics were heavily biased in favour of the majority class. In our study area Humo-Ferric Podzols and Ferro-Humic Podzols accounted for 35% and 29% of the training area, respectively (Figure 3.2). For a better understanding of each individual class, the sampling methods, *i*, were ranked for each soil taxonomic unit for *i* =1, …, *I* based on the overall agreement for each taxonomic unit where the mean rank for each sample method, R_i, could be computed as

$$\text{Eq. (3.9)} \qquad R_i = \frac{1}{I} \sum_{i=I}^{I} r_{ij} \text{ ,}$$

where r_{ij} is the rank of the *i*th sampling method for the *j*th taxonomic unit. Furthermore, the standard deviation, SD_i, of ranks was calculated as:

$$\text{Eq. (3.10)} \qquad SD_i= \left[\frac{1}{I} \sum_{j=1}^{I} (r_{ij}- R_i)^2\right]^{0.5}.$$

The sampling method that results in the highest consistency with soil survey for the various taxonomic units should have a low mean rank and standard deviation of ranks (Laslett et al., 1987; Odeh et al., 1994).

In Figure 3.3, the mean ranks were plotted against the standard deviation of ranks for the four models that had the highest consistencies with soil surveys (CART with bagging, RF, SVM-RBF, and *k*NN), where sampling methods that are closest to the origin of the plot represent the best prediction for the most soil Great Groups and orders. Area-weighted sampling with ROS resulted in the highest consistency for the most number of classes with a minimal decrease in consistency to the majority classes when using CART with bagging, RF or *k*NN, where RF and *k*NN appear to have benefited the most from ROS. For both these models, ROS resulted in a higher ranking for both majority and minority classes when compared to area-weighted sampling; furthermore, both models resulted in the best overall model consistency when ROS was used (Table

3.3). For CART with bagging and SVM-RBF, there appeared to be a trade-off between the consistencies within individual classes and the overall consistency.

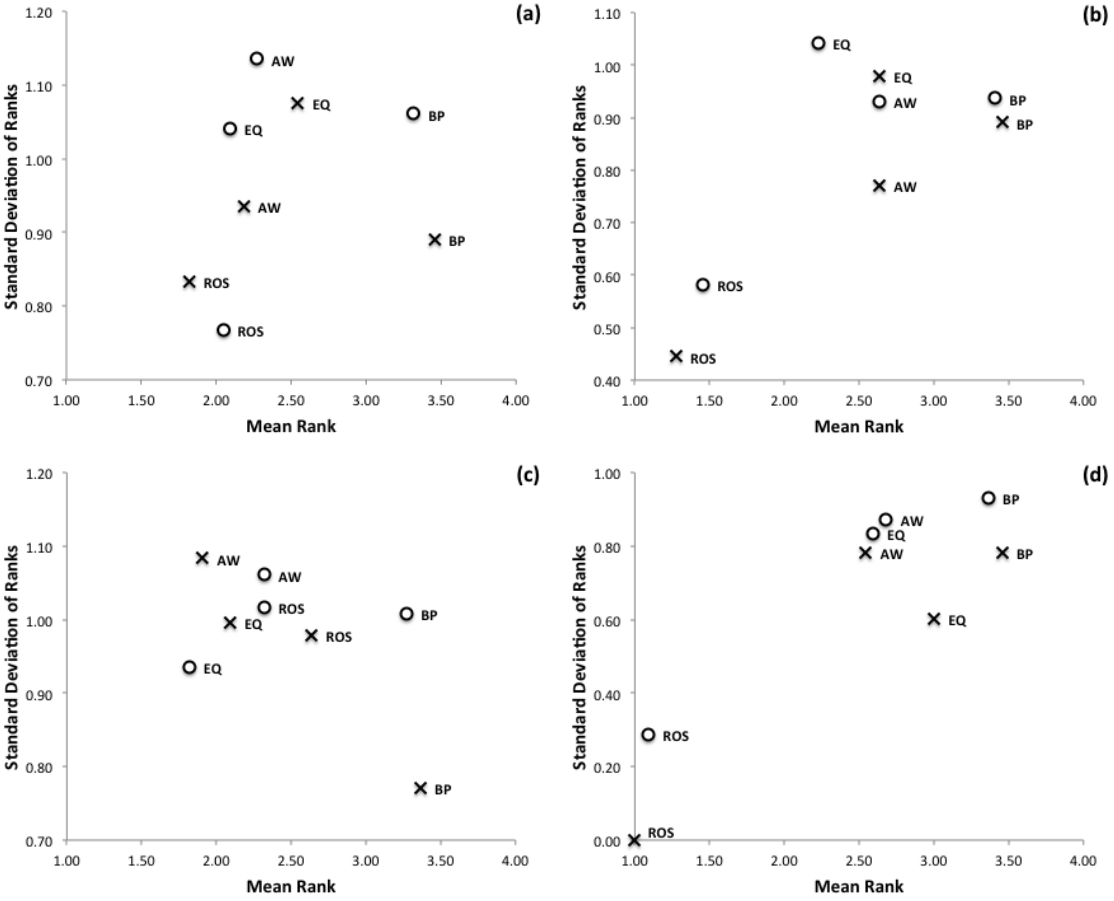


Figure 3.3. Mean rank plotted against the standard deviation of ranks for individual soil Great Groups (X) and soil Orders (O) based on predictions from (a) CART with bagging, (b) Random Forest, (c) support vector machine with radial basis function, and (d) *k*-nearest neighbour classifiers models using various soil survey sampling methods. *AW* = area-weighted sampling; *EQ* = equal-class sampling; *BP* = by-polygon sampling; and *ROS* = area-weighted with random oversampling.

3.4.3. Visual Assessment

Overall, the spatial patterns of the soil Great Groups were consistent with our understanding of soils in the study area. The key features include the presence of hydromorphic soils such as Humic Gleysols or Gleysols located at low elevations of the Lower Fraser Valley in close proximity to the Fraser River and the delta formed by the river; Humo-Ferric Podzols located at mid-elevations where there is improved drainage of the soils; and Ferro-Humic Podzols at higher-elevations along the Coast Mountains where the climate is cooler. Directions to accessing the high-resolution soil Great Group and soil Order maps for all 10 models and four sample approaches are provided in Section 3.7 “Supplementary Figures”. In general, the map units from the soil survey indicated a much greater diversity in soil classes along the Lower Fraser Valley and less diversity at higher elevations along the Coastal Mountain where this was partly due to the different map scales that these two regions of the study area were originally mapped at. As such, the map units were generally smaller within the valley where there was a greater diversity in soil classes with similar soil-environmental conditions; as a result, the model outputs appeared to have the greatest differences within the valley where simple models such as CART, MLR, and NSC were unable to capture the subtle differences in the feature space between the predicted classes. In the case of the CART models, for example, it was observed that spatial patterns produced by the models did not adhere to the physical features of the landscape – especially for Humic Gleysols.

A visual comparison of results using the area-weighted sampling approach is shown in Figure 3.4 for a part of the study area that has a high diversity of soil Great Groups. In general, CART with bagging (Figure 3.4A), LMT (Figure 3.4D), RF (Figure 3.4H), and SVM-RBF (Figure 3.4J) were effective in producing outputs with soil patterns and diversity of classes that were most similar to the training areas (Figure 3.4K). In comparison, models such as NSC (Figure 3.4G) and CART (Figure 3.4B) were not particularly effective because only 4 and 5 out of the potential classes were present in the outputs, respectively. Other models such as MLR (Figure 3.4E), ANN (Figure 3.4F), and SVM-Lin (Figure 3.4I) also did not reproduce the entire set of classes.

Results from using *k*NN (Figure 4.4C; Figure 4.5) were similar to the initial training areas; however, where the topography flattens along the Fraser River, the soil patterns become 'blotchy' or 'speckled' in appearance. The reason for this appearance may be over-fitting the model to the training data. In particular, the steps taken to optimize the parameters for *k*NN resulted in the selection of *k* values close to 1, where the decision boundaries between classes were effectively generated around a small number of training points.

Methodologies used to generate the training data may result in drastically different outputs (Figure 3.5; Section 3.7). When the equal-class sampling was used, the complexity in the predicted soil patterns also increased due to the increased presence of minority classes within the training data; however, the increased complexity in the predicted patterns was not necessarily consistent with the known soil patterns of the area. For example, organic soil Great Groups (Fibrisols, Mesisols, Humisols, and Folisols) were predicted in locations where they are not found in the study area.

3.4.4. Validation with Point Data

Overall, the machine-learner and sample design that resulted in the highest overall accuracies included RF using area-weighted; RF using area-weighted with ROS; and CART with bagging using an area-weighted design – all of which had C = 58% for R = 0 validation for Great Group predictions (Table 3.4). When compared to the cells that were within a 1-cell radius (R = 1) of a validation point, *k*NN and SVM-RBF using area-weighted with ROS resulted in the highest overall agreement of C = 72% (Figure 3.6). When the predictions were aggregated to soil Orders, CART with bagging and RF both, with the area-weighted sample design, had the highest overall agreement with C = 70% using R = 0 validation; however, SVM-RBF with area-weighted sampling (and with ROS) and RF with area-weighted and ROS all produced results with a similar overall agreement of C = 68%. Using R = 1 validation, *k*NN and SVM-RBF with area-weighted and ROS produced the highest overall agreement of C = 80%. The following subsections describe specific comparisons between individual machine-learners and training data development methodologies.

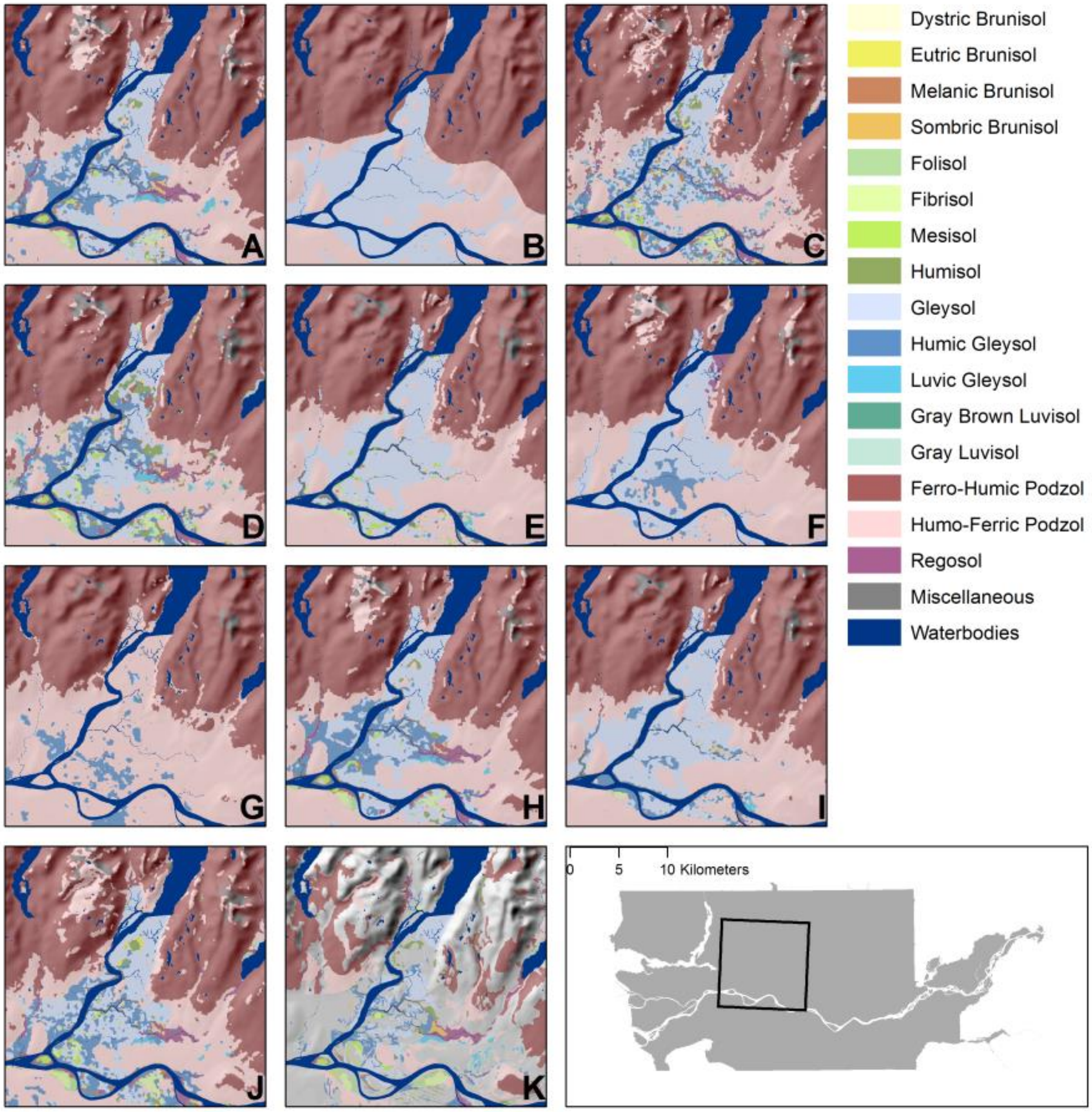


Figure 3.4. Close-up maps of soil Great Group predictions derived from the area-weighted training dataset using (A) CART with bagging, (B) CART, (C) *k*-nearest neighbour, (D) logistic model tree, (E) multinomial logistic regression, (F) artificial neural network, (G) nearest shrunken centroid, (H) Random Forest, (I) linear support vector machine, and (J) support vector machine with radial basis function learners. Training areas derived from single-component map units are shown in (K). Miscellaneous classes such as bedrock, recent alluvium, rock outcrops, and talus slopes are included.

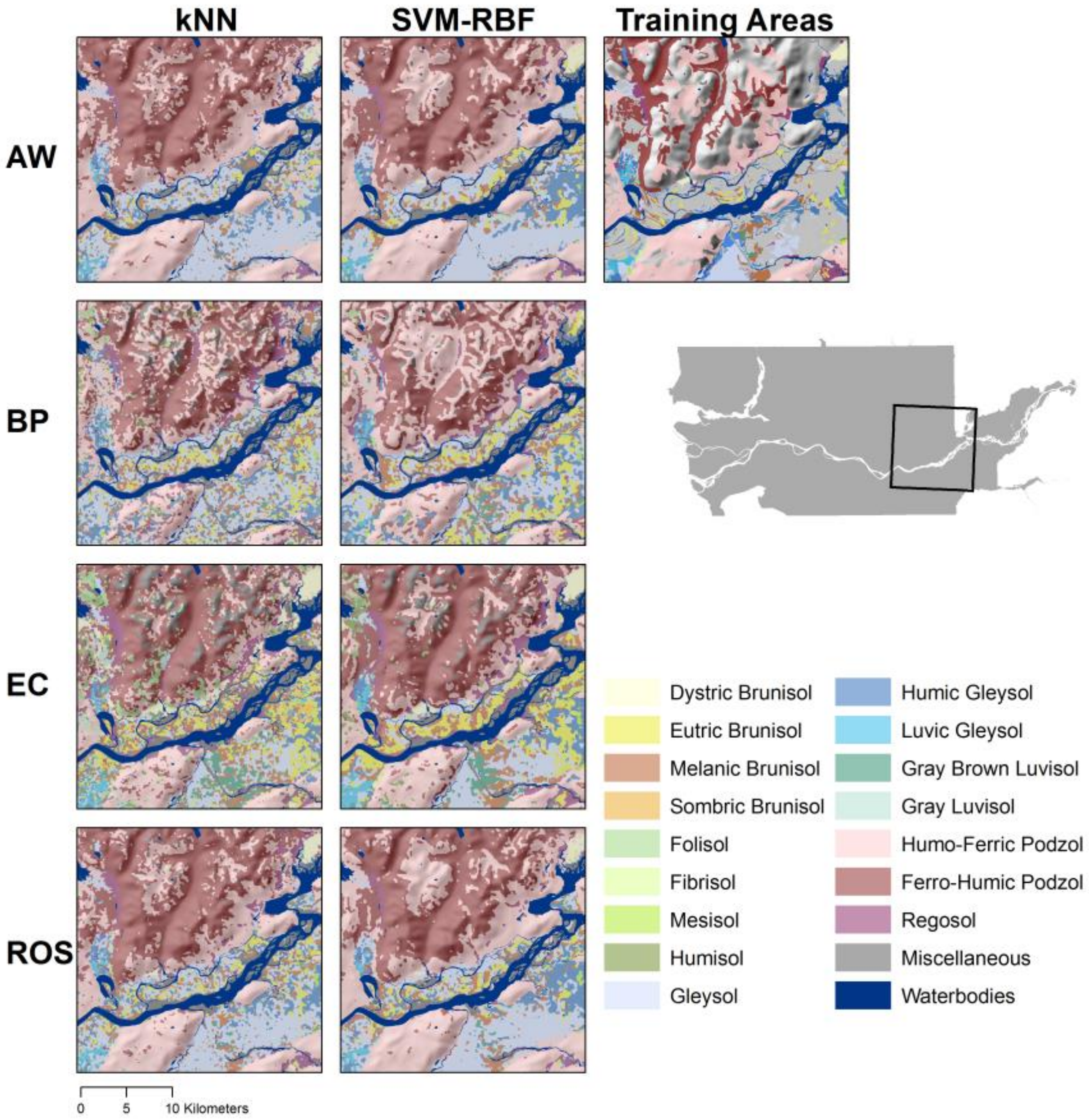
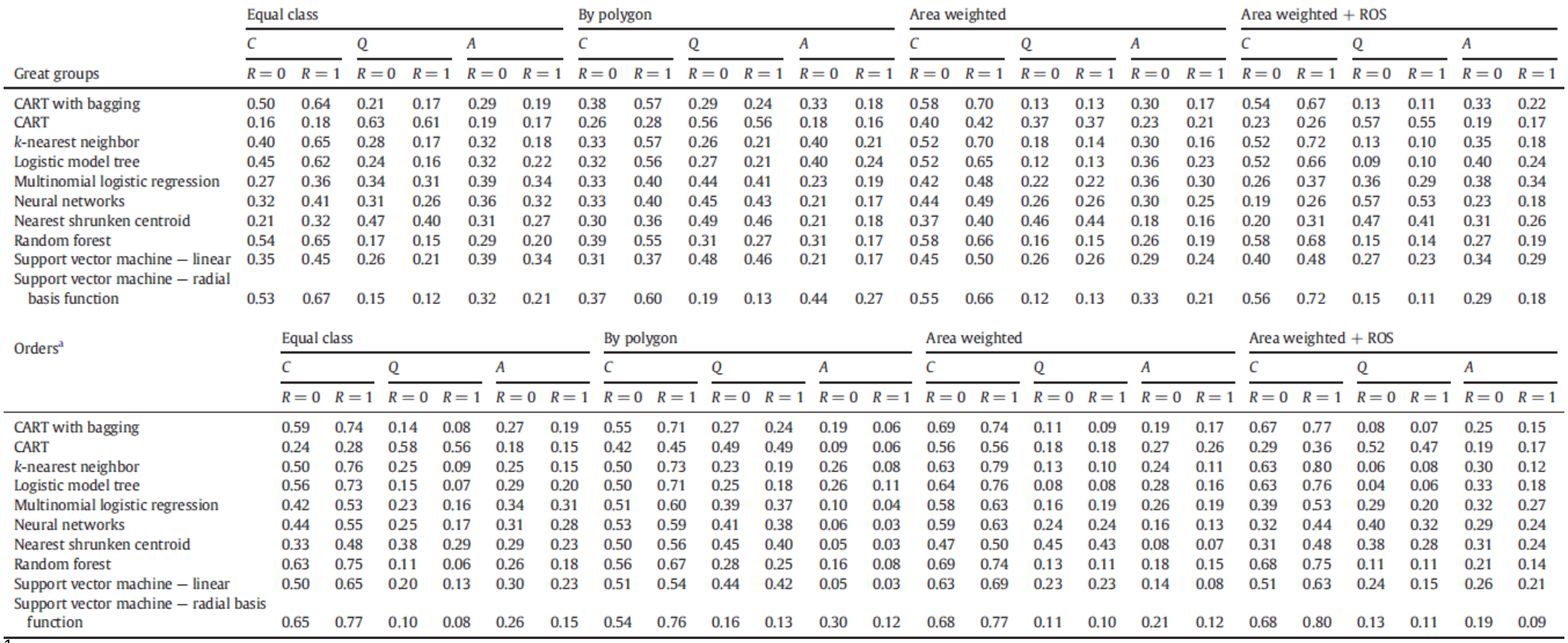


Figure 3.5. **Close-up maps of soil Great Groups derived from area-weighted (AW), by-polygon (BP), equal-class (EC), and area-weighted with random over sampling (ROS) training datasets using *k*-nearest neighbour (kNN) and support vector machine with radial basis function (SVM-RBF) learners. Miscellaneous classes such as bedrock, recent alluvium, rock outcrops, and talus slopes are included.**

Table 3.4. Classification accuracy metrics using overall agreement (*C*), quantity disagreement (*Q*), and allocation disagreement (*A*) using *n* = 262 validation points with *r* = 0 and *r* = 1-cell validation distances.



	Equal class						By polygon						Area weighted						Area weighted + ROS					
	C		*Q*		*A*		*C*		*Q*		*A*		*C*		*Q*		*A*		*C*		*Q*		*A*	
Great groups	$R = 0$	$R = 1$	$R = 0$	$R = 1$	$R = 0$	$R = 1$	$R = 0$	$R = 1$	$R = 0$	$R = 1$	$R = 0$	$R = 1$	$R = 0$	$R = 1$	$R = 0$	$R = 1$	$R = 0$	$R = 1$	$R = 0$	$R = 1$	$R = 0$	$R = 1$	$R = 0$	$R = 1$
CART with bagging	0.50	0.64	0.21	0.17	0.29	0.19	0.38	0.57	0.29	0.24	0.33	0.18	0.58	0.70	0.13	0.13	0.30	0.17	0.54	0.67	0.13	0.11	0.33	0.22
CART	0.16	0.18	0.63	0.61	0.19	0.17	0.26	0.28	0.56	0.56	0.18	0.16	0.40	0.42	0.37	0.37	0.23	0.21	0.23	0.26	0.57	0.55	0.19	0.17
k-nearest neighbor	0.40	0.65	0.28	0.17	0.32	0.18	0.33	0.57	0.26	0.21	0.40	0.21	0.52	0.70	0.18	0.14	0.30	0.16	0.52	0.72	0.13	0.10	0.35	0.18
Logistic model tree	0.45	0.62	0.24	0.16	0.32	0.22	0.32	0.56	0.27	0.21	0.40	0.24	0.52	0.65	0.12	0.13	0.36	0.23	0.52	0.66	0.09	0.10	0.40	0.24
Multinomial logistic regression	0.27	0.36	0.34	0.31	0.39	0.34	0.33	0.40	0.44	0.41	0.23	0.19	0.42	0.48	0.22	0.22	0.36	0.30	0.26	0.37	0.36	0.29	0.38	0.34
Neural networks	0.32	0.41	0.31	0.26	0.36	0.32	0.33	0.40	0.45	0.43	0.21	0.17	0.44	0.49	0.26	0.26	0.30	0.25	0.19	0.26	0.57	0.53	0.23	0.18
Nearest shrunken centroid	0.21	0.32	0.47	0.40	0.31	0.27	0.30	0.36	0.49	0.46	0.21	0.18	0.37	0.40	0.46	0.44	0.18	0.16	0.20	0.31	0.47	0.41	0.31	0.26
Random forest	0.54	0.65	0.17	0.15	0.29	0.20	0.39	0.55	0.31	0.27	0.31	0.17	0.58	0.66	0.16	0.15	0.26	0.19	0.58	0.68	0.15	0.14	0.27	0.19
Support vector machine – linear	0.35	0.45	0.26	0.21	0.39	0.34	0.31	0.37	0.48	0.46	0.21	0.17	0.45	0.50	0.26	0.26	0.29	0.24	0.40	0.48	0.27	0.23	0.34	0.29
Support vector machine – radial basis function	0.53	0.67	0.15	0.12	0.32	0.21	0.37	0.60	0.19	0.13	0.44	0.27	0.55	0.66	0.12	0.13	0.33	0.21	0.56	0.72	0.15	0.11	0.29	0.18

Orders[a]	Equal class						By polygon						Area weighted						Area weighted + ROS					
	C		*Q*		*A*		*C*		*Q*		*A*		*C*		*Q*		*A*		*C*		*Q*		*A*	
	$R = 0$	$R = 1$	$R = 0$	$R = 1$	$R = 0$	$R = 1$	$R = 0$	$R = 1$	$R = 0$	$R = 1$	$R = 0$	$R = 1$	$R = 0$	$R = 1$	$R = 0$	$R = 1$	$R = 0$	$R = 1$	$R = 0$	$R = 1$	$R = 0$	$R = 1$	$R = 0$	$R = 1$
CART with bagging	0.59	0.74	0.14	0.08	0.27	0.19	0.55	0.71	0.27	0.24	0.19	0.06	0.69	0.74	0.11	0.09	0.19	0.17	0.67	0.77	0.08	0.07	0.25	0.15
CART	0.24	0.28	0.58	0.56	0.18	0.15	0.42	0.45	0.49	0.49	0.09	0.06	0.56	0.56	0.18	0.18	0.27	0.26	0.29	0.36	0.52	0.47	0.19	0.17
k-nearest neighbor	0.50	0.76	0.25	0.09	0.25	0.15	0.50	0.73	0.23	0.19	0.26	0.08	0.63	0.79	0.13	0.10	0.24	0.11	0.63	0.80	0.06	0.08	0.30	0.12
Logistic model tree	0.56	0.73	0.15	0.07	0.29	0.20	0.50	0.71	0.25	0.18	0.26	0.11	0.64	0.76	0.08	0.08	0.28	0.16	0.63	0.76	0.04	0.06	0.33	0.18
Multinomial logistic regression	0.42	0.53	0.23	0.16	0.34	0.31	0.51	0.60	0.39	0.37	0.10	0.04	0.58	0.63	0.16	0.19	0.26	0.19	0.39	0.53	0.29	0.20	0.32	0.27
Neural networks	0.44	0.55	0.25	0.17	0.31	0.28	0.53	0.59	0.41	0.38	0.06	0.03	0.59	0.63	0.24	0.24	0.16	0.13	0.32	0.44	0.40	0.32	0.29	0.24
Nearest shrunken centroid	0.33	0.48	0.38	0.29	0.29	0.23	0.50	0.56	0.45	0.40	0.05	0.03	0.47	0.50	0.45	0.43	0.08	0.07	0.31	0.48	0.38	0.28	0.31	0.24
Random forest	0.63	0.75	0.11	0.06	0.26	0.18	0.56	0.67	0.28	0.25	0.16	0.08	0.69	0.74	0.13	0.11	0.18	0.15	0.68	0.75	0.11	0.11	0.21	0.14
Support vector machine – linear	0.50	0.65	0.20	0.13	0.30	0.23	0.51	0.54	0.44	0.42	0.05	0.03	0.63	0.69	0.23	0.23	0.14	0.08	0.51	0.63	0.24	0.15	0.26	0.21
Support vector machine – radial basis function	0.65	0.77	0.10	0.08	0.26	0.15	0.54	0.76	0.16	0.13	0.30	0.12	0.68	0.77	0.11	0.10	0.21	0.12	0.68	0.80	0.13	0.11	0.19	0.09

[1] Results for soil Orders are aggregated from predictions made from the Great Group level of soil taxonomy.

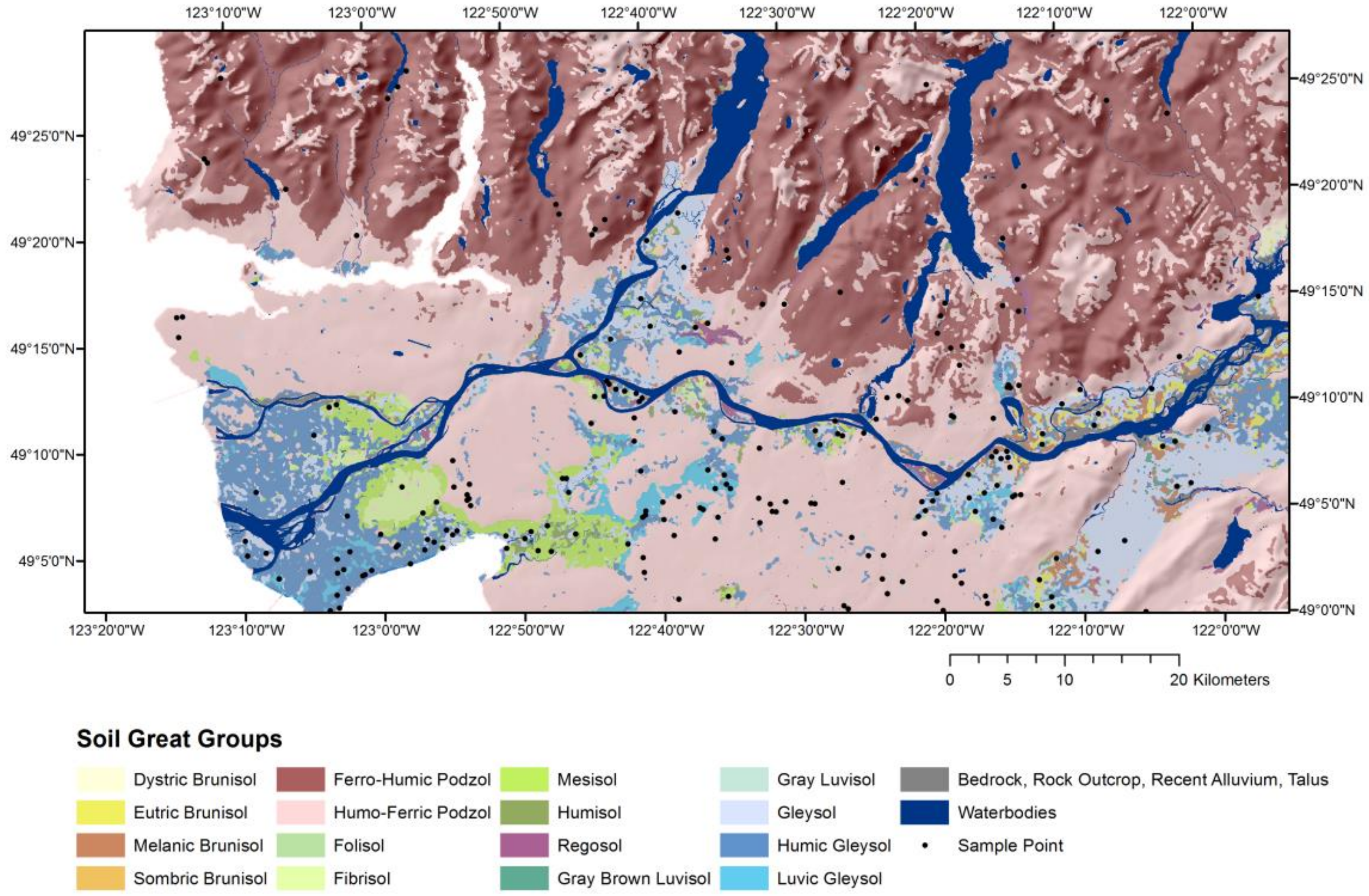


Figure 3.6. Soil Great Group map using support vector machine with radial basis function at a 100 m spatial resolution with underlying hill-shade and overlying sample points for the Langley-Vancouver Map Area, British Columbia.

Comparison of Machine-Learners

When the soil Great Group maps were compared to the validation points (Table 3.4), the average overall accuracies, based on all sample designs, were between 26% and 52% for cases where the validation points matched the predicted map at the exact spatial location (R = 0). Among the different learners, RF outperformed all other models regardless of the sample design with an average overall accuracy of 52% while CART with bagging and SVM-RBF both produced results with a slightly lower average overall accuracy of 50%. When using R = 1 validation, there was a marked improvement in the average overall accuracy of the results where learners such as CART with bagging, *k*NN, LMT, RF, and SVM-RBF all had similar average accuracies ranging between 62% and 66%. The increased accuracy was partly attributed to having many validation points located near the boundaries of cells that were classified differently from the validation point. In terms of overall agreement, *k*NN benefited the most from using R = 1 validation; however, that substantial increase in agreement was probably related to the speckling of the *k*NN maps, where cells that were considered to be valid may have occurred due to chance.

Decomposing the overall error into the quantity disagreement (Q) and allocation disagreement (A) helped explain the reasons why some models performed better or worse than others. For instance, both CART and NSC performed the worst with average accuracy rates of only 29% and 35%, respectively, when R = 1 validation procedure was used. The reasons for the low accuracy rates were attributed to high Q values of 0.52 and 0.43 for the CART and NSC learners, respectively. The high Q values were due to the CART and NSC learners producing results with only 4 out of the 22 potential classes in the training data when using the area-weighted or by polygon sampling designs, which then caused large differences in the proportion of individual classes between prediction and validation datasets.

When the Great Group results were aggregated into orders, the average increase in accuracy was 13%, regardless of whether or not R = 0 or R = 1 validation was used. The difference in overall accuracies between Great Groups and orders are the result of the subtle differences between different taxonomic units that occur at the

Great Group level of the hierarchical system that might not have been detectable by the models given the set of environmental covariates used and the spatial resolution of the results. For instance, the difference between a Eutric Brunisol and a Dystric Brunisol is a pH threshold of 5.5 – a differentiation that was difficult to model with the environmental covariates used in our study.

The results were similar in terms of the relationship between model complexity and overall agreement in the results. The application of ensemble learning resulted in improved overall agreement when CART was compared to CART with bagging or RF; the implementation of a model tree structure was an improvement for MLR models without a hierarchical variable structure; and the use of nonlinear separation planes was an improvement when SVM-RBF was compared to SVM-Lin.

When compared to other model comparison studies, the findings here were similar to those found in Brungard et al. (2015) where models such as CART, MLR, and NSC performed poorly. This is in contrast to the findings of authors such as Taghizadeh-Mehrjardi et al. (2014, 2015) and Bourennane et al. (2014), where single decision-tree models were shown to produce accurate results. Although it is difficult to explain the reasons for these differences, it may be speculated that models such as CART and single decision-trees might be more suitable when predicting only a few soil classes. For instance, Taghizadeh-Mehrjardi et al. (2014) predicted 6 soil classes, Taghizadeh-Mehrjardi et al. (2015) predicted 5 soil classes, and Bourennane et al. (2014) predicted 3 land surface types; however when the number of categories increases, in the cases of this study and Brungard et al. (2015), single decision-tree models were less effective. This would suggest that there may be a limit to the number of classes where single decision-tree models are an effective learner. Although this hypothesis is drawn from a rather small population of model comparison studies, the relationship between the number of categories and the effectiveness of machine-learners warrants further research in the future.

In terms of the models that performed reasonably well, such as SVM-RBF and RF, our findings are consistent with those of Brungard et al. (2015) while in the case of Taghizadeh-Mehrjardi et al. (2015), the accuracy of predictions made by RF performed competitively against other models despite not being ranked as the best. This would seem to suggest that SVM-RBF and RF are fairly effective as machine-learning techniques when using either legacy soil survey data or soil point data as training data; however, a direct comparison of these approaches should still be performed.

Influence of Training Data Development

Analysis of how sampling designs influenced the prediction of individual classes was not possible due to some classes having only a single or few sample points. As such, this section focuses on the overall performance of the sample designs and limits the analysis of individual classes to a visual assessment in Section 3.4.3.

On average, the area-weighted sampling design resulted in predictions that had the highest overall agreement using $R = 0$ and $R = 1$ validation for soil Great Groups and orders. When ROS was applied to the area-weighted design, the results showed either a small improvement in overall agreement, such as a 5% improvement in Great Group prediction using $R = 1$ validation, or it decreased the overall accuracy by up to 23% in the case of using ANN. An interesting observation was that the by-polygon approach consistently had higher Q values for all predictions, indicating that the by-polygon approach was not effective in representing the class proportions between the predicted and validation datasets. A likely reason for this was that having a set number of sample points within each polygon did not account for the differences in the areal extent of each individual polygon. As a result, the method would not be truly representative of the total area occupied by each class in the training data; furthermore, the feature space occupied by large polygons would not be well represented in the training dataset. Similar observations were made in Moran and Bui (2002) and Heung et al. (2014). This was in contrast to the area-weighted and area-weighted with ROS where the areas of the polygons for each class were aggregated first and then followed with random oversampling; therefore, the likelihood of sampling from a large polygon was inherently higher than the likelihood of sampling from a small one. Overall, efforts to balance the initial training dataset resulted in little to no improvement to the overall accuracies of the predictions; however, the relationship between class imbalance and the accuracies of minority classes was still inconclusive due to the limited number of validation points representing the minority classes.

3.5. Conclusions

This study compared 10 machine-learning techniques for mapping soil Great Groups and soil Orders using 4 approaches for developing training datasets from soil survey data. Key findings may be summarized as follows:

1. SVM-RBF and *k*NN produced results with the highest accuracy for R = 1 validation. In the case of SVM-RBF however, parameterization and the amount of time required to apply the model to new dataset was a challenge; hence, the use of SVM-RBF might be problematic for larger datasets. Although *k*NN also resulted in a high accuracy and was relatively efficient in terms of the parameterization process, the *k*NN learner might produce 'speckled' results that are difficult to interpret – potentially due to over-fitting. Alternatives to these models would include the use of CART with bagging, LMT, or RF – all of which had similar accuracies and could be parameterized efficiently.

2. The sampling method used to extract training data from soil surveys can greatly influence the resulting predictions – the area-weighted approach resulted in the highest overall accuracy. In an attempt to address the issue of class imbalance, it was observed that the equal-class sampling approach resulted in decreased accuracy when compared to the area-weighted approach. Furthermore, the resulting maps showed soil patterns that were inconsistent with our understanding of the soils of the study area. Area-weighted with ROS approach resulted in minor improvements for some models; however, the amount of additional processing time required for parameterizing models makes the use of ROS unfeasible for larger datasets.

3. An area of future research could lie in the extension of existing machine algorithms with techniques in ensemble learning through the use of bagging and randomized variable selection – as is the case with CART with bagging and RF. Further research could also be carried out in developing and testing model-tree learners that hybridize linear models with a tree-based structure in order to account for non-linear and hierarchical variable relationships – as is the case with the LMT learner that performed comparably reasonably well with other learners.

Different machine-learning algorithms resulted in drastically different outputs. Future research in DSM, as well as in geomorphic or predictive ecosystem modeling, should not be restricted to a single learner or a small selection of them. Model comparison packages in *R*, such as the *caret* package, greatly increase the efficiency and ease of using and comparing multiple learners.

3.6. Acknowledgements

The authors are thankful for support from an Alexander Graham Bell award from the Natural Sciences and Engineering Research Council of Canada granted to B. Heung.

3.7. Supplementary Figures

Supplementary content consists of 80 maps of soil Great Groups and Orders using various machine-learning techniques and methods for training data development. Figures have been published in Geoderma and may be accessed via Heung et al. (2016).

3.8. References

Aitkenhead, M.J., Coull, M., Hudson, T.G., Black, H.I.J., 2013. Prediction of soil characteristics and colour using data from the National Soils Inventory of Scotland. Geoderma 200-201, 99-107.

Barsi, J.A., Barker, J.L., Schott, J.R., 2003. An atmospheric correction parameter calculator for a single thermal band earth-sensing instrument. *In* Geoscience and Remote Sensing Symposium, 2003. IGARSS'03, IEEE International 5, 3014-3016.

Beven, K.J., Kirkby, M.J., 1979. A physically based, variable contributing area model of basin hydrology. Hydrologic Sciences Bulletin 24, 43-69

B.C. Ministry of Sustainable Resource Management, 2002. Gridded Digital Elevation Model Product Specification, 2nd edition. Base Mapping and Geomatics Services Branch, Ministry of Sustainable Resource Management, Victoria.

Behrens, T., Zhu, A-X., Schmidt, K., Scholten, T., 2010. Multi-scale digital terrain analysis and feature selection for digital soil mapping. Geoderma 155, 175-185.

Bourennane, H., Couturier, A., Pasquier, C., Chartin, C., Hinschberger, F., Macaire, J-J., Salvador-Blanes, S., 2014. Comparative performance of classification algorithms for the development of models of spatial distribution of landscape structures. Geoderma 219-220, 136-144.

Breiman, L., 1996. Bagging predictors. Machine Learning 24, 123-140.

Breiman, L., 1998. Arcing classifiers (with discussion). The Annals of Statistics 26, 801-849.

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

Breiman, L., Friedman, J., Stone, C.J., Olshen, R.A., 1984. Classification and Regression Trees. CRC Press LLC, Boca Raton, FL.

Brungard, C.W., Boettinger, J.L, Duniway, M.C., Wills, S.A., Edwards, T.C., 2015. Machine learning for predicting soil classes in three semi-arid landscapes. Geoderma 239-240, 68-83.

Cavazzi, S., Corstanje, R., Mayr, T., Hannam, J., Fealy, R., 2013. Are fine resolution digital elevation models always the best choice in digital soil mapping?. Geoderma 195-196, 111-121.

Chang, D-H., Islam, S., 2000. Estimation of soil physical properties using remote sensing and artificial neural networks. Remote Sensing of Environment 74, 534-544.

Coll, C., Galve, J.M., Sanchez, J.M., Caselles, V., 2010. Validation of Landsat-7/ETM+ thermal-band calibration and atmospheric correction with ground-based measurements. IEEE Transactions on Geoscience and Remote Sensing 48, 547-555.

Florinsky, I.V., 1998. Accuracy of local topographic variables derived from digital elevation models. International Journal of Geographical Information Science 12, 47-61.

Collard, F., Kempen, B., Heuvelink, G.B.M, Saby, N.P.A., Richer de Forges, A.C., Lehmann, S., Nehlig, P., Arrouays, D., 2014. Refining a reconnaissance soil map by calibrating regression models with data from the same map (Normandy, France). Geoderma Regional 1, 21-30.

Gallant, J.C., Dowling, T.I., 2003. A multiresolution index of valley bottom flatness for mapping depositional areas. Water Resources Research 39, 1347-1359.

Gao, B.C., 1996. NDWI – a normalized difference water index for remote sensing of vegetation liquid water from space. Remote Sensing of Environment 58, 257-266.

Häntzchel, T., Goldberg, V., Bernhofer, C., 2005. GIS-based regionalization of radiation temperature and coupling measures in complex terrain for low mountain ranges. Meteorological Applications 12, 33-42.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed. Springer, New York, NY, 734 pp.

Hectares, B.C., 2012. Hectares BC. Available at http://hectaresbc.org/app/habc/HaBC.html (verified 28 October 2014)

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.

Heung, B., Ho, H.C., Zhang, J., Knudby, A., Bulmer, C.E., Schmidt, M.G., 2016. An overview and comparison of machine-learning techniques for classification purposes in digital soil mapping. Geoderma 265, 62-77.

Ho, H.C., Knudby, A., Sirovyak, P., Xu, Y., Hodul, M., Henderson, S.B., 2014. Mapping maximum urban air temperature on hot summer days. Remote Sensing of Environment 154, 38-45.

Howley, T., Madden, M.G., O'Connell, M.-L., Ryder, A.G., 2006. The effect of principal component analysis on machine learning accuracy with high-dimensional spectral data. Knowledge-Based Systems 19, 363-370.

Jenny, H., 1941. Factors of Soil Formation: A System of Quantitative Pedology. McGraw-Hill, NY.

Kempen, B., Brus, D.J., Heuvelink, G.B.M., Stoorvogel, J.J. 2009. Updating the 1:50,000 Dutch soil map using legacy soil data: a multinomial logistic regression approach. Geoderma 51, 311-326.

Kenney, E., Frank, G., 2010. Creating a seamless soil dataset for the Okanagan Basin, British Columbia. Proceedings of the Western Regional Cooperative Soil Survey, Las Vegas, NV.

Kovačevic, M., Bajat, B., Gajić, B., 2010. Soil type classification and estimation of soil properties using support vector machines. Geoderma 154, 340-347.

Kuhn, M., 2008. Building predictive models in R using the caret package. Journal of Statistical Software 28: 1-26.

Laslett, G.M., McBratney, A.B., Pahl, P.J., Hutchinson, M.F., 1987. Comparison of several spatial prediction methods for soil pH. Journal of Soil Science 38, 513-532.

Ließ, M., Glaser, B., Huwe, B,. 2012. Uncertainty in the spatial prediction of soil texture: comparison of regression tree and Random Forest models. Geoderma 170, 70-79.

Luttmerding, H.A., 1981. Soils of the Langley-Vancouver Map Area. Report No. 15, British Columbia Soil Survey. Research Branch, Canada Department of Agriculture, Ottawa, ON, Canada.

MacMillan, R.A., 2005. A new approach to automated extraction and classification of repeating landform types. Frontiers in Pedometrics, Naples, FL.

Mansuy, N., Thiffault, E., Paré, D., Bernier, P., Guindon, L., Villemaire, P., Poirier, V., Beaudoin, A., 2014. Digital mapping of soil properties in Canadian managed forests at 250 m of resolution using the *k*-nearest neighbor method. Geoderma 235-236, 59-73.

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

McKenzie, N., Austin, M.P., 1993. A quantitative Australian approach to medium and small scale surveys based on soil stratigraphy and environmental correlation. Geoderma 57, 329-355.

McKenzie, N., Ryan, P., 1999. Spatial prediction of soil properties using environmental correlation. Geoderma 89, 67-94.

Moore, I.D., Grayson, R.B., Ladson, A.R., 1991. Digital terrain modelling: a review of hydrological, geomorphological, and biological applications. Hydrological Processes 5, 3-30.

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.

Moran, C.J., Bui, E.N., 2002. Spatial data mining for enhanced soil map modeling. International Journal of Geographical Information Systems 16, 533-549.

Odeh, I.O.A., McBratney, A.B., and Chittleborough, D.J., 1994. Spatial prediction of soil properties from landform attributes derived from a digital elevation model. Geoderma 63, 197-214.

Odgers, N.P., Sun, W., McBratney, A.B., Minasny, B., and Clifford, D., 2014. Disaggregating and harmonising soil map units through resampled classification trees. Geoderma 214, 91-100.

Pojar, J., Klinka, K., Demarchi., 1991. Ecosystems of British Columbia. British Columbia Ministry of Forests and Range, 95-111.

Pontius, R.G. Jr., Millones, M., 2011. Death to Kappa: birth of quantity disagreement and allocation disagreement for accuracy assessment. International Journal of Remote Sensing 32, 4407-4429.

Priori, S., Bianconi, N., Constantini, E.A.C., 2014. Can γ-radiometrics predict soil textural data and stoniness in different parent materials? A comparison of two machine-learning methods. Geoderma 226-227, 354-364.

R Development Core Team, 2012. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available at http://www.R-project.org (verified 28 October 2014).

Riley, S.J., DeGloria, S.D., Elliot, R., 1999. A terrain ruggedness index that quantifies topographic heterogeneity. Intermountain Journal of Science 5, 23-27.

SAGA Development Team, 2011. System for Automated Geoscientific Analyses (SAGA). Available at http://www.saga-gis.org/en/index.html (verified 28 October 2014).

Schmidt, K., Behrens, T., Scholten, T., 2008.Instance selection and classification tree analysis for large spatial datasets in digital soil mapping. Geoderma 146, 138-146.

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

Shawe-Taylor, J., Cristianini, N., 2004. Kernel Methods for Pattern Analysis. Cambridge University Press, Cambridge, UK.

Silveira, C.T., Oka-Fiori, C., Santos, L.J.C., Sirtoli, A.E., Silva, C.R., Botelho, M.F., 2013. Soil prediction using artificial neural networks and topographic attributes. Geoderma 195-196, 165-172.

Soil Classification Working Group, 1998. The Canadian System of Soil Classification, 3rd ed. Research Branch, Agriculture and Agri-Food Canada, Ottawa, ON.

Sondheim, M., Suttie, K., 1983. User Manual for the British Columbia Soil Information System, 1. BC Ministry of Forests Publication R28-82053, Victoria, BC.

Subburayalu, S.K., Jenhani, I., Slater, B.K., 2014. Disaggregation of component soil series on an Ohio County soil survey map using possibilistic decision trees. Geoderma 213, 334-345.

Subburayalu, S.K., Slater, B.K., 2013. Soil series mapping by knowledge discovery from an Ohio County soil map. Soil Science Society of America Journal 77, 1254-1268.

Taghizadeh-Mehrjardi, R., Nabiollahi, K., Minasny, B., Triantafilis, J., 2015. Comparing data mining classifiers to predict spatial distribution of USDA-family soil groups in Baneh region, Iran. Geoderma 253-254, 67-77.

Taghizadeh-Mehrjardi, R., Sarmadian, F., Minasny, B., Triantafilis, J., 2014. Digital mapping of soil classes using decision tree and auxiliary data in the Ardkan region, Iran. Arid Land Research and Management 213, 15-28.

Van de Griend, A.A., Owe, M., 1993. On the relationship between thermal emissivity and the normalized difference vegetation index for natural surfaces. International Journal of Remote Sensing 14, 1119-1131.

Van Hulse, J., Khoshgoftaar, T., 2009. Knowledge discovery from imbalanced and noisy data. Data & Knowledge Engineering 68, 1513-1542.

Van Hulse, J., Khoshgoftaar, T.M., Napolitano, A., 2007. Experimental perspectives on learning from imbalanced data. Proceedings of the 24th Annual International Conference on Machine Learning (ICML 2007), Corvalis, OR, 935-942.

Vasques, G.M., Demattê, J.A,M., Viscarra Rossel, R.A., Ramírez-López, L., Terra, F.S., 2014. Soil classification using visible/near-infrared diffuse reflectance spectra from multiple depths. Geoderma 223-225, 73-78.

Warrens, M.J., 2015. Properties of the quantity disagreement and the allocation disagreement. International Journal of Remote Sensing 36, 1439-1446.

Wilson, J.P., Gallant, J.C., (eds.), 2000. Terrain Analysis: Principles and Applications. John Wiley & Sons, New York, NY.

Witten, I.H., Frank, E., 2005. Data Mining: Practical Machine Learning Tools and Techniques, 2nd ed. Elsevier, San Francisco, CA, 525 pp.

Chapter 4.

Comparing the Use of Legacy Soil Pits and Soil Survey Polygons as Training Data for Mapping Soil Classes[1]

4.1. Abstract

Machine-learners used for digital soil mapping are generally trained using either data derived from field-observed soil pits or from soil survey polygons - although no direct comparison of the accuracy resulting from the two methods has yet to be undertaken. This study examined such a comparison over the Okanagan Valley and Kamloops region of British Columbia where good quality soil pit and soil survey data were available. A standard set of environmental variables including vegetative, climatic, and topographic indices were used to predict soil Great Groups in accordance with the Canadian System of Soil Classification. The pit-derived training dataset was developed using *n* = 478 points from the British Columbia Soil Information System while the polygon-derived training dataset was developed through random sampling of single-component soil survey map units based on an area-weighted approach. In both cases, the training points were intersected with a suite of 18 environmental covariates, reduced from 21 covariates using principal component analysis, and submitted to a machine-learner for predictions at a 100 m spatial resolution. Four single-model learners (CART, *k*-nearest neighbour, multinomial logistic regression, and logistic model tree) and five ensemble-model learners (CART with bagging, *k*-nearest neighbour with bagging, multinomial logistic regression with bagging, logistic model trees with bagging, and

[1] A version of his chapter has been published in "Heung, B., Hodúl, M., Schmidt, M.G., 2017 Comparing the use of legacy soil pits and soil survey polygons as training data for mapping soil classes. Geoderma 290, 51-68."

Random Forest) were compared. Surfaces of prediction uncertainty were produced using ignorance uncertainty and results were validated using a 5-fold cross-validation procedure. Predictions made using polygon-derived training data were consistently higher in accuracy across all models where the Random Forest model was the most effective learner with *C* = 61% accuracy when using pit-derived training data and *C* = 68% accuracy when using polygon-derived training data. Comparing single-model and ensemble-learner models, the bagging algorithm resulted in a 2-11% increase in accuracy when using pit-derived training data. Ensemble-models allowed for the visualization of prediction uncertainty. This study provides further insight into the use of legacy soil data and the development of training data for digital soil mapping.

4.2. Introduction

The soil-environmental variables identified in Jenny (1941) codified the concept of soil-environmental relationships, where easily measurable environmental properties could be used to predict soil properties. In digital soil mapping (DSM), the environmental-correlation concept (McKenzie and Ryan 1996), later formalized within the *scorpan* model (McBratney et al., 2003), takes spatial soil data and co-locates it to readily available environmental data such as digital elevation models (DEM) and remotely sensed data in order to form the training dataset for a type of supervised learning. The relationships between soil and environmental conditions are correlated through the fitting of a model using machine-learning and/or geostatistical techniques, where the soil-environmental relationships are then used to predict the soil properties for areas that have not been sampled. Furthered with increasing computational power, advancing remote sensing and GIS technologies, and the availability of accurate soil-environmental data, the application of the environmental-correlation concept has been applied for the mapping of soil classes and attributes over progressively larger spatial extents and data sizes (Chaney et al., 2016; Hengl et al., 2014, 2015; McBratney et al., 2003; Mulder et al., 2011, 2016).

Within the DSM literature, training data for mapping categorical soil properties has typically come from one of two sources: soil pit data or soil polygon data that has been digitized from conventional soil survey maps (Brungard et al., 2015; Heung et al., 2016). When using soil pit data for mapping soil taxonomic units, geolocated soil profile information is either recovered from a legacy soil database (Bui et al., 2006; Hengl et al., 2014) or based on field data that were collected for specific studies (Brungard et al., 2015; Rad et al., 2014). The use of pit data is particularly useful for situations when there is limited soil survey data available, when there is an existing database of field observations, or when the spatial resolution of existing soil surveys is too coarse.

When using polygon data for model training purposes, the generic procedure typically involves the generation of training points within polygons where values from environmental covariates are extracted. This method has been used to map properties such as surficial geology and soil parent material (e.g. Bui and Moran, 2001) but has most commonly been used to map soil taxonomic units (e.g. Bui and Moran, 2001, 2003; Collard et al., 2014; Grinand et al., 2008; Odgers et al., 2014; Subburayalu and Slater, 2013; Subburayalu et al., 2014). The methods for generating training points have varied amongst studies – some of which included an area-weighted approach where the number of randomly generated sample points for each class were proportional to the class' areal extent (e.g. Moran and Bui, 2002); a by-polygon approach where a set number of training points were randomly generated within each polygon (e.g. Odgers et al., 2014); equal class sampling where the number of randomly generated training points for each class were equal (e.g. Moran Bui, 2002); and a sampling approach that integrated expert knowledge in the selection of points (e.g. Bulmer et al., 2016). Studies that have compared some of these methods have typically identified an area-weighted approach to produce more accurate predictions, relative to other methods, as the spatial extent and variability of the largest classes were better represented within the training data (Moran and Bui, 2002; Heung et al., 2014, 2016).

The main advantages of the polygon method include the ability for users to select an arbitrarily large sample size, which is beneficial for capturing more of the landscape's variability and the multivariate feature space of a categorical variable (Moran and Bui, 2002; Heung et al., 2014, 2016); furthermore, this approach has also been shown to be effective for the refinement and improvement of existing maps through the disaggregation of complex map units (Collard et al., 2014; Häring et al., 2012; Holmes et al., 2015; Odgers et al., 2014; Subburayalu et al., 2014). A concern with this approach has typically been related to the issue of map scale and the variability and purity within individual map units at given scales (Lin et al., 2005). For instance, in Heung et al. (2016), it was visually observed that as the map scale decreased from one region of the study area to another, there was a noticeable decrease in the diversity of soils that were predicted. Furthermore, soils developed from local-scale colluvial and fluvial processes were poorly predicted. The relationship between soil survey scale and the accuracy of predictions has also been observed in studies such as Bui and Moran (2003); in addition, that study also identified that the accuracy of predictions varied greatly even when soil surveys that were mapped at similar scales were used as training data due to differing survey methods and the time given to complete the survey.

Although these two approaches have commonly been used in the DSM literature, studies such as Brungard et al. (2015), Heung et al., (2016), and Lacoste et al. (2011) have identified a potential research gap where these approaches have yet to be directly compared using the same suite of machine-learners and environmental covariates over a study area. As such, the primary objective of this study was to address the comparison between pit-derived and polygon-derived data as training data for predicting soil classes at the Great Group level of the taxonomic hierarchy, based on the Canadian System of Soil Classification (Soil Classification Working Group, 1998), for the Okanagan-Kamloops region of British Columbia. Here, the pit-derived training data were obtained from legacy soil pit data taken from the British Columbia Soil Information System (BCSIS) (Sonheim and Suttie, 1983) and the polygon-derived training data were derived from legacy soil survey polygons using the framework provided in Heung et al. (2014). An identical suite of 27 environmental covariates and nine machine-learning algorithms for classification were tested on each of the two training datasets and as such, differences in the results could be constrained to the differences in training data. Secondary objectives included the comparison of nine machine-learning algorithms: four single-model learners and five ensemble-model learners. Validation of the predictions was performed using soil pit data and a cross-validation procedure.

4.3. Methodology

4.3.1. Study Area

The study area was chosen due to the availability of both soil pit data and soil survey polygon data of high reliability and spatial quality, as well as a relatively high sampling density in the case of pit data. The study area represents a 47,350 km^2 portion of south-central British Columbia (Figure 4.1) and is located at approximately 49.0° N to 51.1° N latitude, 117.5° W to 120.8° W longitude, with an elevation range of 280-2720 m.

There is a great diversity of ecosystem types within the study area with the Interior Douglas Fir (IDF) and Ponderosa Pine (PP) biogeoclimatic zones making up much of the valleys and the Bunchgrass (BG) zone at the lowest elevations. The IDF is the largest zone in the valleys, with a mean annual temperature of 1.6-9.5°C, and 300-750 mm of precipitation, 15-40% of which falls as snow (Hope et al., 1991b). The zone is largely covered by mature stands of Douglas Fir (*Pseudotsuga menziesii*), although grasslands occur in some places. Soils here are primarily Luvisols and Brunisols, with Chernozems occurring in the grasslands. Due to the basic volcanic parent material and low leaching rates in the arid environment, the soils are considered to have a high nutrient status (Hope et al., 1991b). The PP zone occurs below the IDF zone, and is the driest and warmest forested zone in British Columbia, with a mean annual temperature of 4.8-10°C and 280-500 mm of precipitation. Soils here are much the same as in the IDF zone, consisting mostly of Chernozems and Brunisols. The lowest elevations, along valley bottoms of major rivers in the region, are occupied by the BG zone and is characterized by its warm, dry climate with sparse shrubs and grass cover, and Chernozemic soils (Hope et al., 1991a).

Higher elevations are characterized by the forested Montane Spruce (MS) and Interior Cedar Hemlock (ICH) zones, with Engelmann Spruce Subalpine Fir (ESSF) and Interior Mountain Heather Alpine (IMA) zones located at the highest points (Ketcheson et al., 1991). The ICH zone has a mean annual temperature of 2.0-8.7°C, and 500-1200 mm of precipitation of which 25-50% falls as snow. Humo-Ferric Podzols dominate at drier areas while Ferro-Humic Podzols and Gleysols occur in wetter areas. The MS zone occurs at slightly higher elevations, leading to lower mean annual temperatures of 0.5-4.7°C, and 380-900 mm of precipitation. The soils of the MS zone are mostly of the Brunisolic and Luvisolic orders formed from clayey volcanic parent material; however, Humo-Ferric Podzols can be found in areas that are moist with coarse parent materials (Hope et al., 1991c). The ESSF and IMA zones occur only at the highest elevations in the northeast portion of the study area, and represent only a small proportion of its total area.

4.3.2. Environmental Covariates

27 environmental variables were derived from remote sensing, climate and digital elevation model (DEM) data (Table 4.1). In order to decrease multi-collinearity between the variables and computational demand, principal component analysis was performed on the topographic and vegetation data. The analysis resulted in a total of 18 covariates, which were then scaled in order to convert the covariate values into distributions with similar ranges – a procedure that is recommended for machine-learners (such as *k*-nearest neighbors) where the decision boundaries for classes are defined based on the distance in feature space between observed and unobserved points.

Topographic Indices

Topographic indices were derived from a 100 m spatial resolution DEM of the study area, obtained from HectaresBC.org – a provincial repository of freely available environmental data. Consecutive smoothing was applied to the DEM in order to minimize the effects of spatially non-correlated noise on the calculation of topographic indices, in the form of three consecutive mean filters of 3 x 3, 3 x 3, and 5 x 5 pixels (Heung et al., 2014). All indices were calculated using the System for Automated Geoscientific Analysis (SAGA) suite of topographical analysis tools (SAGA Development Team, 2011). The topographic variables were selected based on their ability to represent local scale (e.g. elevation, slope, aspect, and curvature) and landscape scale (multi-resolution ridge top flatness and valley bottom flatness; slope height and position; and valley depth) morphometric characteristics, hydrological characteristics (e.g. stream power index and wetness index), and landscape exposure (e.g. sky view factor and terrain view factor). To reduce computational time and improve prediction accuracy

(Howley et al., 2006), the initial 19 indices were reduced to their first 12 principal components, which accounted for 96.1% of the cumulative variance.

Vegetative Indices

Vegetative indices were derived from six Landsat satellite images with minimal cloud coverage. The Landsat images were obtained from the Landsat Surface Reflectance-Derived Spectral Indices dataset (Masek et al., 2006) as part of the USGS Landsat Higher Level Science Data Products. Spectral indices were calculated from surface reflectance images, which were calculated from Landsat data by applying atmospheric correction using the Landsat Ecosystem Disturbance Adaptive Processing System (LEDAPS; Schmidt et al. 2013).

Here, the Normalized Difference Vegetation Index (NDVI) was used to provide an indication of the amount of healthy vegetation and was formulated as follows:

Eq. (4.1) $$\text{NDVI} = (\rho_{\text{NIR}} - \rho_{\text{R}}) / (\rho_{\text{NIR}} + \rho_{\text{R}}),$$

where ρ_{NIR} and ρ_{R} represent the surface reflectance of the near-infrared and visible red wavelengths respectively. In order to adjust for the effects of soil moisture and color, the Soil Adjusted Vegetation Index (SAVI; Huete, 1988) and the Modified Soil Adjusted Vegetation Index (MSAVI; Qi et al., 1994) were calculated where MSAVI has been shown to be more useful in areas with sparse vegetation coverage in comparison to the SAVI (Rondeaux et al. 1996). The SAVI was calculated as:

Eq. (4.2) $$\text{SAVI} = (1 + L)\,(\rho_{\text{NIR}} - \rho_{\text{R}}) / (\rho_{\text{NIR}} + \rho_{\text{R}} + L),$$

where $L = 0.5$ was selected as the adjustment factor for vegetation density while MSAVI was calculated as:

Eq. (4.3) $$\text{MSAVI} = \{2\rho_{\text{NIR}} + 1 - [(2\rho_{\text{NIR}} + 1)^2 - 8(\rho_{\text{NIR}} - \rho_{\text{R}})]^{0.5}\}/2.$$

The Enhanced Vegetation Index (EVI; Jiang et al., 2008), was used as a vegetative index that has been shown to minimize the effects of soil and atmosphere on satellite imagery and was calculated as:

Eq. (4.4) $\mathrm{EVI} = 2.5(\rho_{\mathrm{NIR}} - \rho_{\mathrm{R}}) / (\rho_{\mathrm{NIR}} + 2.5\rho_{\mathrm{R}} + 1)$.

The Normalized Difference Water Index (NDWI) was calculated in order to provide information on the size and health of vegetation, soil moisture, and evapotranspiration and was formulated as follows (Gao, 1996):

Eq. (4.5) $\mathrm{NDWI} = (\rho_{\mathrm{NIR}} - \rho_{\mathrm{MIR}}) / (\rho_{\mathrm{NIR}} + \rho_{\mathrm{MIR}})$,

where ρ_{MIR} represents the surface reflectance of the mid-infrared wavelengths. The five indices were reduced to the first three principal components, which accounted for 96.0% of the cumulative variance.

Climatic Indices

Climate data was taken from Climate BC (Wang et al., 2012), and accessed through HectaresBC.org. Three climate variables were used: mean annual temperature, mean annual precipitation, and number of frost free days.

4.3.3. Training Data

This study compared the results produced training data derived from legacy soil pit data and training data derived from soil survey polygons. In both cases, training points were co-located with the environmental covariates and the soil-environmental matrix was submitted to the machine-learners for model training.

Training Points using Legacy Soil Pit Data

Pit data were obtained from the BCSIS database (Sondheim and Suttie, 1983), which is comprised of soil data from provincial and federal agencies and compiled from biophysical inventories, terrestrial ecosystem mapping, soil surveys, and habitat monitoring projects. A total of 478 individual field-observed points with a known soil Great Group were used for this project. Of the 18 classes, the majority classes included Eutric Brunisols, Dystric Brunisols, and Gray Luvisols, which consisted of 53% of the training points (Figure 4.2).

Training Points using Legacy Soil Survey Polygons

The digitized soil survey data used for this study were obtained from the BC Soil Information Finder Tool (BC Ministry of Agriculture and BC Ministry of Environment, 2016). The digitized data consisted of four soil surveys of the region: the Okanagan and Similkameen Valleys mapped at 1:20,000 scale (Wittneben, 1986); the Ashcroft Area mapped at 1:50,000 scale (Young et al., 1992); the Princeton Area mapped at 1:125,000 scale (Green and Lord, 1979); and the Tulameen Area mapped at 1:50,000 scale (Lord and Green, 1974).

To simplify the soil survey legend, the soil Series defined by the original soil surveys were reduced to soil Great Groups. In order to reduce the uncertainty in classification from within the training data, only single-component mapping units (pure polygons), based on soil Great Groups, were used as training areas. Of the study area, the coverage by single-component polygons occupied 53.2% (25,278 km^2) of the entire extent (Fig. 4.3). First, polygons were rasterized to a 100 m spatial resolution where the training points selected were based on an area-weighted random sample of the pixels. The number of training points per class was selected from 5% of the areal extent of each class, which resulted in 129,465 training points. The 5% sampling density was selected in order to balance the need for an optimal sample size with concerns related to computational time when using datasets containing >100,000 training points (Heung et al., 2016). The majority classes included Humo-Ferric Podzols located in the forested, high-elevation regions of the study area; Dystric and Eutric Brunisols located in drier regions; and Gray Luvisols located along the transition zones between the grassland and forested areas (Fig. 4.2).

4.3.4. Machine-Learners

This study tested two classes of machine-learners for classification: single-model learners and ensemble-model learners. Single-model learners included classification and regression trees (CART), *k*-nearest neighbour (*k*NN), multinomial logistic regression (MLR), and logistic model trees (LMT) while ensemble-model learners extended the single-model learners by coupling them with a bagging algorithm and thus included CART with bagging (CART+), *k*NN with bagging (*k*NN+), MLR with bagging (MLR+), LMT with bagging (LMT+), and Random Forest (RF). Although this section will briefly summarize each model, more detail of the models are provided in Heung et al. (2016), where the models have been reviewed and compared. Model parameters were optimized using the cross-validation procedure described in Heung et al. (2014). All the modeling was done using the *R* statistical software (R Development Core Team, 2012) and the *caret* package, which included all the tested models (Kuhn, 2008).

Single-Model Learners

The *k*NN learner may be classified as a distance-based learner where an unobserved location is classified based on the distance to the nearest neighbouring point(s) within feature space using a distance function. As such, the main parameter of the *k*NN learner is in the selection of *k*, which determines the number of training points used to make a classification; as a result, when $k = 1$, the predicted location is assigned the class of the nearest neighbouring point in feature space and when $k > 1$, the location is assigned based on a majority vote from multiple training points (Hastie et al., 2009).

Logistic regression models are one of the few linear models that have been applied in DSM and are used to represent dichotomous variables in probabilistic terms ranging from 0 (low probability of occurrence) to 1 (high probability of occurrence). In order to extend this model for multiclass purposes, individual logistic regression models are constructed for each class and generalized into the following multinomial model (Debella-Gilo and Etzelmüller, 2009; Kempen et al., 2009):

Eq. (4.6)
$$p_i = \frac{\exp(p_i)}{\exp(p_1) + \exp(p_2) + \ldots + \exp(p_n)},$$

where p_i is the probability of a class (i) occurring at a location and the denominator represents the sum of occurrence probabilities for n classes. Class assignment is based on the class with the highest probability of occurrence.

The classification and regression tree (CART) is a hierarchal modeling approach that consists of nodes and leaves, where at each node, it consists of an *if-then* statement that partitions the training data by maximizes the within-node homogeneity and the between-node heterogeneity based on a node-splitting rule (Breiman, 1984). The advantages of tree-based models are in their ability to capture the non-linear relationship between predictor and response variables as well as the interactions between predictor variables. Although the CART model has been shown to underperform compared to others (Brungard et al., 2015; Heung et al., 2016), the model was used because of its shared characteristics with LMT and RF models.

The LMT (Landwehr et al., 2005) classifier is a relatively new learner that has not often been used in the DSM or the machine-learning literature; however, in Heung et al. (2016) the model showed promise with results comparable to the predictions made by support vector machines and RF. Model trees, in general, are a type of hybridized model that combine linear models with tree-based models and as such, they minimize the risk of over-fitting and under-fitting the data in the case of tree-based models and linear models, respectively. In the prediction of quantitative variables, examples of model trees include the Cubist and M5 model trees, where the terminal nodes (leaves) of a tree-based model consist of a linear regression model (Quinlan, 1992). In the case of LMT, the terminal nodes consist of individual partial logistic regression models that have been iteratively fitted using a *LogitBoost* algorithm (Friedman et al., 2000), where the tree is reduced in size using the CART pruning procedure (Breiman, 1996).

Ensemble-Model Learners

Ensemble-model learners use multiple models that are integrated into a single predictive model with the intention of improving predictions when compared to single-model learners (Rokach, 2010). In the machine-learning literature, the improvements in accuracy are particularly effective when a single-model learner is sensitive to perturbations in the training data (Breiman, 1996). Four key components form an ensemble-model: the training data; base inducer (single-model classifier); the diversity generator that incorporates perturbations into the modeling process; and a combining function that integrates the predictions of the individual single-models (Rokach, 2010).

For this study, the base inducers tested included the CART, MLR, *k*NN, and LMT learners – as described previously. In order to generate a diverse set of learners, an independent ensemble approach was used where the predictions made by each single-model were independent of each other – as such, the bootstrap aggregation (bagging) method was applied (Breiman, 1996, 2001). When using the bagging method, single-models are trained using a random bootstrap sample, with replacement, from the entire training dataset and the results are aggregated using a majority-vote combination function based on 100 iterations of the single-model learner. A bagging algorithm was chosen due to its effectiveness in reducing the impact of classifiers with a high variance such as CART and *k*NN (depending on the choice of *k*). As a result, CART+, MLR+, *k*NN+, and LMT+ ensemble learners were tested. In addition to CART+, the RF learner (Breiman, 2001) was also tested and differs to CART+ in that it includes an additional diversity generator that uses a subset of randomly selected predictor variables that are tested in making each node splitting rule. Although ensemble-modeling techniques result in a higher computational demand, processing time was minimized through the use of parallel processing.

4.3.5. Assessment of Predictions

Predictions were evaluated using the BCSIS soil pits located within the study area. In order to account for validation points that were located along pixel boundaries, pixels were considered to be valid if the validation point matched a pixel within a radius of 1 pixel (Heung et al., 2014, 2016). To assess the consistency between soil survey polygons and the predicted results, single-component polygons were rasterized and accuracy metrics were calculated based on map comparisons. In addition, visual assessments were performed in order to evaluate the predictions based on expert-knowledge of the soil-environmental relationships of the region.

Accuracy metrics included overall agreement, *C*, which represents the percentage of correctly matched cases between prediction and validation datasets. In addition to the overall agreement, accuracy metrics included the calculation of the quantity disagreement (*Q*) and allocation disagreement (*A*) that could be calculated using the confusion matrices produced from the validation procedure (Pontius and Millones, 2011; Warrens, 2015). Both disagreement values are the result of decomposing the total disagreement, *D,* as follows:

Eq. (4.7) $D = 1 – C = Q + A.$

Here, Q represents the amount of disagreement in the proportion of each soil class between the validation and prediction datasets and is calculated as:

Eq. (4.8) $Q = \frac{1}{2}\sum_{i=1}^{j}\left|p_{i+} - p_{+i}\right|,$

where p_{i+} and p_{+i} represent the row and column totals of the confusion matrix, expressed as proportions of the population, for ith class for j number of soil classes. Values close to 0 represent high agreement in the proportions of coverage for each class while values close to 1 represent high disagreement between class proportions. A represents the amount of disagreement in the spatial allocation of classes between the validation and prediction datasets and is calculated as:

Eq. (4.9) $A = \left[\sum_{i=1}^{j}\min(p_{i+}, p_{+i})\right] - C,$

where values close to 0 represent high agreement while values close to 1 represent high disagreement in spatial allocation for each class.

To assess the predictions made using soil pit data, a 5-fold cross-validation (CV) procedure was applied. Training data points were randomly partitioned into five folds where 80% of the data were used to train the model used for the pit-derived predictions. The remaining 20% of the point data were reserved for validating the pit-derived results. In the CV procedure, each fold was used once for validation, which resulted in five sets of validation metrics which were then aggregated into an overall accuracy value. The CV procedure was necessary in order to make the best use of the limited sample size without the need for additional sampling while insuring independent validation. The polygon-derived results were also validated using the BCSIS point dataset in order to make a fair comparison between the results produced from the two different training datasets. In addition to using the soil pit data for validation, model results were compared to the single-component polygons in order to provide an indication as to how similar the soil pit-derived predictions were to the conventional soil survey maps.

4.3.6. Prediction Uncertainty

The use of ensemble-modeling techniques has the added benefit of being able to estimate the classification uncertainty from the individual models of the ensemble. For each pixel, a prediction of a soil Great Group was made by each member of the ensemble-model and the number of votes for which a Great Group was predicted by the member models was totalled into vote count surfaces that represent certainty for each class. Because each model was reiterated 100 times, values close to 100 would represent higher certainty of a particular Great Group occurring at that pixel. To represent overall uncertainty, ignorance uncertainty measures how evenly distributed the vote counts are for each class. When vote counts are more evenly distributed across the classes, the uncertainty is greater (Leung et al., 1993; Zhu 1997). In order to measure ignorance uncertainty, the information statistic, or entropy measure, *H,* was used to describe the degree to which the members of the ensemble-model concentrate their predictions to a particular class. *H* is calculated as (Zhu 1997):

Eq. (4.10) $$H(x)=\frac{1}{\ln n}\sum_{k=1}^{n} P_k\,(x)\ln P_k\,(x),$$

where P_k is the proportion of instances where pixel x is classified as soil class k and where n is the number of members in the ensemble-model. The values of H range from 0 to 1 with values close to 0 representing low uncertainty and values close to 1 representing high uncertainty in classification. Although the use of entropy is most commonly used in fuzzy inference classification approaches (e.g. Goodchild et al., 1994; Leung et al., 1993; Zhu 1997), its extension for visualizing uncertainty in ensemble-modeling has been shown to be appropriate (Kempen et al., 2009).

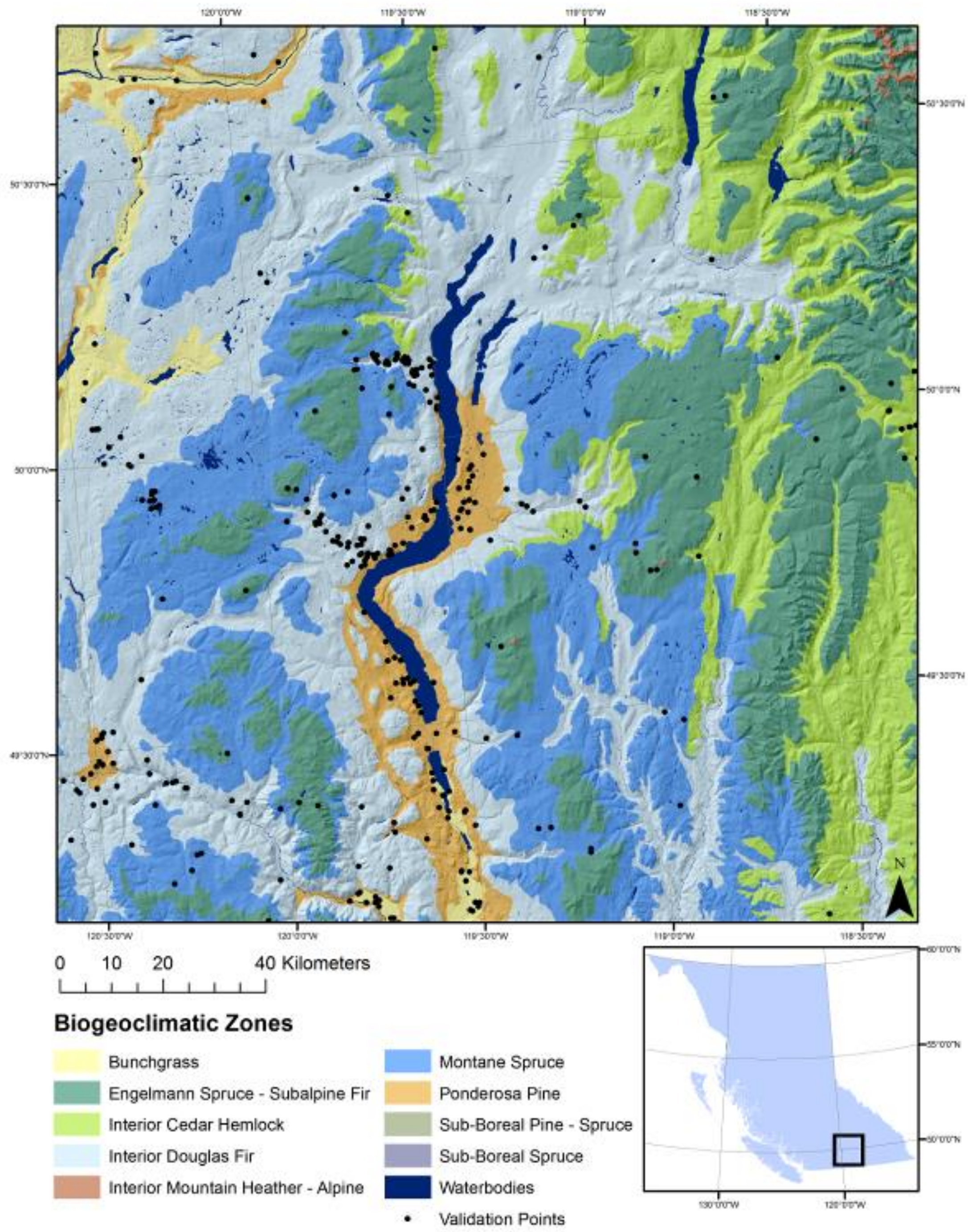


Figure 4.1. **Biogeoclimatic zones of the Okanagan-Kamloops region. Inset: study area in relation to the province of British Columbia, Canada.**

Table 4.1. **Environmental covariates derived from a 100 m spatial resolution DEM and 30 m Landsat imagery.**

Representation	Environmental Covariate
Local Scale Morphometry	Elevation Plan curvature Profile curvature Slope Terrain ruggedness index (Riley et al., 1999) Transformed aspect - eastness Transformed aspect - northness
Landscape Scale Morphometry	Multi-resolution ridge top flatness index (Gallant and Dowling, 2003) Multi-resolution valley bottom flatness index (Gallant and Dowling, 2003) Mid-slope position Normalized height Slope height Valley depth
Hydrologic Characteristics	Slope length factor (Moore et al., 1993) Stream power index (Moore et al., 1991) Topographic wetness index (Beven and Kirkby, 1979)
Landscape Exposure	Sky view factor (Häntzschel et al., 2005) Terrain view (Häntzschel et al., 2005) Visible sky
Climatic Indices	Mean annual precipitation Mean annual temperature Number of frost free days
Vegetative Indices	Enhanced vegetation index (Jiang et al., 2008) Modified soil-adjusted vegetation index (Qi et al., 1994) Normalized difference water index Normalized difference vegetation index Soil-adjusted vegetation index (Huete, 1988)

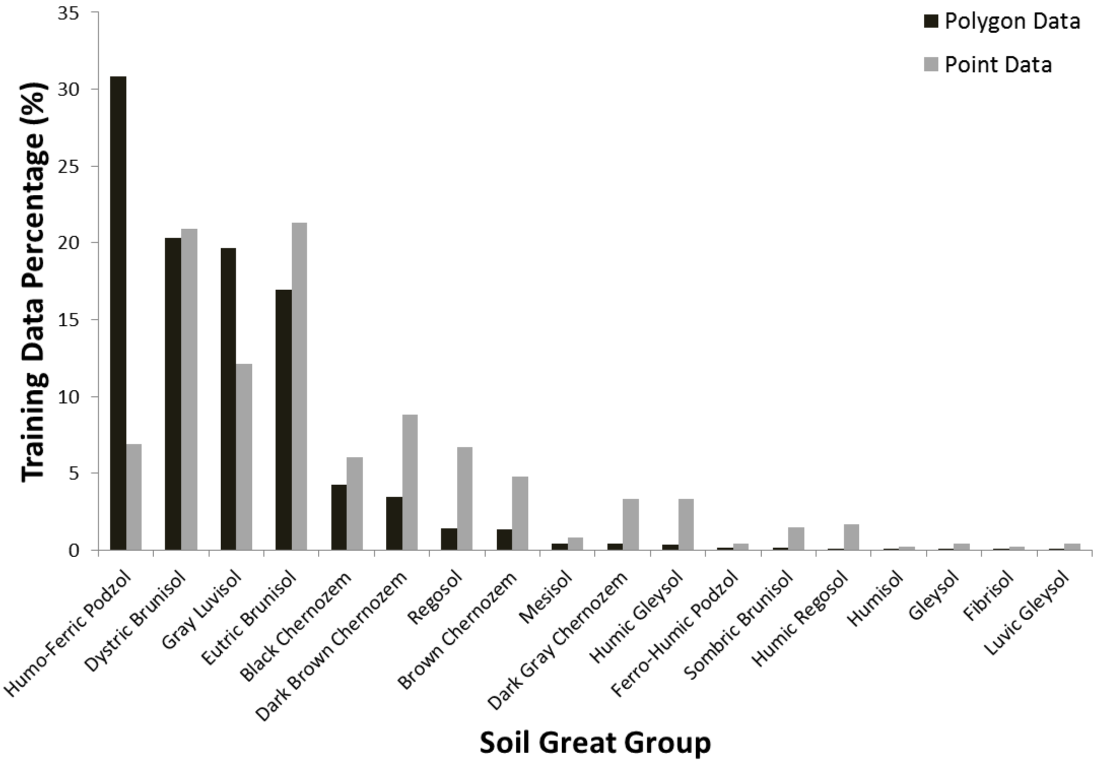


Figure 4.2. **Distribution of training points for soil Great Groups generated from polygon data (n = 129,456) and field collected pits from BCSIS (n = 478).**

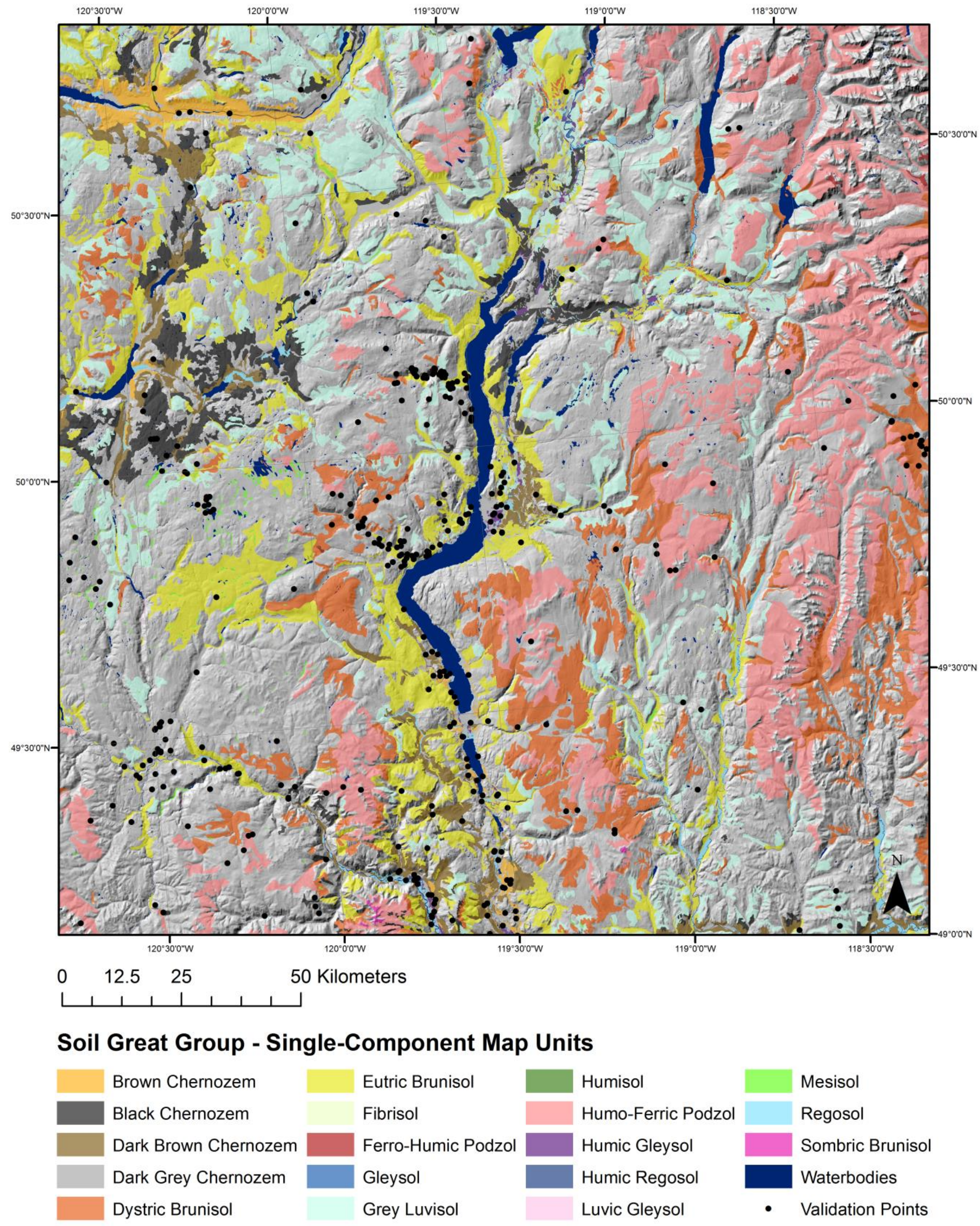


Figure 4.3. Single-component soil Great Group map units from the Okanagan-Kamloops region of British Columbia.

4.4. Results & Discussions

4.4.1. Accuracy Assessment

The mapping accuracies for all models trained with soil pit data and soil survey polygon data are shown in Table 4.2. Accuracies for results using pit-derived training data ranged from 47 to 61%, with the RF model having the highest accuracy; in comparison, results using polygon-derived training data ranged from 50 to 70%, where *k*NN+ had the highest accuracy. The accuracy rates were consistent with our expectation that a larger training dataset would better capture the feature spaces occupied by each soil Great Group and would therefore have higher accuracies in comparison to the results produced using pit-derived training data. The need to adequately represent feature space is especially important for the prediction of minority classes using pit-derived training data, because for soil Great Groups such as Mesisols, Ferro-Humic Podzols, Humisols, Gleysols, Fibrisols, and Luvic Gleysols – soils that are uncommon in drier environments – it was unreasonable to expect them to be predicted well, since the number of training points were often no more than three for each of those classes.

By decomposing error rates into quantity disagreement and allocation disagreement, it was observed that, with the exception of the CART learner, quantity disagreement values were similar regardless of which training dataset was used and ranged between 11 and 17%. When the CART learner was used, quantity disagreement was markedly higher with values of 17% and 27% when comparing pit-derived and polygon-derived training datasets, respectively. The higher disagreement was the result of the CART model not predicting the entire range of different soil Great Groups. Due to the general consistency in quantity disagreement values, the differences between prediction and validation datasets were due to consistently higher allocation disagreements and variability of allocation disagreement amongst the various learners. Finally, it was observed that on average, allocation disagreement was higher when using pit-derived training data, which would indicate that the predictions made using polygon-derived results had a higher spatial accuracy. The higher spatial accuracy may have been the result of the larger training dataset obtained from the polygon data.

In terms of implementing an ensemble-modeling technique, bagging generally benefitted predictions made using the pit-derived training dataset by decreasing the allocation disagreement, while accuracies using the polygon-derived training dataset were similar with the exception of the CART model. The reason for the increased accuracy from the ensemble-models, when using pit-derived training data, was likely due to the improved stability of the models when training with a small number of training points. With a large number of training points, when considering the polygon-derived training data, the models became less sensitive to perturbations due to the size of the training data where the folds in the CV procedure were all similar in terms of the feature space occupied by each fold. As a result, the variance of single-learner models would be initially low and the application of ensemble learning would not further decrease the model variance when that large training dataset was used (Brain and Webb, 1999). In contrast, the pit-derived training dataset was small and, as a result, the effects of decreasing model variance using ensemble-learning techniques became much more apparent as demonstrated by a 2 to 11% increase in accuracy.

When considering the accuracies of the results produced using both training datasets, it was determined that the RF model performed the best overall (Figs. 4.4 and 4.5). Even though *k*NN+ had the highest accuracy when using polygon-derived training data, RF performed comparatively well with an accuracy of 68%. Using a confusion matrix for the RF results (Table 4.3), it was determined that Regosolic soils were predicted particularly poorly with accuracies of 38% and 19% for polygon-derived and pit-derived predictions, respectively. The poor prediction of Regosolic soils was likely due to the models being unable to capture the environmental disturbances and processes that caused the formation of these poorly developed soils. Gray Luvisols were another soil that was often misclassified as Dark Brown Chernozems or Eutric Brunisols in both prediction results, where we suspected that an inclusion of geological information and better precipitation data would have improved the separation of these classes in feature space rather than the reliance on topographic data.

The effectiveness of RF was also corroborated by model comparison studies such as Brungard et al. (2015), which used pit-derived training data to predict soil classes over three semi-arid study areas located in Utah, New Mexico, and Wyoming;

similarly, Taghizadeh-Mehrjardi et al. (2015) showed RF to perform well despite not being ranked the best model. In terms of using polygon-derived training data, Heung et al. (2016) showed the effectiveness of RF for mapping soil Great Groups for the Lower Fraser Valley of British Columbia. Overall, these studies seem to collectively suggest that RF is a consistently effective classifier regardless of whether or not the training data are derived from soil pit data or polygon data.

4.4.2. Comparison with Legacy Soil Surveys

In addition to assessing the accuracy of the polygon-derived predictions, the single-component polygons from the original soil survey were also validated. The single-component polygons had an accuracy of 65% based on the 285 soil pits that intersected with the polygons (Table 4.4). To investigate the influence of map scale on the soil survey accuracy, large-scale surveys had a higher accuracy of 67% for the 1:20,000 and 1:50,000 scales in comparison to the reconnaissance mapping created at a 1:125,000 scale where the accuracy was 48%. The discrepancy in accuracies may be partly due to the small number (*n* = 25) of points located on the 1:125,000 polygons; however, the more likely reason is due to the loss of detail and the greater amount of inclusions when mapping at smaller scales where the subtle variations in environmental covariates become generalized. The average overall accuracies for all models using either training dataset were similar (64%) to the accuracy of the single-component map units (66%). Most noticeable, however, was that even though single-component map units were used to derive one of the training datasets, the resulting predictions using the CART+, RF, *k*NN, and *k*NN+ models all had higher accuracies than the original soil survey. Improvements in accuracy were most noticeable for the CART+ and RF models in areas for which the 1:125,000 polygons were mapped with an increase by up to 9%. When predictions were made using pit-derived training data, their accuracies were not higher than the accuracy of the single-component polygons that were mapped at a 1:125,000 scale. These results were surprising because, similar to Collard et al. (2014) that used an existing soil survey at a 1:250,000 scale to calibrate a model, their predictions had higher accuracies in comparison to the original soil map using MLR, RF, and classification tree models.

Another interesting finding comes from the consistency metrics obtained by comparing the results produced by soil pit data and the single-component polygons of the soil survey where the RF model had an agreement of *C* = 55% in spite of the low sample density of 0.01 samples/km^2 (Table 4.2). When the accuracies of the pit-derived predictions that coincided with single-component polygons and the accuracies of the polygons were compared, the results seemed to suggest that DSM approaches using models such as CART+, RF, *k*NN, and *k*NN+ models could potentially produce outputs that are more accurate than maps produced using conventional mapping approaches when using the same point dataset (Table 4.4). In Kempen et al. (2012), a conventional soil survey map was compared to a DSM where it was observed that there was a 69% discrepancy between the two maps; furthermore, they also observed an increased accuracy using DSM methods where the accuracy increased as sampling density increased. These findings are promising because with higher sample densities and a sample design that effectively captures the feature space of the environment, such as a conditioned Latin hypercube design (e.g. Brungard et al., 2015; Minasny and McBratney, 2006; Rad et al., 2014), pit-derived training data may potentially produce maps that are similar or better than conventional soil survey maps without the reliance on expert knowledge and manual delineation of map units – both of which are costly and subjective processes. The advantage of an automated mapping approach using a machine-learner is that the subjective process of correlating soil types to the environment becomes an objective process where the soil-environmental relationships may then be quantified consistently; whereas in the case of conventional soil surveys, the subjective mental models that describe the soil-environmental relationships are often lost or not recorded.

Although this study has found similarities to Collard et al. (2014) and Kempen et al. (2012), the extent of these similarities and our analyses were limited to where single-component polygons have been mapped in the conventional soil surveys.

Table 4.2. **Classification accuracy and consistency metrics using overall agreement (C), quantity disagreement (Q), and allocation disagreement (A). Average accuracy calculated from 5-fold cross-validation procedure using n = 478 sample points. Consistency metrics calculated based on single-component soil survey polygons for pit-derived results.**

	Accuracy						**Consistency with Soil Survey**		
	Pit-Derived			**Polygon-Derived**			**Pit-Derived**		
	C	***Q***	***A***	***C***	***Q***	***A***	***C***	***Q***	***A***
	(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)
CART	50	17	33	50	27	22	45	22	32
CART + Bagging	60	12	28	67	15	18	48	9	43
Random Forest	61	15	24	68	14	18	55	8	37
Logistic Model Tree	50	7	43	62	11	27	40	17	44
Logistic Model Tree + Bagging	58	12	30	61	12	26	41	16	43
Multinomial Logistic Regression	47	10	43	53	14	33	41	14	46
Multinomial Logistic Regression + Bagging	54	14	32	56	13	31	40	16	44
k Nearest Neighbor	50	11	38	68	12	20	43	11	46
k Nearest Neighbor + Bagging	52	12	36	70	14	16	45	11	44

Table 4.3. **Confusion matrix between 478 observation points and predicted soil Great Group using a Random Forest model. Bold values represent the diagonal of the confusion matrix and the number of correctly classified pixels for each class.**

		Actual Great Group																	
		BC	**BLC**	**DBC**	**DGC**	**DB**	**EB**	**F**	**FHP**	**G**	**GL**	**H**	**HFP**	**HG**	**HR**	**LG**	**M**	**R**	**SB**	**Total**
Predicted Great Group: Polygon-Derived Prediction	**BC**	**18**	0	2	0	0	4	0	0	0	0	1	0	0	0	0	1	0	0	26
	BLC	1	**18**	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	21
	DBC	0	2	**31**	2	0	0	0	0	0	2	0	0	2	0	0	0	0	0	39
	DGC	0	0	0	**1**	0	0	0	0	0	1	0	0	0	0	0	1	0	0	3
	DB	2	3	2	0	**88**	7	1	0	1	6	0	3	2	0	0	0	6	2	123
	EB	2	3	3	8	4	**84**	0	0	0	8	0	0	2	2	2	0	9	0	127
	F	0	0	0	0	1	0	**0**	0	0	0	0	0	0	0	0	0	0	0	1
	FHP	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0
	G	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0
	GL	0	3	1	4	3	3	0	0	0	**33**	0	1	0	0	0	0	2	1	51
	H	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0
	HFP	0	0	0	0	6	1	0	1	0	5	0	**28**	1	1	0	2	1	4	50
	HG	0	0	0	0	0	0	0	0	1	0	0	0	**7**	1	0	0	2	0	11
	HR	0	0	0	1	0	0	0	0	0	0	0	0	0	**3**	0	0	0	0	4
	LG	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0
	M	0	0	0	0	1	0	0	0	0	2	0	0	0	0	0	**0**	0	0	3
	R	0	0	1	0	0	3	0	0	0	0	0	0	1	1	0	0	**12**	0	18
	SB	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	**0**	1
	Total	23	29	42	16	103	102	1	1	2	57	1	33	15	8	2	4	32	7	

Legend: Brown Chernozem (BC); Black Chernozem (BLC); Dark Brown Chernozem (DBC); Dark Gray Chernozem (DGC), Dystric Brunisol (DB); Eutric Brunisol (EB); Fibrisol (F); Ferro-Humic Podzol (FHP); Gleysol (G); Gray Luvisol (GL); Humisol (H); Humo-Ferric Podzol (HFP); Humic Gleysol (HG) Humic Regosol (HR); Luvic Gleysol (LG); Mesisol (M); Regosol (R); Sombric Brunisol (SB).

Table 4.3 (cont.)

Predicted Great Group: Pit-Derived Prediction	Actual Great Group																		
	BC	**BLC**	**DBC**	**DGC**	**DB**	**EB**	**F**	**FHP**	**G**	**GL**	**H**	**HFP**	**HG**	**HR**	**LG**	**M**	**R**	**SB**	**Total**
BC	**10**	0	3	0	0	1	0	0	0	0	0	0	0	1	0	0	1	0	16
BLC	0	**20**	4	2	0	1	0	0	0	0	0	0	0	0	0	1	1	0	29
DBC	4	3	**21**	1	0	2	0	0	0	4	0	0	1	1	0	0	1	0	38
DGC	0	0	0	**7**	0	2	0	0	0	0	0	0	0	0	0	0	1	0	10
DB	1	3	1	0	**79**	5	1	0	1	14	0	3	2	1	1	0	5	3	120
EB	6	2	12	3	9	**82**	0	0	0	10	0	0	4	0	0	0	11	0	139
F	0	0	0	0	1	0	**0**	0	0	0	0	0	0	0	0	0	0	0	1
FHP	0	0	0	0	0	0	0	**0**	0	1	0	0	0	0	0	0	0	0	1
G	0	0	0	0	0	0	0	0	**1**	0	0	0	0	0	0	0	1	0	2
GL	2	1	0	2	8	4	0	0	0	**25**	0	0	0	0	0	0	2	0	44
H	0	0	0	0	0	0	0	0	0	0	**1**	0	0	0	0	0	0	0	1
HFP	0	0	0	0	6	1	0	1	0	3	0	**30**	1	1	0	2	1	3	49
HG	0	0	0	0	0	2	0	0	0	0	0	0	**6**	1	0	0	1	0	10
HR	0	0	0	1	0	0	0	0	0	0	0	0	1	**2**	0	0	1	0	5
LG	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**1**	0	0	0	1
M	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	**1**	0	0	2
R	0	0	1	0	0	1	0	0	0	0	0	0	0	1	0	0	**6**	0	9
SB	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**1**	1
Total	23	29	42	16	103	102	1	1	2	57	1	33	15	8	2	4	32	7	

Legend: Brown Chernozem (BC); Black Chernozem (BLC); Dark Brown Chernozem (DBC); Dark Gray Chernozem (DGC), Dystric Brunisol (DB); Eutric Brunisol (EB); Fibrisol (F); Ferro-Humic Podzol (FHP); Gleysol (G); Gray Luvisol (GL); Humisol (H); Humo-Ferric Podzol (HFP); Humic Gleysol (HG) Humic Regosol (HR); Luvic Gleysol (LG); Mesisol (M); Regosol (R); Sombric Brunisol (SB).

Table 4.4. Classification accuracy of soil Great Groups using soil pit observations coinciding with single-component map units and separated by map scale for polygon and pit-derived predictions and soil survey.

Polygon-Derived Prediction Accuracy (%)

Map Scale	CART	CART+	RF	LMT	LMT+	MLR	MLR+	*k*NN	*k*NN+	Average
20,000	46	70	73	66	66	52	58	71	70	63
50,000	58	72	73	66	65	55	57	76	76	67
125,000	43	57	57	43	43	48	57	52	52	50
Overall	53	69	71	64	63	53	57	72	72	64

Pit-Derived Prediction Accuracy (%)

Map Scale	CART	CART+	RF	LMT	LMT+	MLR	MLR+	*k*NN	*k*NN+	Average
20,000	46	70	73	66	66	52	58	71	70	63
50,000	58	74	75	68	66	56	59	77	77	68
125,000	36	48	48	36	36	40	48	44	44	42
Overall	53	71	72	65	63	54	58	72	72	64

Single-Component Accuracy

	Correct	Total	Accuracy
	(*n*)	(*n*)	(%)
SCALE			
20,000	53	79	67
50,000	122	181	67
125,000	12	25	48
Overall	187	285	66

Legend: Classification and regression tree (CART); CART with bagging (CART+); Random Forest (RF); logistic model tree (LMT); LMT with bagging (LMT+); multinomial logistic regression (MLR); MLR with bagging (MLR+); *k*-nearest neighbors (*k*NN); *k*NN with bagging (*k*NN+).

Table 4.5. Map comparison confusion matrix between polygon-derived and pit-derived predictions for soil Great Groups using a Random Forest model. Bold values represent the diagonal of the confusion matrix and the number of correctly classified pixels for each class.

		Polygon-Derived Predictions (%)																		
		BC	**BLC**	**DBC**	**DGC**	**DB**	**EB**	**F**	**FHP**	**G**	**GL**	**H**	**HFP**	**HG**	**HR**	**LG**	**M**	**R**	**SB**	**SUM**
Pit-Derived Predictions (%)	**BC**	**0.20**	0.00	0.06	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.29
	BLC	0.03	**1.66**	0.54	0.02	0.19	0.49	0.00	0.00	0.00	0.32	0.00	0.01	0.00	0.00	0.00	0.01	0.01	0.00	3.28
	DBC	0.25	0.25	**0.96**	0.02	0.00	0.34	0.00	0.00	0.00	0.04	0.00	0.00	0.01	0.00	0.00	0.00	0.05	0.00	1.93
	DGC	0.02	0.13	0.23	**0.01**	0.02	0.16	0.00	0.00	0.00	0.09	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.66
	DB	0.00	0.00	0.00	0.00	**8.81**	0.73	0.00	0.00	0.00	4.04	0.00	5.60	0.00	0.00	0.00	0.07	0.29	0.00	19.55
	EB	0.28	0.69	0.59	0.07	1.45	**9.57**	0.00	0.00	0.00	2.54	0.01	0.14	0.07	0.00	0.00	0.00	1.03	0.00	16.44
	F	0.00	0.00	0.00	0.00	0.00	0.00	**0.00**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	FHP	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**0.00**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	G	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**0.00**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	GL	0.01	0.23	0.06	0.01	6.96	4.52	0.02	0.00	0.00	**14.39**	0.01	1.95	0.00	0.00	0.00	0.10	0.06	0.00	28.31
	H	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**0.00**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	HFP	0.00	0.02	0.00	0.00	1.81	0.26	0.00	0.01	0.00	0.86	0.00	**25.28**	0.00	0.00	0.00	0.04	0.00	0.09	28.36
	HG	0.02	0.00	0.03	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	**0.06**	0.00	0.00	0.00	0.19	0.00	0.31
	HR	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	**0.00**	0.00	0.00	0.01	0.00	0.03
	LG	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**0.00**	0.00	0.00	0.00	0.00
	M	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	**0.00**	0.00	0.00	0.02
	R	0.01	0.01	0.02	0.00	0.19	0.05	0.00	0.00	0.00	0.14	0.00	0.01	0.03	0.01	0.00	0.00	**0.20**	0.00	0.67
	SB	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.13	0.00	0.00	0.00	0.00	0.00	**0.00**	0.13
	SUM	0.81	2.99	2.49	0.12	19.44	16.15	0.02	0.01	0.00	22.41	0.02	33.14	0.17	0.02	0.00	0.23	1.86	0.09	

Legend: Brown Chernozem (BC); Black Chernozem (BLC); Dark Brown Chernozem (DBC); Dark Gray Chernozem (DGC), Dystric Brunisol (DB); Eutric Brunisol (EB); Fibrisol (F); Ferro-Humic Podzol (FHP); Gleysol (G); Gray Luvisol (GL); Humisol (H); Humo-Ferric Podzol (HFP); Humic Gleysol (HG) Humic Regosol (HR); Luvic Gleysol (LG); Mesisol (M); Regosol (R); Sombric Brunisol (SB).

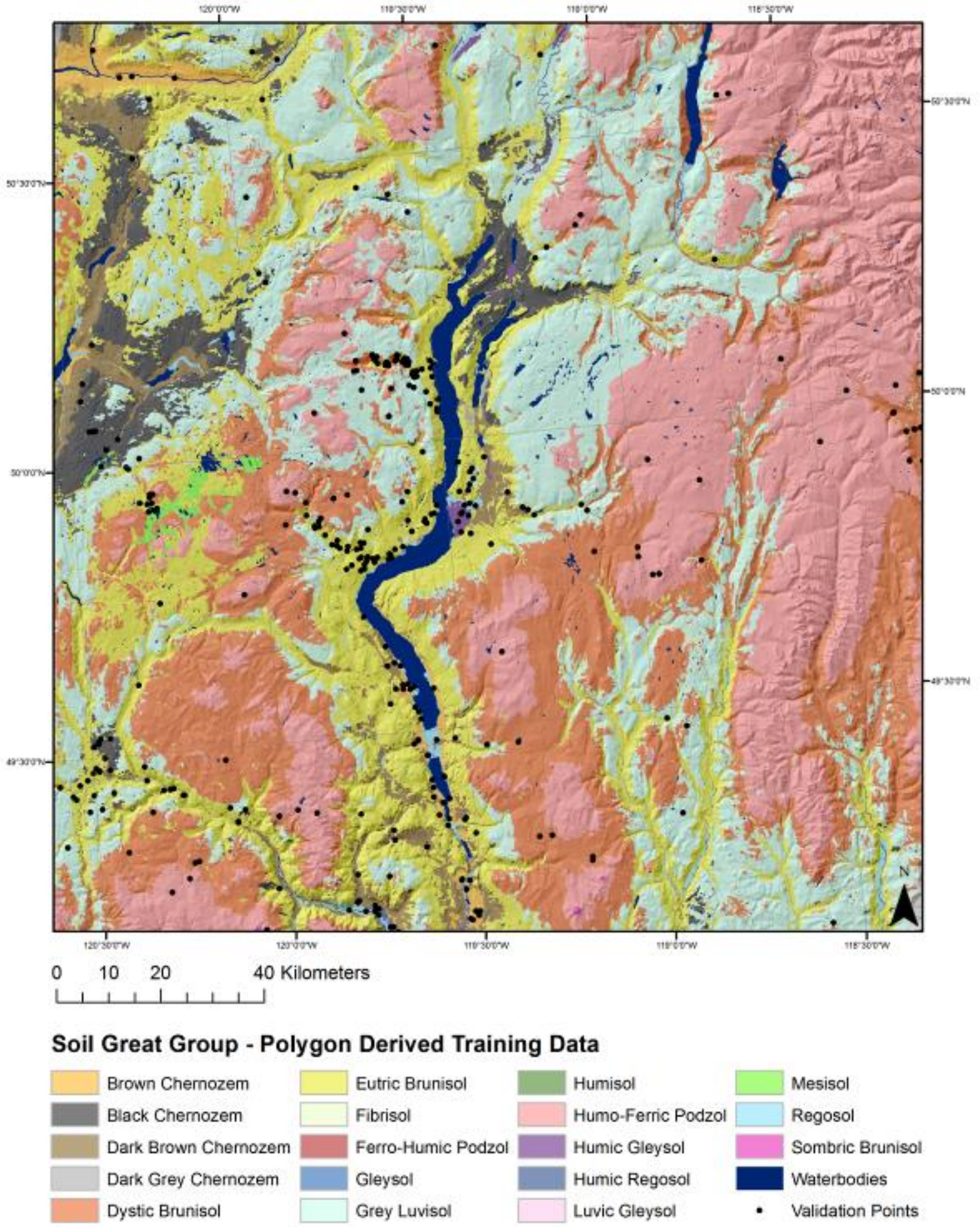


Figure 4.4. **Soil Great Group map using a Random Forest classifier with polygon-derived training data at a 100 m spatial resolution. Map is shown with underlying hill-shade and overlying validation points for the Okanagan-Kamloops region of British Columbia.**

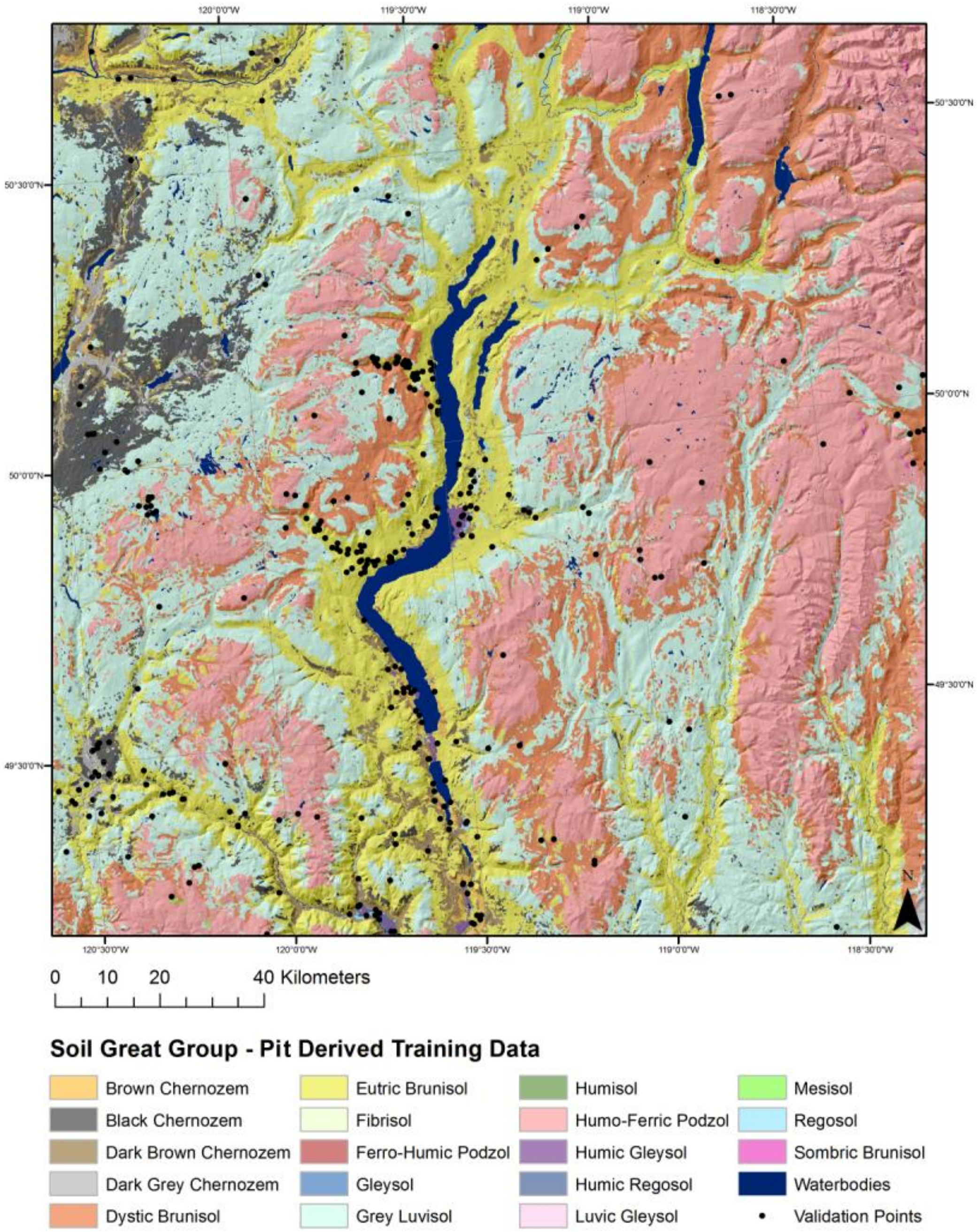


Figure 4.5. Soil Great Group map using a Random Forest classifier with pit-derived training data at a 100 m spatial resolution. Map is shown with underlying hill-shade and overlying validation points for the Okanagan-Kamloops region of British Columbia.

4.4.3. Prediction Uncertainty

Because the RF model resulted in the best overall performance when using pit-derived and polygon-derived training data, combined, uncertainty analysis was done on the RF predictions. In Figs 4.6 and 4.7, the vote count surfaces of the most frequently occurring Great Groups are represented. When comparing the vote count surfaces produced from polygon-derived training data (Fig. 4.6) and the pit-derived training data (Fig. 4.7), the general patterns were similar; however, there was far less certainty in predictions when using the pit-derived training data as evidenced by the lower vote counts.

The overall prediction uncertainty was represented using ignorance uncertainty surfaces in Figures 4.8 and 4.9. When using the polygon-derived training data (Fig. 4.8), values of uncertainty ranged from 0 to 0.78 with a mean of 0.31 and a standard deviation of 0.15. Ignorance uncertainty patterns showed large areas of low uncertainty at high elevations of the eastern region of the study area as a result of the mapping of large polygons representing Humo-Ferric Podzols. In general, the uncertainty was low for regions with high relief or along the side of hillslopes where the topographic covariates would likely have the greatest separation in the feature spaces of each Great Group. This was in contrast to the valley bottoms where the topography is flat and homogenous and would therefore lead to uncertainty because the soil Great Groups located in those regions (e.g. Chernozems, Gleysols, and Regosols) all occupy similar feature spaces. It is also necessary to point out that the uncertainties from the polygon-derived predictions are underestimated not just due to the large number of points used in that training dataset, but additionally that the training data was derived from a generalized representation of soil patterns in reality. Consequently, the localized and subtle variations in topography were not captured in the conventional mapping process – especially when the maps were produced at small-scales.

The ignorance uncertainty surface produced from the results using pit-derived training data (Fig. 4.9) showed noticeably higher uncertainty with the range in values from 0 to 0.84, a mean of 0.57, and a standard deviation of 0.11. The spatial patterns of the uncertainty values were unclear; however, values appeared lower along the sides of hillslopes – similar to Figure 4.8. Although there appeared to be lower uncertainty in areas that had clusters of sample points, especially along Lake Okanagan, the relationship between the intensity of sampling and uncertainty was not clear for the entire study area. Again, a higher sample density with a design that was optimized for capturing the feature space of the entire study area may have provided a solution for decreasing model uncertainty and increasing overall accuracy.

4.4.4. Visual Assessment

Visual assessment and comparisons between the results produced from both training datasets were performed on predictions made using RF due to the model having the highest combined accuracy. In both cases, the distributions of soil Great Groups were consistent with their theoretical distributions within the landscape. Visually, the gradational transitions between soil Great Groups were most noticeable from the vote count surfaces (Figs. 4.6 and 4.7) that showed the dominance of Chernozemic soils at low elevations and in valley bottoms followed by a transition to Brunisols at mid-elevations, and finally Gray Luvisols and Podzols at the highest elevations where the landscape receives greater precipitation and was forest covered.

Where the grassland Chernozemic soils occur, predictions made by both training datasets showed the occurrence of Brown Chernozems at the lowest elevations (northwest corner of the study area) and regions with higher temperatures (south central region of the study area). As the temperature decreases at higher elevations, the Brown Chernozems transition to Dark Brown Chernozems and followed by Black Chernozems. The transition of Chernozemic soils observed was likely the result of decreased decomposition rates with decreasing temperatures, which would then cause the darkening of the soil due to organic matter accumulation (Pennock et al., 2011; Van Ryswyk et al., 1966). As the grasslands transition to a forested cover, Brunisolic soils became more frequently predicted with Eutric Brunisols most commonly found on lower slopes and valley bottoms where they may occur as complexes with Chernozemic soils. At mid-elevations and mid-slopes, Dystric Brunisols were predicted more prevalently where precipitation is higher and thus causing greater leaching of base cations and therefore decreasing the pH (Smith et al., 2011). The accumulation of acidic forest litter and sufficient precipitation results in the enhanced formation of leached horizons due to the process of lessivage and the formation of Luvisolic soils at the higher elevations of the landscape (Lavkulich and Arocena, 2011). Finally, with increased precipitation at the highest elevations, Humo-Ferric Podzols were the most frequently predicted soil Great Group because with sufficient precipitation, both cations and clays are depleted which would allow for the process of podzolization to occur.

4.4.5. Map Comparison

Although the results produced from the training datasets both showed general soil patterns that were consistent with the literature, the predicted extents of each Great Group were noticeably different. A quantitative map comparison between Figs. 4.4 and 4.5 resulted in only a 60% match between the predictions in spite of their similar accuracies (Table 4.2). To further examine these differences a confusion matrix was created for a quantitative comparison between the RF results (Table 4.5). Between the two maps, the greatest confusion occurred between the Dystric Brunisolic and Gray Luvisolic soils, which accounted for 10% of the dissimilarity between the predictions. Typically, these differences occurred on mid-elevation regions of the study area where Gray Luvisols were more prominently predicted in the polygon-derived results whereas the pit-derived results had Dystric and Eutric Brunisols predicted. It is reasonable to expect these differences because those soil Great Groups are all common in forested landscapes and the differences between them are influenced more so by precipitation and soil pH rather than topography. At high elevations, polygon-derived predictions had a greater distribution of Humo-Ferric Podzols, where 5.60% and 1.95% of the study area were classified as Dystric Brunisols and Gray Luvisols, respectively, in the pit-derived predictions. Similar to the mid-elevation regions, the differences in these soils are caused by precipitation and less so by topography. It is also possible that some of these differences may have to do with the difference in mapping scale that the polygon-derived training data were derived from, where at mid- and high elevation and forested terrain, soil surveys were mapped at a 1:125,000 scale, whereas, agriculturally intensive regions and low elevations were mapped at a 1:20,000 scale.

Major differences were also observable for the valleys of the study area, and in particular, the distribution of the Chernozemic Great Groups. When using the pit-derived training data, the resulting predictions showed a greater occurrence of Black Chernozems with minor instances of the other types of Chernozems. In contrast, polygon-derived training data results showed a greater occurrence of Brown and Dark Brown Chernozems in the valley bottoms of the north-western region of the study area. North of Lake Okanagan, the predictions using pit-derived training data failed to detect the occurrence of Black Chernozems, which was possibly due to the lack of sample points that identified the soil Great Group in that area. Again, the differences between the results were likely due to the pit-derived training data capturing a narrower range of covariate values and therefore limiting the prediction of the Brown and Dark Brown Chernozems.

4.4.6. General Discussion

Whereas this study only sampled from the single-component map units of a soil survey, it was one of the many different approaches for sampling soil surveys that may be found in the DSM literature; for instance, in Bui and Moran (2003), the polygons were assigned the dominant soil type. In another example such as the DSMART algorithm used in the POLARIS project (Chaney et al., 2016), all polygons were sampled; however, the SSURGO dataset used for that study consisted of primarily multi-component map units where component labels were assigned using a random allocation that was weighted by the estimated proportion of occurrence from the components in each map unit. In the testing of the DSMART algorithm, complete random allocation and targeted allocation approaches were also tested on multi-component polygons (Odgers et al., 2014). Whereas, similar to this study, the approach of selecting only single-component polygons has been used in Smith et al. (2016). The rationale for sampling only single-component polygons in this study stemmed from the intention to minimize attribute/class noise and uncertainty in the training dataset, which may potentially be added through weighted-random allocation or dominant-soil allocation. Although direct comparisons were not made between these different class-allocation approaches, machine-learning literature (e.g. Van Hulse and Khoshgoftaar, 2009) has shown that through the artificial introduction of noise into a training dataset for imbalanced datasets, the impacts of noise drastically varied amongst different machine-learning approaches and their prediction accuracies. Furthermore, the potential introduction of noise through the sampling of multi-component polygons could cause the over-fitting of the models. Given that single-component map units were quite extensively mapped for this study area (53.2%), their usage seemed appropriate in order to attempt to maximize the separation amongst the classes within feature space. In comparison, for situations where single-component map units are less extensive, alternative methods would need to be considered.

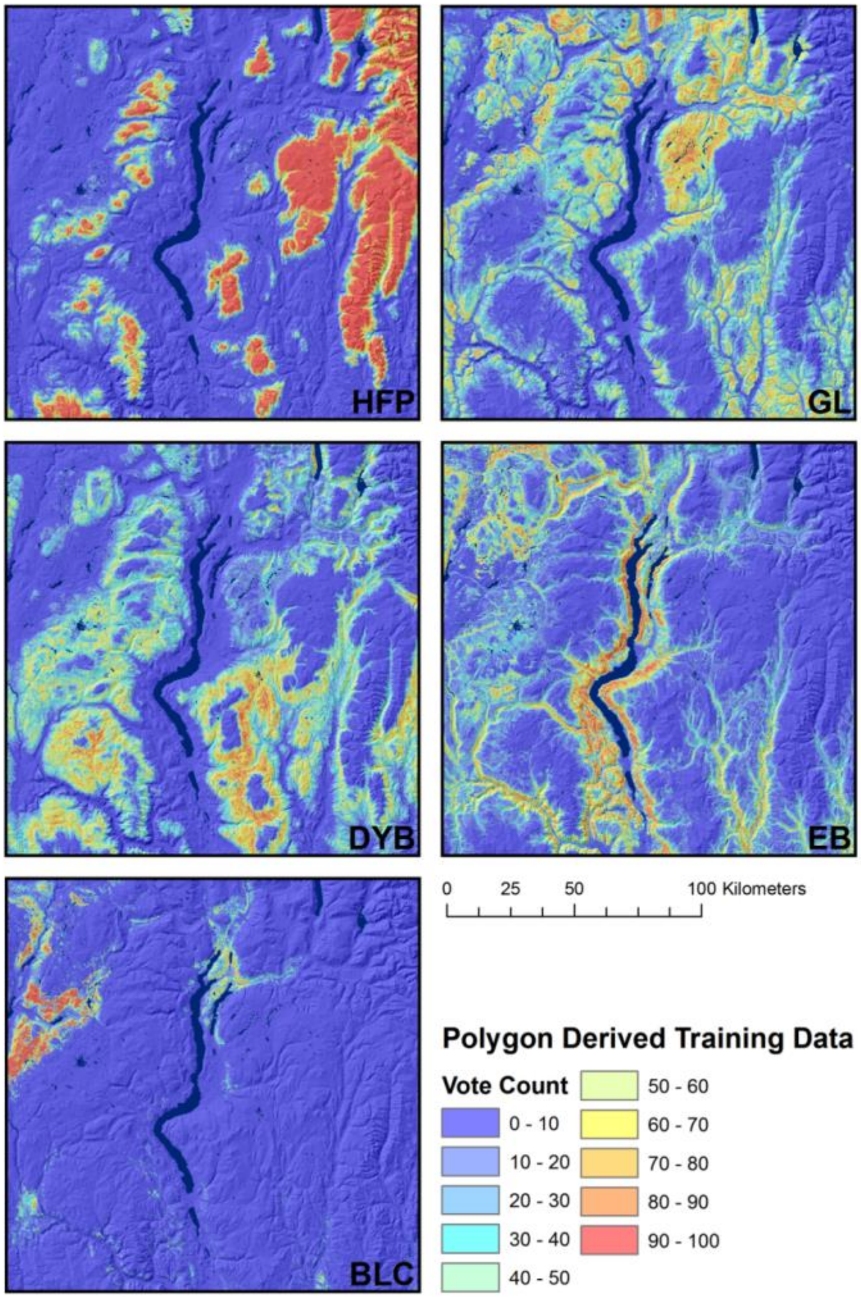


Figure 4.6. **Vote count surfaces based on 100 decision trees of the Random Forest model using polygon-derived training data at a 100 m spatial resolution for the Okanagan-Kamloops region of British Columbia. Most frequently occurring soil Great Groups include Humo-Ferric Podzols (HFP), Gray Luvisols (GL), Dystric Brunisols (DYB), Eutric Brunisols (EB), and Black Chernozems (BLC).**

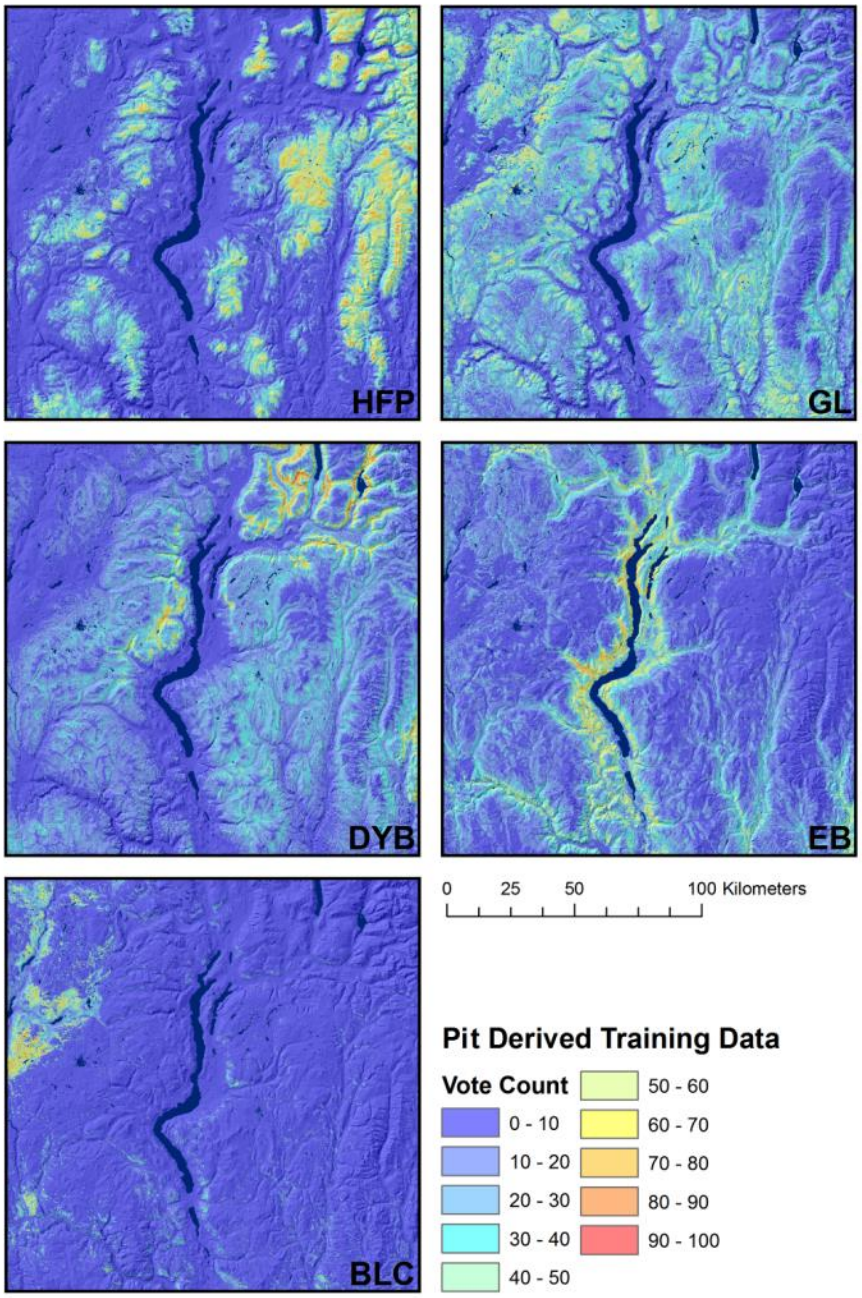


Figure 4.7. Vote count surfaces based on 100 decision trees of the Random Forest model using pit-derived training data at a 100 m spatial resolution for the Okanagan-Kamloops region of British Columbia. Most frequently occurring soil Great Groups include Humo-Ferric Podzols (HFP), Gray Luvisols (GL), Dystric Brunisols (DYB), Eutric Brunisols (EB), and Black Chernozems (BLC).

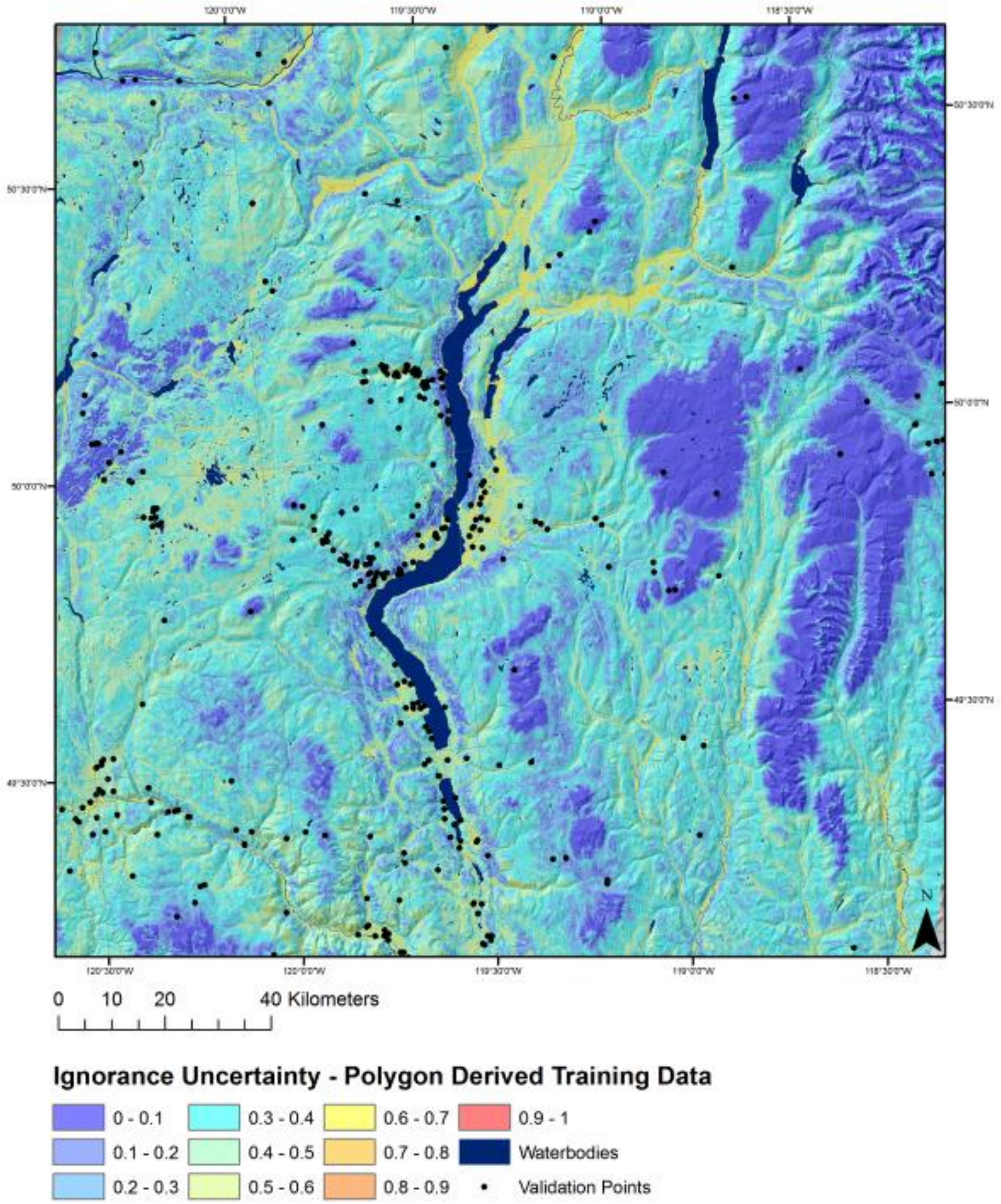


Figure 4.8. Ignorance uncertainty surface based on Random Forest model using polygon-derived training data produced at a 100 m spatial resolution for the Okanagan-Kamloops region of British Columbia.

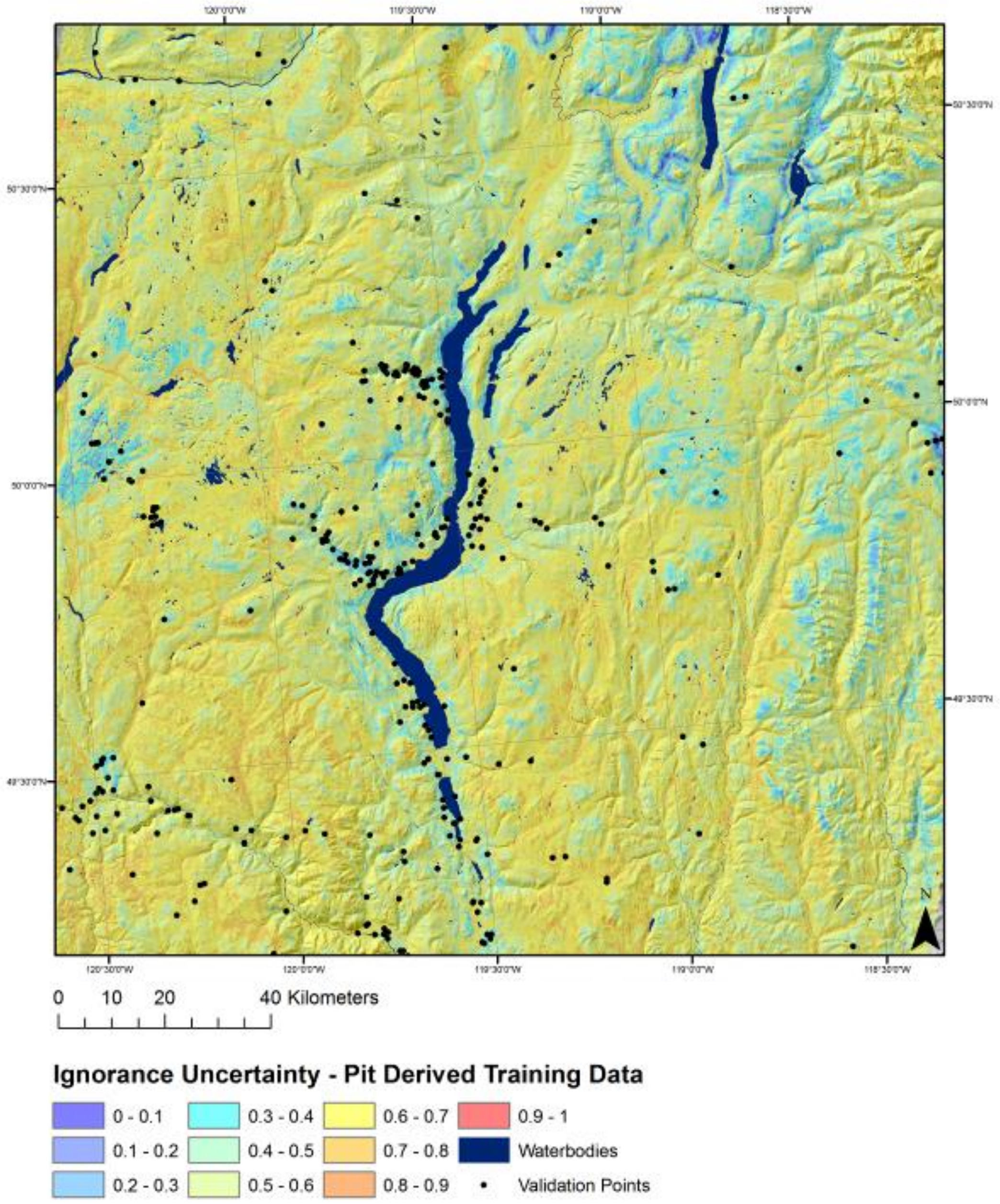


Figure 4.9. **Ignorance uncertainty surface based on Random Forest model using pit-derived training data produced at a 100 m spatial resolution for the Okanagan-Kamloops region of British Columbia.**

4.5. Conclusions

The primary objective of this study was to compare the development of training data derived from legacy soil pit data and soil survey polygons. Secondary objectives included the comparison of 9 machine-learning techniques; and the comparison of single-model and ensemble-model learners. In addition, ensemble-modeling techniques were used to visualize prediction uncertainty. Key findings are summarized as follows:

1. Accuracies were consistently higher for predictions using polygon-derived training data - regardless of the model used. This was likely the result of the higher density of training data, which improved the representation of the feature spaces for the various soil Great Groups. In spite of the differences, the RF model performed reasonably well regardless of the training dataset used where it was ranked the highest, in terms of accuracy, when pit-derived training data were used while having an accuracy that was comparable to the kNN+ model when polygon-derived training data were used.

2. Although the RF models had similar accuracy rates when using either training datasets, there were major differences between the results where both maps shared a 60% similarity. Based on a visual assessment of the results, predictions of the distribution of soil Great Groups were consistent with their theoretical distributions found in the literature; however, the specific extents of each soil Great Group differed considerably.

3. Ensemble-modeling approaches were beneficial when predictions were made using pit-derived training data. Ensemble techniques likely resulted in greater model stability when using small sample sizes - as identified by machine-learning literature. Ensemble techniques were not particularly beneficial when using polygon-derived training data because of the inherent stability of models when using a large number of training data points and thereby decreasing the model variance.

4. When the accuracy of the single-component polygons were compared to the accuracy of the CART+, RF, kNN, and kNN+ predictions made using either training datasets, the DSM approaches were more accurate in comparison to conventional soil mapping approaches. The study suggested that similar to Collard et al. (2014), a conventional soil map may be improved using machine-learning and similar to Kempen et al. (2012), DSM approaches may produce predictions that are similar or more accurate than the maps produced from conventional approaches.

5. The use of ensemble-modeling approaches had the additional benefit in producing raster surfaces of model uncertainty for individual soil Great

Groups as well as overall uncertainty based on ignorance uncertainty, which could be used to aid the visualization of model uncertainty.

This study is the first to illustrate the differences between using legacy soil pit data and soil survey polygons as training data for the prediction of soil classes. Although additional comparative studies on different environments may be required to make generalized statements about these differences, this study provides a better insight into how training data could be used in future DSM studies.

4.6. Acknowledgements

The authors are thankful for support from a Canadian Graduate Scholarship award by the Natural Sciences and Engineering Research Council of Canada granted to B. Heung.

4.7. References

BC Ministry of Agriculture and BC Ministry of Environment, 2016. BC Soil Information Finder Tool. (Available at http://www2.gov.bc.ca/gov/content/environment/air-land-water/land/soil-information-finder, verified 14 November, 2016).

Beven, K.J., Kirkby, M.J., 1979. A physically based, variable contributing area model of basin hydrology. Hydrologic Sciences Bulletin 24, 43-69

Brain, D., Webb, G.I., 1999. On the effect of data set size on bias and variance in classification learning. *In* Proceedings of the 4th Australian Knowledge Acquisition Workshop, Sydney, NSW, pp. 117-128.

Breiman, L., 1996. Bagging predictors. Machine Learning 24, 123-140.

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

Breiman, L., Friedman, J., Stone, C.J., Olshen, R.A., 1984. Classification and Regression Trees. CRC Press LLC, Boca Raton, FL.

Brungard, C.W., Boettinger, J.L, Duniway, M.C., Wills, S.A., Edwards, T.C., 2015. Machine learning for predicting soil classes in three semi-arid landscapes. Geoderma 239-240, 68-83.

Bui, E.N., Henderson, B.L., Viergever, K., 2006. Knowledge discovery from models of soil properties developed through data mining. Ecological Modelling 191, 431-446

Bui, E.N., Moran, C.J., 2003. A strategy to fill gaps in soil survey over large spatial extents: an example from the Murray-Darling basin of Australia. Geoderma 111, 21-44.

Bulmer, C.E., Schmidt, M.G., Heung, B., Scarpone, C., Zhang, J., Filatow, D., Finvers, M., Smith, C.A.S., 2016. Improved soil mapping in British Columbia, Canada, with legacy soil data and Random Forest. *In* Digital Soil Mapping Across Paradigms, Scales and Boundaries. Springer Environmental Science and Engineering, pp. 291-303.

Chaney, N.W., Wood, E.F., McBratney, A.B., Hempel, J.W., Nauman, T.W., Brungard, C.W., Odgers, N.P., 2016. POLARIS: A 30-meter probabilistic soil series map of the contiguous United States. Geoderma 274, 54-67.

Collard, F., Kempen, B., Heuvelink, G.B.M, Saby, N.P.A., Richer de Forges, A.C., Lehmann, S., Nehlig, P., Arrouays, D., 2014. Refining a reconnaissance soil map by calibrating regression models with data from the same map (Normandy, France). Geoderma Regional 1, 21-30.

Debella-Gilo, M., Etzelmüller, B., 2009. Spatial prediction of soil classes using digital terrain analysis and multinomial logistic regression modeling integrated in GIS: examples from Vestfold County, Norway. Catena 77: 8-18.

Gallant, J.C., Dowling, T.I., 2003. A multiresolution index of valley bottom flatness for mapping depositional areas. Water Resources Research 39, 1347-1359.

Gao, B.C., 1996. NDWI – a normalized difference water index for remote sensing of vegetation liquid water from space. Remote Sensing of Environment 58, 257-266.

Green, A.J., Lord, T.M., 1979. Soils of the Princeton Area of British Columbia. Report No.14. Research Branch Agriculture Canada, Agriculture Canada, Ottawa, ON, Canada.

Grinand, C., Arrouays, D., Laroche, B., Martin, M.P., 2008. Extrapolating regional soil landscapes from an existing soil map: sampling intensity, validation procedures, and integration of spatial context. Geoderma 143, 180-190.

Friedman, J., Hastie, T., Tibshirani, R., 2000. Additive logistic regression: a statistical view of boosting. The Annals of Statistics 38, 337-374.

Gallant, J.C., Dowling, T.I., 2003. A multiresolution index of valley bottom flatness for mapping depositional areas. Water Resources Research 39, 1347-1359.

Goodchild, M.F., Chin-Chang, L., Leung, Y., 1994. Visualizing fuzzy maps. In: Visualization in Geographical Information Systems. John Wiley & Sons, NY, pp. 158-167.

Häntzchel, T., Goldberg, V., Bernhofer, C., 2005. GIS-based regionalization of radiation temperature and coupling measures in complex terrain for low mountain ranges. Meteorological Applications 12, 33-42.

Häring, T., Dietz, F., Osenstetter, S., Koschitzki, T., Schröder, B., 2012. Spatial disaggregation of complex soil map units: a decision-tree based approach in Bavarian forest soils. Geoderma 185-186, 37-47.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed. Springer, New York, NY, 734 pp.

Hengl, T., de Jesus, J.M., MacMillan, R.A., Batjes, N.H., Heuvelink, G.B.M., Riberio, E., Samuel-Rosa, A., Kempen, B., Leenaars, J.G.B., Walsh, M.G., Gonzalez, M.R., 2014. SoilGrids1km – Global soil information based on automated mapping. PLoS ONE 9.

Hengl, T., Heuvelink, G.B.M., Kempen, B., Leenaars, J.G.B., Walsh, M.G., Sheperd, K.D., Sila, A., MacMillan, R.A., Mendes de Jesus, J., Tamene, L., Tondoh, J.E., 2015. Mapping Soil Properties of Africa at 250 m Resolution: Random Forests Significantly Improve Current Predictions PLoS ONE 10.

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.

Heung, B., Ho, H.C., Zhang, J., Knudby, A., Bulmer, C.E., Schmidt, M.G., 2016. An overview and comparison of machine-learning techniques for classification purposes in digital soil mapping. Geoderma 265, 62-77.

Holmes, K.W., Griffin, E.A., Odgers, N.P., 2015. Large-area spatial disaggregation of a mosaic of conventional soil maps: Evaluation over Western Australia. Soil Research 53, 865-880.

Hope, G.D., Lloyd, D.A., Mitchell, W.R., Erickson, W.R., Harper, W.L., Wikeem, B.M., 1991a. Ponderosa Pine Zone. In: Ecosystems of British Columbia, British Columbia Ministry of Forests.

Hope, G.D., Mitchell, W.R., Lloyd, D.A., Erickson, W.R., Harper, W.L., Wikeem, B.M., 1991b. Interior Douglas-fir Zone. In: Ecosystems of British Columbia, British Columbia Ministry of Forests.

Hope, G.D., Mitchell, W.R., Lloyd, D.A., Harper, and W.L., Wikeem, B.M., 1991c. Montane Spruce Zone. In Ecosystems of British Columbia, British Columbia Ministry of Forests.

Howley, T., Madden, M.G., O'Connell, M.-L., Ryder, A.G., 2006. The effect of principal component analysis on machine learning accuracy with high-dimensional spectral data. Knowledge-Based Systems 19, 363-370.

Huete, A.R., 1988. A soil-adjusted vegetation index (SAVI). Remote Sensing of Environment 25, 295-309.

Jiang, Z., Huete, A.R., Didan, K., Miura, T., 2008. Development of a two-band enhanced vegetation index without a blue band. Remote Sensing of Environment 112, 3833-3845.

Kempen, B., Brus, D.J., Heuvelink, G.B.M., Stoorvogel, J.J., 2009. Updating the 1:50,000 Dutch soil map using legacy soil data: a multinomial logistic regression approach. Geoderma 51, 311-326.

Kempen, B., Brus, D.J., Stoorvogel, J.J., Heuvelink, G.B.M., de Vries, F., 2012. Efficiency comparison of conventional and digital soil mapping for updating soil maps. Soil Science Society of America Journal 76: 2097-2115.

Ketcheson, M.V., Braumandl, T.F., Meidinger, D. Utzig, G., Demarchi, D.A., and Wikeem, B.M., 1991. Interior Cedar-Hemlock Zone. In Ecosystems of British Columbia, British Columbia Ministry of Forests.

Kuhn, M., 2008. Building predictive models in R using the caret package. Journal of Statistical Software 28: 1-26.

Lacoste, M., Lemercier, B., Walter, C., 2011. Regional mapping of soil parent material by machine learning based on point data. Geomorphology 133, 90-99.

Landwehr, N., Hall, M., Frank, E., 2005. Logistic model trees. Machine Learning 59, 161-205.

Lavkulich, L.M., Arocena, J.M., 2011. Luvisolic soils of Canada: genesis, distribution, and classification. Canadian Journal of Soil Science 91, 781-806

Leung, Y., Goodchild, M.F., Lin, C.C., 1993. Visualization of fuzzy scenes and probability fields. In: Computing Science and Statistics, Volume 24: Graphics and Visualization. Proceedings of the 24th Syposium on the Interface. Fairfax Station, VA, pp. 416-422.

Lin, H.S., Wheeler, D., Bell, J., Wilding, L., 2005. Assessment of soil spatial variability at multiple scales. Ecological Modelling 182, 271–290.

Lord, T.M., Green, A.J., 1974. Soils of the Tulameen Area of British Columbia. Report No. 13. Research Branch, Canada Department of Agriculture, Ottawa, ON, Canada.

Masek, J.G., Vermote, E.F., Saleous N.E., Wolfe, R., Hall, F.G., Huemmrich, K.F., Gao, F., Kutler, J., and Lim, T-K., 2006. A Landsat surface reflectance dataset for North America, 1990–2000. IEEE Geoscience and Remote Sensing Letters 3, 68-72.

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

McKenzie, N., Ryan, P., 1999. Spatial prediction of soil properties using environmental correlation. Geoderma 89, 67-94.

Minasny, B., McBratney, A.B., 2006. A conditioned Latin hypercube method for sampling in the presence of ancillary information. Computers and Geosciences 32, 1378-1388.

Moore, I.D., Grayson, R.B., Ladson, A.R., 1991. Digital terrain modelling: a review of hydrological, geomorphological, and biological applications. Hydrological Processes 5, 3-30.

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.

Moran, C.J., Bui, E.N., 2002. Spatial data mining for enhanced soil map modeling. International Journal of Geographical Information Systems 16, 533-549.

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, 1-19.

Mulder, V.L., Lacoste, M., Richer-de-Forges, A.C., Martin, M.P., Arrouays, D., 2016. National versus global modelling the 3D distribution of soil organic carbon in mainland France. Geoderma 263, 16-34.

Odgers, N.P., Sun, W., McBratney, A.B., Minasny, B., Clifford, D., 2014. Disaggregating and harmonising soil map units through resampled classification trees. Geoderma 214, 91-100.

Pennock, D., Bedard-Haughn, A., Viaud, V., 2011. Chernozemic soils of Canada: genesis, distribution, and classification. Canadian Journal of Soil Science 91, 719-747.

Pontius, R.G. Jr., Millones, M., 2011. Death to Kappa: birth of quantity disagreement and allocation disagreement for accuracy assessment. International Journal of Remote Sensing 32, 4407-4429.

Qi, J., Chehbouni, A., Huete, A.R., Kerr, Y.H., Sorooshian, S., 1994. A modified soil adjusted vegetation index. Remote Sensing of Environment 48, 119-126.

Quinlan, J., 1992. Learning with continuous classes. In: Adams, A., Sterlin, L. (Eds.), Proceedings AI’92, 5th Australian Conference on Artificial Intelligence. World Scientific, Singapore, pp. 343-348.

R Development Core Team, 2012. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available at http://www.R-project.org (verified 26 February 2016).

Rad, M.R.P., 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 loess derived soils of northern Iran. Geoderma 232, 97-106.

Riley, S.J., DeGloria, S.D., Elliot, R., 1999. A terrain ruggedness index that quantifies topographic heterogeneity. Intermountain Journal of Science 5, 23-27.

Rokach, L., 2010. Ensemble-based classifiers. Artificial Intelligence Review 33, 1-39.

Rondeaux, G., Steven, M., and Baret, F., 1996. Optimization of Soil-Adjusted Vegetation Indices. Remote Sensing of Environment 55, 95-107.

SAGA Development Team, 2011. System for Automated Geoscientific Analysis (SAGA). Available at http://www.saga-gis.org/en/index.html (verified 28 October 2014).

Schmidt, G.L., Jenkerson, C.B., Masek, J., Vermote, E., and Gao, F.Schmidt, G.L., Jenkerson, C.B., Masek, J., Vermote, E., and Gao, F., 2013, Landsat ecosystem disturbance adaptive processing system (LEDAPS) algorithm description: U.S. Geological Survey Open-File Report 2013–1057, 17 p. 2013.

Smith, C.A.S., Webb, K.T., Kenney, E., Anderson, A., Kroetsch, D., 2011. Brunisolic soils of Canada: genesis, distribution, and classification. Canadian Journal of Soil Science 91, 695-717.

Smith, S., Neilsen, D., Frank, G., Flager, E., Daneshfar, B., Lelyk, G., Kenney, E., Bulmer, C., Filatow, D., 2016. Disaggregatopm of legacy soil maps to produce a digital soil attribute map for the Okanagan Basin, British Columbia, Canada. *In* Digital Soil Mapping Across Paradigms, Scales and Boundaries. Springer Environmental Science and Engineering, pp. 305-317.

Sondheim, M., Suttie, K., 1983. User Manual for the British Columbia Soil Information System, 1. BC Ministry of Forests Publication R28-82053, Victoria, BC.

Subburayalu, S.K., Jenhani, I., Slater, B.K., 2014. Disaggregation of component soil series on an Ohio County soil survey map using possibilistic decision trees. Geoderma 213, 334-345.

Subburayalu, S.K., Slater, B.K., 2013. Soil series mapping by knowledge discovery from an Ohio County soil map. Soil Science Society of America Journal 77, 1254-1268.

Taghizadeh-Mehrjardi, R., Nabiollahi, K., Minasny, B., Triantafilis, J., 2015. Comparing data mining classifiers to predict spatial distribution of USDA-family soil groups in Baneh region, Iran. Geoderma 253-254, 67-77.

Van Hulse, J., Khoshgoftaar, T., 2009. Knowledge discovery from imbalanced and noisy data. Data & Knowledge Engineering 68, 1513-1542.

Van Ryswyk, A., McLean, A., Marchand, L.S., 1966. Climate, native vegetation and soils of some grasslands at different elevations in British Columbia. Canadian Journal of Plant Science 46, 35-49.

Wang, T., Hamann, A., Spittlehouse, D., Murdock, T. N., 2012. ClimateWNA - High-resolution spatial climate data for western North America. Journal of Applied Meteorology and Climatology 61, 16-29.

Warrens, M.J., 2015. Properties of the quantity disagreement and the allocation disagreement. International Journal of Remote Sensing 36, 1439-1446.

Wittneben, U., 1986. Soils of the Okanagan and Similkameen Valleys. Report No. 52. Surveys and Resource Mapping Branch, BC Ministry of Environment, Victoria, BC, Canada.

Young, G., Fenger, F.A., Luttmerding, H.A., 1992. Soils of the Ashcroft Map Area. Report No. 26. Integrated Management Branch, BC Environment, Victoria, BC, Canada.

Zhu, A.X., 1997. Measuring uncertainty in class assignment for natural resource maps using fuzzy logic. Photogrammetric Engineering & Remote Sensing 63, 1195-1202.

Chapter 5.

Conclusions

5.1. General Conclusions & Research Contributions

This research focussed on the development of methodologies that would facilitate the production of digital soil maps using legacy soil data. Aspects of the modeling process were extensively tested throughout the chapters of the dissertation. Outcomes of this dissertation included the development of a framework for extracting soil information from legacy soil surveys as well as a comprehensive review and testing of machine-learning techniques for mapping soil taxonomic units for the Lower Fraser Valley. In addition, the dissertation also provides a comparison of the use of soil polygon and soil pit derived training data for the prediction of soil Great Groups for the Okanagan-Kamloops region of BC. Although the study was designed with the premise of facilitating the development of DSMs for BC, the themes explored in this dissertation are applicable to the broader DSM literature. This is especially the case with Chapter 3 where the main messaging of its contents had the intention of alerting practitioners of pedometrics on the importance of performing comprehensive model comparison studies.

Chapter 2 of this dissertation proposed a framework for DSM using legacy soil data from single-component map units of a detailed soil survey (Objective 1) where three sampling methods were tested for mapping soil parent materials. Here, it was determined that developing training data using an area-weighted approach was most effective in predicting soil parent material classes when the RF classifier was used. Additional objectives tested the parameter optimization when using RF, where it was concluded that optimization resulted in a minimal increase in accuracy; furthermore, variable reduction also had limited influence on accuracy. Due to the effectiveness of the proposed framework, it was later applied in the subsequent chapters of the dissertation.

As far as the DSM literature is concerned, the development of a specific framework may not be the main point of interest because there are numerous frameworks that have been developed by other researchers in the field. This framework was developed mainly within the context of using BC's legacy soil data and data structure; however, it is also recognized that much of the legacy soil data around the world share a similar structure by way of having mixtures of single-component and multi-component polygons in legacy soil surveys and having a repository of legacy soil pit data. When considering the broader contribution of this chapter to the DSM literature, Heung et al. (2014) provided one of the first examples where RF was used specifically for classification purposes. Numerous recent studies, subsequent to its publication, have used RF. For example, Brungard et al. (2015), Kempen et al. (2015), and Taghizadeh-Mehrjardi et al. (2015) for the prediction of soil classes using pit-derived training data; Collard et al. (2014) and Chaney et al. (2016) in the refinement of existing soil survey maps; Gambill et al. (2016) for classifying soils using soil quantitative variables; and Wang et al. (2015) for assessing flood hazard risk. Furthermore, the methodologies presented in the chapter were scaled up for the production of a soil parent material map for BC (Bulmer et al., 2016) while similar approaches for the development of training data from legacy soil surveys have subsequently been used for mapping soil types over regions of France (Collard et al. 2014; Vincent et al., 2016) and Cyprus (Camera et al., 2017).

Chapter 3 presented a model comparison study that addressed Objective 2 of the dissertation. Here, it was concluded that there were major implications in the choice of model when predicting categorical data where predictions using the same training data, but different models, would produce drastically different results. Secondly, the different methods of extracting training data from soil survey data, using the same model, also produce drastically different results. By comparing 10 machine-learning models, it was determined that models such as CART with bagging, LMT, and RF were the most useful models when factors such as parameterization time, computational time, and the interpretability of results were taken into account in addition to accuracy. Furthermore, when testing the effectiveness of balancing the training data, it was concluded, in conjunction with the results of Chapter 2, that using the imbalanced area-weighted approach consistently produced higher accuracies in comparison to equal-class and by-polygon approaches. When ROS was used, the improvements in accuracies were minimal and not worth the additional computational demand. Although it was noted from the literature that numerous methodologies have involved a by-polygon approach for developing a training dataset (e.g. Odgers et al., 2014; Chaney et al., 2016); here, it was recommended that area-weighted sampling from soil survey polygons is used in order to best capture the feature space of the individual classes.

Chapter 3 along with the overview of machine-learning techniques for classification purposes in Section 1.1.3 of the dissertation was presented in Heung et al. (2016). The publication provided several important contributions to the DSM literature by compiling the first comprehensive overview of different types of machine-learning techniques specifically for DSM purposes. With the notable exceptions of Brungard et al. (2015) and Taghizadeh-Mehrjardi et al. (2015) that performed model comparison studies within the context of using soil pit-derived training data, Heung et al. (2016) was the first study that used training data from a conventional soil survey. The main message of the publication was to stress the importance of performing model comparison studies in DSM especially when numerous machine-learning models exist; yet, the majority of studies in the literature have limited their choice of model to either a few or a single model(s). Furthermore, the process of comparing numerous models was facilitated through the use of *R* packages such as *caret*, which greatly increased the efficiency of parameter optimization and developing *R* scripts in a standardized way – as such, model comparison should be adopted as ‘best practice’ for DSM.

Chapter 4 compared the accuracies of soil Great Group predictions when using training data derived from legacy soil pit data against training data derived from soil survey polygons – Objective 3 of the dissertation. Heung et al. (2017) provided a comparison of 9 machine-learning techniques, the results consistently showed that the use of polygon-derived training data resulted in higher accuracies when compared to soil pit-derived training data due to a better representation of the feature space of the various soil classes when using polygon data. The study also determined that the RF model performed reasonably well regardless of the training dataset used. Furthermore, single-model learners were compared to ensemble-model learners, where it was determined that ensemble techniques improved predictions when soil pit-derived training data were used. Improvements were likely due to greater model stability for small sample sizes. Finally, this study proposed the use of ensemble-modeling approaches as a way for assessing model uncertainty where the vote-count surfaces assisted in the visual interpretation of the results.

The motivation behind this chapter was based on the identified need in Brungard et al. (2015) and Heung et al. (2016) for a comparative study between the use of polygon-derived and soil pit-derived training data for DSM where such comparisons had not been made, previously. Another contribution made by this chapter was in the novel development of several ensemble-models that have never been tested in DSM (e.g. MLR with bagging, LMT with bagging, and *k*NN with bagging). Finally, this study also made the unique observation that similar to Collard et al. (2014), DSM approaches have the potential to improve existing conventional soil maps. Furthermore, similar to Kempen et al. (2012), the accuracy of DSM approaches, using soil pit data, could produce predictions that are similar or more accurate than the maps produced from conventional approaches.

In summation, this dissertation mainly contributes to the fields of pedometrics and soil science; however, the themes explored here may also contribute to other disciplines such as geomorphology and the modeling of surficial materials; hydrological modelling; and predictive ecosystem and resource mapping. Furthermore, the methodologies presented may be extended to provincial- or national-scale DSM initiatives.

5.2. Research Limitations

Throughout the course of this research, a number of limitations were identified with regards to the original mapping scale of the soil surveys; the choice of environmental covariates; quality of the point data; and the computational limitations of the modelling process.

The mapping scales of the conventional soil surveys used to develop the training data would have undoubtedly influenced the amount of detail in the resultant maps as well as the diversity of classes for those maps. For the agriculturally-intensive parts of the Lower Fraser Valley and the Okanagan-Kamloops regions, soils were mapped with greater detail – as reflected by the larger number of smaller map units. As a result, a visual assessment of the results for those areas showed soil patterns that were closely associated to the subtle changes in the landscape. In contrast, the soil surveys were performed at a larger map scale for the Coast Mountain region of the Lower Fraser Valley and forested landscapes of the Okanagan-Kamloops area. The map units for these regions were generally larger and unable to capture soil types that were the result of local-scale soil processes. As a consequence, localized occurrences of colluvial material were not mapped as effectively as the other parent material classes for the Lower Fraser Valley and where, similarly, the occurrence of hydromorphic soils adjacent to local streams may not have been captured by the training data or predicted by the models.

In terms of the environmental covariates used for this project, Chapter 2 used only topographic indices derived from a DEM, while Chapters 3 & 4 included some climatic and vegetative indices in addition to a large number of topographic indices. For the purposes of mapping soil parent materials, the explicit use of topographic indices was justifiable because landforms are the product of geomorphic processes that transport and deposit materials across the landscape and thus give rise to the topographic features that may be characterized from a DEM. Other studies have previously used gamma-radiometric data using passive remote sensing techniques in order to quantify potassium, thorium, and uranium abundances of bedrock and weathered materials (Wilford and Minty, 2007); however, the use of such data for glaciated landscapes, where the geochemical properties of the parent materials is highly heterogeneous, is unclear. For mapping soil classes, topographic indices may be used as a proxy-variable to represent the climatic variability for local-scale mapping where the long-range trends in temperature and precipitation are assumed to be constant while local-scale trends are encapsulated within topographic indices such as elevation, slope position, exposure, and aspect. However, to extend DSM projects to regional-, national-, and global-scales, there would be a greater reliance on climatic indices derived from satellite imagery and climate model data (e.g. Hengl et al., 2015, 2017) – all of which were used to a limited extent for this study.

An inherent challenge with the use of legacy soil point data is in their spatial accuracy as well as their distribution over a study area. With respect to the BCSIS dataset, where soil data were collected since the 1970s for various government projects, the coordinates of sample sites were typically determined from an aerial photograph or a topographic map. Another issue with the BCSIS dataset, especially for the Okanagan-Kamloops study area, related to the clustering of sampling locations due to the variability in study extents of the individual government projects that make up the BCSIS dataset. Generally, smaller projects resulted in greater clustering of sample points whereas larger projects resulted in a greater dispersion of sample points over the landscape. However, implementing an ideal sampling scheme, using approaches such as a conditioned Latin Hypercube sample design (Minasny and McBratney, 2006), was unfeasible in terms of cost and time given the size of the study area for this project. Another limitation in using the BCSIS dataset was due to the class imbalance issue, where the development of a training and validation datasets could have been improved with a greater number of sample points that represented the minority classes.

In order to extend the mapping approaches used in this dissertation to larger spatial-extents or higher-resolutions, an inherent challenge relates to computational limitations in terms of computer processor speed (CPU), memory (RAM), and storage space. For this study, limited CPU power and RAM hindered the parameter optimization procedure and the model-fitting process in spite of using parallel processing techniques. In Chapter 3, it was noted that even though the SVM-RBF approach resulted in the most accurate soil Great Group map, the computational demand related to parameter optimization, especially for a model that has an infinite combination of model parameters, was not worth the small gain in accuracy in comparison to the RF learner. Similarly, the computational cost of performing ROS in order to address the class imbalance issue on the training data resulted in marginal gains in accuracy. Challenges related to computational limitations, however, will not be a long-term issue due to the greater use of super-computing technology and cloud-computing services in DSM research for the foreseeable future (e.g. Hengl et al., 2014, 2015, 2017; Chaney et al., 2016).

5.3. References

Brungard, C.W., Boettinger, J.L, Duniway, M.C., Wills, S.A., Edwards, T.C., 2015. Machine learning for predicting soil classes in three semi-arid landscapes. Geoderma 239-240, 68-83.

Bulmer, C.E., Schmidt, M.G., Heung, B., Scarpone, C., Zhang, J., Filatow, D., Finvers, M., Smith, C.A.S., 2016. Improved soil mapping in British Columbia, Canada, with legacy soil data and Random Forest. *In* Digital Soil Mapping Across Paradigms, Scales and Boundaries. Springer Environmental Science and Engineering, pp. 291-303.

Camera, C., Zomeni, Z., Noller, J.S., Zissimos, A.M., 2017. A high resolution map of soil types and physical properties for Cyprus: A digital soil mapping optimization. Geoderma 285, 35-49.

Chaney, N.W., Wood, E.F., McBratney, A.B., Hempel, J.W., Nauman, T.W., Brungard, C.W., Odgers, N.P., 2016. POLARIS: A 30-meter probabilistic soil series map of the contiguous United States. Geoderma 274, 54-67.

Collard, F., Kempen, B., Heuvelink, G.B.M, Saby, N.P.A., Richer de Forges, A.C., Lehmann, S., Nehlig, P., Arrouays, D., 2014. Refining a reconnaissance soil map by calibrating regression models with data from the same map (Normandy, France). Geoderma Regional 1, 21-30.

Gambill, D.R., Wall, W.A., Fulton, A.J., Howard, H.R., 2016. Predicting USCS soil classification from soil property variables using Random Forest. Journal of Terramechanics 64, 85-92.

Hengl, T., Heuvelink, G.B.M., Kempen, B., Leenaars, J.G.B., Walsh, M.G., Shepherd, K.D., Sila, A., MacMillan, R.A., de Jesus, J.M., Tamene, L., Tondoh, J.E., 2015. Mapping soil properties of Africa at 250 m resolution: Random Forests significantly improve current predictions. PLoS ONE 10, 26pp.

Hengl, T., de Jesus, J.M., MacMillan, R.A., Batjes, N.H., Heuvelink, G.B.M., Ribeiro, E., Samuel-Rosa, A., Kempen, B., Leenaars, J.G.B., Walsh, M.G., Gonzalez, M.R., 2014. SoilGrids1km – Global soil information based on automated mapping. PLOS ONE 9, 17 pp.

Hengl, T., de Jesus, J.M., Heuvelink, G.B.M., Gonzalez, M.R., Kilibarda, M., Blagotić, Shangguan, W., Wright, M.,N., GEng, X., Bauer-Marschallinger, B., Guevara, M.A., Vargas, R., MacMillan, R.A., Batjes, N.H., Leenaars, J.G.B., Ribeiro, E., Wheeler, I., Mantel, S., Kempen, B., 2017. SoilGrids250m: Global gridded soil information based on machine learning. PLoS ONE 12, 40pp.

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.

Heung, B., Ho, H.C., Zhang, J., Knudby, A., Bulmer, C.E., Schmidt, M.G., 2016. An overview and comparison of machine-learning techniques for classification purposes in digital soil mapping. Geoderma 265, 62-77.

Heung, B., Hodúl, M., Schmidt, M.G., 2017 Comparing the use of legacy soil pits and soil survey polygons as training data for mapping soil classes. Geoderma 290, 51-68.

Kempen, B., Brus, D.J., de Vries, F., 2015. Operationalizing digital soil mapping for nationwide updating of the 1:50,000 soil map of the Netherlands. Geoderma 241-242, 313-329.

Kempen, B., Brus, D.J., Stoorvogel, J.J., Heuvelink, G.B.M., de Vries, F., 2012. Efficiency comparison of conventional and digital soil mapping for updating soil maps. Soil Science Society of America Journal 76: 2097-2115.

Minasny, B., McBratney, A.B., 2006. A conditioned Latin hypercube method for sampling in the presence of ancillary information. Computers and Geosciences 32, 1378-1388.

Odgers, N.P., Sun, W., McBratney, A.B., Minasny, B., and Clifford, D., 2014. Disaggregating and harmonising soil map units through resampled classification trees. Geoderma 214, 91-100.

Taghizadeh-Mehrjardi, R., Nabiollahi, K., Minasny, B., Triantafilis, J., 2015. Comparing data mining classifiers to predict spatial distribution of USDA-family soil groups in Baneh region, Iran. Geoderma 253-254, 67-77.

Vincent, S., Lemercier, B., Berthier, L., Walter, C., 2016 (In Press). Spatial disaggregation of complex Soil Map Units at the regional scale based on soil-landscape relationships. Geoderma, 1-13.

Wang, Z., Lai, C., Chen, X., Yang, B., Zhao, S., Bai, X., 2015. Flood hazard risk assessment model based on Random Forest. Journal of Hydrology 527, 1130-1141.

Appendix A. Co-Authorship Statement

Chapter 1: Section 1.2.3 was initially part of a manuscript that was accepted for publication with Chapter 3 (Heung et al., 2016) where its preparation was assisted by all co-authors.

Chapter 2: This chapter was co-designed under the guidance of C.E. Bulmer and M.G. Schmidt and assisted in manuscript preparation for Heung et al. (2014). In addition, C.E. Bulmer assisted with field work.

Chapter 3: This chapter was initiated and designed by myself where I conducted the research and modeling. H.C. Ho and A. Knudby assisted with the development of satellite imagery-derived covariates and J. Zhang assisted with validation of predictions. All co-authors assisted with manuscript preparation for Heung et al. (2016).

Chapter 4: This chapter was initiated by myself and co-designed with M. Hodúl where preliminary work on the research was carried out by M. Hodúl under my direction. M. Hodúl assisted with the development of environmental covariates and the processing of legacy soil data. Programing of models, development of ensemble-models, validation, and uncertainty mapping was done by me. M.G. Schmidt and M. Hodúl assisted with manuscript preparation for Heung et al. (2017).