Improvement of Soil Property Mapping in Northern Ontario’s Great Clay Belt using Multi-source Remotely Sensed Data

Rory Pittman

A Thesis submitted to
The Faculty of Graduate Studies
In Partial Fulfillment of the Requirements
for the Degree of
Master of Science

Graduate Program in
Earth and Space Science and Engineering

York University
Toronto, Ontario

November 2020

ABSTRACT

The prediction for the soil properties of texture, calcareous substrate reaction to acid, and ELC (Ecological Land Classification) moisture regime from environmental covariates derived from multi-source remotely sensed data was conducted for study areas located in the District of Cochrane in Ontario (49° – 50° N, 81° – 84° W). Random forest (RF) and support vector machine (SVM) approaches were applied to model soil property classifications for 3 adjacent regions: Hearst, Gordon Cosens Forest (GCF) and Abitibi River Forest (ARF). LiDAR (light detection and ranging) data was exploited to derive the detailed vegetation properties of canopy height model (CHM) and gap fraction, as well as a digital elevation model (DEM) for generating topographic covariates. The results indicate that vegetation covariates, particularly LiDAR-derived vegetation properties, were important in the prediction of the soil properties of interest. The most accurate models had accuracy scores greater than 0.7 and Cohen's kappa greater than 0.5.

ACKNOWLEDGEMENTS

I would like to sincerely thank my supervisor, Dr. Baoxin Hu for suggestions and feedback throughout my progression of pursuing this research. I am grateful for her comments and support as I am learning to write manuscripts for publication. I am amazed with her skill in determining the "opportunity" and corresponding "contribution" in a research study, and I hope to one day master that mode of thinking for future research analysis.

I am grateful to Dr. Jiali Shang for her suggestions, kindness and encouragement. I would also like to thank my graduate committee members at York University, Dr. Jianguo Wang and Dr. Kathy Young, for their suggestions and feedback with my thesis.

I thank Dr. Kara Webster for her suggestions for soil sampling methodologies, and for sharing her expertise concerning forest soils. I would like to thank Stephanie, Tom, Holly and Laura of the Great Lakes Forestry Centre in Sault Ste. Marie (Ontario) for teaching me about how to collect soil samples, and how to prepare and pre-process soil samples.

I would like to thank graduate students at York University for company and assistance. Specifically, I thank Wendy and Ima for advice with regards to navigating York University, and I thank Cheryl for working with me to coordinate and exchange physics lab sessions for teaching assistant duties to fit my schedule when I was away from Toronto. I also thank Steven, Mike, Ben and others for many conversations in our lab room.

I am especially thankful to my parents, Wayne & Marilyn, for their unconditional support, as well as my brothers Tyler and Jeremy. Maybe my brothers and I did not have opportunities or privilege growing up in rural Saskatchewan as what is accustomed by many individuals in this nation, but we persevered and made the best of our situation; now I can move on and pursue my dreams.

PUBLICATIONS

Results for this thesis have been published in a journal paper:

Pittman, R., Hu, B., and Webster, K. (2020) "Improvement of Soil Property Mapping in the Great Clay Belt of northern Ontario using Multi-source Remotely Sensed Data". *Geoderma* 381. Accepted for publication on September 24th 2020. https://doi.org/10.1016/j.geoderma.2020.114761

Results for this thesis will be published in a conference paper:

Pittman, R., and Hu, B. (2020) "Improvement of Soil Texture Classification with LiDAR Data". To appear in *Geoscience and Remote Sensing Symposium* (IGARSS), 2020 IEEE International. Accepted March 29th 2020. Presented October 2nd 2020.

Results from the journal paper form the basis of Chapters 3, 4, 5 and 6 of this thesis. The soil data and study areas are discussed in Chapter 3, methodology discussed in Chapter 4, results presented in Chapter 5, and a discussion section noted in Chapter 6. Results from the conference paper are analogous to soil texture classification results in Chapter 5 of this thesis.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF EQUATIONS

LIST OF ABBREVIATIONS

ALOS	Advanced Land Observing Satellite
ARF	Abitibi River Forest
ASCII	American Standard Code for Information Interchange
CHM	Canopy Height Model
CPU	Central Processing Unit
CSV	Comma-Separated Values
CTI	Compound Topographic Index
CV	Cross Validation
DEM	Digital Elevation Model
DN	Digital Number
DSM	Digital Surface Model
ELC	Ecological Land Classification
Esri	Environmental Systems Research Institute
EVI	Enhanced Vegetation Index
FPAR	Fraction of Photosynthetically Active Radiation
FRI	Forest Resource Inventory
GCB	Great Clay Belt
GCF	Gordon Cosens Forest
GeoTIFF	Georeferenced Tagged Image File Format
GIS	Geographic Information System
HCl	Hydrochloric acid
HH	Horizontal transmit and Horizontal receive
JAXA	Japanese Aerospace Exploration Agency
kNN	k Nearest Neighbors
LAI	Leaf Area Index
LAS	LiDAR point cloud data
LAZ	compressed LiDAR point cloud data

LiDAR	Light Detection and Ranging
LST	Land Surface Temperature
ME	Mean Error
MNRF	Ontario Ministry of Natural Resources and Forestry
MODIS	Moderate Resolution Imaging Spectroradiometer
MRRTF	Multi-Resolution Ridge Top Flatness
MRVBF	Multi-Resolution Valley Bottom Flatness
MSAVI	Modified Soil Adjusted Vegetation Index
NAD	North American Datum
NDVI	Normalized Difference Vegetation Index
NDWI	Normalized Difference Water Index
NFI	National Forest Inventory
NRCan	Natural Resources Canada
PALSAR	Phased Array type L-band Synthetic Aperture Radar
PCC	Percent Correct Classification
RADAR	Radio Detection and Ranging
RF	Random Forest
RMSE	Root-Mean-Square Error
ROC	Receiver Operating Characteristic
SAGA	System for Automated Geoscientific Analyses
SAR	Synthetic Aperture Radar
SAVI	Soil Adjusted Vegetation Index
SQL	Structured Query Language
SVM	Support Vector Machine
TRI	Topographic Ruggedness Index
TWI	Topographic Wetness Index
USGS	United States Geological Survey
vP	von Post Scale of Decomposition

Chapter 1

INTRODUCTION

Digital soil mapping, also recognized as digital soil modeling or as pedometric mapping, is the process of fitting models with environmental covariates as predictors to infer soil properties (Nussbaum et al., 2018). The process of digital soil mapping commonly results in the output of prediction maps, which can denote concentrations of soil properties, or the classification of categories relating to soil attributes. A prediction map of soil properties can be applied as an input layer or predictor for subsequent geospatial analysis concerning environmental studies or land management. Direct observation of soil properties may be impractical as vegetation canopies and debris materials can obscure the soil surface, and can even be difficult or impossible from direct observation of bare surfaces. This results in the need of sample collection for even the topmost soil horizons, and subsequently the implementation of soil property modelling. The application of digital soil mapping can be regarded as an improvement from interpolation methods such as kriging or area-averaging techniques, in the sense that digital soil mapping permits the generation of explicitly stated or quantitative models based on theory for soil survey that can be knowledge based (McBratney et al., 2003). Digital soil mapping can pick up on detailed variations while depending upon fewer sampling locations that would not be practical with kriging techniques, since with digital soil mapping the variation within the soil map can be deduced from the environmental covariates. Due to this dependence on environmental covariates, digital soil mapping methods can result in better accuracy at a more local scale if soil properties vary considerably between what was sampled at neighboring locations, that would not be possible to detect with interpolation methods.

The types of modeling processes utilized can be resolved based upon the observational nature of the soil properties, in regards to whether categorial attributes in terms of either nominal or ordinal values, or whether continuous (i.e. rational number) measurements were collected. For categorical attributes, one would have classification problems, whereas continuous values would be modeled in terms of regression. With soil classification problems, commonly modeled soil properties include soil texture classification (Akumu et al., 2015; Brungard et al., 2015; Heung et al., 2017, 2016), indicators for the existence of peat (Minasny et al., 2019; Poggio et al., 2013), or moisture regime ratings (Akumu et al., 2019). For regression modelling of continuous attributes, this is generally applied to concentrations of soil nutrients. Frequently regressed properties include soil carbon (Angelini et al., 2016; Mulder et al., 2016; Wang et al., 2018), or even soil texture in terms of concentrations of clay, sand and silt (Adhikari et al., 2013). Digital soil models can correspond to carbon concentrations and stock (Grimm et al., 2008) for environmental studies, or for agricultural purposes as soil texture (Walters et al., 1992; Yang et al., 2014), acidity (Fageria et al., 2014), and moisture (Misra and Tyler, 1999) can affect soil nutrient status.

Digital soil mapping critically relies on the linkage of environmental covariates to soil formation factors, from which environmental covariates are assumed to relate to aspects of fundamental categories that affect soil. Jenny (1941) was one of the first researchers to systematically hypothesize that soil can be considered as a product of the interaction of soil formation factors, by reasoning with case studies where one soil formation factor was controlled and another varied. The categories of soil formation factors contemplated by Jenny (1941) are abbreviated as CLORPT, which denote climate [CL], organisms [O], relief [R], parent material [P] and time [T]. Analogous acronyms and methodologies include SCORPAN as popularized by McBratney et al. (2003), which denote soil [S], climate [C], organisms [O], relief [R], parent material [P], age [A] and the geospatial component [N]. In general, only factors corresponding to the SCORP categories are considered, as the geospatial component is recorded in the collection of other attributes, and the age factor is outright ignored for point in time studies.

Digital soil mapping has been employed in numerous studies during the past decade, but there is still a need of improvement with modelling accuracies. Accuracies for digital soil mapping models for regression rarely exceed a coefficient of determination (R^2) of 0.7 and are generally reported between 0.3 to 0.6 (Nussbaum et al., 2018). Classification models for digital soil mapping tend to yield only satisfactory accuracies as well; the fraction of soil samples properly classified tend to be less than 70% (Brungard et al., 2015; Heung et al., 2017, 2016). Consequently, the improvement of modeling accuracies for digital soil mapping is an active area of interest. This intention can be improved by either considering different modelling techniques, or by considering a more comprehensive set of environmental covariates that either better capture the underlying aspects of soil formation factors, or correspond to soil formation factors that are larger drivers of the area of interest. Advances in digital soil mapping have also been attained in part by the implementation of machine learning approaches, many techniques which are flexible and applicable on predictors of different scales of measurement for either classification or regression type modelling.

Recently there has been considerable interest in digital soil mapping applications for boreal regions, as environmental stresses compounded by climate change due to global warming have placed an urgency on better understanding of the state of soils in such regions. Many of these boreal locales consist of peatlands which are susceptible to degradation via climate change through the release of stored soil carbon into the atmosphere (Minasny et al., 2019). However, boreal forest study areas can pose additional challenges for digital soil mapping applications. The remoteness and inaccessibility of these regions often results in a dearth of soil sample data being available. Many boreal regions in Canada consist of relatively flat topography with homogenous climate for the geospatial scale of a study area, with uncertain or unavailable data attributes of a finer spatial scale corresponding to the parent bedrock material laying underneath the surface. Therefore, it is hypothesized that vegetation would be the main soil formation factor for such boreal regions, so the utilization of environmental covariates corresponding to vegetation is a priority. Advancements with the availability of remote sensing data during the past few years, provides an opportunity for the generation of environmental covariates from this data as better predictors for digital soil mapping applications.

The purpose of the research for this thesis was to address challenges of digital soil mapping, which in this case was employed for study areas in a boreal biome. The output of this research, through models and prediction maps of soil properties had the intent of increasing knowledge of the state of such properties in a relatively remote region, which could be applied to land-use studies in the future. By the means of employing multi-source remotely sensed data to improve modelling accuracy for a boreal biome study area in the District of Cochrane in northern Ontario, this goal was accomplished through the following research objectives:

1) Identify environmental covariates of higher variable importance for a boreal biome among an expanded set of predictors.

Many contemporary digital soil mapping studies have considered study areas with assorted topography, and subsequently differing climatic zones. This would lead to topographic and climatic predictors having high enough variable significance in order to differentiate soil properties based upon respective covariates correlated to those soil formation factors. Previous digital soil mapping studies for boreal biomes have found topographic and climatic covariates to be the most significant when considering study areas of coarse spatial resolutions over the zonal scale of the continent (Beguin et al., 2017). For relatively flat topographies and homogenous climatic zones as encountered in boreal biomes of less than a few thousand square kilometers, topographic or climatic covariates might not vary widely enough over a study area to provide relevant information to model soil properties. Minasny et al. (2019) wrote that it is important to consider detailed vegetation covariates for peatland environments. Even when considering vegetation covariates, many studies only utilized vegetation covariates that are restricted to optical imagery (Minasny et al., 2019; Mulder et al., 2011) for periods of peak vegetation. Furthermore, many studies outright ignore covariates for parent material. It is anticipated that exploiting a more comprehensive set of environmental covariates corresponding to other classes of soil formation factors will advance the identification of predictors of higher pertinence for soil modeling.

The availability of recent LiDAR (light detection and ranging) data for the District of Cochrane for the study area of interest provided a novel research opportunity, as it allowed the generation of detailed vegetation covariates that have not been utilized in digital soil mapping literature before. Previous studies for digital soil mapping have utilized finer scale digital elevation models (DEMs) computed from LiDAR data for the generation of detailed topographic covariates (Goldman et al., 2020). LiDAR data has not been exploited to calculate vegetation covariates for pedometric modelling outside of small study areas, as LiDAR data has been limited to study areas of less than 400 km^2 (Minasny et al., 2019). Vegetation is expected to be the main soil formation factor for the study area, due in part to the consistency of climate and flatness (i.e. lack of topographical variation) for the study area (McBratney et al., 2003; Minasny et al., 2019). Detailed vegetation covariates can provide useful information pertaining to the vegetation canopy, that might not be captured by optical imagery or other sensor types. Canopy height model (CHM) and gap fraction are detailed environmental covariates corresponding to vegetation that can be derived from LiDAR data. CHM characterizes the height of the canopy layer, whereas gap fraction relates to vegetation cover of the ground beneath the canopy layer. From literature, CHM was found to be significant for regression modelling of soil moisture for a study area of 21 km^2 near Timmins in northeastern Ontario (Southee et al., 2012). To our knowledge, gap fraction has not been previously utilized in digital soil mapping research.

Multi-source remotely sensed data can be applied to generate environmental covariates, as an aggregation of products derived from different remote sensing technologies can lead to better predictions (Minasny et al., 2019). LiDAR data can be exploited to derive a detailed DEM that can be utilized for computing topographic covariates, as well as separately for the calculation of vegetation covariates relating to other features rather than surface reflectance or optical properties. Optical imagery can be utilized to generate covariates corresponding to surface reflectance for multiple seasons throughout a year, or even over the span of years. Synthetic aperture RADAR (SAR) can be utilized to generate covariates corresponding to understory moisture of a canopy. Geophysical attributes such as aeromagnetic surveys can be applied as covariates corresponding to parent material.

2) Systematically investigate advanced modelling techniques to improve accuracies of models trained on predictors based upon multi-source remotely sensed data.

Machine learning approaches have been popular recently for digital soil mapping, and demonstrated to have superior modelling performance in comparison with traditional modelling techniques such as multinomial regression or multilinear regression (Brungard et al., 2015; Heung et al., 2016; Were et al., 2015). Improving classification accuracies for digital soil mapping is an active area of research, as accuracies for the best models rarely exceed 0.7 in correctness scores (Dornik et al., 2018; Heung et al., 2017, 2016). It is expected that machine learning approaches such as random forest (RF) and support vector machine (SVM), combined with a more comprehensive set of environmental covariates for model predictors corresponding to further aspects of soil formation factors, will work to improve modelling accuracies.

3) Assess the spatial relationship between soil properties and land forms for the boreal study areas.

It is probable that there are interrelations between various soil properties. Different biotypes or zones within the boreal forest biome could be discerned from prediction maps of the soil properties. Peatlands, second growth forest or agricultural regions can be correlated to specific soil properties. Uncertainties with predictions maps, in terms of entropy or ignorance uncertainty (Heung et al., 2017) to identify zones with highest or lowest prediction uncertainties, are also of interest in order to attest confidences with predictions (Minasny et al., 2019). The identification of localities associated with certain soil properties will support future environmental studies, although it is not the prime objective of this thesis.

The outline of this thesis is structured into the subsequent sections. Background literature concerning environmental covariates, modelling techniques, number of sampling points and study area sizes are considered in Chapter 2. The targeted study areas for the research presented in this thesis, as well as the respective soil sampling data are discussed in Chapter 3. Methodology for thesis research, including environmental covariates utilized and derived, are discussed in Chapter 4. In Chapter 5 the modelling accuracies and soil prediction maps are presented and reviewed. A discussion section follows in Chapter 6, with the conclusion and closing remarks presented in Chapter 7.

Chapter 2

LITERATURE REVIEW

Digital soil mapping has been popular in soil modelling research for the past couple of decades, with current interest spurred in part by the growing collection and access of data sources for environmental covariates as well as machine learning techniques. The main assumption of digital soil mapping is that soil properties can be modelled based upon a combination of environmental covariates as predictors corresponding to soil formation factors. From literature, recent studies concerning digital soil mapping were examined to determine insights. These matters reviewed included the environmental covariates utilized as predictors for soil models, the soil depths of interest for sampling, targeted soil properties, study area sizes, covariate spatial resolutions, and the number of soil sampling points utilized. Modelling techniques applied, as well as the metrics used for model evaluation were also studied. The purpose of this literature review was to determine commonalities among the various considerations discussed, in order to better setup and conduct the analysis for the soil modelling research contained in this thesis.

2.1 Environmental Covariates

A SCORPAN methodology (Keskin and Grunwald, 2018; McBratney et al., 2003) or analogous are typically incorporated when applying digital soil mapping. Other acronyms for similar methods include CLORPT (Behrens et al., 2018; Mulder et al., 2011) or equivalent (McKenzie and Ryan, 1999; Nussbaum et al., 2018), where typically only predictors from just a few or less of these categories are applied. The commonly utilized soil formation factors correspond to soil, climate, vegetation, topography, and parent material. Other methods will incorporate a time factor and geospatial factor; the geospatial factor is typically included in the collection of environmental covariates for other soil formation factors, so it is usually not outright considered as its own factor. The time factor is usually ignored for digital soil mapping studies, as it can take decades or longer before soil properties change due to a stressor (Jenny, 1941). Properties for soil are often obtained from the site level, so oftentimes soil properties are not considered for predictors as soil properties are typically the target variables. It makes little sense to utilize a soil input obtained from the same site level as a predictor for another soil property as that might defeat the efficacy of why to implement digital soil mapping; the goal of digital soil mapping is to build soil prediction maps that would need to be extrapolated over areas outside soil sampling locations. Environmental covariates for the soil formation factors are obtained usually in the form of either rasterized imagery or vector files.

A few studies did utilize soil properties as covariates for modelling. Adhikari et al. (2013) applied digital soil types maps of 11 classes as an input for a model to predict soil textural components. Mulder et al. (2011) wrote that certain soil properties, such as soil moisture or soil texture could be determined with medium to high feasibility from active Radar instruments. Mulder et al. (2011) also wrote that optical spectroscopy can be utilized to infer soil attributes relating to mineralogy, iron content, soil texture and salinity with medium to high feasibility. However, sensors for the correct spectroscopy settings may not be available for most study areas, so these applications are not practical. Some studies have considered legacy soil property maps as input (Camera et al., 2017; Ellili Bargaoui et al., 2019; Heung et al., 2017) for training. Nonetheless, the accuracy of legacy soil property maps might be questionable, so thus be of limited usage for remote regions.

Topographic covariates are usually generated from a digital elevation model (DEM) and correspond to landscape features. Commonly used topographic covariates allude to local scale morphometry parameters (Adhikari et al., 2013; Dornik et al., 2018; Heung et al., 2017, 2016). Local scale morphometry covariates include elevation (often just denoted by DEM), slope, aspect, and plan or profile curvature (Heung et al., 2016; Ließ et al., 2012). There are other attributes of topography that covariates can represent, including landscape-scale morphometry, hydrologic characteristics and landscape exposure (Heung et al., 2017, 2016). Topographic covariates pertaining to landscape-scale morphometry include slope height, normalized height, valley depth, mid-slope position, multi-resolution ridge top flatness (MRRTF) index, and multi-resolution valley-bottom flatness (MRVBF) index. Hydrologic attributes include topographic wetness indices, stream power index, hydrologic slope positions and distances to nearest bodies of water (lakes or streams) (Brungard et al., 2015; Heung et al., 2016). Earlier papers (McBratney et al., 2003; McKenzie and Ryan, 1999) have applied compound topographic index (CTI) which is based upon slope and upslope area. Landscape exposure of the ground can be incorporated by a sky-view or terrain-view covariates (Heung et al., 2016).

Climatic covariates can include temperature or precipitation, which can correspond to quarterly or annual values, such as mean annual precipitation (Heung et al., 2016) or maximum mean temperature (Mulder et al., 2016). These measurements can be obtained from surface weather stations, and kriged to obtain values for an area, or derived weather product maps can be applied instead. Land surface temperature (LST) averages obtained from winter and summer, or thermal band temperatures, can be utilized. Satellite-derived products are often obtained from MODIS (Moderate Resolution Imaging Spectroradiometer) (Poggio et al., 2013) or Landsat imagery (Heung et al., 2016).

Environmental covariates relating to vegetation are often derived from data obtained from optical satellite imagery. These covariates can be incorporated as surface reflectance of the vegetation cover for various optical bands, ranging from ultraviolet through to visible and to shortwave infrared wavelengths. Normalized difference indices computed from applicable optical imagery bands are popular in literature (Brungard et al., 2015; Dornik et al., 2018; Heung et al., 2017, 2016). Normalized difference vegetation index (NDVI) and normalized difference water index (NDWI) are commonly utilized. Other examples of indices used as vegetation covariates include the soil adjusted vegetation index (SAVI) (Dornik et al., 2018), modified soil adjusted vegetation index (MSAVI) and the enhanced vegetation index (EVI) (Heung et al., 2017), which were each based on combinations of the near-infrared and red optical bands. Landsat, MODIS, and recently Sentinel-2 are common sources for multispectral satellite imagery (Minasny et al., 2019). Land management land-use specifications for vegetation have also been utilized (Mulder et al., 2016).

Attributes for parent material are commonly ignored in digital soil mapping research. If anything is applied, it usually corresponds to the underlying bedrock geology or other lithology. Mulder et al. (2011) wrote that fine spectral resolution imagery can be utilized to determine mineralogy properties from spectral signatures of rock outcrops, but this would only be feasible for exposed parent material at the surface. Both the spectral and spatial resolutions of MODIS and Landsat imagery have been found to be too coarse for such mineralogy analysis (Mulder et al., 2011). A gamma radiometer can be utilized to record the amount of radioactive isotopes in the soil, where its gamma intensity is directly related to the mineralogy and parent material (Minasny et al., 2019).

Innovation with digital soil mapping research has led to integration of data obtained from different remote sensing technologies. Such technologies of recent interest include SAR and LiDAR (Minasny et al., 2019). SAR of lower frequencies (P-band, L-band, and C-band) can penetrate vegetation canopies to retrieve soil characteristics (Minasny et al., 2019). LiDAR data has been utilized to derive finer spatial resolution DEMs for digital soil mapping studies (Akumu et al., 2015; Anderson et al., 2006; Greve et al., 2012; Mohamed, 2020; Mulder et al., 2011), for which detailed topographic covariates can be generated.

There has been a recent interest in generating DEMs and terrain models for wetland environments (O'Neil et al., 2019; Rapinel et al., 2019; Vernimmen et al., 2019). A LiDAR-derived vegetation covariate of CHM was utilized for regression modelling of soil moisture (Southee et al., 2012). However, due to cost and data processing considerations, as LiDAR is often acquired via airborne survey campaigns (Minasny et al., 2019), LiDAR covariates have generally been restricted to relatively small study areas of less than a couple hundred square kilometers.

Topographic covariates have been used extensively in digital soil mapping studies. In many studies, topographic covariates formed the majority of covariates utilized (Adhikari et al., 2013; Brungard et al., 2015; Dornik et al., 2018; Heung et al., 2016; Ließ et al., 2012; Wu et al., 2019). Climatic covariates corresponding to precipitation and temperature for either quarters of the year or annually can be applied in digital soil mapping studies (Beguin et al., 2017; Heung et al., 2017, 2016) for where data or estimates exist. Vegetation covariates derived from multispectral imagery have become increasingly common (Minasny et al., 2019; Mulder et al., 2011). In a review of digital soil mapping for peatlands, Minasny et al. (2019) wrote that approximately 25% of studies utilized only one source for covariates, and that most recent studies used only 2 or 3 sources. Minasny et al. (2019) recommended an integrated approach of combining optical, SAR or other remote sensing data, that is, multi-source remotely sensed data, for digital soil mapping research as it can lead to better predictions.

2.2 Study Scales

Commonly reported target variables include particulate concentrations, soil texture attributes, or soil profile depths (McBratney et al., 2003). Soil samples are usually collected as point measurements retrieved during field campaigns, obtained from depths varying from the surface down to a maximum of 2 m deep. Generally, the first 0-15 cm of the surface is the layer of interest for digital soil mapping applications. Many studies just consider the first 10 cm depth of mineral soil (Mulder et al., 2016). Nussbaum et al.(2018) wrote that soil models usually have higher accuracies when trained on data for topsoil layers when compared with model results for deeper subsoil layers. Carbon stock and nitrogen values tend to be maximum at the surface (Gomez et al., 2008; Grimm et al., 2008), where either the 5-15 cm or 0-10 cm depth layer is what is analyzed. Soil textural properties are also assessed at the surface layers, as these are the layers that interact and are directly affected by agriculture usage and vegetation.

The number of soil sampling locations obtained per study varied generally according to scale of the targeted study areas. These numbers ranged from between 50 to less than 300 sites for study areas of a couple thousand square kilometers or less (Brungard et al., 2015; Heung et al., 2016; Ließ et al., 2012; Were et al., 2015; Yang et al., 2016). Larger study areas had 705 soil sample sites for 30,000 km^2 in the Tibetan plateau of China (Wang et al., 2018), 955 sites for 78,000 km^2 for Scotland (Poggio et al., 2013), and 1958 sites for the whole of Denmark (Adhikari et al., 2013). From reviewing literature, it should be acceptable to utilize soil samples collected from a minimum of a couple hundred different sites for a targeted study area of a couple thousand square kilometers. Minasny et al. (2019) wrote that low sampling densities can be expected for studies based on field survey due to accessibility issues in regards with field observations.

The spatial resolution of analysis for digital soil mapping is often restricted by the spatial resolutions of the common covariates. In general, this will be limited by the spatial resolution of the DEM. The spatial resolution of the DEM (and corresponding topographic predictors) varied from 100 m for mountainous terrain (Heung et al., 2016) to 90 m for Scotland (Poggio et al., 2013) and 30,000 km^2 of the Tibetan plateau (Yang et al., 2016), to 2 m resolution for areas less than a thousand square kilometers (Adhikari et al., 2013; Ließ et al., 2012; Nussbaum et al., 2018). However, commonly utilized optical imagery such as Landsat-8 has spatial resolutions of 30 m, and MODIS is even coarser with spatial resolutions of 250 m to 500 m. Thus, the spatial scale for the analysis of digital soil mapping will be the maximum of the DEM or optical imagery spatial resolution, which is usually a minimum of 30 m. In general, better modelling accuracy can be obtained from utilizing finer spatial resolutions for covariates applied as model predictors (Garcia et al., 2017).

2.3 Modelling Techniques

Machine learning techniques have been dominant lately for digital soil mapping (Brungard et al., 2015; Heung et al., 2016), in part as models can consist of large sets of predictors (Nussbaum et al., 2018). The highest accuracies in digital soil mapping are generally attained by machine learning methods, which tend to also work better than linear regression for regression, or even logistic regression for classification problems (Minasny et al., 2019). Common machine learning approaches for digital soil mapping include random forest (Brungard et al., 2015; Ließ et al., 2012; Nussbaum et al., 2018; Poggio et al., 2013; Were et al., 2015; Yang et al., 2016), as well as regression trees (Ließ et al., 2012; McKenzie and Ryan, 1999; Poggio et al., 2013), and boosted regression trees (Nussbaum et al., 2018; Wang et al., 2018; Yang et al., 2016). Support vector machines, either linear or of radial basis functions, have also been popular (Heung et al., 2016; Wang et al., 2018). Other machine learning methods that have been employed for digital soil mapping include deep-learning (i.e. unsupervised) approaches such as artificial neural networks (Heung et al., 2016; Were et al., 2015). Machine learning approaches tend to be versatile, as they can be applied to predictors recorded at different levels of measurements (i.e. ordinal, categorial, or continuous), and many approaches can be fitted to target variables that are either categorical or continuous. However, machine learning approaches can be more difficult to understand, and can require greater training time and processing requirements. Minasny et al. (2019) wrote that supervised classification approaches work best where there is general site knowledge and available field observations, whereas unsupervised classification is generally utilized for regions with few field observations.

Regression models such as multivariate linear regression have been utilized for continuous target variables with concentrations or amounts (Adhikari et al., 2013; Gessler, P.E., Chadwirk, O. A., Chamran, F., Althouse, L., Holmes, 2000), and multinomial logistic regression has been applied for soil group classifications (Heung et al., 2016). Other approaches utilized for soil classifications include k-nearest neighbors (k-NN) (Brungard et al., 2015; Heung et al., 2016), and linear discriminant analysis (LDA) (Brungard et al., 2015). However, in these studies the accuracies of multinomial logistic regression, k-NN and LDA models were usually lower than those obtained from machine learning approaches.

In regards to specific approaches that obtain the highest accuracies, in general, random forest and support vector machines attained higher accuracies than other modelling approaches (Brungard et al., 2015; Heung et al., 2016). The flexibility of the random forest approach to be applied to both classification and regression-type problems, as well as higher accuracy, render random forest as a first technique of choice for digital soil mapping. SVMs can have comparable accuracies to RFs for classification problems (Heung et al., 2016). Linear SVMs have been established to have competitive accuracies when evaluated against SVMs based on radial basis functions (Brungard et al., 2015). For classification problems, SVMs are also a suitable choice to apply as an approach for digital soil mapping.

The metrics utilized for model evaluation depend on whether classification or regression-type modelling is of interest. For classification modelling, the correctness, also known as the percent correct classification (PCC) (also referred to as percent correctly classified), is the main metric for model assessment. Some studies employing random forests or boosted regression trees have utilized the out-of-bag error (OOB) (Brungard et al., 2015; Camera et al., 2017; Ließ et al., 2012), which is a misclassification rate. The OOB is calculated as the mean prediction error for each bootstrap sample of training observations, computed using predictions from decision trees that did not comprise those training observations in their corresponding bootstrap sample (Genuer et al., 2010; Ließ et al., 2012). For measuring disagreement in regards to inter-rater reliability for categorical predictions, accounting for by-chance agreement, Cohen's kappa statistic can be utilized (Brungard et al., 2015). Regarding accuracy for soil group classifications, from recent literature the most accurate models had correctness scores between 0.5 to 0.8 (Dornik et al., 2018; Heung et al., 2017, 2016). For continuous target variables or regression-type modelling, commonly applied metrics for model evaluation include the coefficient of determination (R^2) (Mulder et al., 2011) with either root mean square error (RMSE) (Adhikari et al., 2013; Mulder et al., 2016) or mean error (ME) (Adhikari et al., 2013; Angelini et al., 2016) reported.

For categorical target variables, confusion matrices are popular for determining what classes of the target variable are better modelled and predicted (Dornik et al., 2018; Heung et al., 2017). Assessment for model predictions can be judged on the basis of entropy maps (Heung et al., 2017, 2016; Minasny et al., 2019), which are graphical output depicting regions within a study area in regards to uncertainty with modelling predictions. Heung et al. (2017) applied entropy maps, here deemed as ignorance calculations, for determining areas of the map that had the higher modelling uncertainties. These calculations were computed based upon the voting, or probabilities, of applying fitted models to the verification data for prediction and extrapolation. Entropy maps provide a visual representation of modelling uncertainties among an area, whereas confusion matrices provide the uncertainty in table form for the specific predicted classes. From entropy maps, in conjunction with soil property prediction maps, it can be ascertained what categories of properties have the lowest or highest uncertainties with modelling. This can contribute to deciphering patterns between soil properties and landforms in order to make interrelations more apparent.

Chapter 3

STUDY AREAS & SOIL DATA

The soil sampling data and associated study areas for this research are presented in this chapter. Details pertaining to the study areas are first discussed, followed by specifics about the soil sites and properties in the subsequent section.

3.1 Study Areas

The study area for this research existed in the District of Cochrane, with latitudes ranging from 49° N to 50° N and longitudes ranging from 81° W to 84° W. This expanse extends from the communities of Hearst in the northwest to Smooth Rock Falls in the southeast, along the Highway 11 corridor in northern Ontario. This region is colloquially referred to as the Great Clay Belt, due to the dominance of heavy clay at depths below 20 cm or so throughout the region. The Great Clay Belt extends eastward into Quebec, but for research purposes only the portion residing in Ontario and accessible by roadways was considered. For soil classification modelling, the study area consisted of three neighboring subregions of Hearst, Gordon Cosens Forest (GCF) and the Abitibi River Forest (ARF), as shown in Figure 1. Note that the coordinate system of the grid for this figure, as well as for subsequent figures presented in this thesis, is the NAD 1983 Lambert Conformal Conic projection. These study areas were selected to correspond to separately clustered and collected soil samples, confined within the bounds of where recently acquired LiDAR data was available. The three study areas together are considered the

Great Clay Belt (GCB) study area, which is the whole region of all soil samples grouped together.

This region is part of the boreal biome and is mostly forested, with some agricultural land, particularly concentrated around the community of Kapuskasing. The indigenous tree species for this area are black spruce [*Picea mariana*], white spruce [*Picea glauca*], balsam fir [*Abies balsamea*], trembling aspen [*Populus tremuloides*], tamarack [*Larix laricina*], white birch [*Betula papyrifera*], balsam poplar [*Populus balsamifera*], jack pine [*Pinus banksiana*] and eastern cedar [*Thuja occidentalis*]. Black spruce is the most dominant tree species throughout the region, followed by balsam fir, white spruce and trembling aspen. Balsam poplar and tamarack are common in open areas and pastures or sections that were previously cleared. Jack pine and eastern cedar are not commonly encountered in this region. Peatlands are prevalent in wetland areas throughout the region.

The climate of this region is continental subarctic, with long cold winters and short warm summers. Long-term climatic records are available for Kapuskasing, which is located in the center of the region in the GCF study area. For Kapuskasing, annual temperature averages around 1° C, with mean January temperatures of -18° C and mean July temperatures of 17° C (Environment Canada). The annual precipitation is 83 cm, with amounts that are fairly regular throughout the year with higher precipitation totals from May through November (Environment Canada). Evaporation for this region is low, and in part wetness is an issue. For agricultural regions, the wetness is of such an issue that field tiling is performed to drain excess moisture from the soils (personal communication with farmers).

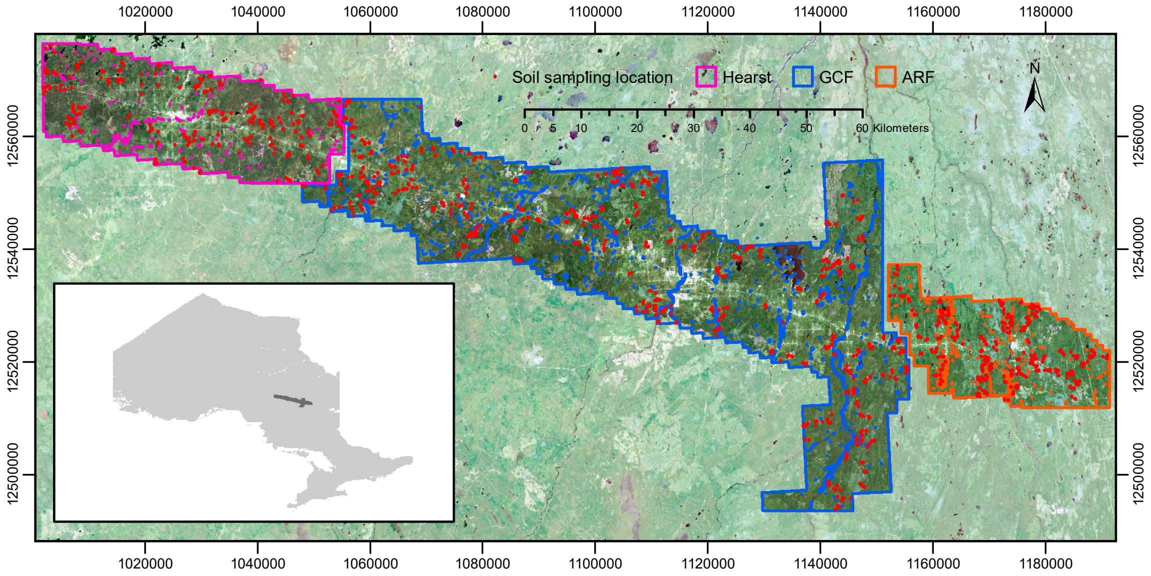


Figure 1 - Study areas for the GCB in the District of Cochrane in northern Ontario. Note that Hearst is outlined in magenta (left), GCF is outlined in blue (center) and ARF is outlined in orange (right). Inset: Location of GCB within the Province of Ontario.

The westernmost region is the Hearst study area (outlined in magenta in Figure 1), which comprises of 897 km^2 centered around the community of Hearst. Latitudes for this study area vary from 49.6° N to 49.8° N, with longitudes ranging from 83.3° W to 84.0° W, with elevations ranging from a minimum of 209 m in the north to a maximum of 283 m in the southwestern portion. The Mattawishkwia River flows through the study area on a northeast direction. The Hearst study area is mostly forested, with black spruce as the dominant tree species, but also with localities with balsam fir and trembling aspen. Compared with the other subregions, this study area averages slightly less annual precipitation at around 77 cm, and average about a degree C colder than the other study areas (WorldClim version 1).

In the center of the region is the Gordon Cosens Forest (GCF) (outlined in blue in Figure 1) which is positioned around Kapuskasing, with an area of 2,577 km^2. The latitudes for this study area range from 49.0° N to 49.7° N, with longitudes ranging from 81.9° W to 83.4° W. Elevations vary from a minimum of 197 m to a maximum of 284 m. In order from east to west, the Groundhog, Kapuskasing and Missinaibi rivers flowed through the study area, each from a predominantly southwest to northeast trajectory. There are agricultural lands surrounding Kapuskasing, but most of the area is forested with black spruce as the dominant tree species. Due to logging activity associated with the long-term presence of a forestry mill in Kapuskasing (Town of Kapuskasing, 2020), there exists substantial areas of second-growth forest as much of this region has been previously cleared. There is more variation with CHM for the vegetation in the GCF when compared to the vegetation for the Hearst and ARF regions.

The easternmost region is the Abitibi River Forest (ARF) (here outlined in orange in Figure 1) which consists of 611 km^2 and is centered around the community of Smooth Rock Falls. Latitudes vary from 49.2° N to 49.4° N, with longitudes varying from 81.4° W to 81.9° W. Elevations vary slightly in this region, with a minimum of 212 m and a maximum of 291 m. The Mattagami river flows northward through the center of this study area. Black spruce is the dominant tree species, but balsam fir is also dominant at a large minority of sites. This study area has the most annual precipitation of the study areas by a slight amount, which is 84 cm (WorldClim version 1).

3.2 Soil Data

Soil data for this study was obtained from the Forest Resource Inventory (FRI) within the Ontario Ministry of Natural Resources and Forestry (MNRF). The data obtained by the FRI was categorical in nature. The FRI data was collected and already separated into the neighboring subregions of Hearst, Gordon Cosens Forest (GCF) and Abitibi River Forest (ARF); this same cluster of soil sample collection was kept for this analysis of digital soil mapping for this study. Soil sampling locations consisted of 157 sites for Hearst, 446 sites for GCF, and 305 sites for ARF, for a total of 908 sites for the GCB. These soil samples were obtained primarily from August to October 2011 for Hearst, from July to November 2011 and October 2012 for the GCF, and from September to November 2012 and August 2013 for the ARF. The soil sample data collected contained information pertaining to soil texture class, calcareous reaction to acid, and ELC (Ecological Land Classification) moisture regime. All information pertaining to the soil data was obtained from a technical report written by Paloniemi (2018). These soil samples were collected primarily from the shallowest depth of mineral soil, which usually corresponded to the 5-15 cm depth layer. The dominant tree species were also noted at these soil sampling locations. Summaries for these soil samples for each targeted study area are reported in Table 1.

	Area	Number of	Soil Texture			Calcareous Substrate*	
Region	[km^2]	Sites	[Peat]	[Loamy]	[Clayey]	[n]	[k]
Hearst	897	157	63	64	30	76	81
Gordon Cosens Forest (GCF)	2577	446	212	164	70	253	193
Abitibi River Forest (ARF)	611	305	139	122	44	153	152
All Regions (GCB)	4085	908	414	350	144	482	426

	Ecological Land Classification (ELC) Moisture Regime									
Region	[0]	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]
Hearst	3	1	32	23	15	11	9	25	37	1
Gordon Cosens Forest (GCF)	3	8	80	53	27	39	24	94	113	5
Abitibi River Forest (ARF)	1	4	50	35	23	26	26	48	91	1
All Regions (GCB)	7	13	162	111	65	76	59	167	241	7

* Reaction to solution of 10% HCl; 'k' denotes fizzing reaction, 'n' denotes no fizzing reaction

Table 1 - Soil summaries for study areas.

Soil texture was firstly classified into categories relating to concentrations of the textural components of sand, silt and clay. Samples corresponding to peat were classified on a van Post (vP) system of decomposition (Malterer et al., 1992; Paloniemi, 2018). Expert assessment in the field was performed to determine the soil texture classification, based upon graininess, dry feel, stickiness, moist cast, ribbon, taste and shine tests. These tests were a combination of finger tests to help determine the soil texture group by following guidelines. Following the texture triangle utilized for the soil texture classification (Paloniemi, 2018), fine loamy soil consisted of at least 25% to a maximum of 50% clay with at most 45% sand, had thin short ribbons of 2.5 to 5 cm for the ribbon test, and had a strong cast for the moist cast test. Coarse loamy soil was composed of a maximum of 30% clay, and had very short and thick ribbons for the ribbon test. For clayey soil, it consisted of at least 35 to 40% clay, had long ribbons of at least 5 cm length for the ribbon test, and for the shine test had a shiny appearance. The von Post scales of decomposition were utilized to categorize peat; here peat ranged from scales of vP1 for completely undecomposed plant structure to vP10 for entirely decomposed plant material. These soil textures were subsequently condensed to the soil texture family categories of peat, fine loamy, coarse loamy and clayey; for this analysis, the fine loamy and coarse loamy categories were combined into just the category of loamy. The most common soil texture categorization for the study areas in general was peat, followed by loamy. Looking at Table 1, for the soil texture frequencies among the samples corresponding to the sites, the 157 sites for Hearst had 63 that were peat, 64 that were loamy, and 30 were clayey. For the GCF, for its 446 sites there were 212 that were peat, 164 were loamy and 70 were clayey. The 305 sites in the ARF corresponded to 139 for peat, 122 for loamy and 44 for clayey.

Calcareous substrate testing was done in the field by exposing the substrate sample to hydrochloric acid (HCl). When the soil substrate was exposed to a solution of 10% HCl, 'k' was noted if a fizzing reaction with the soil substrate occurred, otherwise 'n' was recorded if no reaction was observed. In the Hearst study area, for its 157 sites, 81 sites presented a fizzing reaction, whereas 76 sites did not. For the ARF, the sites were almost evenly split, with 152 sites exhibiting a reaction and 153 sites did not. Incidences for the GCF were skewed with 193 sites presenting a reaction versus 253 that showed no reaction.

Testing for moisture regime was performed using field-derived soil attributes following guidelines from the Ontario Ecological Land Classification (ELC) program (Pokharel and

Dech, 2011). ELC moisture regime is an ordinal classification with scales ranging from 0 for dry, 1, 2 & 3 for fresh, 4 & 5 for moist, 6 for very moist, 7, 8 & 9 for wet, with a higher number or letter indicating wetter saturated soil. A combination of soil texture category, depth to mottles for the mineral soil, and presence of gley for deeper soil depths were utilized to determine the ELC moisture regime classifications. ELC moisture regime values for the subregions in general indicated moist or wet, due in part to the poor drainage, low annual evapotranspiration, and presence of peatlands in the GCB region.

The soil properties of texture family, calcareous substrate reaction to 10% HCl and ELC moisture regime were the dependent variables utilized for this research. It was anticipated that there would be similarities in patterns with prediction maps with these 3 target variables, as calcareous substrate is related to soil group and the ELC moisture regime guidelines considered soil texture grouping for moisture regime classification. The methods for modelling are discussed in the following chapter, and results for the prediction maps are in Chapter 5.

Chapter 4

METHODS

The execution of the digital soil mapping analysis for this thesis research is presented in this chapter. Specifically, the treatment of the LiDAR data is discoursed. The set of environmental covariates utilized as predictors, covariate extraction and layer integration, and modelling techniques for soil property classification are discussed. Also discussed are the metrics for evaluating the models, and how prediction maps and uncertainty measurements were obtained. Formulae for the vegetation covariates derived from LiDAR are presented, as well as the Cohen's kappa for accuracy and entropy calculation utilized to compute the uncertainties in predictions.

4.1 Derivation of Environmental Covariates

A combination of multi-sourced remotely sensed data was considered when implementing the environmental covariates. Environmental covariates were generated from multispectral satellite imagery, LiDAR data, L-band SAR imagery, and aeromagnetic survey data. For the classification modeling with the FRI data, 76 predictors were utilized in the full set as shown in Table 2. These predictors corresponded to a broad set of environmental covariates that relate to the climatic, vegetation, topographic relief and parent material classes of soil formation factors.

Predictor	Spatial Resolution
DEM (derived from LiDAR)	
Aspect	30 m
Convergence Index	30 m
Elevation	30 m
General Curvature	30 m
Hillshading	30 m
Mid-Slope Position	30 m
Multiresolution Ridge Top Flatness (MRRTF)	30 m
Multiresolution Valley Bottom Flatness (MRVBF)	30 m
Plan Curvature	30 m
Profile Curvature	30 m
SAGA Topographic Wetness Index (SAGA TWI)	30 m
Sky View Factor	30 m
Slope	30 m
Slope Height	30 m
Slope Length	30 m
Standardized Height	30 m
Stream Power Index	30 m
Terrain Ruggedness Index (TRI)	30 m
Terrain View Factor	30 m
Topographic Wetness Index (TWI)	30 m
Total Curvature	30 m
Valley Depth	30 m
Visible Sky	30 m
WorldClim V1Bioclim	
Annual Mean Temperature	30 arcsec
Annual Precipitation	30 arcsec
Mean Temperature of Warmest Quarter	30 arcsec
Precipitation Seasonality	30 arcsec
Temperature Annual Range	30 arcsec

Predictor	Spatial	Spectral
	Resolution	Resolution
PALSAR		
HH 2017	30 m	HH Polarization
Distance to Water Bodies	30 m	HH Polarization
LiDAR		
Canopy Height Model (CHM)	10 m	1.064 μm
Gap Fraction	10 m	1.064 μm
Ontario FRI		
Black Spruce Indicator	Site	
Balsam Fir Indicator	Site	
Landsat*		
B1 (Winter, May, Summer, Autumn) 2017	30 m	0.435-0.451 μm
B2 (Winter, May, Summer, Autumn) 2017	30 m	0.452-0.512 μm
B3 (Winter, May, Summer, Autumn) 2017	30 m	0.533-0.590 μm
B4 (Winter, May, Summer, Autumn) 2017	30 m	0.636-0.673 μm
B5 (Winter, May, Summer, Autumn) 2017	30 m	0.851-0.879 μm
B6 (Winter, May, Summer, Autumn) 2017	30 m	1.566-1.651 μm
B7 (Winter, May, Summer, Autumn) 2017	30 m	2.107-2.294 μm
B10 Summer 2017	100 m	10.60-11.19 μm
B11 Summer 2017	100 m	11.50-12.51 μm
NDVI (Winter, May, Summer, Autumn) 2017	30 m	0.636-0.673 μm, 0.851-0.879 μm
NDWI (Winter, May, Summer, Autumn) 2017	30 m	0.851-0.879 μm, 1.566-1.651 μm
NRCan**		
Gravity Anomaly	2 km	
Magnetic Residual	200 m	

* Surface reflectance bands; season were Winter (January, February, March), May, Summer (June, July) & Autumn (Sep. 10 - Oct. 10). For May & Autumn, median surface reflectance was obtained over same period over 5 year period (2015-2019). Landsat-8 imagery was utilized

** First vertical derivative of each attribute was also included

Table 2 - Environmental covariates utilized for digital soil mapping models.

The LiDAR data was obtained via Land Information Ontario (LIO) from the Ontario MNRF. LiDAR point-cloud data was obtained for the District of Cochrane, which was collected from airborne campaigns during autumn 2016 and spring 2017. This LiDAR data for the GCB corresponded to 6085 km^2, and was obtained at an average point density of 8 retrievals per m^2 (Airborne Imaging, 2018). A LiDAR-derived digital elevation model (DEM) with pixels of 30 m resolution was generated. For each pixel, The DEM was calculated from the bottommost LiDAR point-cloud data, and correspondingly a digital surface model was calculated from the topmost LiDAR point-cloud data.

A provincial DEM of 30 m spatial resolution for 2013 (i.e. version 3.0) from the Ontario Ministry of Natural Resources and Forestry (MNRF) was originally utilized to derive topographic covariates. The files for this provincial DEM were downloaded from Ontario GeoPortal, which is maintained by Infrastructure Ontario. A combination of a RADAR-based digital surface model, Ontario base mapping, and digital terrain model (DTM) points and contours (Ontario Ministry of Natural Resources, 2013) were utilized to generate this DEM. The DEM files that the provincial DEM comprise of, corresponded to tiles for MNRF's categorization of both northern and southern Ontario. As the study areas straddled the boundary between the north and south files, tiles from both sections had to be stitched together to generate a DEM for the GCB by means of the Mosaic tool in ArcMap. The comparison of the LiDAR-derived DEM versus the provincial DEM is shown in Figure 2. From this figure, one can see that the LiDAR-derived DEM captures finer details and allows better precision; the provincial DEM only had elevations to within a meter precision. For reasons of higher accuracy and precision, it was decided to utilize the LiDAR-derived DEM, rather than the provincial DEM, for the generation of topographic covariates.

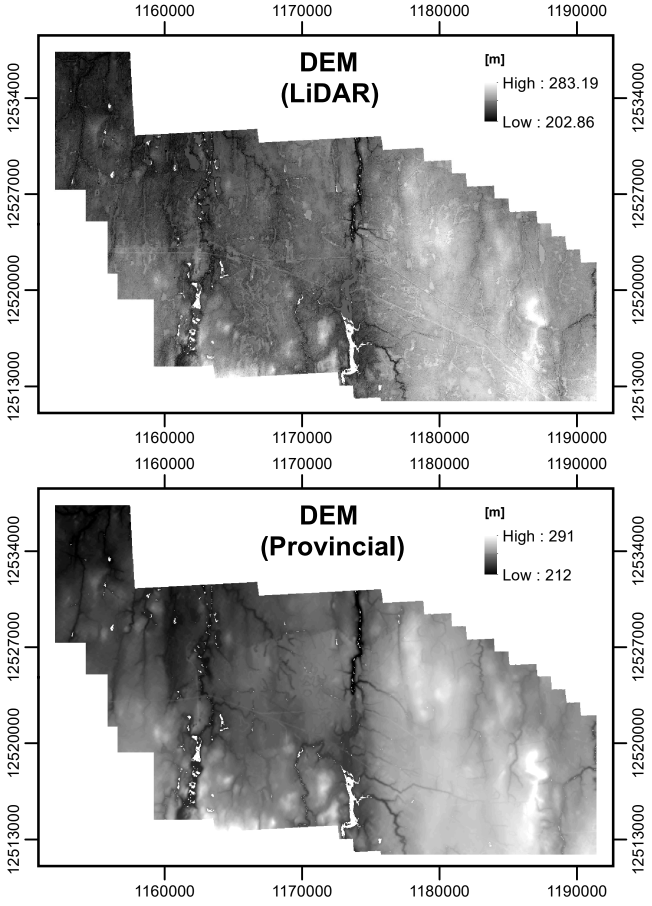


Figure 2 - DEM derived from LiDAR (Top) versus provincial DEM (Bottom) for ARF region.

System for Automated Geoscientific Analyses (SAGA) GIS version 7.6.3 software (SAGA Development Team, 2020) was utilized to generate topographic covariates from the LiDAR-derived DEM. A variety of topographic covariates were computed using the

Terrain Analysis tools. These included covariates corresponding to channels (valley depth), hydrology (topographic wetness index (TWI), SAGA TWI, slope length, stream power index), visibility (hillshading, view factors) and morphometry (convergence index, slope, aspect, curvatures, relative heights and slope positions, terrain ruggedness index (TRI), multiresolution indices of flatness). In total, 23 separate topographic covariates were generated. The full set of topographic covariates is listed in Table 2.

Climatic data for temperature and precipitation were obtained from the WorldClim Climatology version 1 model generated by the University of California at Berkeley (Fick and Hijmans, 2017; Hijmans et al., 2005). These corresponded to climatology values at a spatial scale of 30 arcsec (i.e. roughly 1 km). Mean annual temperature, the mean temperature for the warmest quarter of the year (i.e. summer) and annual temperature range were the temperature attributes utilized. For precipitation, annual precipitation and precipitation seasonality were used. Brightness temperatures for shortwave thermal bands from Landsat-8 for summer of 2017 were also utilized as climatic covariates as these were measured in Kelvin; the B10 (10.60 – 11.19 um) and B11 (11.50 – 12.51 um) bands were utilized. All these climatic covariates were downloaded as rasterized imagery in GeoTIFF format from Google Earth.

Environmental covariates corresponding to vegetation cover were obtained from surface reflectance from Landsat optical imagery provided by USGS (United States Geological Survey, 2019) as downloaded from Google Earth. Detailed vegetation covariates from

LiDAR data were also derived, but will be discussed in the subsequent section. Median surface reflectance values were calculated from cloud-free scenes (i.e. determined as less than 1% cloud cover for total scene). The median surface reflectance values were computed for 4 separate periods of the year: winter (January to March), spring (May), summer (June and July) for peak-vegetation, and then autumn (September 10th to October 10th). In order to obtain sufficient cloud-free imagery for spring and autumn scenes, imagery for a 5 year period was utilized (i.e. 2015-2019). Landsat-8 imagery was utilized for the most scenes (i.e. after 2012) and had a spatial resolution of 30 m. The specific bands utilized, with corresponding spectral resolutions, are listed in Table 2. The normalized difference vegetation index (NDVI) and normalized difference water index (NDWI) were calculated from the corresponding bands of the Landsat surface reflectance. For the NDVI, the calculation involved measurements of surface spectral reflectance (ρ) for the near infrared (NIR) and red (R) visible light wavelengths, whereas for NDWI the near infrared and shortwave infrared (SWIR) wavelengths were applied. These spectral bands for the Landsat-8 surface reflectance corresponded to B4 (0.636 – 0.673 um) for red, B5 (0.851 – 0.879 um) for near infrared, and B11 (1.566 – 1.651 um) for shortwave infrared wavelengths, respectively. The NDVI was calculated as

$$NDVI = \frac{Near\ Infrared - Red}{Near\ Infrared + Red} = \frac{NIR - R}{NIR + R} = \frac{\rho_{B5} - \rho_{B4}}{\rho_{B5} + \rho_{B4}}. \qquad \text{Eq. 4-1}$$

Similarly, the NDWI was calculated as

$$NDWI = \frac{Near\ Infrared - Shortwave\ Infrared}{Near\ Infrared + Shortwae\ Infrared} = \frac{NIR - SWIR}{NIR + SWIR} = \frac{\rho_{B5} - \rho_{B6}}{\rho_{B5} + \rho_{B6}}. \qquad \text{Eq. 4-2}$$

Tree species indicators for black spruce and balsam fir were also utilized as environmental covariates. This information of where these tree species were each dominant was recorded at the site level in the FRI data. There can be strong associations between tree species type and soil group and nutrients (Jenny, 1941), so such data can be of substantial relevance to vegetation soil formation factors. Black spruce was the most dominant tree species for the sites throughout the GCB region, with balsam fir commonly as either a secondary tree species or even as the most dominant species.

L-band SAR imagery obtained from the JAXA ALOS PALSAR (Shimada et al., 2014) was utilized as covariates for understory and surface moisture. This imagery was obtained from Google Earth for the HH-polarization for the year 2017. The L-band HH-polarization SAR 16-bit digital number (DN) imagery for the ARF region is shown in Figure 3. Wetter regions show up as darker, where the darkest intensities correspond to water bodies. Topographic covariates for distances to bodies of water were also derived in part from the L-band SAR imagery, by means of the Euclidean Distance and Line Density tools of the Spatial Analyst Tools toolkit of Esri ArcMap software.

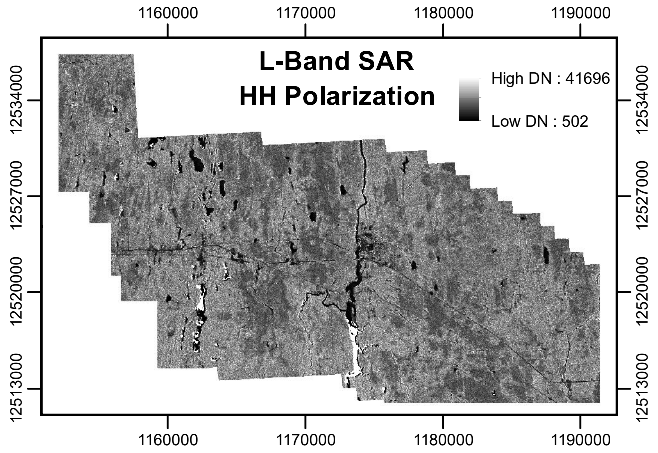


Figure 3 - L-band SAR HH-polarization imagery for ARF region for 2017.

Geophysical data in the form of aeromagnetic surveys obtained from Natural Resources Canada via the Canadian Aeromagnetic Data Base were utilized as covariates for parent material. These included airborne gravity anomaly at 2 km resolution for 2016, and aeromagnetic residual total field at 200 m resolution for the Earth's magnetic field for November 2018. Vertical derivatives for gravity anomaly and aeromagnetic residual total field were also obtained and utilized. For parent material, a bedrock geology vector file was obtained from Natural Resources Canada, which was subsequently rasterized. From this rasterized file, binary variables for the 5 bedrock categories (glaciofluvial ice-contact deposits, bedrock, till, glaciolacustrine deposits, and organic deposits) were applied as covariates.

4.1.1 LiDAR Data Analysis for Covariates

As mentioned earlier, the novelty for digital soil mapping for this research was the application of detailed LiDAR-derived vegetation attributes as LiDAR retrievals were available for the study areas. The bounds for the LiDAR data available roughly correspond to the bounds of the study areas in Figure 1. A total of 6612 files of LiDAR point-cloud data were uncompressed from .LAZ to .LAS format using LASzip version 1.0.0.1 software, which when uncompressed corresponded to 1.7 TB of LiDAR data. Each one of these LiDAR point-cloud data files corresponded to roughly 1 km^2, specified using the NAD 1983 UTM Zone 17N coordinates. Scripts written in Python version 3.7.2 read in each file one at a time, which were processed in daily batches of a few hundred files on an Intel Core i7-7700 CPU 3.60 GHz processor with 8 GB RAM over the course of a few weeks. SQL (structured query language) was implemented via the sqlite3 version 2.6.0 module to perform the queries for computing the LiDAR-derived covariates. Pandas version 0.24.2 in Python was utilized to stack the SQL query lists into matrices corresponding to values for easting (x) and northing (y) coordinates, which were then outputted as ASCII tables which were then inputted and mosaiced together in ArcMap version 10.6 software. These detailed vegetation covariates were first calculated for a cell size of 10 m, which were then later resampled to 30 m spatial resolution to match the resolution of the optical imagery. Canopy height model (CHM) was calculated from the LiDAR data as the difference between digital surface model and digital elevation model.

For each area (originally 10 m by 10 m), gap fraction was calculated as

$$Gap\ Fraction = \frac{Number\ of\ Returns\ with\ Only\ 1\ Return}{Total\ Number\ of\ Returns}. \qquad \text{Eq. 4-3}$$

The covariate layer maps for CHM and gap fraction for the ARF region are shown in Figure 4 as examples.

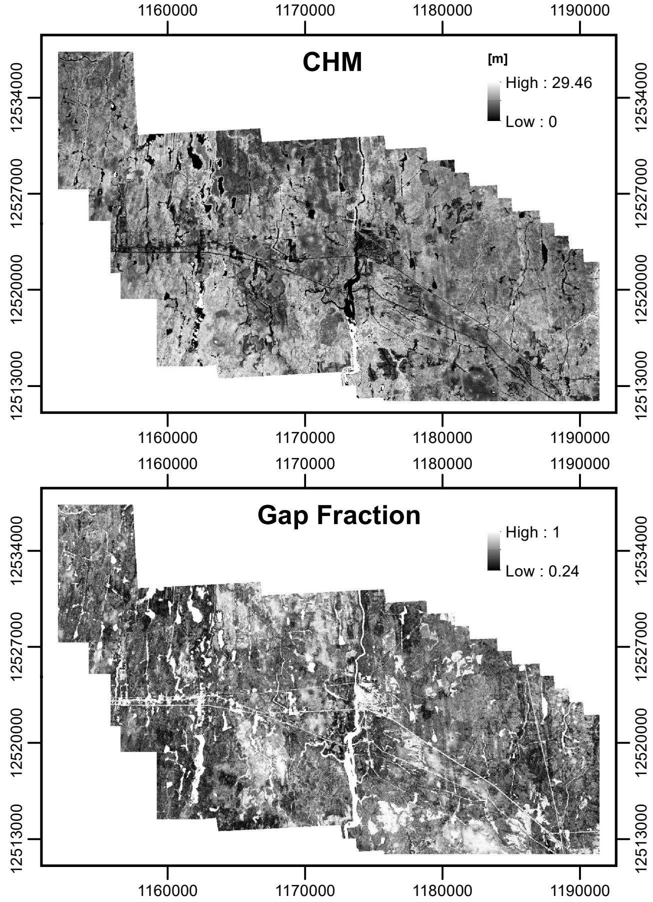


Figure 4 - Canopy height model (CHM) and gap fraction for ARF region.

From the CHM map, one can discern the peatland areas which have shallower tree heights (mostly restricted to stunted black spruce), as well as water bodies (i.e. minimal to no canopy height). From the gap fraction map, one can see that the areas with highest gap fractions correspond to the regions of shallowest canopy height, that is the regions where the overwhelming majority (if not all) LiDAR retrievals had only one return.

4.2 Modelling

For data preparation, different predictor layers of the covariates, as well as the target variable soil data layers, were each reprojected to the NAD 1983 Lambert Conformal Conic projection. This projection specifies horizontal (x) and vertical (y) offsets in terms of false easting and false northing, respectively, so as to have x and y coordinates denoted in meters. Once transformed, each layer was resampled to 30 m spatial resolutions for both x and y directions, with the cell sizes set to align at common bounds. These resampled layers were then clipped to polygon vector files of the boundaries of the targeted study areas, so that pixels for each predictor layer aligned exactly with pixels in the other layers. Each layer was then outputted as an ASCII table.

Pixels for the soil sampling locations were also pinpointed, which were utilized for the target variable layers corresponding to the soil properties. All of pixels for the soil property layers outside of the sampling locations were set as no-data, which allowed the identification of only pixels corresponding to the soil samples. A script was written in Python to load each ASCII table for each predictor, and restack the corresponding matrices into long arrays. An SQL query was implemented to intersect and merge all layers, those corresponding to the soil properties and environmental covariates based upon the northing and easting coordinates of the soil sampling. This query was applied separately to each targeted study region, where the resulting table with all the predictor and target variable data for each soil sampling site was then exported as a CSV file. The

CSV files for each region were then combined, which was then the data table inputted into R for model fitting.

Since RF and SVM models tend to yield the highest accuracy for both classification problems for soil groupings (Brungard et al., 2015; Heung et al., 2016), it was decided to apply RF and SVM approaches. Both linear SVM and SVM with radial basis functions were employed, as linear SVM can have competitive accuracies to radial-basis SVMs (Brungard et al., 2015). An overview of these modelling approaches is discussed in the subsequent paragraphs.

The RF approach works by using an ensemble of decision trees that each consist of nodes and leaves, where usually at least a minimum of a few hundred to a thousand decision trees are utilized for the RF in order to obtain stability with results. For each decision tree of a RF, each node corresponds to a randomized bootstrap sample of training data, where node-splitting decisions determined from a randomized subset of predictors aim to maximize within-node homogeneity and between-node heterogeneity (Heung et al., 2016). Decisions corresponding to the dependent variable are attained at the leaves, which are the terminal nodes of the decision trees. The number of randomly selected predictors variables to be considered for each node must be specified, and is prompted by the need for each tree to be different so to make an ensemble of trees beneficial (Sage et al., 2020). For classification, the RF ensembles the decision tree results and outputs the class category that has the maximum votes (i.e. the mode) among the individual decision trees (Kulkarni et al., 2014).

The SVM classifier works by optimally splitting the various classes of a feature space by means of hyperplanes (Wang et al., 2018). A linear SVM uses a linear model to determine the class boundaries in the feature space by means of optimizing hyperplanes so that there are maximum gaps between classes. Radial-basis SVM apply a kernel trick by utilizing radial basis expansions to transform feature space into a higher dimensional space, by which a linear model is applied to determine the class boundaries (Heung et al., 2016). To prevent the SVM from over-fitting on training data with a hard margin, a cost parameter is considered so as to control the complexity of the decision boundary (Kuhn, 2008) and relax the SVM fitting and allow a soft margin in order to reduce variance. Radial-basis SVMs are fitted with a kernel corresponding to a Gaussian exponential function, and hence require an extra kernel width (sigma) parameter as a scale function to be specified (Kuhn, 2008).

The caret package for classification and regression training (Kuhn, 2008) in R (R Core Team, 2012) was used for fitting the RF and SVM models. It is worth mentioning that the randomForest package (Liaw and Wiener, 2002) and the e1071 package of miscellaneous statistical functions (Meyer et al., 2019) for SVM models in R were also employed as comparison. Similar results were obtained between the caret package and the respective R packages for each modelling approach. The results contained in this thesis were obtained from the caret package. RFs were fitted with the number of trees (n_{tree}) specified set to 1000, with the number of randomly selected predictor variables at each node (m_{try}) set to 10. The m_{try} was set to approximately the square root of the number of columns (predictors) in the training dataset as suggested in randomForest package documentation for classification modelling (Breiman et al., 2018). Were et al. (2015) used a m_{try} set to 12, and Heung et al. (2016) used m_{try} set to either of 9 or 12. The default setting for fitting a random forest is 500 trees in the caret package, but to ensure stability of results it is recommended to use 1000 trees for training (Were et al., 2015). The radial-basis SVMs were fitted with the penalty parameter (cost) ranging from 0.25, 0.5, 1, 2, 4, 8, 16, 32, 64 and 128, with the kernel width parameter (sigma) ranging from 0.005 to 0.05 as what was comparably done by Wang et al. (2018). Linear SVMs were fitted with the tuning cost parameter fixed at the constant default value of 1 (Kuhn et al., 2020; Meyer et al., 2019).

The soil data was split using stratified random sampling for each soil target variable into model training and validation sets using a 80:20 ratio respectively (Heung et al., 2016), where a 10-fold cross validation with 3 repeats was implemented for model training. This implementation of cross validation is commonly utilized for the selection of model parameters for classification problems (Kuhn et al., 2020). Separate models for each region were independently fitted. A regional model for the GCB, comprising of all sites for Hearst, GCF and ARF, was also implemented. It was anticipated that a model consisting of all 3 regions would be warranted, due to the homogeneity of landforms and climate for the regions, and that the distance between corresponding sites did not exceed more than a couple hundred kilometers (see Figure 1).

Variable importance was determined for each region in order to identify environmental covariates of greater significance for each target variable. This form of feature selection was implemented since the environmental covariates may not contribute equally to the modelling of soil properties. The variable importance was determined from the respective RF models. For determining the variable importance of a RF using the caret package (Kuhn, 2008), first the prediction accuracy of the out-of-bag portion of data is calculated for each tree of the RF (Kuhn, 2007). Afterwards, each predictor variable is permuted, and the same prediction accuracy calculation as done before is performed. The difference between these two prediction accuracies are then computed for each tree of the RF, and then are subsequently averaged and normalized by the standard error.
Prediction accuracy can be based on the Gini index, which is a measure of node impurity for the nodes of trees for the RF (Sage et al., 2020). Splits are performed on the nodes such as to decrease the node impurity, that is increase homogeneity among the leaves of trees. The mean decrease in Gini is based on the total decrease in node impurity, where higher variable importance for a predictor variable is indicated by a corresponding higher mean decrease in Gini.

Variable importance from the RF models was utilized to generate reduced sets of predictors, which were subsequently also applied to the SVM models, due to the lack of a model-dependent method of determining variable importance for SVMs (Kuhn et al., 2020). A similar approach was applied by Brungard et al. (2015), which used a covariate selection determined by RF models to generate reduced sets of predictors which were applied to other machine learning approaches, including SVMs. Models were initially fitted with a larger set (i.e. 76) of covariates relating to multiple classes of soil formation factors, and then models were generated from corresponding reduced sets comprising of the most important covariates. The narrower selection of predictors was decided with the objective of reducing multi-collinearity among the predictors, as well as to improve model accuracy (Nussbaum et al., 2018). In total, there were 12 separate reduced sets of predictors generated; one for each target variable (soil texture class, calcareous substrate testing to 10% HCl, ELC moisture regime) for each study area (Hearst, GCF, ARF, and GCB).

The metrics for assessing the models utilized were the percent correct classification (PCC) as the accuracy score, and Cohen's kappa statistic. The PCC is the fraction of predictions that are correct, which varies from 0 to 1, with scores of 1 denoting perfect prediction and scores of 0 meaning complete disagreement. Cohen's kappa statistic, also known as just the kappa score, takes into account by-chance agreement; it considers the chance of misclassification with prediction (Cohen, 1960; Simon, 2006; Warrens, 2011). The kappa score, here denoted as κ, was calculated as

$$\kappa = \frac{p_o - p_c}{1 - p_c}\,. \qquad \text{Eq. 4-4}$$

Here p_o is the observed probability agreement (which can be taken as PCC), and p_c is probability due to by-chance agreement. The by-chance agreement probability (Simon, 2006; Warrens, 2011) can be calculated as

$$p_c = \sum_{k=1}^{n} \frac{q_{k,o}}{N} \frac{q_{k,p}}{N} . \qquad \text{Eq. 4-5}$$

The number of classification categories is denoted by n, and N is the number of observations which in this case would correspond to the number of soil sampling locations for the verification data set. The integer $q_{k,o}$ corresponds to the number of observations for category k, and $q_{k,p}$ corresponds to the number of predictions for category k. Cohen's kappa is bounded between 0 and 1, with values of 0 when obtained agreement equals by-chance agreement, and a maximum of 1 obtained when there are no disagreements (Cohen, 1960). Brungard et al. (2015) wrote that kappa scores between 0.4 and 0.8 demonstrate moderate agreement, whereas scores less than 0.4 would signify poor agreement.

4.3 Prediction & Uncertainty Maps

Minasny et al. (2019) wrote in a review for digital soil mapping of peatlands that there is a need for the uncertainty of estimates for corresponding digital soil maps and products. The purpose of the prediction maps was to determine patterns in regards to the predicted soil properties for the study areas. Prediction maps were obtained in R by applying fitted models to the predictor layers corresponding to ASCII text files for the whole study area. These ASCII files were first inputted as matrices which were then subsequently restacked into long arrays to work as input vectors for the corresponding predict function in R. Once the fitted models were applied to the predictor arrays, the predictions were restacked from long arrays back into matrices, which were then exported as text files.

To determine the uncertainty with the prediction maps, entropy scores were calculated to determine a measure of how much a model concentrates predictions to specific categories (Heung et al., 2017; Roulston and Smith, 2002; Zhu, 1997). The entropy, denoted by H, was calculated for each pixel as

$$H = \frac{-1}{\ln(n)} \sum_{k=1}^{n} p_k \ln(p_k) \, , \qquad \text{Eq. 4-6}$$

where the probabilities p_k for categorial assignment (i.e. voting) for class k out of total of n different classes were determined from the models. The probabilities p_k were generated when the models were fitted by specifying the appropriate probabilities argument flag, which resulted in the training function also returning a data frame with the probability columns for each class (Kuhn, 2008). These probabilities were calculated for the RF models as the portion of the total number of trees that voted for the respective category (Sage et al., 2020). For SVMs, a secondary Platt model was utilized to generate these probabilities (Kuhn, 2013), which corresponded to posterior probabilities approximated by a sigmoid function (Lin et al., 2007; Platt, 2000). Note that for each pixel that the combined probabilities for all n different classes will sum to 1, with the individual class probabilities p_k varying between 0 and 1. The entropy, which has also been called ignorance uncertainty (Heung et al., 2017) is bounded between 0 and 1, where higher values indicate more uncertainty. This can be seen from the entropy equation, where equal probabilities among all category predictions would lead to an entropy of 1, which would mean the greatest modelling uncertainty since there is no discernment of prediction made. At the other extreme of complete certainty, with a probability of one being assigned to one class (and consequently other classes having probabilities of zero) would result with an entropy of 0.

Chapter 5

RESULTS

5.1 Important Covariates

Variable importance among the model covariates were resolved for the prediction of soil texture, calcareous substrate reaction to hydrochloric acid, and ELC moisture regime, respectively. The variable importance were determined from the RF models fit on the full set of predictors (i.e. 76 predictors) by means of the caret package in R (Kuhn, 2007). Plots for variable importance in the prediction of the three soil parameters for the ARF region are shown in Figure 5. Variable importance plots for other study areas (Hearst, GCF and GCB) are shown in Appendix A (see Figure 12). Due to the similarities of results, it was decided to just discuss variable importance results for the ARF region. The results for only the ARF are shown because this was the region that had the highest density of sampling locations to study area (see Table 1) which later were shown to have the highest modelling accuracies (which will be discoursed in the following section). For all the regions, variable importance indicated vegetation covariates to have the highest variable importance. In general, NDVI for the summertime, tree species indicator for black spruce, and CHM had among the highest variable importance. Topographic covariates were also important for the Hearst study area.

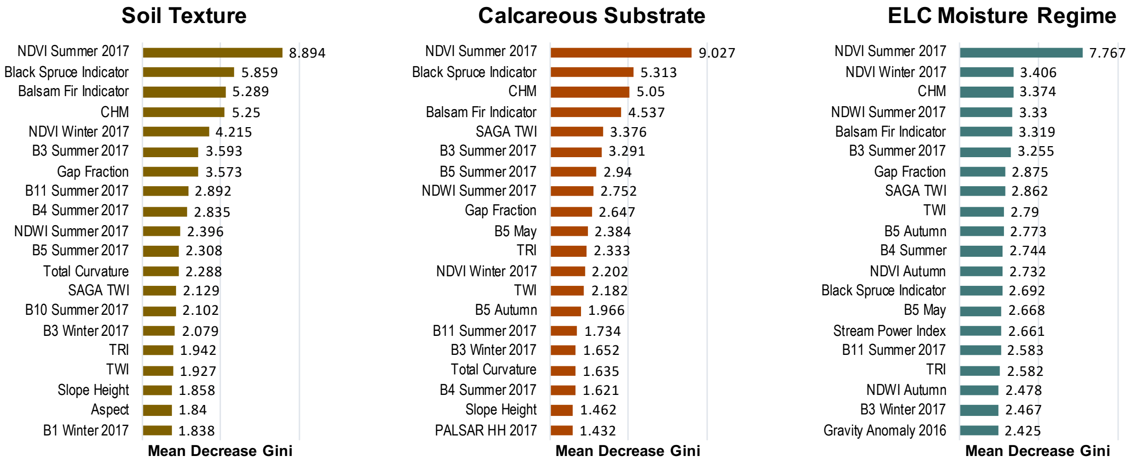


Figure 5 - Variable importance for models for ARF region.

Examining the results from Figure 5, there are many similarities in the set of important covariates between the three soil properties. Consistently, one can see that environmental covariates corresponding to vegetation had the highest variable importance. The NDVI for the summertime had the highest variable importance for all the soil property models for the ARF. Indicator variables for black spruce and balsam fir also had high variable importance. The detailed vegetation covariates derived from LiDAR had high variable importance. One can see that CHM had among the 3rd to 4th highest variable importance, and gap fraction was in the top 10 important variables. Optical imagery for the summertime, specifically NDVI, and the channels B3, B4, B5 and B11 had high variable importance as well. The TWI, SAGA TWI, TRI and total curvature had the highest variable importance among the topographic covariates. The top 20 predictors were sufficient for capturing the bulk of information necessary for modelling, as the variable importance tapered off after the first 10 or so most important variables.

5.2 Model Accuracies

The accuracies for model evaluation of soil texture class for all study areas and approaches are reported in Table 3. As can be seen, the accuracies are fairly consistent between models fitted on the full set and those on the reduced sets, with slight improvements with accuracies on the reduced sets for some models. In general, RFs had the highest accuracies, whereas linear SVMs had the lowest accuracies, though the differences in accuracies between the approaches usually did not vary by more than 0.05 with the few exceptions. Lowest modelling accuracies were obtained for the Hearst region, with highest modelling accuracies obtained for the ARF. The second most accurate models were for the GCB, which is all sites considered together. Among the reduced set of predictors, both the ARF and GCB had modelling accuracies greater than 0.7 and kappa scores greater than 0.4, which demonstrate good modelling accuracy for digital soil mapping (Nussbaum et al., 2018).

Model Accuracies From Full Set of Predictors	Soil Texture Class [Peat / Loamy / Clayey]							
	Hearst		Gordon Cosens Forest (GCF)		Abitibi River Forest (ARF)		All Regions (GCB)	
	Accuracy	Kappa	Accuracy	Kappa	Accuracy	Kappa	Accuracy	Kappa
Random Forest (RF)	0.53	0.23	0.61	0.31	0.78	0.61	0.67	0.42
Support Vector Machine Linear-basis (SVM Linear)	0.53	0.26	0.57	0.30	0.64	0.44	0.58	0.29
Support Vector Machine Radial-basis (SVM Radial)	0.37	0.04	0.57	0.24	0.76	0.58	0.67	0.42

Model Accuracies From Reduced Sets of Predictors	Soil Texture Class [Peat / Loamy / Clayey]							
	Hearst		Gordon Cosens Forest (GCF)		Abitibi River Forest (ARF)		All Regions (GCB)	
	Accuracy	Kappa	Accuracy	Kappa	Accuracy	Kappa	Accuracy	Kappa
Random Forest (RF)	0.50	0.18	0.61	0.34	0.78	0.62	0.67	0.44
Support Vector Machine Linear-basis (SVM Linear)	0.57	0.31	0.58	0.28	0.76	0.58	0.67	0.42
Support Vector Machine Radial-basis (SVM Radial)	0.50	0.19	0.61	0.31	0.75	0.55	0.67	0.42

Table 3 - Modelling accuracies for soil texture.

The modelling accuracies for calcareous substrate reaction to 10% HCl for modelling approaches and study areas are reported in Table 4. There were comparable accuracies between models fitted on the full set, versus the corresponding reduced sets of predictors. In terms of modelling approaches, all three approaches had fairly consistent accuracies among them, and none appeared to dominate over the others for all regions. As with the results for soil texture models, the Hearst region had the lowest modelling accuracies, with the ARF having the highest accuracies. The GCB (i.e. all sites considered together) had the second highest accuracies. Both the ARF and GCB had models for all approaches which had kappa greater than 0.4, indicating moderate agreement, and accuracies greater than 0.7.

Model Accuracies From Full Set of Predictors	Calcareous Substrate [n / k]							
	Hearst		Gordon Cosens Forest (GCF)		Abitibi River Forest (ARF)		All Regions (GCB)	
	Accuracy	Kappa	Accuracy	Kappa	Accuracy	Kappa	Accuracy	Kappa
Random Forest (RF)	0.48	0.02	0.64	0.22	0.82	0.63	0.73	0.46
Support Vector Machine Linear-basis (SVM Linear)	0.48	0.03	0.64	0.26	0.72	0.43	0.69	0.38
Support Vector Machine Radial-basis (SVM Radial)	0.48	0.01	0.67	0.30	0.82	0.63	0.70	0.39

Model Accuracies From Reduced Sets of Predictors	Calcareous Substrate [n / k]							
	Hearst		Gordon Cosens Forest (GCF)		Abitibi River Forest (ARF)		All Regions (GCB)	
	Accuracy	Kappa	Accuracy	Kappa	Accuracy	Kappa	Accuracy	Kappa
Random Forest (RF)	0.48	0.04	0.74	0.45	0.75	0.50	0.71	0.43
Support Vector Machine Linear-basis (SVM Linear)	0.55	0.10	0.72	0.40	0.85	0.70	0.73	0.46
Support Vector Machine Radial-basis (SVM Radial)	0.58	0.15	0.66	0.27	0.82	0.63	0.70	0.41

Table 4 - Modelling accuracies for calcareous substrate reaction to 10% HCl. Note that 'k' denotes a fizzing reaction, whereas 'n' denotes no fizzing reaction.

Results for the ELC moisture regime models for the study areas and modelling approaches are summarized in Table 5. There was a slight improvement with accuracies when considering the reduced sets of predictors, which can likely be attributed to a decrease of multi-collinearity and lack of noisy covariates that would impact accuracy (Nussbaum et al., 2018) for models formulated from the reduced predictor sets. RF models tended to have the highest accuracies, followed by the radial-basis SVMs, with linear SVMs having the lowest accuracies. The ARF region had the most accurate models, followed by the GCF, then Hearst and then the GCB on the whole. For exact category matching, the modelling accuracies were not high, but when considered for within one ordinal category of precision (e.g. ELC moisture regime rating of 4 considered accurate if predicted as either 3, 4 or 5) then accuracies grossly improved. Furthermore, slight improvements in accuracies were gained when considered for within 2 ordinal categories of precision (e.g. ELC moisture regime rating of 4 is considered accurate if predicted as either 2, 3, 4, 5, or 6). The ELC moisture regime ratings correspond to wetness factors, where 0 corresponds to dry and 9 wet within an ordinal ranging in between. For within 1 ordinal category prediction, all models trained on the reduced sets of predictors had accuracies exceeding 0.58 and kappa exceeding 0.48, which indicates moderate agreement.

Model Accuracies From Full Set of Predictors		ELC Moisture Regime [0 / 1 / 2 / 3 / 4 / 5 / 6 / 7 / 8 / 9]					
		Exact Category Match		Out by at most 1 Category		Out by at most 2 Categories	
		Accuracy	Kappa	Accuracy	Kappa	Accuracy	Kappa
Random Forest (RF)	Hearst	0.32	0.14	0.68	0.60	0.79	0.73
	Gordon Cosens Forest (GCF)	0.30	0.09	0.57	0.45	0.62	0.52
	Abitibi River Forest (ARF)	0.43	0.28	0.71	0.64	0.81	0.77
	All Regions (GCB)	0.31	0.12	0.61	0.51	0.70	0.62
Support Vector Machine Linear-basis (SVM Linear)	Hearst	0.25	0.10	0.64	0.56	0.79	0.74
	Gordon Cosens Forest (GCF)	0.24	0.09	0.50	0.41	0.65	0.59
	Abitibi River Forest (ARF)	0.38	0.25	0.66	0.58	0.76	0.70
	All Regions (GCB)	0.26	0.13	0.54	0.46	0.65	0.58
Support Vector Machine Radial-basis (SVM Radial)	Hearst	0.21	0.03	0.57	0.47	0.68	0.60
	Gordon Cosens Forest (GCF)	0.35	0.14	0.60	0.48	0.64	0.54
	Abitibi River Forest (ARF)	0.45	0.29	0.71	0.63	0.78	0.72
	All Regions (GCB)	0.37	0.20	0.59	0.48	0.70	0.63

Model Accuracies From Reduced Sets of Predictors		ELC Moisture Regime [0 / 1 / 2 / 3 / 4 / 5 / 6 / 7 / 8 / 9]					
		Exact Category Match		Out by at most 1 Category		Out by at most 2 Categories	
		Accuracy	Kappa	Accuracy	Kappa	Accuracy	Kappa
Random Forest (RF)	Hearst	0.46	0.33	0.82	0.78	0.86	0.82
	Gordon Cosens Forest (GCF)	0.39	0.24	0.65	0.57	0.75	0.69
	Abitibi River Forest (ARF)	0.45	0.30	0.74	0.68	0.81	0.77
	All Regions (GCB)	0.35	0.18	0.61	0.52	0.70	0.63
Support Vector Machine Linear-basis (SVM Linear)	Hearst	0.32	0.18	0.61	0.53	0.75	0.70
	Gordon Cosens Forest (GCF)	0.36	0.20	0.58	0.48	0.67	0.59
	Abitibi River Forest (ARF)	0.48	0.35	0.78	0.73	0.90	0.87
	All Regions (GCB)	0.37	0.21	0.62	0.53	0.71	0.65
Support Vector Machine Radial-basis (SVM Radial)	Hearst	0.39	0.22	0.71	0.64	0.75	0.68
	Gordon Cosens Forest (GCF)	0.38	0.20	0.63	0.53	0.68	0.59
	Abitibi River Forest (ARF)	0.52	0.39	0.76	0.70	0.88	0.85
	All Regions (GCB)	0.37	0.19	0.62	0.52	0.70	0.63

Table 5 - Modeling accuracies for ELC moisture regime.

5.3 Prediction Maps

Prediction maps were generated for soil properties of each region. It was decided not to present prediction maps for study areas other than the ARF in the main section of the thesis, as the ARF had the highest modelling accuracies. Analogous structures and patterns can be ascertained from prediction maps of soil properties for the other regions, hence another reason for focusing on just the ARF region. Predictions maps for other study areas (Hearst, GCF) are presented in Appendix B.

The prediction results for soil texture classifications from the models for the ARF are shown in Figure 6. On the left hand column are the predictions for soil texture classification based upon the regional ARF model, whereas on the right column are predictions for the ARF study area based upon the model trained for the entire GCB region. Prediction results for the RF models are at the top, the linear SVMs are in the middle, and the radial-basis SVM predictions are at the bottom. From these models, one can see that predictions for areas of peat are consistent among the different models. There are differences between predictions for areas with clayey and loamy soils, in particular, the linear SVM trained just for the ARF appears to overpredict clayey soils, whereas the radial-basis SVM trained for the whole GCB underpredicts clayey soil areas. The linear SVM approach had the poorest accuracies from Table 3 for soil texture classification for the ARF and GCB, so it likely here that the linear SVM did not perform adequately when compared to the SVM based upon radial basis functions or the RF approaches. Areas of clayey soils should appear closer to Ontario Highway 11 corridor (center) and in settled regions. Overall, the RF models give the most consistent prediction maps of soil textures, where it was ascertained that RF models had the highest accuracies for model evaluation (see Table 3).

Predictions for areas with calcareous substrate reaction to 10% HCl are shown in Figure 7. As with the layout of the figure for the predictions of soil texture classification, the left hand column are predictions from the regional ARF model, whereas the right hand column are the predictions from the GCB model fitted on all regions. Inspecting the various maps, the areas predicted for the binary classification of where there are or are no reactions of calcareous substrate to acid are similar among the different approaches. The most consistency is seen between the RF models. One can note that the areas that are predicted having no reaction to 10% HCl for calcareous substrate testing are the same regions as predicted to have peat in Figure 6. This result relates to peat typically being acidic in substance (Belova et al., 2006; Vitt et al., 2009), so applying the 10% HCl would not yield a fizzing reaction; whereas calcareous soils are basic (Jalali and Moradi, 2020; Rogovska and Blackmer, 2009) and hence demonstrate a fizzing reaction when exposed to hydrochloric acid.

Results of predictions using the ELC moisture regime for the ARF are shown in Figure 8. The areas predicted for the various moisture regime ratings are fairly consistent. The left column shows predictions based upon the regional ARF models, whereas the right column shows results based upon the GCB models. For the ARF region, the linear SVMs had the best accuracies (though RF and radial-basis SVMs were comparable) as seen in Table 5. The linear SVM prediction results were similar for both the ARF and GCB trained models, and appeared to use the most variation of moisture regime categories; which is, not as large of areas were just assigned to one rating. Comparing the results of prediction maps of ELC moisture regime with those for soil texture in Figure 6, one can see that areas having the highest ELC moisture regime ratings (i.e. are wet) correspond to peat, which are the wetland areas.

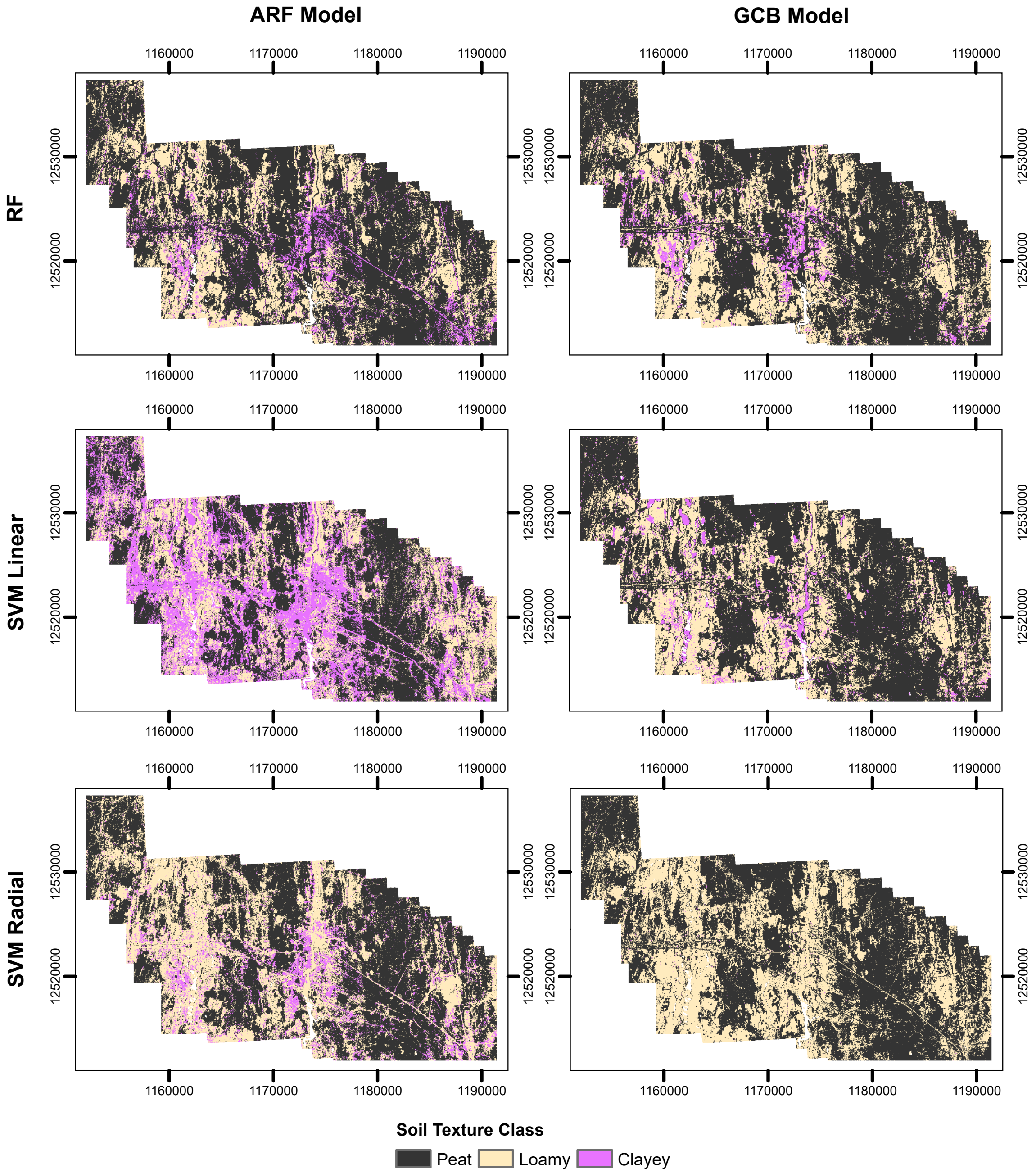


Figure 6 - Prediction maps of soil texture class for the ARF region.

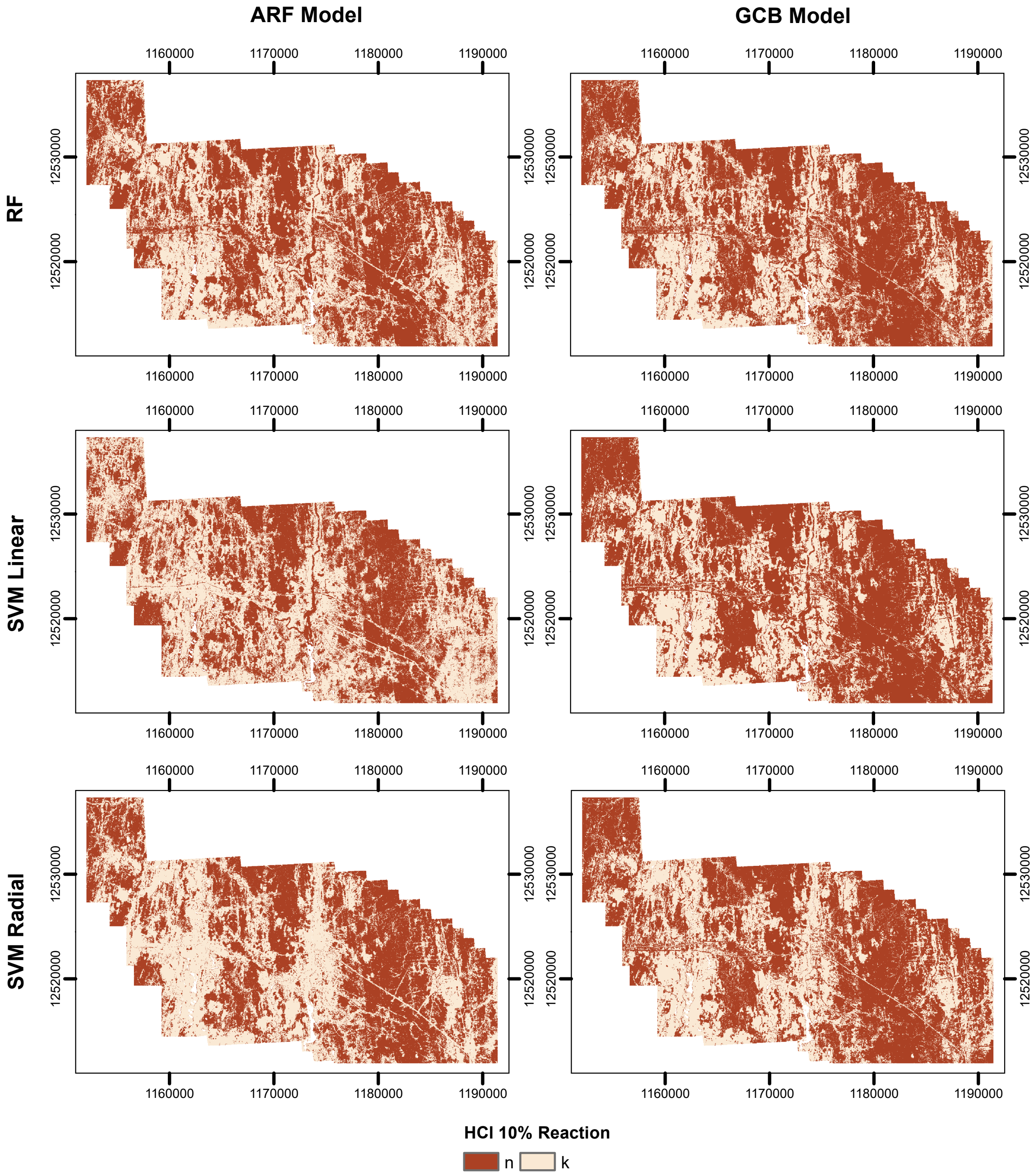


Figure 7 - Prediction maps for calcareous substrate reaction for the ARF region.

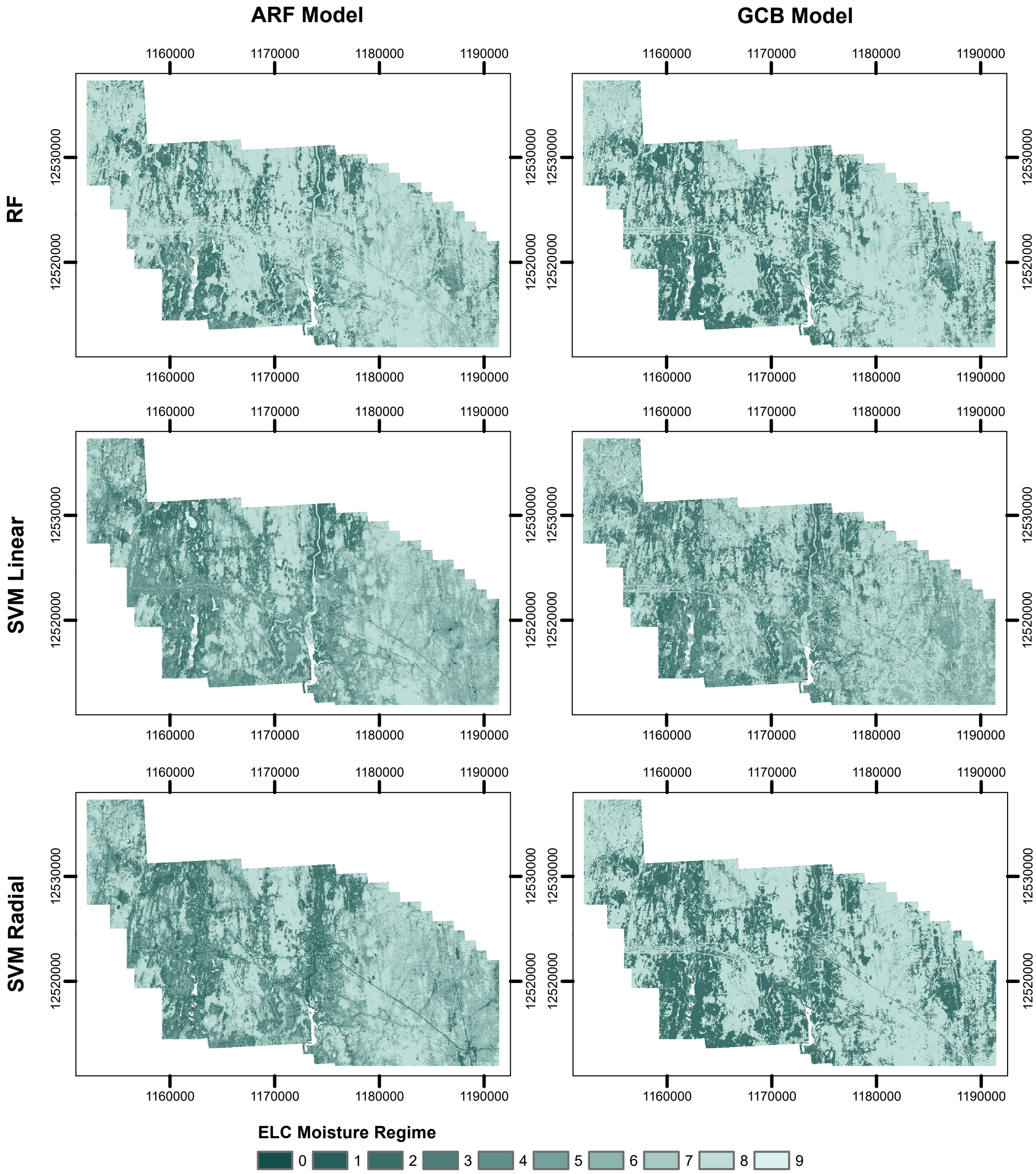


Figure 8 - Prediction maps for ELC moisture regime for the ARF region.

Entropy maps for the uncertainty of predictions for all modelling approaches were also calculated for the ARF region based upon the respective ARF models, as shown in Figure 9. Similar patterns can be recognized between the different target variables. It can be seen that the lowest modelling uncertainties were in regions of peat (see Figure 6), which also had no reaction to 10% HCl reaction for the calcareous substrate (see Figure 7). From Figure 6, it was noted that all approaches predicted essentially the same areas with peat, which corroborate the results in Figure 9.

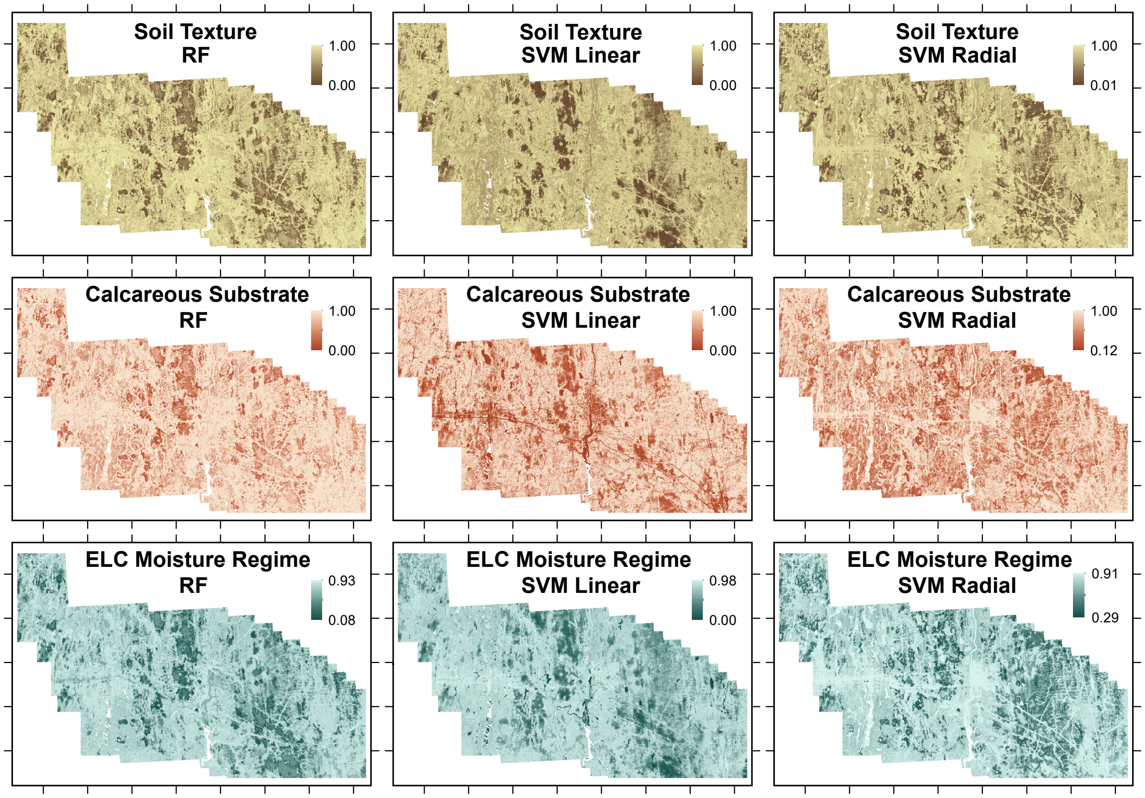


Figure 9 - Entropy maps for uncertainty of predictions for models for ARF region.

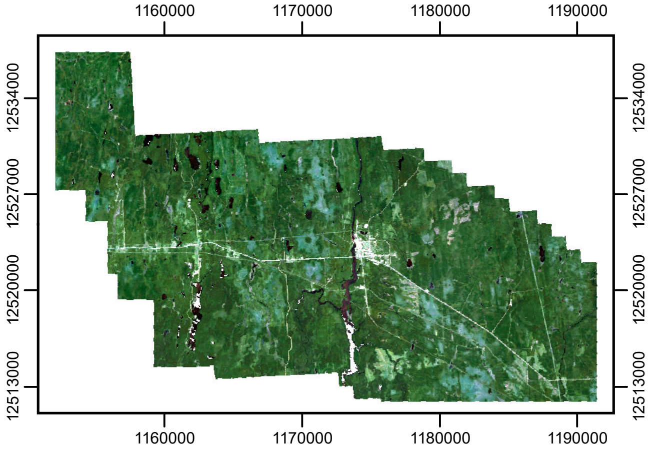


Figure 10 - True-color composite from Landsat-8 imagery for the summer of 2018 for the ARF region.

A true-color composite of the ARF study area is shown in Figure 10. This composite was generated from Landsat-8 imagery and is composed of median surface reflectance from cloud-free days from June and July of 2018. For this true-color composite image, red corresponds to band 4 (0.636-0.673 µm), green corresponds to band 3 (0.533-0.590 µm) and blue corresponds to band 2 (0.452-0.512 µm). Similar patterns can be seen in this image as predicted for Figures 6, 7, 8 & 9. The peatlands correspond to the areas of less dense vegetation, which would typically have peat for soil texture. Evidently, the peatland regions also have the wetter moisture regimes.

All in all, the results from the prediction maps from the various soil properties appear consistent across the different modelling approaches, and similar prediction patterns in terms of areas for the properties are apparent. These results are logical, as it would make sense for areas with peat to have shorter vegetation due to the stunted growth of black spruce trees (which are generally the only trees that can grow in wetlands for the GCB region). It is also rational that the areas with peat will have the highest ELC moisture regime ratings which correlate to wet soils, which is obvious for a wetland. The areas with peat tested no reaction to hydrochloric acid to calcareous substrate, which is explainable as distinctive soil texture groups will be correlated differently to calcareous substrate. All modeling approaches consistently had the lowest prediction uncertainties for the peatlands, which conveys a greater degree of confidence in prediction for those regions. Correspondingly, areas with loamy soil exist outside the wetlands (and thus have drier ELC moisture regimes), and hence throughout the forested regions with taller canopies. The regions predicted as clayey soils had the greatest modelling uncertainties, which explain the variation in predictions for clayey soil among the difference modelling approaches. One possible reason for the larger modelling uncertainties for clayey soil is that clayey soil dominates the deeper soil horizons (below 20 to 25 cm) throughout the GCB region, so sites that were sampled as clayey could have had shallower or absent top soil horizons.

Chapter 6

DISCUSSION

6.1 Modelling

Results from this research obtained higher accuracy scores than those reported in other recent studies. RF and SVM approaches for the classification of soil type have either comparable or lower reported accuracies (Brungard et al., 2015; Dornik et al., 2018; Heung et al., 2017, 2016), or lower kappa scores (Brungard et al., 2015). However, it might not be justified to directly compare studies against each other, as the context of comparison would depend on crucial factors. Such considerations could include the distribution and concentration of sampling points, the allocation of target classes among a modelled soil property, or even the physical environment or biome for which the digital soil mapping was undertaken. For peatland environments, Minasny et al. (2019) wrote that machine learning approaches can reveal spatial correlation among environmental covariates, where although correlation does not mean causation, nonetheless insights can be gained concerning the distribution and formation of different soil types. The higher the accuracy attained by models for digital soil mapping applications, the better the understanding and knowledge about soils that can be achieved.

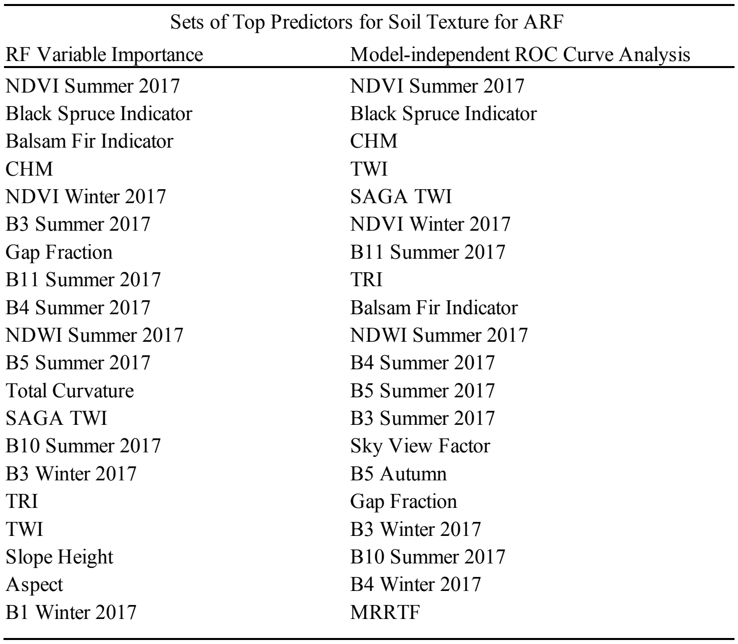
Local versus regional models were investigated for the digital soil mapping classification analysis contained in this thesis, as can be seen from the results of local models (Hearst, GCF, ARF) being compared to that for the GCB region which comprises of all 3 subregions, in Chapter 5. From the results in Tables 3, 4 & 5, some regions had better modeling accuracy than others. In general, the ARF locality had the highest modelling accuracies, followed by the GCB region on the whole. The least accurate models were for Hearst. From the comparison of the results, and the homogeneity of climate and landcover of the subregions, it can be argued that a regional model for just the GCB would be sufficient. Brungard et al. (2015) considered digital soil mapping on 3 separate study areas separated from each by at least a few hundred kilometers, with the study areas existing in the states of New Mexico, Utah and Wyoming. These study areas and sites were not combined into just one study area, as the variation and distance between the separate study areas were too great. From Figure 1, it can be seen that the maximum distance between all FRI sites for the GCB are less than a couple hundred kilometers primarily along the same latitude. Thus, the inclusion of all sites together into one targeted study area on the basis of similarity of climate and biome can be justified. The separation of sites into the groupings used was present when the soil data was obtained from the FRI, and the ARF region did have a higher density of sites sampled per area, which likely supported the improved modelling accuracies for that region.

From the prediction maps for the classifications, as shown in Figures 6, 7 & 8, similar patterns can be ascertained. These patterns should be expected, as ELC moisture regime is based in part on soil texture classification (Paloniemi, 2018). The ELC classifications correspond to different types of biotypes or habitats in northern Ontario (Pokharel and

Dech, 2011), which could be of interest to different habitats or regions of specific vegetation cover within the boreal biome. The fact that the ELC moisture regime here had a scale of 10 categories, versus 3 for soil texture or the binary calcareous substrate reaction to acid, allowed for greater specificity. For that reasoning, if only one of these soil properties were to be investigated, probably ELC moisture regime would be the best choice. However, the ELC moisture regime classification would only be relevant to within Ontario, which would limit its comparability to moisture regime classifications for other jurisdictions.

6.2 Feature Selection

A model-independent approach for feature selection using receiver operating characteristic (ROC) curve analysis (Kuhn, 2007) was also performed. For each class pair of multi-class outcomes, the class was determined from a series of cutoffs applied on each predictor. The area under the ROC curve was calculated, where variable importance was determined from the maximum area under the curve among corresponding pairs. For each binary set of class pairs of the multi-class outcomes, cutoffs were applied to each predictor to determine class from the class pairs. For each cutoff, both the proportions of true positives and true negative classifications, also denoted as sensitivity and specificity, were correctly predicted. The ROC curve was then generated from these sensitivity and specificity values, from which the area under the ROC curve was computed for each class pair. From among the pair-wise areas, variable importance was then determined from the ROC curve with the maximum area under the curve. Compared with the variable importance determined from the RF models, similar reduced sets of covariates were determined from the model-independent ROC curve analysis approach. As an example, these reduced sets are presented for soil texture for the ARF region in Table 6. The application of the reduced sets determined from the RF models for the SVMs were deemed satisfactory, as the SVMs had comparable accuracies to the RF models as shown in Tables 3, 4 & 5.



Sets of Top Predictors for Soil Texture for ARF	
RF Variable Importance	Model-independent ROC Curve Analysis
NDVI Summer 2017	NDVI Summer 2017
Black Spruce Indicator	Black Spruce Indicator
Balsam Fir Indicator	CHM
CHM	TWI
NDVI Winter 2017	SAGA TWI
B3 Summer 2017	NDVI Winter 2017
Gap Fraction	B11 Summer 2017
B11 Summer 2017	TRI
B4 Summer 2017	Balsam Fir Indicator
NDWI Summer 2017	NDWI Summer 2017
B5 Summer 2017	B4 Summer 2017
Total Curvature	B5 Summer 2017
SAGA TWI	B3 Summer 2017
B10 Summer 2017	Sky View Factor
B3 Winter 2017	B5 Autumn
TRI	Gap Fraction
TWI	B3 Winter 2017
Slope Height	B10 Summer 2017
Aspect	B4 Winter 2017
B1 Winter 2017	MRRTF

Table 6 - Top predictors listed by order for soil texture for ARF, as determined from RF model (left) and model-independent ROC curve analysis approach (right).

6.3 Environmental Covariates

From the most significant covariates as determined by the variable importance in Figure 5, it can be ascertained that vegetation was the biggest soil formation factor for these study areas for this region. Poggio et al. (2013) found that covariates corresponding to vegetation had the highest importance for predicting areas with peat, as corroborated by Minasny et al. (2019) in that optical imagery is of prime important for the prediction of peatlands in Canada. The overwhelming importance of vegetation covariates, when compared with covariates corresponding to other soil formation factors, justifies the usage of multi-source remotely sensed data. Remote sensing technologies can convey details about the vegetation cover on the surface, and a wider array of different sensor types can provide further information corresponding to vegetation.

It was attempted to incorporate multispectral satellite imagery over a multi-year period, as so to involve a temporal component for soil formation. Landsat imagery from 1984, 1995, 2005 and 2017 were applied for environmental covariates; images separated by approximately 10 years each over a 30 year timeline. Landsat-8 imagery was used for 2017 (and for 2015-2019 for May and autumn scenes), whereas Landsat-5 was utilized for the earlier imagery obtained for 1984, 1995 and 2005. Both Landsat-5 and Landsat-8 imagery had spatial resolutions of 30 m. The earliest available Landsat imagery of 30 m spatial resolution for the GCB that could be acquired from Google Earth corresponded to 1984, hence the choice for including imagery for that year. Surface reflectance was obtained from various years as an attempt to try modeling a temporal scale for the digital soil mapping analysis. It was hypothesized that earlier scenes of optical imagery would have higher significance for present soil properties, as previous scenes could correspond to factors that would result in gradually changing soil properties at a later date. This postulation was not able to be determined in this study. It was recognized that if optical imagery is obtained for either abnormally dry or wet years, the vegetation within the scenes would respond accordingly with either less vegetation or more vegetation due to such precipitation stresses. When establishing variable importance for a model, this fluctuation of stressing could result in differing levels of significance among predictors, or potentially ascertain other soil formation factors; this issue should be contemplated when collecting environmental covariates for a digital soil mapping study. Imagery from different seasons of the year were implemented, which was warranted as there were differences in predictor significance between times of peak vegetation and dormancy. The scenes corresponding to peak vegetation had the higher variable importance, but in some instances surface reflectance from winter scenes had among the highest variable importance. This could possibly be due to a correlation between certain soil properties and evergreen conifers rather than deciduous trees, that would be more apparent when deciduous trees can be better differentiated from the evergreen conifers by the deciduous leaves losing green color, or even by the leaves being off the trees.

Different gap fractions corresponding to the second and third returns of LiDAR retrievals, respectively, were also tried. The calculations were obtained originally at pixel-scale (originally 10 m by 10 m) and calculated respectively as

$$Gap\ Fraction\ (Second\ Returns) = \frac{Number\ of\ Second\ Returns}{Total\ Number\ of\ Retrievals}, \qquad \text{Eq. 6-1}$$

$$Gap\ Fraction\ (Third\ Returns) = \frac{Number\ of\ Third\ Returns}{Total\ Number\ of\ Retrievals}. \qquad \text{Eq. 6-2}$$

These gap fractions for the ARF region are shown in Figure 11. Inspecting this figure, one can see that areas with a low portion of second returns correspond to areas where the canopy layer is shorter. In general, these areas correspond to the wetlands. One can see that there are similarities between the gap fractions in Figure 4 and Figure 11, except that now the pattern is for the second returns to be inverted. For this LiDAR data, dense canopies will generally have more than one return per retrieval, whereas shallow canopies will usually have only one return (i.e. first return is only return). Incidences with a third return are low (i.e. maximum portion of only 0.307), but areas with higher portions are more spread out, yet do occur in the wetland areas. It is speculated that the taller structures that appeared in a wetland were more likely to have more than 2 returns per retrievals, whereas this would had been more difficult to occur for a denser canopy. These values would be contingent on the point density of the LiDAR, which should be considered when deriving environmental covariates from LiDAR data. Different point densities could lead to dissimilar results for derived covariates, particularly for covariates defined based on counts or ratios of returns for LiDAR retrievals.

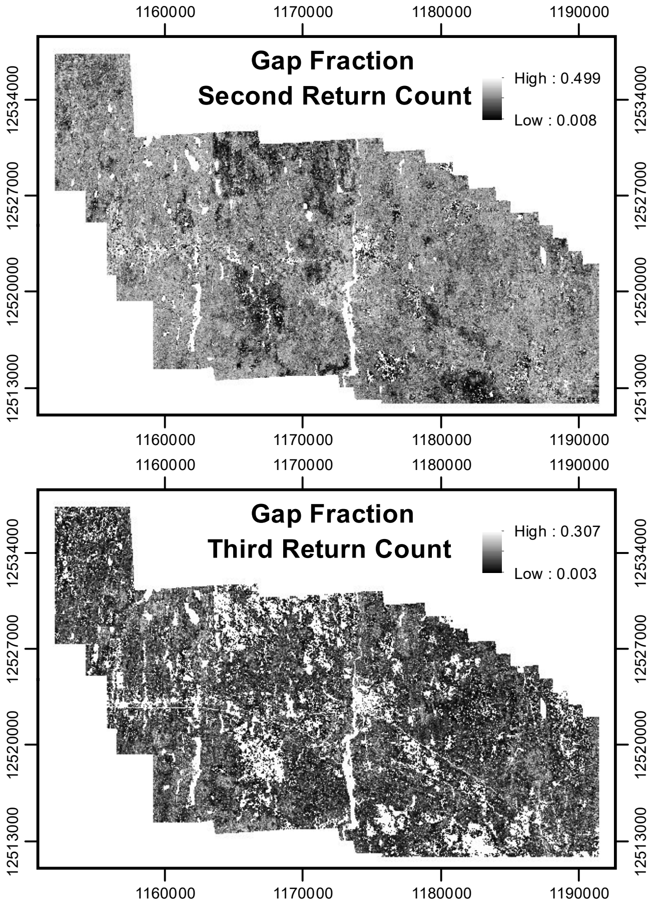


Figure 11 - Gap fractions for second and third returns, respectively, for the ARF region.

MODIS-derived products for vegetation covariates such as leaf area index (LAI) and fraction of absorbed photosynthetically active radiation (FPAR) (Myneni et al., 2002) were considered as predictors. However, these did not have high variable importance for the digital soil mapping models. The spatial resolution of 500 m for these predictors could had been too coarse for the digital soil mapping analysis undertaken, which could likely explain their lack of variable importance.

The covariates derived from L-band SAR imagery did not have the higher variable importance that was expected. It was anticipated that L-band SAR imagery would be able to capture the understory or surface moisture, which was expected to be strongly correlated to moisture regime or certain soil texture classes (such as peat which tends to be wetter than other soil types). A possible reason why this was not the case was that the L-band SAR imagery utilized was compiled for a whole year (i.e. 2017) where a lot of the differences between digital number intensities would be more of interest for just the summer months. Minasny et al. (2019) wrote that C-band SAR in HH-polarization is most sensitive for detecting soil moisture for peatlands. SAR imagery does seem to be relevant to the GCB region, as it can be difficult to obtain cloud-free imagery there for many periods of the years; this would not be an issue for the all-weather penetrating capability of SAR imagery. The implementation of SAR imagery could be useful for similar studies in the future.

Tree species prediction maps from the National Forestry Inventory (NFI) were tried as predictors in the digital soil mapping models. However, none of those attributes had high variable importance for any of the models, which was unexpected as site-level indicators for black spruce and balsam fir had high variable importance. The tree species maps from NFI were created from k-nearest neighbor (kNN) models applied to MODIS imagery (Beaudoin et al., 2018, 2014), which had an output spatial resolution of 250 m. When the tree species recorded from the site-level FRI data was compared with predictions from the NFI tree species maps, the incidence of black spruce was over-estimated by the NFI prediction maps; over 90% of pixels for the sites indicated black spruce as dominant, whereas in reality just over half the FRI sites sampled were reported as black spruce dominant. A likely explanation for this could be that on a larger scale of the spatial resolution of the NFI prediction maps (i.e. 250 m) black spruce is the most common tree species, whereas within those spacings sites could had been selected by the FRI to be more representative of different tree species types. For digital soil mapping applications on a larger scale, such as for the boreal forest across Canada, Mansuy et al. (2014) did find NFI tree species maps to be of significance. The spatial scale of analysis for digital soil mapping is an important factor to consider when utilizing tree species prediction maps.

Chapter 7

CONCLUSION & CLOSING REMARKS

The development and implementation of a framework for digital soil mapping analysis corresponding to classification modelling as discussed in Chapter 4 was conducted for the GCB region of Ontario. In this analysis, the most important environmental covariates from an expanded set of predictors corresponding to various soil formation factors for modelling the classification of soil property classifications to soil texture, reaction of calcareous substrate to hydrochloric acid, and ELC moisture regime were identified. It was able to be established in Section 5.1 that vegetation covariates had the highest variable importance, which suggests that vegetation was the main soil formation factor of this region. The results demonstrate the importance of LiDAR data, as the derivation and computation of detailed covariates derived from that LiDAR data were able to improve model accuracy. Prediction maps for the soil properties of soil texture family, calcareous reaction to hydrochloric acid and ELC moisture regime were generated in Section 5.3, with peatland environments being ascertained with a high degree of confidence. The research conducted in this thesis will provide useful steps for identifying peatland environments for supplementary environmental studies in the future.

The LiDAR-derived vegetation covariates of CHM and gap fraction were useful additions, as they had amongst the highest variable importance for all the soil property classifications, particularly for the ARF region. Detailed topographic covariates were also generated from a DEM derived from the LiDAR data, which reinforces the usefulness of utilizing LiDAR data for digital soil mapping purposes if it is available. Including aeromagnetic survey data as a covariate corresponding to the underlying parent material was also a suitable addition, as this provided an environmental covariate for parent material which could be better ascertained than prespecified and inferenced vector files corresponding to underlying geology. Surface reflectance from multispectral satellite imagery also had high variable importance, in particular the NDVI for the summertime which would correspond to the period of peak vegetation for the study area. The implementation of multi-source remotely sensed data for environmental covariates was beneficial, as it allowed covariate representation for a variety of different soil formation factors. The representation of finer scale environmental covariates corresponding to different attributes of the soil formation factors, in particular vegetation and topography, but also with representation from climatic and parent material, is a necessity for digital soil mapping research for a boreal biome.

The machine learning approaches of RF and SVM provided modelling accuracies for the best models for each targeted soil property that exceeded 0.7 in correctness and 0.5 in Cohen's kappa score, as seen in the tables of Section 5.2. Model accuracies were able to be attained from RF and SVM approaches that matched or exceeded the accuracies of the best models in comparable studies (Brungard et al., 2015; Dornik et al., 2018; Heung et al., 2017, 2016). From the results in this thesis, RF and SVM approaches had similar accuracies, but in general the RF models had slightly higher accuracies more times than what the SVM models achieved. The RF and SVM approaches were able to satisfactorily model the soil classification properties of texture, calcareous substrate reaction to hydrochloric acid and ELC moisture regime, and thus are great approaches to try for digital soil mapping classifications.

Spatial patterns within the prediction maps and uncertainty maps of the soil properties were able to be resolved, as seen in Section 5.3. The RF and SVM approaches generated similar spatial output for predictions of the soil properties of soil texture, calcareous substrate reaction to hydrochloric acid, and ELC moisture regime. Areas consisting of peat had no fizzing reaction to hydrochloric acid for calcareous substrate acid testing, and had the highest ELC moisture regime scores denoting wet ground conditions. The areas consisting of loamy soil tended to denote a fizzing reaction for calcareous substrate acid testing, and had drier ELC moisture regime ratings than that for peat. Clayey soils were predicted for settled areas of agriculture land and roadways where one would expect soil compaction, which corresponded to the areas of the driest ELC moisture regime ratings. From the uncertainty maps, it was demonstrated that predictions for peatlands had the highest modelling certainties due to corresponding lower entropy scores, whereas modelling for areas with clayey soil at the surface with drier moisture regimes had the lowest modelling certainties. These results will assist in distinguishing the wetland environments from areas more suitable for agricultural purposes, which could be beneficial for policymakers for land management concerning the GCB.

For future work, it is advisable to note:

- In this study the RF models were used to determine the feature selection of environmental covariates. A model-independent feature selection approach based upon ROC curve analysis was investigated, which resulted in analogous sets of important covariates as what were resolved from the RF variable importance method. However, it is recommended to implement a model-independent feature selection approach in order to determine the most important predictors, so as to have greater confidence in the variable importance of predictors. This corresponding reduced set of predictors would be utilized as the predictors for the digital soil mapping models, with the same reduced set applied regardless of modelling approach.

- Apply C-band SAR with HH-polarization for environmental covariates, as L-band SAR was not as important of a predictor as what was anticipated. It might be worthwhile to obtain C-band and L-band SAR imagery specifically associated with certain times of the year, such as midsummer, that could possibly be better spatially correlated with soil properties relating to moisture regime.

- Utilize finer spatial resolution tree species predictor maps, or other covariates relating to vegetation canopy (specifically if spatial resolutions for those products are less than 250 m). Tree species predictors maps obtained from the NFI were not able to have significance for the soil property models as their spatial resolutions were too coarse (i.e. 250 m or greater).

- Multispectral imagery from earlier years could be useful, particularly if acquired from drier years versus wetter years (to capture the response of the resulting stress or forcing on vegetation). This could elucidate the synergetic relationship between vegetation and soil, to determine how vegetation drives soil formation.

REFERENCES

Adhikari, K., Kheir, R.B., Greve, M.B., Bøcher, P.K., Malone, B.P., Minasny, B., McBratney, A.B., Greve, M.H., 2013. High-Resolution 3-D Mapping of Soil Texture in Denmark. Soil Sci. Soc. Am. J. 77, 860–876. https://doi.org/10.2136/sssaj2012.0275

Airborne Imaging, 2018. Final Report For Project Cochrane LiDAR. Calgary, Alberta.

Akumu, C.E., Johnson, J.A., Etheridge, D., Uhlig, P., Woods, M., Pitt, D.G., McMurray, S., 2015. GIS-fuzzy logic based approach in modeling soil texture: Using parts of the Clay Belt and Hornepayne region in Ontario Canada as a case study. Geoderma 239–240, 13–24. https://doi.org/10.1016/j.geoderma.2014.09.021

Anderson, E.S., Thompson, J.A., Crouse, D.A., Austin, R.E., 2006. Horizontal resolution and data density effects on remotely sensed LIDAR-based DEM. Geoderma 132, 406–415. https://doi.org/10.1016/j.geoderma.2005.06.004

Angelini, M.E., Heuvelink, G.B.M., Kempen, B., Morrás, H.J.M., 2016. Mapping the soils of an Argentine Pampas region using structural equation modelling. Geoderma 281, 102–118. https://doi.org/10.1016/j.geoderma.2016.06.031

Beguin, J., Fuglstad, G.A., Mansuy, N., Paré, D., 2017. Predicting soil properties in the Canadian boreal forest with limited data: Comparison of spatial and non-spatial statistical approaches. Geoderma 306, 195–205. https://doi.org/10.1016/j.geoderma.2017.06.016

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

Belova, S.E., Pankratov, T.A., Dedysh, S.N., 2006. Bacteria of the genus Burkholderia as a typical component of the microbial community of Sphagnum peat bogs. Microbiology 75, 90–96. https://doi.org/10.1134/S0026261706010164

Breiman, L., Cutler, A., Liaw, A., Wiener, M., 2018. Package “randomForest”: Breiman and Cutler’s Random Forest for Classification and Regression.

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, 68–83. https://doi.org/10.1016/j.geoderma.2014.09.019

Camera, C., Zomeni, Z., Noller, J.S., Zissimos, A.M., Christoforou, I.C., Bruggeman, A., 2017. A high resolution map of soil types and physical properties for Cyprus: A digital soil mapping optimization. Geoderma 285, 35–49. https://doi.org/10.1016/j.geoderma.2016.09.019

Cohen, J., 1960. A Coefficient of Agreement for Nominal Scales. Educ. Psychol. Meas. XX, 37–46.

Dornik, A., Dragut, L., Urdea, P., 2018. Classification of Soil Types Using Geographic Object-Based Image Analysis and Random Forests. Pedosphere 28, 913–925. https://doi.org/10.1016/S1002-0160(17)60377-1

Ellili Bargaoui, Y., Walter, C., Michot, D., Saby, N.P.A., Vincent, S., Lemercier, B., 2019. Validation of digital maps derived from spatial disaggregation of legacy soil maps. Geoderma 356, 113907. https://doi.org/10.1016/j.geoderma.2019.113907

Environment Canada, 2020. Canadian Climate Normals 1981-2010 Station Data: KAPUSKASING A [WWW Document]. URL ttps://climate.weather.gc.ca/climate_normals/results_1981_2010_e.html?searchType=stnName&txtStationName=Kapuskasing&searchMethod=contains&txtCentralLatMin=0&txtCentralLatSec=0&txtCentralLongMin=0&txtCentralLongSec=0&stnID=4157&dispBack=0 (accessed 7.2.20).

Fick, S.E., Hijmans, R.J., 2017. WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. Int. J. Climatol. 37, 4302–4315. https://doi.org/10.1002/joc.5086

Garcia, M., Saatchi, S., Ferraz, A., Silva, C.A., Ustin, S., Koltunov, A., Balzter, H., 2017. Impact of data model and point density on aboveground forest biomass estimation from airborne LiDAR. Carbon Balance Manag. 12, 1–18. https://doi.org/10.1186/s13021-017-0073-1

Genuer, R., Poggi, J.M., Tuleau-Malot, C., 2010. Variable selection using random forests. Pattern Recognit. Lett. 31, 2225–2236. https://doi.org/10.1016/j.patrec.2010.03.014

Gessler, P.E., Chadwirk, O. A., Chamran, F., Althouse, L., Holmes, K., 2000. Modeling Soil-Landscape and Ecosystem Properties using Terrain Attributes. Soil Sci. Soc. Am. J.

Goldman, M.A., Needelman, B.A., Rabenhorst, M.C., Lang, M.W., McCarty, G.W., King, P., 2020. Digital soil mapping in a low-relief landscape to support wetland restoration decisions. Geoderma 373, 114420. https://doi.org/10.1016/j.geoderma.2020.114420

Gomez, C., Viscarra Rossel, R.A., McBratney, A.B., 2008. Soil organic carbon prediction by hyperspectral remote sensing and field vis-NIR spectroscopy: An Australian case study. Geoderma 146, 403–411. https://doi.org/10.1016/j.geoderma.2008.06.011

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

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

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

Heung, B., Hodúl, M., Schmidt, M.G., 2017. Comparing the use of training data derived from legacy soil pits and soil survey polygons for mapping soil classes. Geoderma 290, 51–68. https://doi.org/10.1016/j.geoderma.2016.12.001

Hijmans, R.J., Cameron, S.E., Parra, J.L., Jones, P.G., Jarvis, A., 2005. Very High Resolution Interpolated Climate Surfaces for Global Land Areas. Int. J. Climatol. 25 1965-1978.

Jalali, M., Moradi, A., 2020. Measuring and simulating pH buffer capacity of calcareous soils using empirical and mechanistic models. Arch. Agron. Soil Sci. 66, 559–571. https://doi.org/10.1080/03650340.2019.1628344

Jenny, H., 1941. Factors of Soil Formation: A System of Quantitative Pedology. McGraw-Hill Book Company Inc., New York.

Keskin, H., Grunwald, S., 2018. Regression kriging as a workhorse in the digital soil mapper's toolbox. Geoderma 326, 22–41. https://doi.org/10.1016/j.geoderma.2018.04.004

Kuhn, M., 2013. Inconsistent results between caret+kernlab versions [WWW Document]. R-help. URL https://r.789695.n4.nabble.com/Inconsistent-results-between-caret-kernlab-versions-td4680500.html (accessed 9.10.20).

Kuhn, M., 2008. Building Predictive Models in R using the caret Package. J. Stat. Softw. 28. https://doi.org/10.18637/jss.v028.i05

Kuhn, M., 2007. Variable Importance Using The caret Package.

Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z., Ziem, A., Scrucca, L., Hunt, T., 2020. Package ’caret ’: Classification and Regression Training.

Kulkarni, V.Y., Petare, M., Sinha, P.K., 2014. Analyzing Random Forest Classifier with Different Split Measures, in: Babu, B., Nagar, A., Deep, K., Pant, M., Chand Bansal, J., Ray, K., Gupta, U. (Eds.), Proceedings of the Second International Conference on Soft Computing for Problem Solving (SocProS 2012), December 28-30, 2012. Springer, New Delhi, pp. 691–699. https://doi.org/https://doi-org.ezproxy.library.yorku.ca/10.1007/978-81-322-1602-5_74

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

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. https://doi.org/10.1016/j.geoderma.2011.10.010

Lin, H.T., Lin, C.J., Weng, R.C., 2007. A note on Platt’s probabilistic outputs for support vector machines. Mach. Learn. 68, 267–276. https://doi.org/10.1007/s10994-007-5018-6

Malterer, T.J., Verry, E.S., Erjavec, J., 1992. Fiber Content and Degree of Decomposition in Peats: Review of National Methods. Soil Sci. Soc. Am. J. 56, 1200–1211.

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

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

McKenzie, N.J., Ryan, P.J., 1999. Spatial prediction of soil properties using environmental correlation. Geoderma 89, 67–94. https://doi.org/10.1016/S0016-7061(98)00137-2

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F., 2019. Package ‘e1071.’

Minasny, B., Berglund, Ö., Connolly, J., Hedley, C., de Vries, F., Gimona, A., Kempen, B., Kidd, D., Lilja, H., Malone, B., McBratney, A., Roudier, P., O'Rourke, S., Rudiyanto, Padarian, J., Poggio, L., ten Caten, A., Thompson, D., Tuve, C., Widyatmanti, W., 2019. Digital mapping of peatlands – A critical review. Earth-Science Rev. 196, 102870. https://doi.org/10.1016/j.earscirev.2019.05.014

Mohamed, M.A., 2020. Classification of landforms for digital soil mapping in urban areas using LiDAR data derived terrain attributes: A case study from Berlin, Germany. Land 9. https://doi.org/10.3390/LAND9090319

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. https://doi.org/10.1016/j.geoderma.2010.12.018

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. https://doi.org/10.1016/j.geoderma.2015.08.035

Myneni, R.B., Hoffman, S., Knyazikhin, Y., Privette, J.L., Glassy, J., Tian, Y., Wang, Y., Song, X., Zhang, Y., Smith, G.R., Lotsch, A., Friedl, M., Morisette, J.T., Votava, P., Nemani, R.R., Running, S.W., 2002. Global products of vegetation leaf area and fraction absorbed PAR from year one of MODIS data. Remote Sens. Environ. 83, 214–231. https://doi.org/10.1016/S0034-4257(02)00074-3

Nussbaum, M., Spiess, K., Baltensweiler, A., Grob, U., Keller, A., Greiner, L., Schaepman, M.E., Papritz, A., 2018. Evaluation of digital soil mapping approaches with large sets of environmental covariates. Soil 4, 1–22. https://doi.org/10.5194/soil-4-1-2018

O'Neil, G.L., Saby, L., Band, L.E., Goodall, J.L., 2019. Effects of LiDAR DEM Smoothing and Conditioning Techniques on a Topography-Based Wetland Identification Model. Water Resour. Res. 55, 4343–4363. https://doi.org/10.1029/2019WR024784

Ontario Ministry of Natural Resources, 2013. Provincial Digital Elevation Model Technical Specifications, v3.0. Peterborough, Ontario.

Paloniemi, J., 2018. FRI Field Guide. Forest Resources Inventory, Ontario Ministry of Natural Resources and Forestry, Ontario.

Platt, J., 2000. Probabilistic outputs for support vector machines and comparison to regularized likelihood methods, in: Smola A., P., B., B., S., D., S. (Eds.), Advances in Large Margin Classifiers. MIT Press, Cambridge.

Poggio, L., Gimona, A., Brewer, M.J., 2013. Regional scale mapping of soil properties and their uncertainty with a large number of satellite-derived covariates. Geoderma 209–210, 1–14. https://doi.org/10.1016/j.geoderma.2013.05.029

Pokharel, B., Dech, J.P., 2011. An Ecological Land Classification approach to modeling the production of forest biomass. For. Chron. 87, 23–32. https://doi.org/10.5558/tfc87023-1

R Core Team, 2012. R: A Language and Environment for Statistical Computing.

Rapinel, S., Fabre, E., Dufour, S., Arvor, D., Mony, C., Hubert-Moy, L., 2019. Mapping potential, existing and efficient wetlands using free remote sensing data. J. Environ. Manage. 247, 829–839. https://doi.org/10.1016/j.jenvman.2019.06.098

Rogovska, N., Blackmer, A.M., 2009. Remote sensing of soybean canopy as a tool to map high pH, calcareous soils at field scale. Precis. Agric. 10, 175–187. https://doi.org/10.1007/s11119-008-9087-8

Roulston, M.S., Smith, L.A., 2002. Evaluating probabilistic forecasts using information theory. Mon. Weather Rev. 130, 1653–1660. https://doi.org/10.1175/1520-0493(2002)130<1653:EPFUIT>2.0.CO;2

SAGA Development Team, 2020. System for Automated Geoscientific Analysis (SAGA).

Sage, A.J., Genschel, U., Nettleton, D., 2020. Tree aggregation for random forest class probability estimation. Stat. Anal. Data Min. 13, 134–150. https://doi.org/10.1002/sam.11446

Shimada, M., Itoh, T., Motooka, T., Watanabe, M., Shiraishi, T., Thapa, R., Lucas, R., 2014. New global forest/non-forest maps from ALOS PALSAR data (2007-2010). Remote Sens. Environ. 155, 13–31. https://doi.org/10.1016/j.rse.2014.04.014

Simon, P., 2006. Including omission mistakes in the calculation of Cohen's Kappa and an analysis of the coefficient's paradox features. Educ. Psychol. Meas. 66, 765–777. https://doi.org/10.1177/0013164405285548

Southee, F.M., Treitz, P.M., Scott, N.A., 2012. Application of lidar terrain surfaces for soil moisture modeling. Photogramm. Eng. Remote Sensing 78, 1241–1251. https://doi.org/10.14358/PERS.78.11.1241

Town of Kapuskasing, 2020. Forestry [WWW Document]. URL http://www.kapuskasing.ca/en/growing/Forestry.aspx (accessed 10.26.20).

United States Geological Survey, 2019. Landsat Surface Reflectance Data, U.S. Geological Survey Fact Sheet 2015-3034. https://doi.org/https://doi.org/10.3133/fs20153034

Vernimmen, R., Hooijer, A., Yuherdha, A.T., Visser, M., Pronk, M., Eilander, D., Akmalia, R., Fitranatanegara, N., Mulyadi, D., Andreas, H., Ouellette, J., Hadley, W., 2019. Creating a lowland and peatland landscape digital terrain model (DTM) from interpolated partial coverage LiDAR data for Central Kalimantan and East Sumatra, Indonesia. Remote Sens. 11. https://doi.org/10.3390/rs11101152

Vitt, D.H., Wieder, R.K., Scott, K.D., Faller, S., 2009. Decomposition and peat accumulation in rich fens of boreal Alberta, Canada. Ecosystems 12, 360–373. https://doi.org/10.1007/s10021-009-9228-6

Wang, B., Waters, C., Orgill, S., Gray, J., Cowie, A., Clark, A., Liu, D.L., 2018. High resolution mapping of soil organic carbon stocks using remote sensing variables in the semi-arid rangelands of eastern Australia. Sci. Total Environ. 630, 367–378. https://doi.org/10.1016/j.scitotenv.2018.02.204

Warrens, M.J., 2011. Weighted kappa is higher than Cohen’s kappa for tridiagonal agreement tables. Stat. Methodol. 8, 268–272. https://doi.org/10.1016/j.stamet.2010.09.004

Were, K., Bui, D.T., Dick, Ø.B., Singh, B.R., 2015. A comparative assessment of support vector regression, artificial neural networks, and random forests for predicting and mapping soil organic carbon stocks across an Afromontane landscape. Ecol. Indic. 52, 394–403. https://doi.org/10.1016/j.ecolind.2014.12.028

Wu, W., Yang, Q., Lv, J., Li, A., Liu, H., 2019. Investigation of Remote Sensing Imageries for Identifying Soil Texture Classes Using Classification Methods. IEEE Trans. Geosci. Remote Sens. 57, 1653–1663. https://doi.org/10.1109/TGRS.2018.2868141

Yang, R.M., Zhang, G.L., Liu, F., Lu, Y.Y., Yang, Fan, Yang, Fei, Yang, M., Zhao, Y.G., Li, D.C., 2016. Comparison of boosted regression tree and random forest models for mapping topsoil organic carbon concentration in an alpine ecosystem. Ecol. Indic. 60, 870–878. https://doi.org/10.1016/j.ecolind.2015.08.036

Zhu, A.X., 1997. Measuring uncertainty in class assignment for natural resource maps under fuzzy logic. Photogramm. Eng. Remote Sensing 63, 1195–1202.

Appendix A

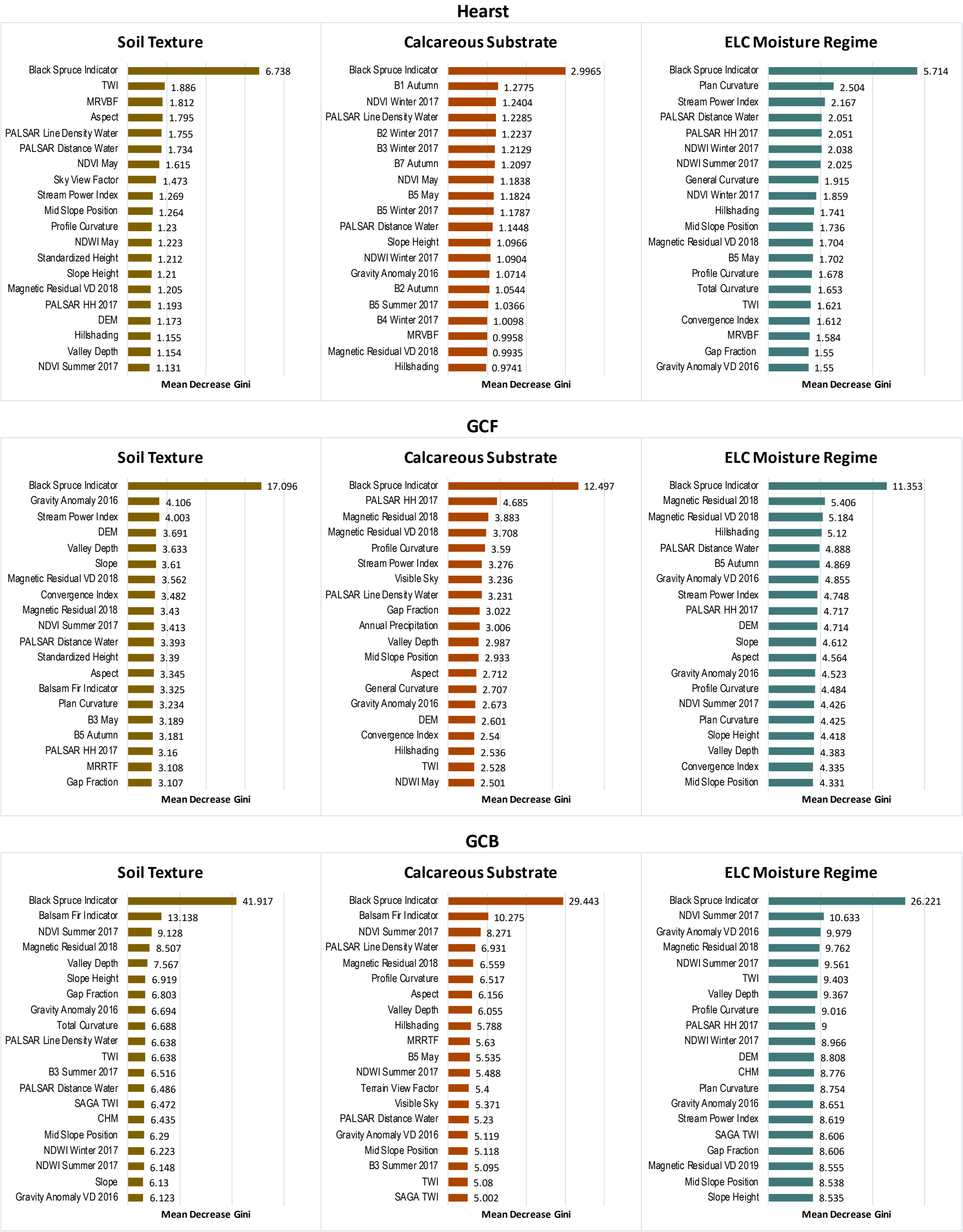


Figure 12 - Variable importance for models for Hearst, GCF and GCB regions.

Appendix B

The bounds for the ASCII tables corresponded to the NAD 1983 Lambert Conformal Conic projection. This coordinate system was applied for all ASCII tables for each targeted study area. The Hearst region corresponded to the extent of 12,576,550 for top, 12,551,740 for bottom, 1,001,800 for left and 1,055,350 for right. The layers for this region correspond to matrices with 1785 columns by 827 rows. For the GCF region, the extents were 12,566,620 for the top, 12,493,630 for the bottom, 1,047,970 for the left and 1,155,790 for the right; this resulted in layers of matrices with 3594 columns by 2433 rows. The ARF region had extents for 12,537,340 for top, 12,511,930 for bottom, 1,152,070 for left and 1,191,400 for right, which corresponded to layers with matrices of 1311 columns by 847 rows. Only pixels within the boundary polygon confines contained relevant information; pixels outside the boundaries were specified as no-data.

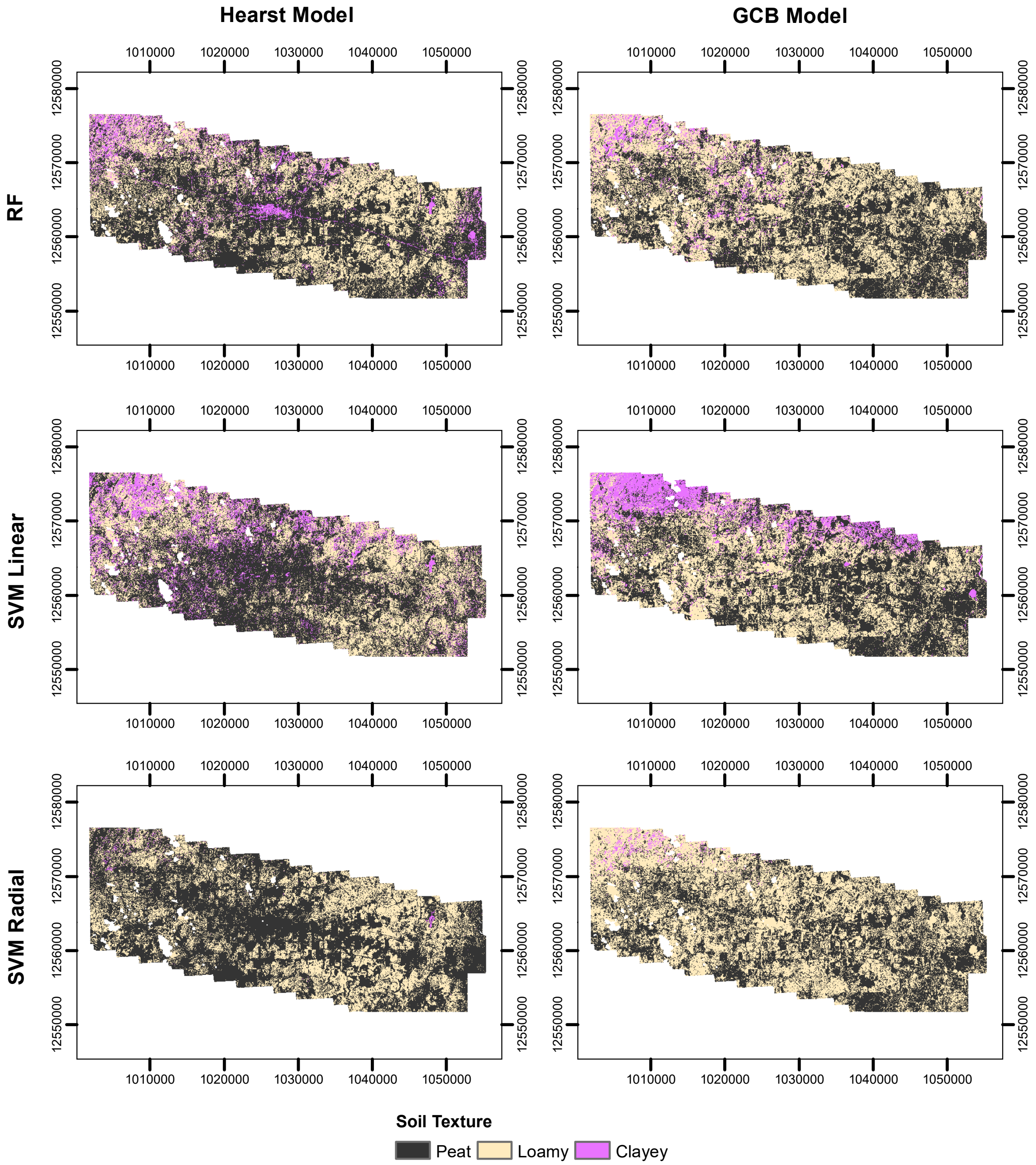


Figure 13 - Prediction maps for soil texture for the Hearst region.

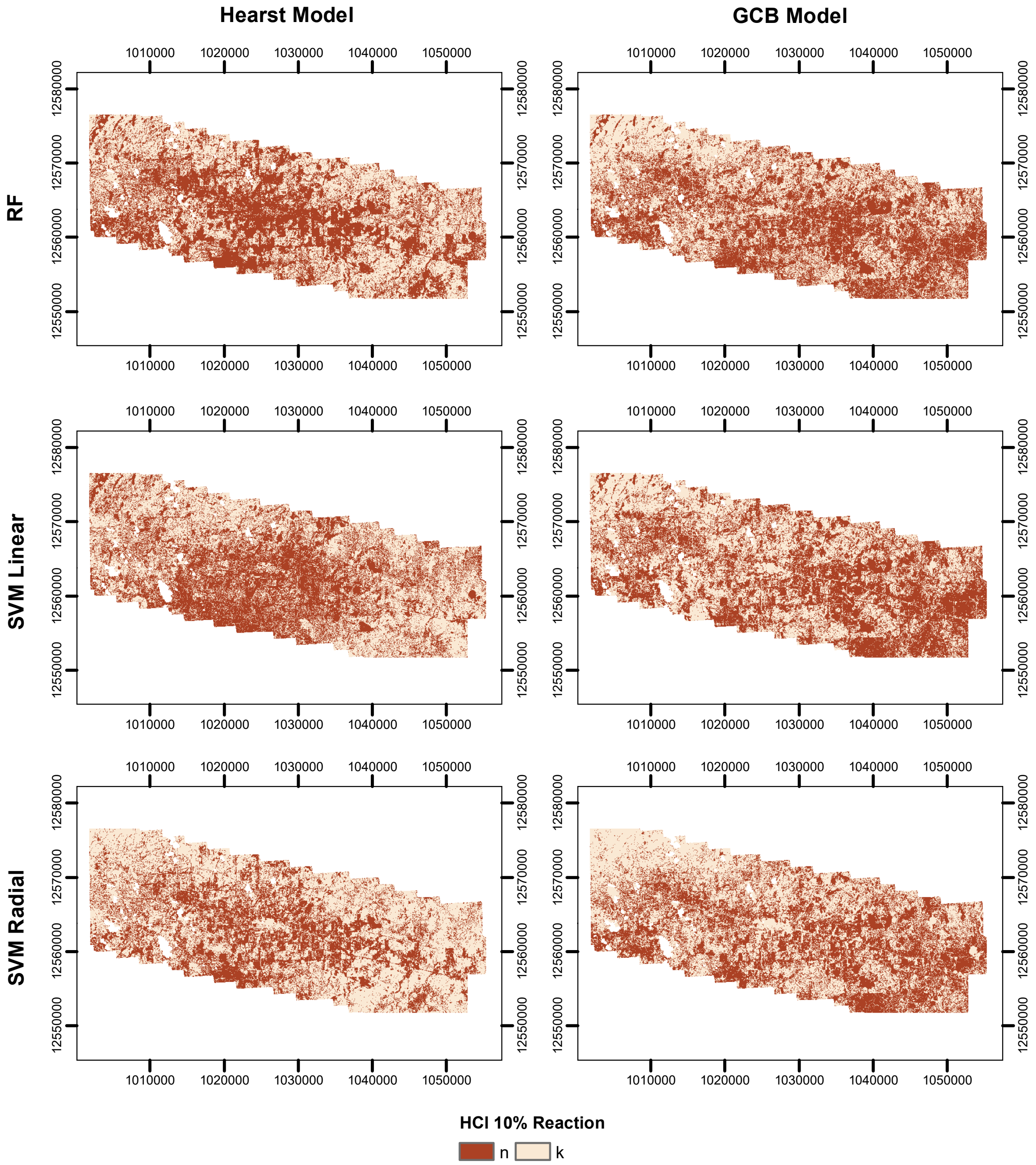


Figure 14 - Prediction maps for calcareous substrate reaction for the Hearst region.

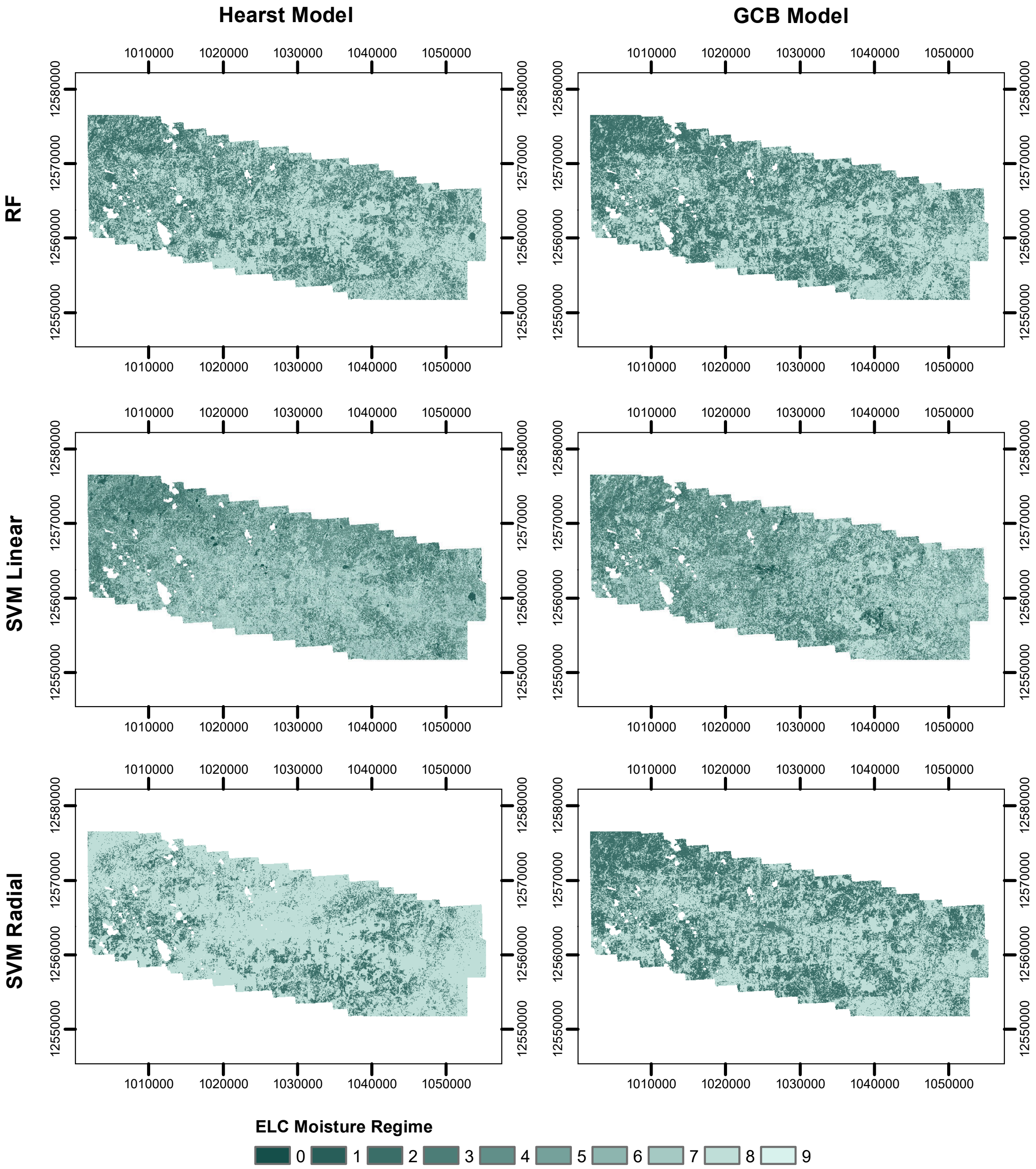


Figure 15 - Prediction maps for ELC moisture regime for the Hearst region.

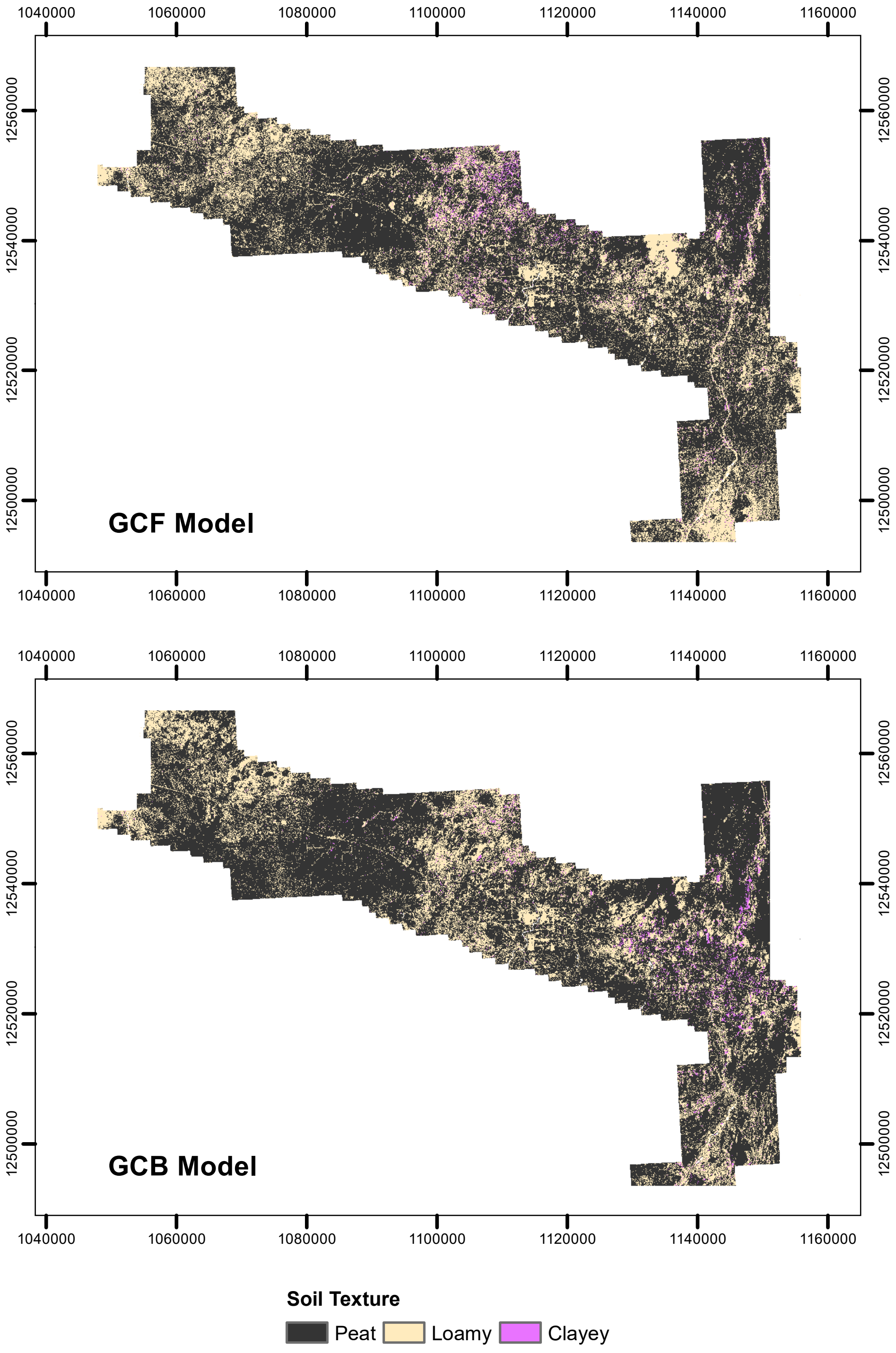


Figure 16 - Prediction maps of soil texture for RF models for the GCF region.

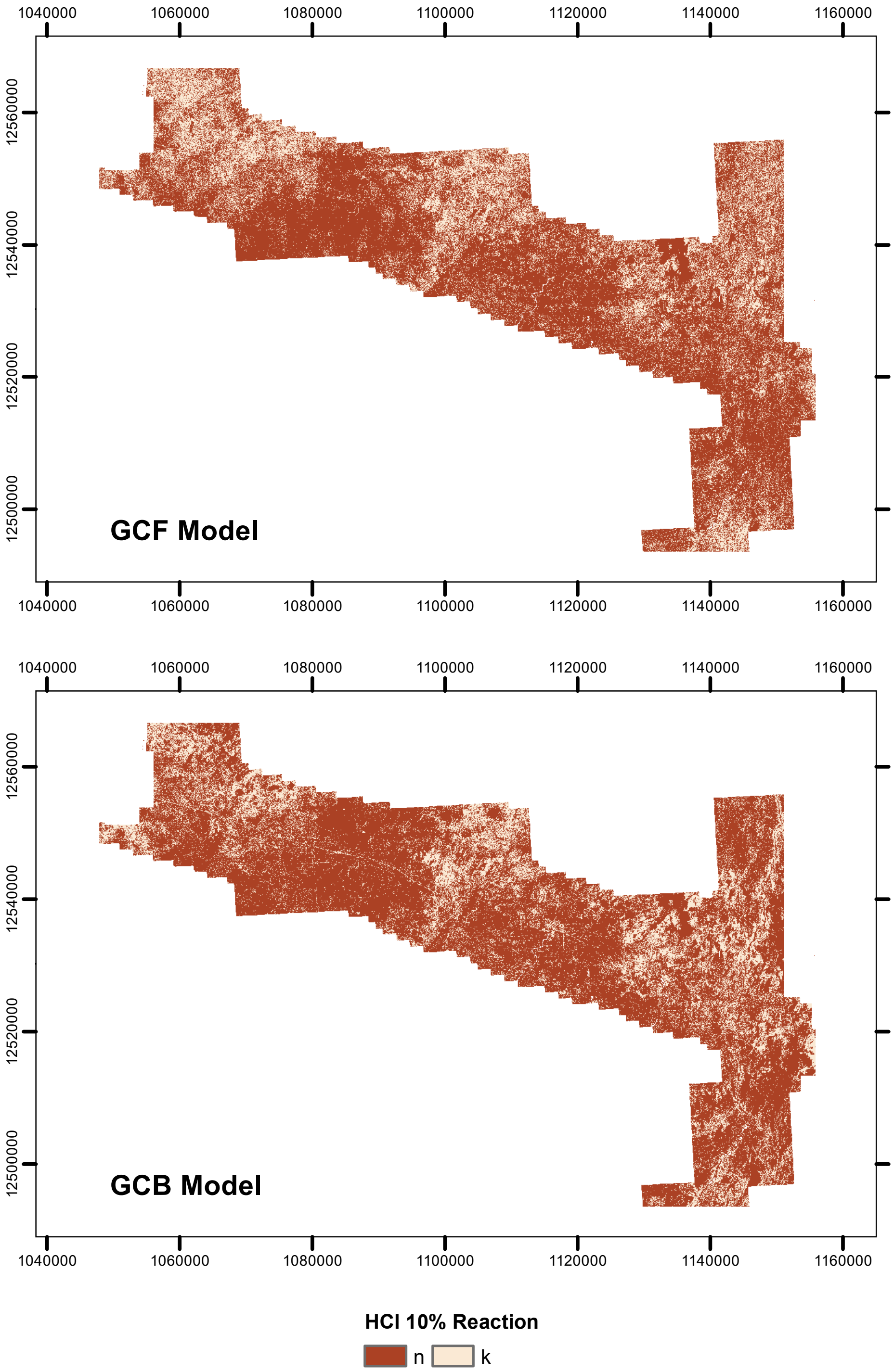


Figure 17 - Prediction maps of calcareous substrate reaction for RF models for the GCF region.

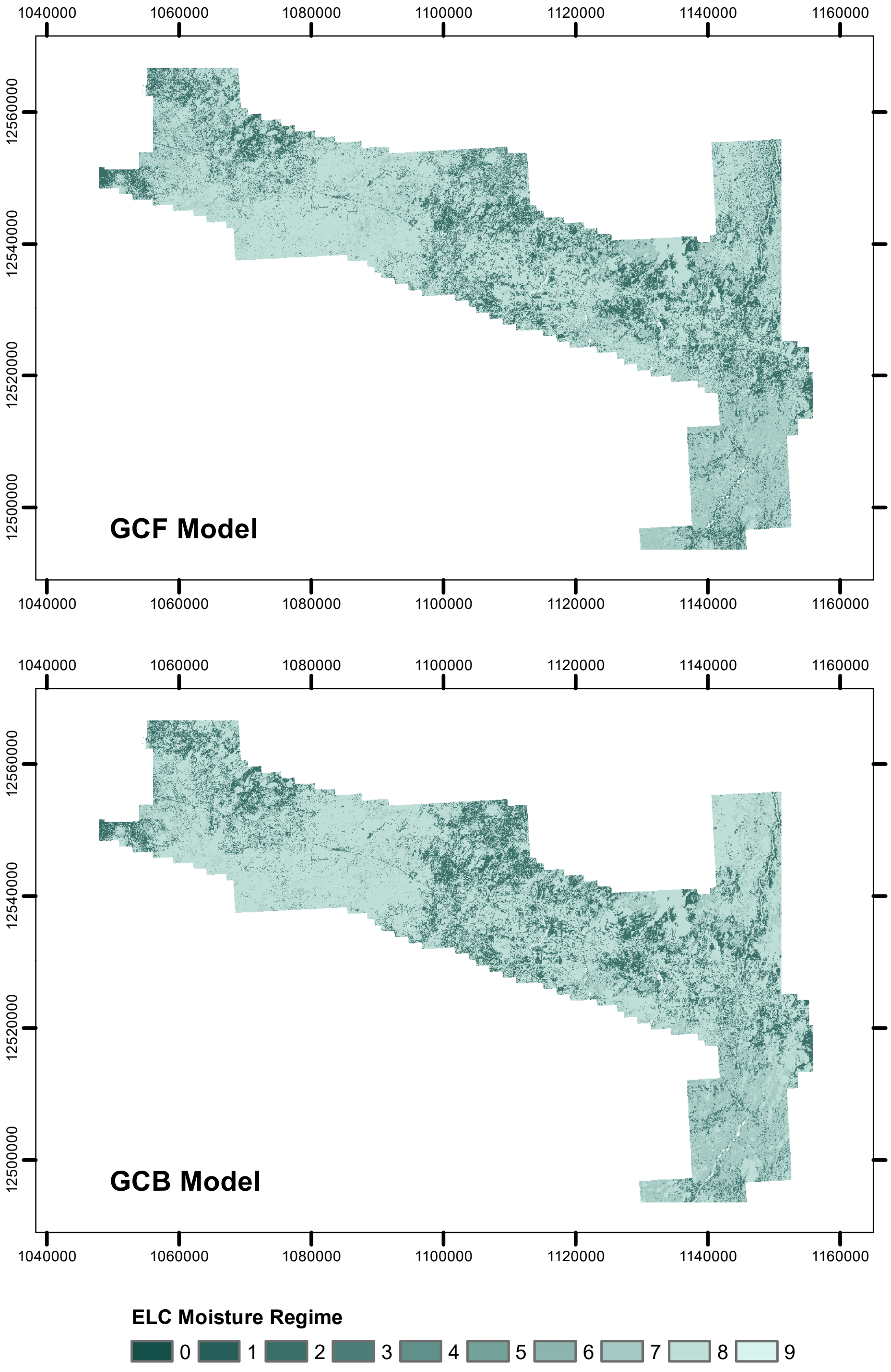
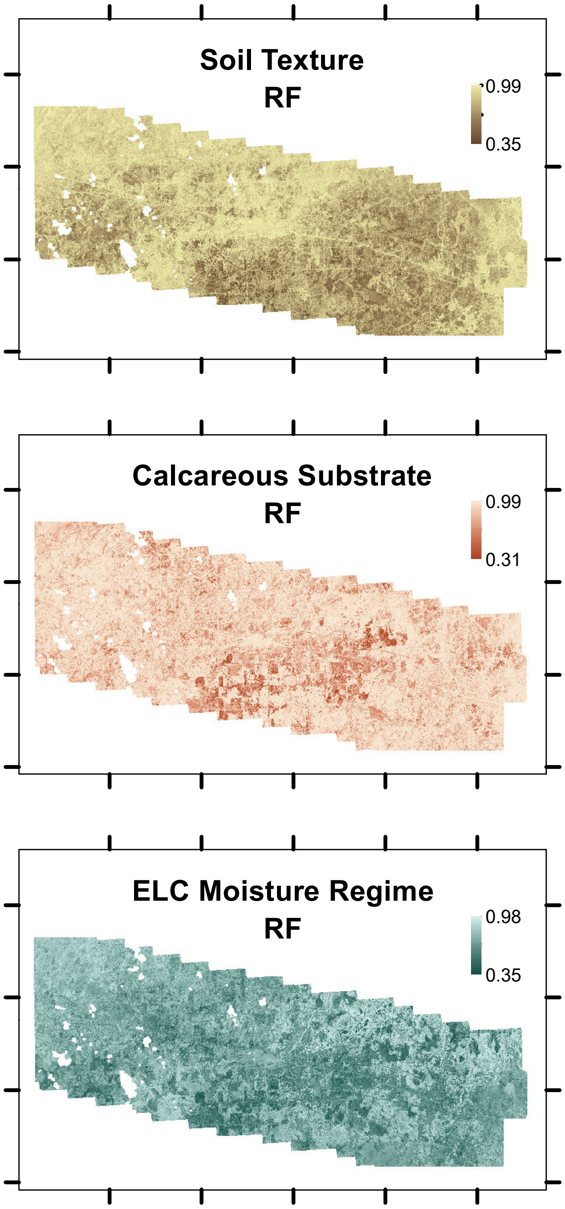


Figure 18 - Predictions maps of ELC moisture regime for RF models for the GCF region.



Soil Texture
RF

Calcareous Substrate
RF

ELC Moisture Regime
RF

Figure 19 - Entropy maps for uncertainty of predictions for RF models for the Hearst region.

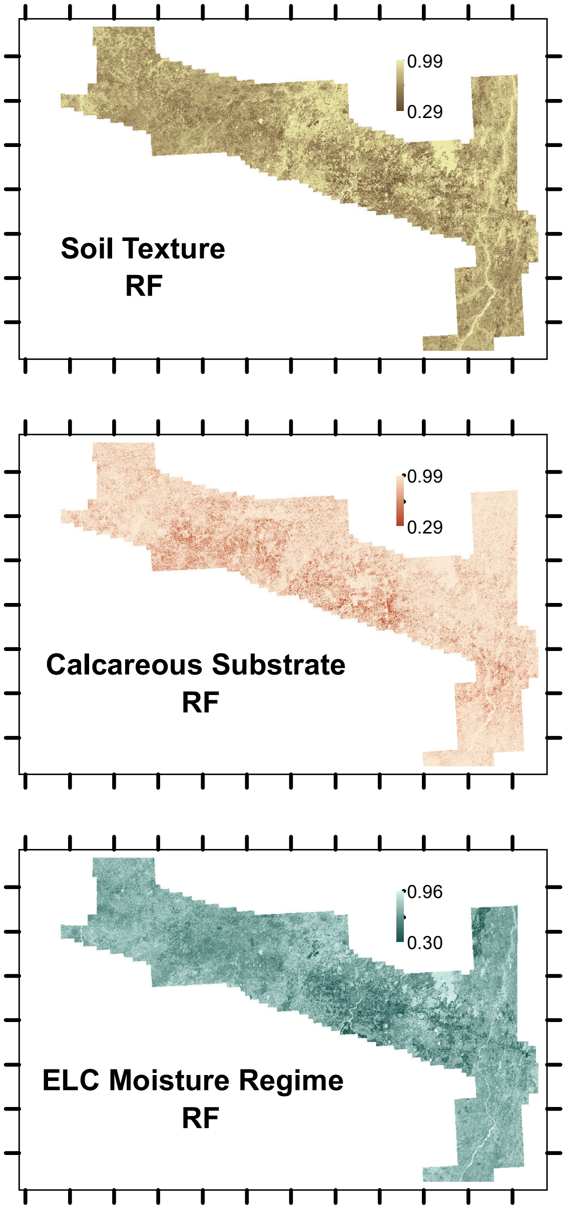


Figure 20 - Entropy maps for uncertainty of predictions for RF models for the GCF region.

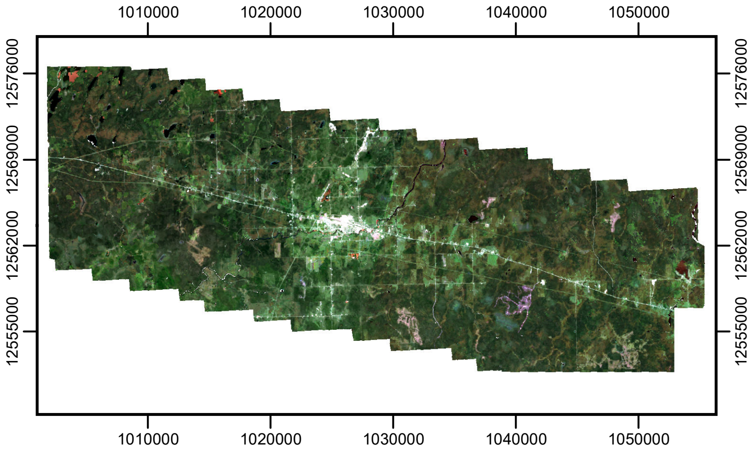


Figure 21 - True-color composite from Landsat-8 imagery for the summer of 2018 for the Hearst region.

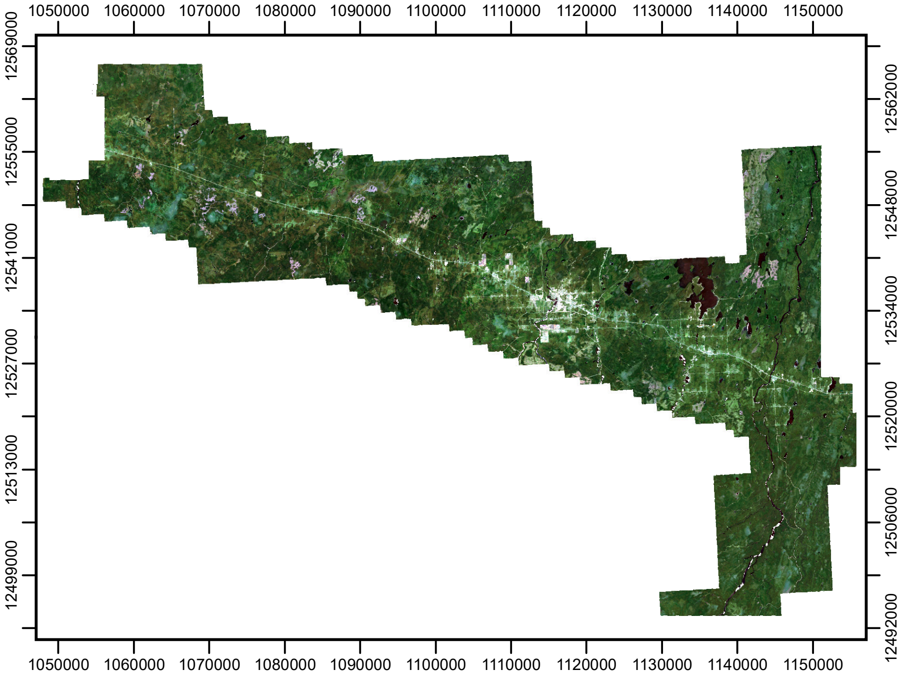


Figure 22 - True-color composite from Landsat-8 imagery for the summer of 2018 for the GCF region.