ADVANCEMENTS IN MODELING SOIL CARBON USING REMOTELY SENSED DATA

RORY CLIFFORD PITTMAN

A DISSERTATION SUBMITTED TO
THE FACULTY OF GRADUATE STUDIES
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

GRADUATE PROGRAM IN EARTH & SPACE SCIENCE & ENGINEERING
YORK UNIVERSITY
TORONTO, ONTARIO

November, 2025

Abstract

This research investigated the resolving of soil carbon and interconnected properties in conjunction with vegetation and landform attributes for a boreal context within the Great Clay Belt region of northern Ontario, Canada. Objectives entailing sophisticated statistical inference of extracted soil samples, an endemic tree species classification and enhanced digital soil mapping of carbon were explored, yielding research contributions and practicable results.

Statistical inference was achieved by adapting generalized estimating equations (GEEs) to estimate and assess refined mean differences of soil carbon between specific land cover types and tree species groupings. Furthermore, the feasibility of GEEs for digital soil mapping was demonstrated by predicting carbon concentrations with associated confidence intervals. Regarding boreal vegetation modeling, a pixel-based tree species classification was developed for natural settings at the remarkable spatial resolution of 2 m over an area of 100 km^2. Afterwards, digital soil mapping of carbon was implemented by employing data-driven approaches via encoder-decoder (ED) neural networks formulated for relatively smaller data sets. Using EDs, modeling accuracy for soil carbon was augmented with coefficients of determination (R^2) increasing to over 0.5. Finally, knowledge-based approaches to unravel linkages between soil carbon and interrelated properties were accommodated through structural equation modeling (SEM). A framework was devised to effectively facilitate SEM for digital soil mapping. The SEM quantified impacts on soil properties from covariates relating to soil formation factors, supporting the discernment of vegetational and environmental drivers for bulk density, carbon and carbon-to-nitrogen ratio.

Uncertainty with prediction was also ascertained. A normalized entropy metric constituted from the top two contending groupings was conceived, which better encapsulated prediction uncertainty when compared to conventional entropy. Quantile mapping was also incorporated to uncover insights regarding prediction uncertainty from model ensembles. Structured query language (SQL) was harnessed to efficiently derive and generate rasterized covariates from LiDAR point cloud data. This novel computation consisted of a more

detailed digital terrain model (DTM), as well as covariates pertaining to vegetational structure with canopy height model (CHM) and a gap fraction. These covariates were successfully exploited as predictors for modeling with both digital soil mapping and tree species classification.

Acknowledgements

There are many people I want to show my gratitude and thank regarding my aspiration of pursuing a Ph.D. program at York University.

Firstly, I want to thank my supervisory committee for their support and feedback for this research. I want to thank my supervisor, Dr. Baoxin Hu, for excellent guidance and support that has greatly assisted me in conducting my research studies. I also want to heartily thank my committee members, Dr. Jianguo Wang and Dr. Kara L. Webster, as well as Dr. Jiali Shang, for their support, advice and sharing knowledge pertaining to statistics, soil science, and forestry. I also want to thank all members of my examination committee, Dr. Richard Bello, Dr. Peter Taylor and Dr. Jonathan Li.

For the collection of soil samples, I want to thank York University graduate students in the Earth and Space Science and Engineering (ESSE) department for assistance with the 2018 and 2019 field campaigns. I want to thank Stephanie Wilson and Laura Hawdon of the Great Lakes Forestry Centre (GLFC) in Sault Ste. Marie, Ontario, for support and training with extracting and processing of soil samples. I want to thank landowners in the District of Cochrane for permission to obtain soil samples from their land properties.

I also want to thank York University graduate students, former and current, for camaraderie and advice relating to requirements and access-related information relevant for graduate studies. This appreciation is expressed to Dr. Wendy Zhang, Dr. Ima Ituen, Cheryl Li, Steven Chen, Mike Ma and Taewoong Ham. I wish everyone the best for their future endeavors.

I want to thank my brothers, Tyler and Jeremy, and my parents, Marilyn and Wayne.

With respect to funding sources sustaining portions of this research, I want to acknowledge the Ontario Ministry of Agriculture, Food and Rural Affairs (OMAFRA) for a grant that supported fieldwork and soil sample processing for chemistry analysis. Moreover, I was fortunate to receive an Ontario Graduate Scholarship (OGS) for two separate years of my Ph.D. program studies at York University.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

17 N	zone 17 north (for UTM projection)
1-D	one-dimensional
2-D	two-dimensional
%	percentage
&	and
°C	degree Celsius
° N	degree north (latitude)
° W	degree west (longitude)
A&L	A&L Canada Laboratories Incorporated
ANN	artificial neural network
ANOVA	analysis of variance
ARF	Abitibi River Forest
ASCII	American Standards Code for Information Interchange
AVG	average
B1 – B8	multispectral bands (vary by sensor)
BD	bulk density
C	total carbon
C:N	carbon-to-nitrogen ratio
C-band	SAR with wavelengths around 5.6 cm
CFI	comparative fit index
CHM	canopy height model
CI	confidence interval
CNN	convolutional neural network
CP	carbon pool
CPU	central processing unit
Cw	eastern white cedar
DEM	digital elevation model
DNN	dense neural network

DSM	digital surface model
DTM	digital terrain model
ED	encoder-decoder neural network
ED-CNN	encoder-decoder composed of CNN layers
ED-DNN	encoder-decoder composed of DNN layers
ESSE	Earth and Space Science and Engineering
FCN	fully connected network
FRI	Ontario Forest Resources Inventory
g	gram (kg for kilogram)
GB	gigabyte (MB for megabyte)
GEE	generalized estimating equation
GCB	Great Clay Belt
GCF	Gordon Cosens Forest
GIS	geographic information system
GLFC	Great Lakes Forestry Centre
HH	horizontal transmit, horizontal receive polarization for SAR
HV	horizontal transmit, vertical receive polarization for SAR
Hz	Hertz (gHz for gigahertz)
JAXA	Japan Aerospace Exploration Agency
k-NN	k nearest neighbors
KW test	Kruskal-Wallis test
L-band	SAR with wavelengths around 15 – 30 cm
La	larch (tamarack)
LAS	LASer file format (for point cloud data)
LAZ	compressed LAS file
LiDAR	light detection and ranging
LIO	Land Information Ontario
LOESS	locally estimated scatterplot smoothing
m	meter (cm for centimeter, km for kilometer)
MAE	mean absolute error

MDPI	Multidisciplinary Digital Publishing Institute
Mix	mixed (no tree species dominant)
Mix-Bf	mixed with balsam fir dominant (50 – 70% plot composition)
Mix-Sb	mixed with black spruce dominant (50 – 70% plot composition)
MLP	multilayer perceptron network
MNDWI	modified normalized difference water index
MNR	Ontario Ministry of Natural Resources
MODIS	Moderate Resolution Imaging Spectroradiometer
MRRTF	multi-resolution ridge top flatness
MRVBF	multi-resolution valley bottom flatness
N	total nitrogen
NAD	North American Datum
NDBI	normalized difference built-up index
NDVI	normalized difference vegetation index
NFI	National Forest Inventory (Canada)
NIR	near infrared
NN	neural network
OGS	Ontario Graduate Scholarship
OMAFRA	Ontario Ministry of Agriculture, Food and Rural Affairs
ON	Ontario
P	populus tree species (trembling aspen and balsam poplar)
PA	producer's accuracy
PALSAR	Phased Array L-band Synthetic Aperture Radar
PCC	percent correct classification
Ph.D.	Doctor of Philosophy
PLS	partial least squares regression
PP	percentage point
R	R programming language
R^2	coefficient of determination
RADAR	radio detection and ranging

RAM	random access memory
ReLU	rectified linear unit
RF	random forest
RMSE	root mean square error
RMSEA	root mean square error of approximation
ROC	receiver operating characteristic
SAGA	System for Automated Geoscientific Analyses
SAR	synthetic aperture radar
Sb	black spruce
SEM	structural equation modeling
SQL	structured query language
SR	surface reflectance
SRMR	standardized root mean squared residual
STDDEV	standard deviation
SVM	support vector machine
SVM Radial	support vector machine with radial-basis functions
Sw	white spruce
TIN	triangulated irregular network
TRI	terrain ruggedness index
TWI	topographic wetness index
UA	user's accuracy
UAV	unmanned aerial vehicle
USGS	United States Geological Survey
UTM	Universal Transverse Mercator coordinate system
VH	vertical transmit, horizontal receive polarization for SAR
VV	vertical transmit, vertical receive polarization for SAR
WV2	WorldView-2 satellite
YU	York University (YorkU)

Chapter 1

Introduction

Land cover conversion due in part to global warming is likely to transpire across the southern transition zones of the boreal forest within Canada (Boulanger et al. 2017; Jiang et al. 2023). Moreover, in many localities, economic prospects pertaining to forestry and agriculture have hasten land cover changes relating to those activities. Vast quantities of carbon are currently contained within the soil and vegetation of boreal regions across the northern hemisphere (Deighton, Bell, and Lindo 2025; Menichetti et al. 2025; Minasny et al. 2019). The concern with land cover conversion, is that oxidation into the atmosphere will materialize for carbon currently stored within vegetation and surface soil layers if the existing forest cover is transformed. This release of carbon can further compound climate change, ultimately stress boreal environments and lead to additional land cover conversion. For that reason, assessing the impact of land cover and land use conversion with respect to carbon is of paramount importance. However, gauging the magnitudes of carbon content within boreal ecosystems is difficult, particularly for resolving the stocks of carbon within the ground. Presently, retrieved soil samples correspond to the verifiable measurements for ascertaining soil carbon, and many challenges arise regarding the extracting of soil samples. Consequently, modeling and related approaches need to be adopted for estimating carbon contents throughout the terrestrial domain of study areas.

Digital soil mapping has become a dominant methodology ascribed for modeling targeted soil properties. The premise of digital soil mapping asserts with predictors conforming to soil formation factors (McBratney, Mendonça Santos, and Minasny 2003; Suleymanov, Arrouays, and Savin 2024). To obtain predictor data covering the geographical extents of study areas, remote sensing technologies are typically employed to generate environmental covariates (Liang et al. 2020; Mulder et al. 2011). These environmental covariates are then utilized as predictors for digital soil mapping. Improvements with modeling approaches and the application of more functional environmental covariates have advanced accuracies attained with digital soil mapping research. Nevertheless, modeling accuracies remain relatively low for modeling soil carbon within boreal regions (Minasny et al. 2019). For the boreal forest of northern Ontario, substantial variations for soil carbon predominate across land cover type, as well as in reference to other soil properties. The interconnection of environmental covariates with soil properties relates to soil formation factors, hence resolving these relations are of interest for subsequent modeling.

To advance digital soil mapping for the boreal context, the following issues should be addressed. Within the boreal forest of northern Ontario, knowledge gaps persist pertaining to the variation of soil carbon with land cover, as well as the interconnection between soil carbon and other associated soil properties. In addition, there is a connection between soil properties and tree species (Mueller et al. 2012; Zhang et al. 2025) that can be better established for boreal study areas. This involves attributes or covariates relating to certain tree species, that can constitute as predictors for digital soil mapping. The recent advancement and availability of remotely sensed data for boreal regions imparts an opportunity for deriving environmental covariates. Accordingly, dynamics between

environmental covariates and soil properties need to be explored, considering both spatial and temporal scales. Uncertainty with modeling also needs to be determined. Furthermore, insights can be acquired from research progress into these subject matters.

The goal of this research was to enhance and formulate methods for resolving soil carbon concentrations and expounding connections with vegetation and other physical factors for the southern transition zone of the boreal forest in northern Ontario, Canada. To fulfill this goal, the following objectives were investigated. Research pertaining to these objectives conforms to the focus of this dissertation.

(1) Initiate comprehensive and proper statistical analysis for the assessment of soil properties from field measurements.

Considerable variations exist for soil properties regarding land cover type within boreal regions (Hyväluoma et al. 2024; Laganière et al. 2013; Vesterdal et al. 2013). The main impetus is in ascertaining concentrations of soil carbon, but also other properties such as clay component, carbon-to-nitrogen ratio and bulk density. Statistical analysis appropriate for skewed non-normal distributions and for potentially clustered samples should be implemented and can be conducted as a two-step process. Firstly, statistical testing can be enacted to establish whether dissimilarities of soil properties persist across land cover and vegetation types. This assessment can confirm whether sufficient heterogeneity has been obtained from the collected soil samples, regarding the utilization of the soil data as target data for soil modeling. Secondly, significances and estimates of refined differences for soil carbon between specific cover types can be resolved. These results yield can support for the

selection of more suitable environmental covariates from remote sensing technologies to better differentiate between certain land cover, to use as predictors for digital soil mapping.

For evaluating diverseness across cover types for each soil property investigated, the Kruskal-Wallis test (Bargagliotti and Greenwell 2015; Johnson 2022) can be adopted. This can be followed by generalized estimating equations (GEEs) (Jung 2008; Salazar et al. 2016) to determine estimates and significances of pairwise differences between groupings, separately by land cover and tree species type. Afterwards, GEEs can be employed to resolve estimates with confidence intervals for soil carbon, based on other soil data, land usage or vegetation information available at sites.

(2) Improve methods for classification of endemic tree species using multi-source remotely sensed data, devising an enhanced quantification of prediction uncertainty.

Vegetation has been identified as one of the most important soil formation factors for boreal regions (Minasny et al. 2019). Thus, the computations of intricate tree species maps are of interest as predictors regarding digital soil mapping. The retrieval of LiDAR (light detection and ranging) data for districts in northern Ontario (Airborne Imaging 2018) presents an opportunity for the generation of more detailed LiDAR-derived covariates. This permits the calculation of covariates at a finer spatial scale to cohere with individual trees in natural settings, so to obtain representation for tree species of lesser occurrences. A more accurate and precise digital elevation model (DEM) can be constructed from the LiDAR data, which can be utilized to generate a suite of detailed topographic covariates relating to various attributes of topography (Franklin 2020). In addition to refined classification for characteristic boreal settings, uncertainty with prediction can be investigated. Specifically, a

more relevant entropy metric for tree species classification can be devised to provide insights when multiple categories are existent.

This classification for endemic tree species can be accomplished by random forest methods (Immitzer, Atzberger, and Koukal 2012; Quan et al. 2023), with comparison to results from support vector machines. Finer resolution environmental covariates can also be applied as predictors for the modeling. Furthermore, entropy metrics can be computed to gauge uncertainty with prediction. A normalized entropy metric (Kumar, Kumar, and Kapur 1986; Wu, Shih, and Tu 2021) just based upon top contending categories can also be formulated for better quantifying prediction uncertainty when compared to the conventional entropy calculation.

(3) Exploit data-driven methods to construct models for augmenting modeling accuracy of soil carbon.

This objective corresponds to attaining the best modeling accuracies possible for soil carbon. Modeling accuracies for soil modeling of carbon within boreal regions in Canada tend to remain relatively low even for more recent studies (Beguin et al. 2017; Mansuy et al. 2014). Machine learning approaches with respect to random forest and support vector machines methodologies have been commonly adopted for digital soil mapping (Minasny et al. 2019; Suleymanov, Arrouays, and Savin 2024). However, there remains a need for further accuracy gain. Deeper machine learning modeling (Ke, Ren, and Yin 2024) compatible for smaller data sets as typical with regional digital soil mapping can be implemented to advance modeling accuracy for soil carbon.

ED neural networks (Ke, Ren, and Yin 2024; Q. Li et al. 2022) suitable for smaller data sets can be formulated and trained, consisting of both dense neural network (DNN) and convolutional neural network (CNN) structures. The ED neural networks can potentially boost modeling accuracy when compared to other machine learning or regression methods, to attain the best possible predictions for soil carbon. Quantile mapping (Maraun 2013) with respect to deciles has been investigated and provides additional insights regarding land covers and localities conforming to the highest uncertainties with modeling prediction. This can involve the quantification of prediction uncertainty as derived from an ensemble modeling approach from the various assessed models.

(4) Develop a framework to integrate knowledge-based methods to improve digital soil mapping of carbon and associated properties.

It is imperative to better discern physical drivers impacting soil properties. A variety of environmental covariates can be utilized as predictors (Chen et al. 2022; Nussbaum et al. 2018) within digital soil mapping, with many of these covariates having been obtained from remote sensing technologies (Mulder et al. 2011). Numerous environment covariates can be utilized, even if ultimately only a few of those predictors fundamentally contribute to the soil modeling (Minasny et al. 2019). Consequently, there is a need to resolve the most important features in advance of soil modeling, particularly for integrated models so to elucidate interconnections between predictors and targeted soil properties.

A framework can be devised to facilitate SEM (Angelini et al. 2016; Rappaport, Amstadter, and Neale 2020) for digital soil mapping. This can involve a multiapproach for establishing variable importance by reducing bias and calculating pairwise correlations between these

identified predictors so to minimize multicollinearity (Gokmen, Dagalp, and Kilickaplan 2022). SEM can be implemented via a stepwise process, retaining only statistically significant predictors for the regression modeling of soil properties. Goodness of fit indices (Chen 2007) should be evaluated so to keep the SEM statistically sound. Finally, path analysis with SEM can determine the relative contributions of predictors in conjunction for an integrated model of soil properties.

(5) Devise a method for processing vast quantities of LiDAR data for the construction of environmental covariates.

The recent collection of LiDAR point cloud data presents prospects for deriving detailed environment covariates, whether relating to topography or forest structure (Takeshige et al. 2025). However, the generation of covariates over the entire domain of study areas can be formidable challenge. There is an impetus for streamlined processing of point cloud data, to summarize essential information and compute rasterized covariates within relatively realistic timeframes.

Structured query language (SQL) (Sumathi and Esakkirajan 2007) can be deployed to derive as well as calculate rasterized environmental covariates from LiDAR point cloud data. Such novel processing can generate a digital terrain model (DTM) or DEM that is more precise than the provincially available DEM for the corresponding study region. Moreover, a canopy height model (CHM) and a gap fraction can also each be computed for the overlying vegetation. These covariates can be used as predictors for both tree species classification and digital soil mapping.

All these aforementioned objectives are linked for fulfilling the goal of advancing methods for resolving the status of soil carbon in relation to vegetation and other attributes within the southern transition zone of northern Ontario's boreal forest. The inferential analysis on the soil properties established significances of soil properties across cover types (An et al. 2022; Wu et al. 2023), supporting the prospect of this soil data constituting as ground truth for digital soil mapping work. In addition, certain tree species and land covers were associated with higher soil carbon concentrations, yielding support for tree species classification as vegetation is an important soil formation factor for soil properties (Minasny et al. 2019). Improved knowledge and prediction maps for specific tree species and vegetation covariates were incorporated as predictors with the digital soil mapping work, leading to improves with modeling accuracies. These covariates also improved accuracies for data-based methods, as well as comprehension of interconnection with predictors and soil properties for knowledge-drive methods. To generate detailed vegetation and topographic covariates from LiDAR as necessary for the tree species classification and digital soil mapping work, SQL (Gaffney et al. 2022) was utilized to effectively generate these covariates for large study areas.

In achieving these research objectives, firstly sufficient quantities of soil samples had to be retrieved from the ground via fieldwork. As a dearth of soil data with chemistry analysis measurements ensued for the study region, consequently soil samples had to be extracted and subsequently processed and analyzed. The soil samples were collected following a soil sampling methodology (Pittman and Hu 2024), which has been briefly explained and referenced in the chapter comprising the statistical inferential work for the soil samples. This soil sample data corresponded to ground truth measurements for the bulk of this research in this dissertation. Information pertaining to tree species was acquired from the Ontario Forest

Resources Inventory (FRI), to be regarded as observation data for the tree species classification.

Various types of modeling were investigated for this research. This included inferential statistical approaches for soil properties, to machine learning classifiers for tree species classification. Furthermore, a variety of regression approaches were examined for soil property modeling. These consisted of conventional regression modeling methods, as well as more basic machine learning and deeper learning approaches. Regarding the processing for vast quantities of point cloud data, structured query language was implemented.

This dissertation is organized as follows into seven chapters. The first chapter (this section) presents the motivation for this research concerning soil carbon with respect to land cover for a boreal study region, as well as discoursing the research goal and objectives. Chapter two pertains to the inferential statistical analysis regarding cover type for the collected soil samples. The tree species classification work for boreal settings is featured in the third chapter. Advancements with modeling accuracy for soil carbon as achieved by encoder-decoder (ED) neural networks, with details on modeling formulations, are expanded on in the fourth chapter. Research pertaining to digital soil mapping as has been investigated through structural equation modeling (SEM) is explained in chapter five. Chapter six provides an overview of structured query language (SQL) with regard to deriving as well as computing rasterized environmental covariates from LiDAR point cloud data. Lastly, chapter seven summarizes the conclusions and highlights the contributions in reference to the research objectives. Limitations for the research work are also mentioned, and recommendations of best practices for digital soil mapping have also been imparted.

Suggestions for future work are discussed concerning improvements for both tree species classification and digital soil mapping within the boreal context.

Research conducted during this Ph.D. program resulted in the following publications:

Journal Papers

- Pittman, R. and B. Hu, 2025. “Enhanced Prediction of Soil Carbon via Encoder-Decoder Neural Networks for a Boreal Study Area in Northern Ontario.” *Sensors* 25(8): 1–23. https://doi.org/10.3390/s25082583.
- Pittman, R., Hu., B., Pittman, T., Webster, K.L., Shang, J., Nelson, S.A., 2025. “Inferential Approach for Evaluating the Association Between Land Cover and Soil Carbon in Northern Ontario.” *Earth* 6(1): 1–21. https://doi.org/10.3390/earth6010001.
- Pittman, R. and B. Hu., 2023. “Contribution of Topographic Features and Categorization Uncertainty for a Tree Species Classification in the Boreal Biome of northern Ontario.” *GIScience and Remote Sensing* 60(1): 1–21. https://doi.org/10.1080/15481603.2023.2214994.

Conference Papers

- Pittman, R. and B. Hu, 2024. “Constructing Rasterized Covariates from LiDAR Point Cloud Data via Structured Query Language.” *Proceedings* 110(1): 1–8. https://doi.org/10.3390/proceedings2024110001.
- Pittman, R. and B. Hu, 2022. “Modelling Gap Probability for a Right Canonical Canopy.” *IGARSS 2022 – 2022 IEEE International Geoscience and Remote Sensing Symposium*: 6244–6247. https://doi.org/10.1109/IGARSS46834.2022.9883848.

- Pittman, R. and B. Hu, 2022. “Improving the Binary Classification of Peat Localities from Multi-source Remotely-sensed Data using CNN.” *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*: 983–988. https://doi.org/10.5194/isprs-archives-XLIII-B3-2022-983-2022.
- Pittman, R., Hu., B. and G. Sohn, 2021. “Determination of Regions Suitable for Agriculture in the Gordon Cosens Forest of Ontario by Means of Analytical Hierarchy Process with Fuzzy Logic Inference.” *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*: 623–629. https://doi.org/10.5194/isprs-archives-XLIII-B3-2021-623-2021.

Resubmitted Journal Paper

- Pittman, R., Hu., B., Webster, K.L., Wang, J., 2025. “An Enhanced Implementation for Structural Equation Modeling of Bulk Density, Carbon and Carbon-to-Nitrogen Ratio within a Boreal Study Region in Northern Ontario, Canada.”

Chapter 2

Inferential Approach for Evaluating the Association between Land Cover and Soil Carbon

Research presented in this chapter has been published in the journal paper:

[1] Pittman, R., Hu., B., Pittman, T., Webster, K.L., Shang, J., Nelson, S.A., 2025. "Inferential Approach for Evaluating the Association Between Land Cover and Soil Carbon in Northern Ontario." *Earth* 6(1): 1–21. https://doi.org/10.3390/earth6010001.

[1] I want to thank the publisher MDPI and the coauthors who have permitted me to reuse this article in my dissertation.

2.1. Background

Resolving carbon stocks in boreal regions is of great importance for researchers, especially for addressing impacts of carbon regarding land cover conversion regarding carbon sequestration (Boulanger et al. 2017; Fradette et al. 2021). Within boreal regions, it has been established that greater reserves of carbon can exist with the ground that within overlying vegetation (Minasny et al. 2019). Consequently, there is a need to determine the contents of carbon within soil and thus resolve the status of carbon by land cover type. As soil samples correspond to the ground truth of carbon, consequently, to measure carbon there needs to soil samples collected. In addition, associations between carbon with other soil properties typically exist (McBratney, Mendonça Santos, and Minasny 2003; Mulder et al. 2011), but these relationships can vary from study area to study area and may not be well-resolved for boreal regions. For those reasons, it can be advantageous in studying other soil properties in addition to carbon, to better understand characteristics pertaining to soil carbon.

Many factors need to be considered for collecting an adequate amount of soil samples for digital soil mapping. Most importantly, enough samples should be extracted to represent the heterogeneity of soil properties within the study region. Additionally, there is an interest in resolving differences between soil properties between certain cover types. This information can be of relevance for the selection of environmental covariates to utilize as predictors for a subsequent digital soil mapping. Most of these predictors conform to environmental covariates that are primarily retrieved from remote sensing technologies (Minasny et al. 2019; Mulder et al. 2011). Specifically, remote sensors can generally distinguish between land cover types (Mulder et al. 2011).

Inferential analysis by means of hypothesis testing can provide support in a statistical context of differences of soil carbon across various cover types, whether corresponding to land cover usage or dominant tree species. Typically, analysis of variance (ANOVA) has been employed for evaluating statistical difference between groupings (Fang et al. 2018; Simfukwe et al. 2011; Suratno, Seielstad, and Queen 2009). However, certain assumptions are presumed for ANOVA, namely that a targeted property with respect to a population contends to a normal distribution (Driscoll 1996; Lantz 2013). This assumption is generally not valid for soil samples retrieved via field campaigns, as certain soil properties can be skewed in distributions or are collected from subplots that are clustered. Hence, there is a need for the implementation of more appropriate statistical approaches for hypothesis testing with soil samples data. The Kruskal-Wallis (KW) test (Bargagliotti and Greenwell 2015; Vargha and Delaney 1998) is a suitable alternative to ANOVA that can be adopted for clustered non-normally distributed data. Specifically, the KW test is nonparametric and comparable to one-way ANOVA testing (Bargagliotti and Greenwell 2015). The KW test assesses for differences with the median, without requiring data to be normally distributed (Vargha and Delaney 1998). As the KW test evaluates for significance with median rather than mean values, consequently it is less impacted by outliers in data (Lung-Yut-Fong et. al 2011) as therefore less likely to reject the null hypothesis and inflate significances (Bargagliotti and Greenwell 2015). can be the case with soil sampling data. All in all, the KW test is suitable for soil sampling data as the distributions of soil properties can be skewed and are generally unknown.

For refined statistical analysis, generalized estimating equations (GEEs) (Jung 2008; Salazar et al. 2016) can resolve further inferences between distinct cover types without requiring

assumptions about normality for the underlying distributions of data. GEEs have been applied for assessing clinical trials within the medical sciences (Dahmen and Ziegler 2004; Salazar et al. 2016) for obtaining estimates for targeted properties. Advantages of GEEs include that the correct specification of distribution of the response variable need not be specified or known (Jung 2008, Dahmen and Ziegler 2004) and can be used on nested data (Zeger et al. 1988). To best available knowledge, GEEs have not been previously employed for inferential testing for data within soil science research. Furthermore, GEEs enable the possibility of more refined statistical testing between specific cover types, permitting more in-depth analysis for determining what specific cover types correspond to greater concentrations of soil carbon. On that account, the KW test and GEEs can be deployed for complementary statistical testing for corroborating inferences of soil properties regarding cover type.

2.2. Study Region

The study region of interest corresponds to what is colloquially referred to as the Great Clay Belt (GCB) of northern Ontario, situated within the District of Cochrane in Ontario, Canada. For this expanse, the longitudes range from approximately 80°40’ W – 84° W, over the latitudes of 49° N – 50° N. This region has been studied due to the risk of future land cover conversion within the southern transition zone of the boreal forest, whether through economic activities or climate change (Boulanger et al. 2017). Forestry and agriculture are important economic activities for this region, which has resulted in land cover changes. Furthermore, this study region is all situated within the GCB, which has unique sublevel soil characteristics differing from neighboring regions (Kroetsch et al. 2011). Study areas are focused along the

Ontario highway 11 corridor, and from the northwest to the southeast consist of Hearst, the Gordon Cosens Forest (GCF), Kapuskasing within the GCF, and Cochrane. The delineation of the study areas within the Great Clay Belt study area are depicted in Figure 2.1, with reference to the water bodies within the study region.

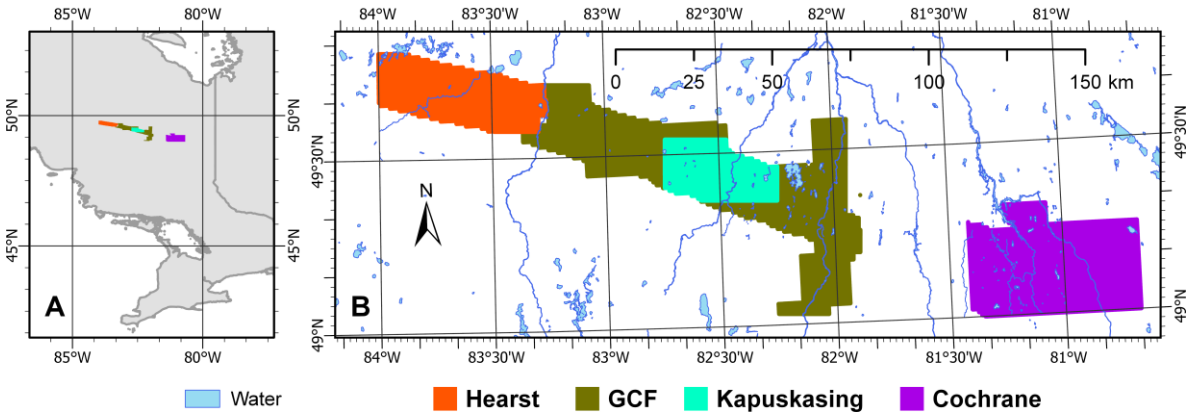

Figure 2.1. Designated study areas within the Great Clay Belt (GCB) region. A: Reference of location for the study areas in the context of northern Ontario, Canada. B: Delineated study areas, showing more detailing regarding water bodies and coordinates.

Geology of this region conforms to the Canadian Shield, with the land profile dotted with numerous lakes and rivers. The rivers flow principally on northeast trajectories towards James Bay in the Artic Ocean. Topography for this region is relatively flat at the landscape scale. Elevations vary from around 120 m in the northeast to a maximum of 350 m northwest of the community of Cochrane, with a gradual transition to elevations approximately 100 m higher across western portions. Regarding soil family for the deeper soil horizon layers, grayish luvisols are present throughout the study region (Kroetsch et al. 2011); hence the attribution to this region as the Great Clay Belt of northern Ontario. Other soils present include gleysols and mesisols within wetland localities (Kroetsch et al. 2011).

This area consists of a mixture of forested and agricultural land cover types, situating across the southern transition zone of the boreal forest. These include mature and secondary forests, as well as abandoned pastures with regrowth of trees, and wetlands intrinsically with lower canopies. Agricultural land tends to have no trees present and principally correspond to hay fields and croplands. Prevalent tree species comprise of black spruce (*Picea mariana*), balsam fir (*Abies balsamea*) and trembling aspen (*Populus tremuloides*). Other tree species present throughout this region consist of white spruce (*Picea glauca*), tamarack (*Larix laricina*) and balsam poplar (*Populus balsamifera*). Trees that are rare for this region include eastern white cedar (*Thuja occidentalis*) due to northern confines with climate, and jack pine (*Pinus banksiana*) which is impacted due to the presence of heavier clayey soil at sublayer depths. For agriculture, hay fields generally consist of brome grass (*Bromus*) mixed with either vetch (*Vicia sativa*) or red clover (*Trifolium pratense*). Typical crops that are seeded for the croplands are oats (*Avena sativa*) and barley (*Hordeum vulgare*), with varieties of canola (*Brassica rapa*, *B. napus* and *B. juncea*) being more commonly grown during the past few years.

2.3. Soil Data

Soil samples were collected from a variety of representative cover types within delineated study areas for the study region. To assess variation of samples within a sampling site, samples were obtained from 3 subplots for each site. Each site corresponded to one cover type, with the same dominant tree species noted for the subplots. These subplots were located at distances of 4.5 m, 7.5 m and 9.5 m at bearings of 0°, 120°, and 240° respectively, from the site center. In total, soil samples were extracted for 144 sites corresponding to 431

subplots. These samples were retrieved during three separate field campaigns in September 2018 (12 sites), August 2019 (28 sites) and August–September 2021 (104 sites). For the study areas, this corresponded to 32 sites near Ryland for Hearst, 42 sites near Kapuskasing and 3 sites near Fauquier-Strickland for the Gordon Cosens Forest (GCF), and 67 sites for Cochrane, respectively. As these soil properties were extracted within a three-year period and were collected over various cover types for each year, temporal differences between the sampling years were negligible. Soil chemistry properties can take decades or longer before major changes in compositions occur as soil properties likely correspond to logarithmic or square root of time relationships (McBratney, Mendonça Santos, and Minasny 2003). Furthermore, many studies consider soil samples collected over multiple years with legacy soil data (Minasny et al. 2019), so temporal differences in soil carbon measurements were of minor concern. Depicted in Figure 2.2 are the locations of the soil sampling sites, here with the backgrounds of the study areas corresponding to a canopy height model (CHM).

For extraction, soil samples for chemistry properties were obtained at the mineral soil depths of 0–5 cm, 5–15 cm, 15–30 cm, 30–45 cm and 90–105 cm for the 2018 and 2019 field campaigns, but only for the 0–5 cm and 5–15 cm depth layers during the 2021 field campaign. Bulk density (BD) samples were extracted for all subplots for the 0–5 cm and 5–15 cm depth layers. At each subplot, pits were dug to the depth of around 30 cm for collecting samples. The BD samples were extracted via BD rings, with careful attention heeded for fully and consistently filling the BD rings. A soil auger was utilized for obtaining soil samples from the 30-45 cm and 90-105 cm depth layers.

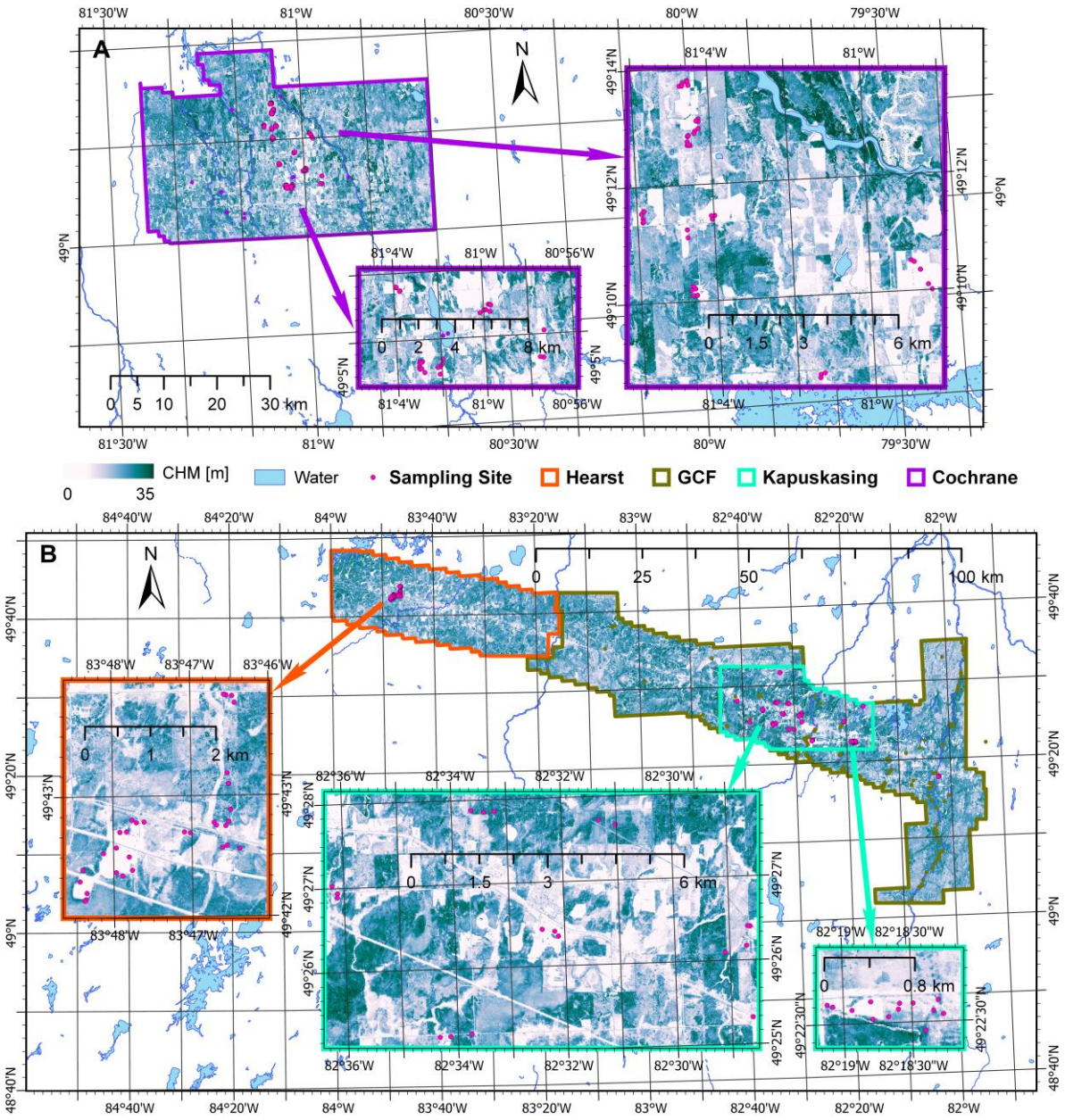

Figure 2.2. Study areas and location of sampling sites, with canopy height model (CHM) [m] depicted within backgrounds of delineated study areas. Water bodies within region are also displayed. A: Cochrane study area, with insets focusing on clusters of sampling sites. B: Hearst, Gordon Cosens Forest (GCF) and Kapuskasing (within GCF) study areas, with insets focusing on clusters of sampling sites.

The property investigated for soil C corresponded to total carbon in mineral soil (Gao et al. 2018; Laganière et al. 2013), which comprised of both organic carbon and inorganic carbon. Carbon concentration was of interest for the soil C measurements, as these were the C resolved from the laboratory analysis. Analysis for carbon stocks was avoided as carbon stocks required assumptions with profile depths and masses (Minasny et al. 2019) over deeper soil horizons. Due to the presence of roots for biological relevance (Gao et al. 2018) and general maximum carbon contents, the 5–15 cm depth layer was the focus of analysis for soil properties with this research. The profile of C concentration by mineral soil depth is shown in Figure 2.3, as obtained from the summary Table A.1 of all soil properties from Appendix A.

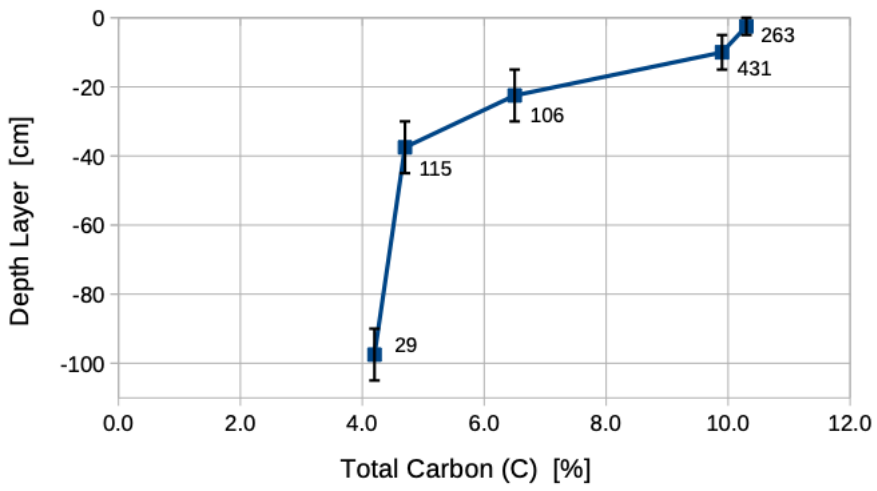

Figure 2.3. Plot of total carbon (C) concentration [%] by depth. Note that vertical bars denote the depth layers (0–5 cm, 5–15 cm, 15–30 cm, 30–45 cm, and 90–105 cm). Number labels indicate the number of soil samples processed for C measurements at each depth layer.

Once collected, all the soil samples had to be processed before analysis for measurements. Firstly, the samples were sufficiently air-dried for the duration of a couple of months. The

BD samples were further thermal dried in an oven at 110 °C for a minimum of 48 h. Once completely dry, these BD samples were pulverized and weighed, with coarse fragments of rocks and root material removed, so to calculate the mass per volume for the soil material. Regarding chemistry soil samples, these were processed to assess carbon (C) and nitrogen (N) contents by means of combustion methods (Gao et al. 2018). Textural components of clay, silt and sand were also resolved. Measurements for BD were reported in gram per cubic centimeter [g/cm^3], whereas C, N and clay component were each noted in percentage points [%]. Carbon-to-nitrogen ration (C:N), here defined as C divided by N, was unitless. All the BD samples were processed in the Great Lakes Forestry Centre (GLFC) (Sault Ste. Marie, ON, Canada). The chemistry soil samples were processed and analyzed in the GLFC for samples collected during the 2018 field campaign, but all other samples were assessed by A&L Canada Laboratories Inc. (London, ON, Canada).

To analyzing the soil data by cover types, the dominant tree species and land cover types were observed and noted for all the sampling sites. The purpose of this analysis was to perform analysis using only soil samples and site attributes observed and recorded within fieldnotes, prior to and independent of digital soil mapping work. Categories for the tree species groupings included balsam fir, balsam poplar, black spruce, larch, trembling aspen, white spruce, as well as a category for none. Groupings for land cover type consisted of abandoned field, crop field, hay field or pasture, old yard site or clearing, recently cleared land, secondary forest or woodlot, mature forest, and wetland.

2.4. Statistical Analysis

A Two-step inferential analysis was implemented. The first step entailed the KW test to substantiate median differences of soil properties across the various land cover types and dominant tree species at site. This was followed by a second step involving GEEs to resolve refined estimates of mean differences of soil properties between specific cover types. For a final analysis, GEEs were also employed to estimate soil C content and associated confidence intervals based on clay content, land cover and tree species information.

2.4.1. Kruskal-Wallis Test

The Kruskal-Wallis (KW) test assessed for median differences among cover types, and the formula adopted for the KW test statistic (Johnson 2022; De Neve and Thas 2015) was

$$K = \frac{12}{n(n+1)}\left[\sum_{i=1}^{g} n_i \left(\bar{R}_i - \frac{n+1}{2}\right)^2\right]. \quad (2.1)$$

Here, n corresponded to the total number of observations, with g the number of groupings with n_i observations for grouping i, with $\bar{R}_i$ the average rank of observations corresponding to grouping i. Average rank was computed as the arithmetic mean of the index ranking for all observations per grouping. For calculating the index ranking, all observations were arranged in increasing order from smallest to largest. The observation corresponding to the smallest value was assigned an index ranking of 1, with averages of respective rankings when there were two or more observations of the same value. Within the formula, the $(n + 1)/2$ term was regarded as an overall average of the assigned ranks (Johnson 2022). Hence, the KW test statistic was maximized when the absolute difference between the average rank of observations per grouping with respect to the overall average rank was greatest. Regarding

the null hypothesis, the KW test statistic approximated a χ^2_{g-1} (chi-squared) distribution with $g-1$ degrees of freedom (Johnson 2022).

The null hypothesis corresponded to the medians of the target soil property being the same across the various cover types. Contrarily, the alternative hypothesis conformed to the median of the target soil property of at least one cover type being different from the medians for other cover types. A hierarchical KW text was implemented for nested inferences accounting for the subplots. The nested null hypothesis was that for each cover type the sites had the same median value for the target soil property. Correspondingly, the nested alternative hypothesis was that this median value for at least one site was different.

For the statistical analysis, the stats package in R (R Core Team 2012) was utilized for evaluating the KW test. The reportRmd package (Avery et al. 2023) was adopted for the hierarchical KW tests, for implementing both the unnested and nested assessments. Reporting thresholds for the level of significance with the *p*-values were marked with asterisks; 0.05 significance was noted with one asterisk, 0.01 significance with two asterisks, and 0.001 significance with three asterisks, accordingly. In addition, effect size regarding the eta-squared effect size (Cohen 1988; Richardson 2011) was reported for the unnested evaluations. The effect size assessed the strength of relationships regarding to practical significance that were less subjected to sample size (Hojat and Xu 2004). Eta-squared effect size was bounded from 0 to 1, with a greater practical significance imparted by a larger value (Cohen 1973; Richardson 2011). For interpretation (Cohen 1988; Richardson 2011), eta-squared effect sizes of 0.01 conformed to a small effect, around 0.06 corresponded to a medium effect, and greater than 0.14 indicated a large effect.

2.4.2. Generalized Estimating Equations

A GEE was adopted to obtain unbiased and consistent parameter estimates for a generalized linear model (Dahmen and Ziegler 2004; Zeger and Liang 1986), which was a more general form of a regression model. Brief background following the notation (Jung 2008) was presented. For response observations $\boldsymbol{y}_i = \left(y_{i1}, \dots, y_{in_i}\right)^T$ with n_i measurements for subject $i = 1, \dots, M$, the covariate matrix was of form $\boldsymbol{X}_i = \left(\boldsymbol{X}_{i1}, \dots, \boldsymbol{X}_{in_i}\right)^T$ with dimension $n_i \times p$ for p covariate variables. In the context of soil samples, the sites corresponded to subjects and measurements to subplots as each subplot had a unique observation per target soil property. The means were $\mathrm{E}\left(y_{ij}\right) = \mu_{ij}$ and the variances were of form $\mathrm{var}\left(y_{ij}\right) = \phi v\left(\mu_{ij}\right)$ with ϕ a scale parameter and v a variance function. The goal was to calculate a $p \times 1$ vector of regression parameters $\boldsymbol{\beta} = \left(\beta_1, \dots, \beta_p\right)^T$ corresponding to covariate variables of the generalized linear model. For expressing the estimating equation (Jung 2008) conforming to M subjects the form was

$$\sum_{i=1}^{M} \left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}\right)^T [\boldsymbol{V}_i(\boldsymbol{\alpha})]^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu}_i) = \boldsymbol{0}_p. \tag{2.2}$$

Here, $\boldsymbol{V}_i = \phi \boldsymbol{A}_i^{1/2} \boldsymbol{R}_i(\boldsymbol{\alpha}) \boldsymbol{A}_i^{1/2}$ was the variance matrix and $\boldsymbol{A}_i$ was a diagonal matrix of dimension $n_i \times n_i$ with $v\left(\mu_{ij}\right)$ for each *j*-th diagonal element. The working correlation matrix was $\boldsymbol{R}_i(\boldsymbol{\alpha})$ for the *i*-th subject (Zeger and Liang 1986), which consisted of hypothesized correlations between the response variables regarding an association parameter vector $\boldsymbol{\alpha}$ (Li 1997). A correct parametrization for the working correlation matrix did not need to be presumed (Li 1997; Salazar et al. 2016) as GEEs exhibited robustness to incorrect specifications of an initial correlation matrix (Salazar et al. 2016). Iteration by means of

weighted least-squares methods was exploited to solve the estimating equation, implementing moment estimators of ϕ and $\boldsymbol{\alpha}$ (Jung 2008) to ultimately solve for $\boldsymbol{\beta}$. Univariable and multivariable formulations with either correlated or uncorrelated error structures were specified for approximating a likelihood function (Yan and Fine 2004) which was possible to remain unknown (Jung 2008). Regarding statistical inference for the fit of a GEE, the Wald test was utilized with the test statistic approximating an *F*-distribution for the null hypothesis (Fan and Zhang 2014).

GEEs were utilized to evaluate for mean differences of a target soil property between two separate groupings. To assess for mean differences between cover types, Tukey's honest significance test was used with test statistic for the null hypothesis corresponding to a studentized range distribution (Goeman and Solari 2022). For each pairwise difference, the null hypothesis conformed to the corresponding β parameter for the mean difference equaled to zero; the alternative hypothesis being that this mean difference was not equaled to zero.

For the analysis with GEEs, the geepack package (Højsgaard et al. 2006) in R was employed, which was facilitated by the reportRmd package (Avery et al. 2023). Univariable GEEs were adopted for the statistical inference of mean differences with soil C between the cover types. A multivariable GEE was employed to estimate soil C based on land cover, dominant tree species and clay percentage, as the multivariable model controlled for effects from more than one predictor. The coefficient of determination (R^2) was reported for the fitting of the multivariate GEE. A larger R^2 corresponded to a better model, as the R^2 pertained to the variation proportion explicated by the model (Keskin and Grunwald 2018; Smith et al. 2017).

2.5. Results

Tabulation summaries of the target soil properties with respect to dominant tree species for the samples have been depicted in Figure 2.4 using bar plots with standard errors. Further refinement including land cover type for the target soil properties is shown in Figure 2.5. Inspecting Figure 2.4, differences amongst the various dominant tree species are apparent, particularly for C, BD and clay percentage. The bar plots for C:N ratio are more consistent across the various tree species. When looking at the bar plots from Figure 2.5, differences between land cover types are more obvious. There are stark differences in soil C across the land cover types, with large differences also apparent for BD and clay percentage. In particular, land cover types with higher C tend to have lower BD, and vice versa. These appreciable differences between land cover and tree species, are an impetus for establishing significance for these interactions with respect to statistical inference. Specifically, mean differences between land cover and tree species pairings are the subject for the GEE statistical inference.

Statistical significance for each target soil property across the dominant tree species groupings, as well as by land cover type, was established by the KW test in Table 2.1 and Table 2.2, respectively. Note that one decimal place of precision was used for reporting results for C, clay percentage and C:N, as C and clay percentage corresponded to percentages, and C:N was a ratio value ranging from less than 10 to more than 25 as observed in Figure 2.5. However, two decimal place precision was retained for BD results as BD only ranged from 0 to no more than 3 g/cm^3 indicated in Figure 2.5 and Table 2.1 and Table 2.2. None of the nested *p*-values were significant, indicating that no statistical median differences were

inferred for any of the soil properties between subplots. Regarding effect size, all statistical inference had larger effects except for clay percentage with respect to dominant tree species, or for C:N by either dominant tree species or land cover type. Nonetheless, effect sizes were still moderate for all these noted exceptions.

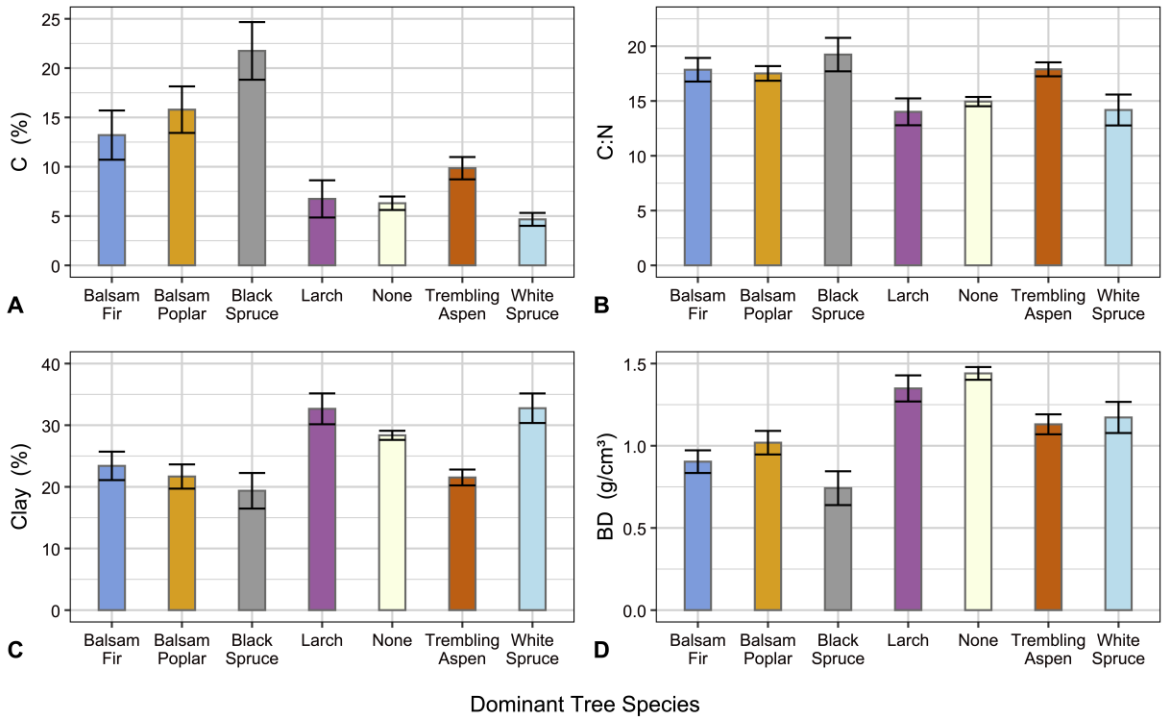

Figure 2.4. Bar plots of mean values soil properties, with standard errors indicated, by dominant tree species present at site for A: total carbon (C), B: carbon-to-nitrogen ratio (C:N), C: clay component, and D: bulk density (BD).

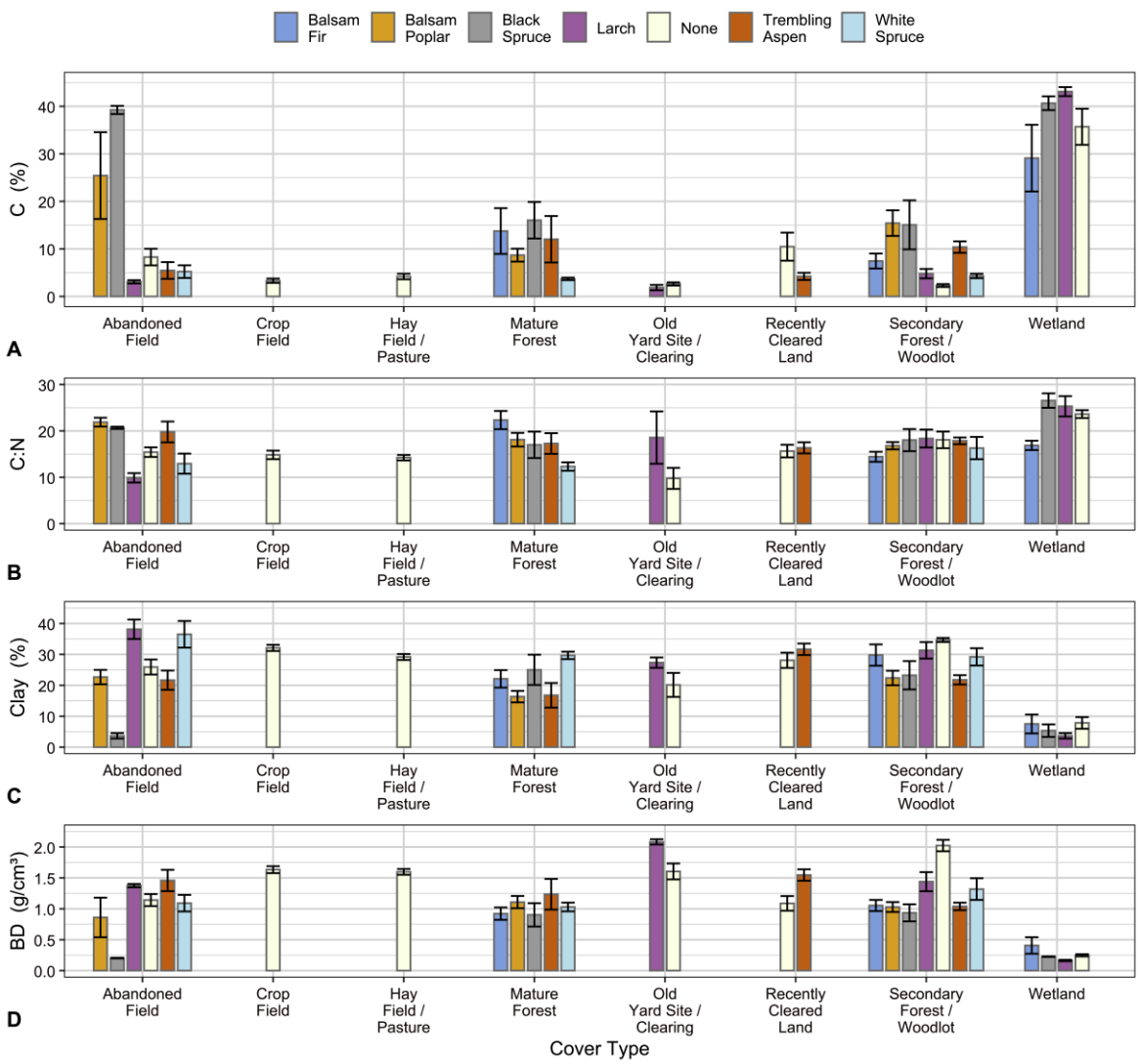

Figure 2.5. Bar plots of mean values soil properties, with standard errors indicated, as summarized by dominant tree species present and land cover for A: total carbon (C), B: carbon-to-nitrogen ratio (C:N), C: clay component, and D: bulk density (BD).

Table 2.1. Summaries of soil properties for median values by dominant tree species present at site. Statistical inference for the unnested *p*-values and effect size and by nested *p*-values are both resolved from the Kruskal-Wallis (KW) test.

Soil Property		Full Sample	Balsam Fir	Balsam Poplar	Black Spruce	Larch	None	Trembling Aspen	White Spruce	Unnested		Nested
		(*n* = 431)	(*n* = 39)	(*n* = 54)	(*n* = 36)	(*n* = 36)	(*n* = 180)	(*n* = 63)	(*n* = 23)	Effect Size	*p*-Value	*p*-Value
C	[%]	4.3	5.8	7.9	23.4	3.2	3.0	7.1	4.2	0.15	**<0.001** ***	0.88
C:N		16.0	17.7	17.8	20.5	13.3	15.1	17.7	13.1	0.07	**<0.001** ***	0.88
Clay	[%]	27.0	25.1	20.0	14.0	29.0	29.0	21.0	32.6	0.11	**<0.001** ***	0.90
BD	[g/cm^3]	1.31	1.01	1.15	0.33	1.37	1.50	1.06	1.22	0.18	**<0.001** ***	0.74

Table 2.2. Summaries of soil properties for median values by land cover type at site. Statistical inference for the unnested *p*-values and effect size and by nested *p*-values are both resolved from the Kruskal-Wallis (KW) test.

Soil Property		Full Sample	Abandoned Field	Crop Field	Hay Field/ Pasture	Mature Forest	Old Yard Site/ Clearing	Recently Cleared Land	Secondary Forest/ Woodlot	Wetland	Unnested		Nested
		(*n* = 431)	(*n* = 74)	(*n* = 42)	(*n* = 75)	(*n* = 48)	(*n* = 9)	(*n* = 24)	(*n* = 138)	(*n* = 21)	Effect Size	*p*-Value	*p*-Value
C	[%]	4.3	4.6	2.6	2.8	5.2	2.4	3.9	6.2	40.6	0.28	**<0.001** ***	0.93
C:N		16.0	15.3	15.7	13.8	17.5	11.4	16.1	16.6	23.6	0.11	**<0.001** ***	0.80
Clay	[%]	27.0	28.0	32.0	29.0	22.5	24.0	28.0	25.0	4.0	0.18	**<0.001** ***	0.75
BD	[g/cm^3]	1.31	1.32	1.71	1.59	1.08	1.91	1.34	1.11	0.22	0.33	**<0.001** ***	0.92

Inferential analysis with GEEs for mean differences in total carbon (C) are reported in Table 2.3 with respect to dominant tree species and Table 2.4 regarding land cover type, respectively. From Table 2.3, there is evidence to support that the presence of black spruce corresponds to larger amounts of soil C than for other tree species. Notably, the estimate of the mean difference with soil C for black spruce minus the no tree present category (i.e. none) is positive and significant with a p-value < 0.05, indicating that black spruce plots contain more soil carbon. The difference of mean soil C for white spruce minus black spruce is negative with a p-value < 0.01 regarding significance, inferring that black spruce conforms to more soil C than white spruce. There is also evidence to suggest that plots with balsam poplar also correspond to more soil C, as the difference of mean soil C for white spruce minus balsam poplar is negative with a p-value < 0.05. For the GEE analysis regarding land cover types in Table 2.4, there are distinct differences in soil C for certain land covers. There exists more soil C in wetlands than in almost any other land cover type, with p-values < 0.001. In addition, each of secondary forest and mature forest have mean differences of soil C between either crop fields or old yard sites with p-values < 0.05 for significance.

Table 2.3. Estimates of mean differences in total carbon (C) between groupings regarding dominant tree species, as resolved by generalized estimating equations (GEEs). Confidence interval (CI) bounds, standard errors and *p*-values for significance are reported.

Comparison	Estimate [%]	Confidence Bound Lower 95% [%]	Confidence Bound Upper 95% [%]	Standard Error	*p*-Value
Balsam Fir — None	8.70	−3.81	21.20	4.34	0.370
Balsam Poplar — None	8.42	−0.66	17.50	3.15	0.088
Black Spruce — None	15.63	2.00	29.26	4.73	**0.014 ***
Larch — None	0.95	−9.03	10.93	3.46	1.000
Trembling Aspen — None	3.06	−1.79	7.91	1.68	0.492
White Spruce — None	−1.07	−5.70	3.56	1.61	0.993
Balsam Poplar — Balsam Fir	−0.27	−15.15	14.61	5.17	1.000
Black Spruce — Balsam Fir	6.93	−11.09	24.95	6.25	0.912
Larch — Balsam Fir	−7.75	−23.19	7.70	5.36	0.744
Trembling Aspen — Balsam Fir	−5.64	−18.38	7.11	4.42	0.841
White Spruce — Balsam Fir	−9.77	−22.43	2.90	4.40	0.248
Black Spruce — Balsam Poplar	7.20	−8.63	23.04	5.50	0.822
Larch — Balsam Poplar	−7.48	−20.31	5.35	4.45	0.589
Trembling Aspen — Balsam Poplar	−5.36	−14.77	4.04	3.27	0.614
White Spruce — Balsam Poplar	−9.49	−18.79	−0.20	3.23	**0.042 ***
Larch — Black Spruce	−14.68	−31.05	1.69	5.68	0.111
Trembling Aspen — Black Spruce	−12.57	−26.42	1.29	4.81	0.102
White Spruce — Black Spruce	−16.70	−30.47	−2.92	4.78	**0.007 ****
Trembling Aspen — Larch	2.11	−8.17	12.39	3.57	0.996
White Spruce — Larch	−2.02	−12.20	8.16	3.53	0.997
White Spruce — Trembling Aspen	−4.13	−9.37	1.11	1.82	0.225

Table 2.4. Estimates of mean differences in total carbon (C) between groupings regarding land cover type, as resolved by generalized estimating equations (GEEs). Confidence intervals (CI) bounds, standard errors and *p*-values for significance are reported.

Comparison	Estimate [%]	Confidence Bound Lower 95% [%]	Confidence Bound Upper 95% [%]	Standard Error	*p*-Value
Abandoned Field — Mature Forest	−4.54	−16.64	7.57	4.17	0.940
Crop Field — Mature Forest	−10.33	−20.53	−0.12	3.51	**0.045** *
Hay Field/Pasture — Mature Forest	−9.21	−19.53	1.12	3.55	0.116
Old Yard Site/Clearing — Mature Forest	−11.40	−21.57	−1.24	3.50	**0.017** *
Recently Cleared Land — Mature Forest	−3.19	−19.64	13.26	5.66	0.999
Secondary Forest/Woodlot — Mature Forest	−3.15	−14.02	7.72	3.74	0.985
Wetland — Mature Forest	22.94	10.49	35.39	4.28	**<0.001** ***
Crop Field — Abandoned Field	−5.79	−12.46	0.88	2.29	0.138
Hay Field/Pasture — Abandoned Field	−4.67	−11.52	2.18	2.36	0.412
Old Yard Site/Clearing — Abandoned Field	−6.87	−13.48	−0.26	2.28	**0.036** *
Recently Cleared Land — Abandoned Field	1.35	−13.18	15.87	5.00	1.000
Secondary Forest/Woodlot — Abandoned Field	1.38	−6.27	9.03	2.63	0.999
Wetland — Abandoned Field	27.48	17.72	37.23	3.36	**<0.001** ***
Hay Field/Pasture — Crop Field	1.12	−0.99	3.23	0.73	0.719
Old Yard Site/Clearing — Crop Field	−1.08	−2.19	0.04	0.38	0.065
Recently Cleared Land — Crop Field	7.14	−5.84	20.12	4.47	0.682
Secondary Forest/Woodlot — Crop Field	7.17	3.17	11.18	1.38	**<0.001** ***
Wetland — Crop Field	33.27	26.00	40.53	2.50	**<0.001** ***
Old Yard Site/Clearing — Hay Field/Pasture	−2.20	−4.12	−0.28	0.66	**0.014** *
Recently Cleared Land — Hay Field/Pasture	6.02	−7.05	19.09	4.50	0.841
Secondary Forest/Woodlot — Hay Field/Pasture	6.05	1.75	10.36	1.48	**0.001** **
Wetland — Hay Field/Pasture	32.15	24.72	39.58	2.56	**<0.001** ***
Recently Cleared Land — Old Yard Site/Clearing	8.22	−4.73	21.17	4.46	0.507
Secondary Forest/Woodlot — Old Yard Site/Clearing	8.25	4.34	12.16	1.35	**<0.001** ***
Wetland — Old Yard Site/Clearing	34.34	27.13	41.55	2.48	**<0.001** ***
Secondary Forest/Woodlot — Recently Cleared Land	0.04	−13.47	13.55	4.65	1.000
Wetland — Recently Cleared Land	26.13	11.32	40.93	5.09	**<0.001** ***
Wetland — Secondary Forest/Woodlot	26.09	17.92	34.27	2.81	**<0.001** ***

GEEs were also employed to estimate soil C using just clay component, dominant tree species and land cover type as predictors. Both univariable and multivariable formulations were utilized, with results reported in Table 2.5. Here, crop field was set as the reference land cover type, with none (i.e. no tree present) set as the reference dominant tree species grouping. The estimate for clay was negative, as in general land cover types with more clay had lower soil C, as seen in Figure 2.3. The multivariable results were less significant than those from the univariable model. Nonetheless, the multivariable GEE fit still attained a R^2 of 0.59, indicating that GEEs feasibly estimated soil C utilizing a model formulation based only on clay component, dominant tree species and land cover type.

Table 2.5. Estimates for total carbon (C) as formulated for generalized estimating equations (GEEs) considering clay component, land cover type and dominant tree species grouping for both univariable and multivariable models. Corresponding 95% confidence intervals (CIs) and *p*-values are noted.

Predictor	Number of Samples (*n*)	Univariable		Multivariable	
		Estimate and 95% CI [%]	*p*-Value	Estimate and 95% CI [%]	*p*-Value
Clay	424	−0.67 (−0.74, −0.60)	**<0.001** ***	−0.48 (−0.59, −0.37)	**<0.001** ***
Land Cover Type	431		**<0.001** ***		**<0.001** ***
- Crop Field	42	Reference		Reference	
- Abandoned Field	74	5.46 (1.28, 9.64)	**0.01** *	2.35 (−1.02, 5.72)	0.17
- Hay Field/Pasture	75	0.88 (−3.28, 5.05)	0.68	−0.59 (−2.71, 1.53)	0.58
- Mature Forest	48	9.57 (5.00, 14.14)	**<0.001** ***	1.85 (−4.11, 7.81)	0.54
- Old Yard Site/Clearing	9	−0.94 (−8.88, 7.01)	0.82	−5.79 (−8.74, −2.85)	**<0.001** ***
- Recently Cleared Land	24	6.39 (0.85, 11.92)	**0.02** *	4.74 (−2.77, 12.25)	0.22
- Secondary Forest/Woodlot	138	7.70 (3.88, 11.51)	**<0.001** ***	2.68 (−1.40, 6.76)	0.2
- Wetland	21	32.98 (27.20, 38.76)	**<0.001** ***	18.00 (11.43, 24.56)	**<0.001** ***
Dominant Tree Species	431		**<0.001** ***		0.16
- None	180	Reference		Reference	
- Balsam Fir	39	6.91 (2.76, 11.06)	**0.001** **	1.17 (−5.08, 7.43)	0.71
- Balsam Poplar	54	9.49 (5.84, 13.14)	**<0.001** ***	2.96 (−3.11, 9.04)	0.34
- Black Spruce	36	15.45 (11.15, 19.74)	**<0.001** ***	7.42 (0.86, 13.99)	**0.03** *
- Larch	36	0.44 (−3.85, 4.73)	0.84	0.71 (−3.12, 4.55)	0.71
- Trembling Aspen	63	3.55 (0.11, 7.00)	**0.04** *	−1.31 (−5.56, 2.95)	0.55
- White Spruce	23	−1.63 (−6.84, 3.57)	0.54	−0.76 (−4.85, 3.33)	0.72

2.6. Discussion

The differences of soil C between the various land cover types in Table 2.4 are apparent. Due to these contrasts with C, environmental covariates relating to land cover types could likely be useful as predictors for digital soil mapping of C. These environmental covariates can be generated from data that has been obtained from remote sensing technologies, to distinguish different land cover types. A canopy height model (CHM) can resolve differences in height between agricultural land with low CHM, versus high CHM for forested tracts. The normalized difference vegetation index (NDVI) can also be important as this index can differentiate between vegetation types. Topographic covariates relating to hydrology can also be of consequence, as noted in Table 2.4 the wetlands contain the greatest C. Thus, the results of the inferential statistical analysis yields support for the selection of environmental covariates to differentiate by land cover type and tree species type, that can be applied as predictors for a related digital soil mapping.

The KW and GEEs conformed to distinct assessments (Vargha and Delaney 1998) in establishing significance for differences of soil properties with respect to cover types. Specifically, the KW test evaluated median differences, whereas mean differences were resolved by GEEs. With regards to the distribution of soil properties, the median was a more appropriate metric than the mean as some soil properties were skewed and consisted of outliers. The variation of soil properties across cover types was obvious from inspecting the bar plots in Figures 2.4 and 2.5, thus assessing for median differences across cover types was more suitable. Regarding refined analysis between specific pairings of cover types, outliers were less of an issue for comparisons between two cover types. GEEs resolved significance

for mean differences of C between pairings of cover types. Thus, establishing significances from both median and mean differences were substantiating, warranting the two-step inferential analysis approach adopted for this research.

GEEs can also be adapted to infer estimates of soil C for legacy soil sample data. For the vicinity of this study region, the Ontario Forest Resources Inventory (FRI) has extracted soil samples for over one thousand sites (Paloniemi 2018). Soil texture family has been resolved for those collected soil samples. A texture ternary diagram (Dexter 2004) can assign soil texture family when the corresponding components of silt, sand and clay are measured. Conversely, bounds for clay component can be determined when soil texture family is known. As GEEs have been fitted to estimate soil C concentrations using clay component, dominant tree species and land cover type as reported in Table 2.5, these GEEs can possibly be extrapolated to estimate C for legacy sites. Even though this sort of modeling can only yield estimates for soil C, nonetheless this approach can facilitate application to an order of magnitude of more sites with soil data.

2.7. Summary

- Statistical analyses via the KW test and GEEs were appropriate for smaller sample sizes with skewed distributions of soil properties.
- Two-step process was practical for resolving statistical inference, firstly with the KW test ascertaining median differences across the cover types, followed by more refined analysis for mean differences between specific cover types as by adapting GEEs.

- KW test established statistical significance for all target soil properties across the various cover types, corroborating that a sufficient number of soil samples were extracted for the purposes of digital soil mapping.
- Statistical significance for nested differences at the subplot level were not established. This warranted support for digital soil mapping at a site level, through the application of environmental covariates of 30 m spatial resolution with one pixel encompassing all subplots for a site.
- Wetlands contain significantly more soil C than other land cover types. Similarly, forested land cover types contain more soil C than non-forested land cover types. For digital soil mapping, this suggests that environmental covariates that differentiate between various land cover types should be considered for soil modeling.
- Evidence from statistical inference that sites with black spruce contain more soil C than other tree species, with lesser statistical significance that sites with balsam poplar also contain more soil C. This corroborates the inclusion of tree species indicators as predictors for digital soil mapping of C.
- GEEs can be employed to estimate soil C for legacy soil sample data where clay percentage (in relation to soil texture family), dominant tree species and land cover type at site are noted.

Chapter 3

Impact of Topographic Features and Prediction Uncertainty for Tree Species Classification

Research presented in this chapter has been published in the journal paper:

[2] Pittman, R. and B. Hu., 2023. "Contribution of Topographic Features and Categorization Uncertainty for a Tree Species Classification in the Boreal Biome of northern Ontario." *GIScience and Remote Sensing* 60(1): 1–21. https://doi.org/10.1080/15481603.2023.2214994.

[2] I want to thank the publisher Taylor & Francis and the coauthor who have permitted me to reuse this article in my dissertation.

3.1 Background

Tree species classification for natural forested areas often requires a sufficiently large study area so to obtain adequate representation of endemic tree species and settings. As a product of tree species classification, tree species maps have been an objective for research studies in the boreal forest in Canada (Beaudoin et al. 2018) and as well as for temperature forests (Modzelewska, Fassnacht, and Stereńczak 2020; Zhu and Liu 2014). Generally, these studies have adopted pixel-based approaches for classification. Object-based approaches have been employed when demarcation or modeling of characteristics of individual trees is of importance (Fassnacht et al. 2016; Wang, Wang, and Liu 2018). This necessitates very high spatial resolution imagery for resolving features (Franklin 2018), which is often collected from unmanned aerial vehicles (UAVs) or drones and consequently are limited to study areas of a few km^2 or less (Natesan, Armenakis, and Vepakomma 2020). Regions of research interest for natural forested domains commonly well exceed that extent, hence object-based approaches are typically not feasible due to data and computational constraints, warranting pixel-based approaches for tree species classification.

Tree species and topography are linked in the context of an autecological concept for the boreal forest in eastern Canada (Denneler, Bergeron, and Bégin 1999; Kulha et al. 2019). Within natural forested areas, local topography can shape exposure considerations and some soil attributes, consequently impacting the overlying vegetation and governing what tree species grow (Denneler, Bergeron, and Bégin 1999). Topography consists of local and landscape scale morphometry, hydrological considerations and landscape exposure (Franklin 2020; Heung et al. 2016). Studies dating back previous decades (Stage 1976; Strahler, Logan,

and Bryant 1978) have investigated the adoption of topographic features for vegetation mapping. Topographic features that have been included in earlier studies for tree species classification involve shaded hillside (Pitkänen 1998), slope characteristics and aspect (Thomas and Anderson 1993), and slope inclination and fetch along with elevation, aspect and exposure (Denneler, Bergeron, and Bégin 1999). Generally, within existing studies, basic topographic features have been utilized when study areas contend with appreciable variations of elevation. Interconnections between tree species and topography are more evident in mountainous terrain (Chiang and Valdez 2019; Dalponte, Bruzzone, and Gianelle 2012). However, more research is required for the application of topographic features for domains that are relatively flat at the landscape scale. Light detection and ranging (LiDAR) data (Airborne Imaging 2018) can be exploited to derive topographic features of finer spatial resolution, facilitating detailed features for localities conforming to various settings of a natural forested area. Moreover, topographic features relating to other morphometry and hydrological considerations rather than just basic features should be applied for tree species classification.

Uncertainty estimates with classification mapping should be investigated (Fassnacht et al. 2016; Minasny et al. 2019). Typically, uncertainty with vegetation classification has been reported by confusion matrices with overall accuracy and Cohen's kappa score (Dalponte et al. 2015; Fassnacht et al. 2016; Sheeren et al. 2016). Nonetheless, metrics should be adopted for gauging classification uncertainty for individual pixels of a prediction map. Entropy from posterior probabilities for all category groupings can be computed as a metric for classification uncertainty (Pittman, Hu, and Webster 2021). However, an entropy that has

been calculated from just the categories with highest posterior probabilities can be formulated, to better assess entropy with regard to the contending categories.

3.2. Materials and Methods

3.2.1. Study Region

The research area corresponds to part of the Abitibi River Forest (ARF) surrounding the community of Smooth Rock Falls, within the District of Cochrane in Ontario, Canada. This area is located between the latitudes of 81.55° W – 81.80° W, and from latitudes of 49.15° N – 49.35° N. In reference to other delineated study areas, the ARF is situated between the Gordon Cosens Forest (GCF) and Cochrane study areas, regarding depiction in Figure 2.1. This area is mostly forested, consisting of representative expanses of boreal forest. Clearings for roadways, a railway and utility lines are present. Topography is relatively flat at the landscape scale, varying from 210 m in the north to around 280 m in southern portions. An assortment consisting of wetlands, low hills and a river valley are present in the 4 delineated study areas, that amount to a combined area of 108 km^2.

Mature tree stands of primary forest are predominant in these study areas, with secondary forest being encountered along the confines of the study areas resulting from previous logging activities. The dominant tree species are black spruce (*Picea mariana*) and balsam fir (*Abies balsamea*). Other endemic tree species include trembling aspen (*Populus tremuloides*), white spruce (*Picea glauca*), balsam poplar (*Populus balsamifera*), tamarack (*Larix laricina*) and eastern white cedar (*Thuja occidentalis*). The wetlands essentially only consist of stunted black spruce, with tamarack along the periphery of wetland expanses.

Taller black spruce is also present in dominant stands in non-wetland environments. Balsam fir is typically interspersed with other tree species such as trembling aspen, with trembling aspen commonly growing nearby balsam poplar in groves. Along river valleys or streams in upslope positions is where white spruce tends to subsist. Eastern white cedar is relatively rare in this region as this corresponds to its northerly bounds within Ontario, Canada.

3.2.2. Field Data

Tree species composition data was acquired from the Ontario Forest Resources Inventory (FRI) (Paloniemi 2018). In total, tree species information was obtained for 200 sites. Each of these sites existed at the position of 90 m into a 200 m long transects, with tree species composition noted at site. This was ascertained by the counting of all the trees over 2 m in height at the site, from which a composition of tree species was recorded (Paloniemi 2018). Only tallies consisting of living trees were considered; dead trees and trees without green foliage or live buds were omitted from the counts (Paloniemi 2018). The delineated study areas along with locations of the FRI sites allocated regarding training and validation, are shown in Figure 3.1. Within this figure, the background of the study areas corresponded to true-color composition WorldView-2 (WV2) imagery acquired for August 10th, 2014.

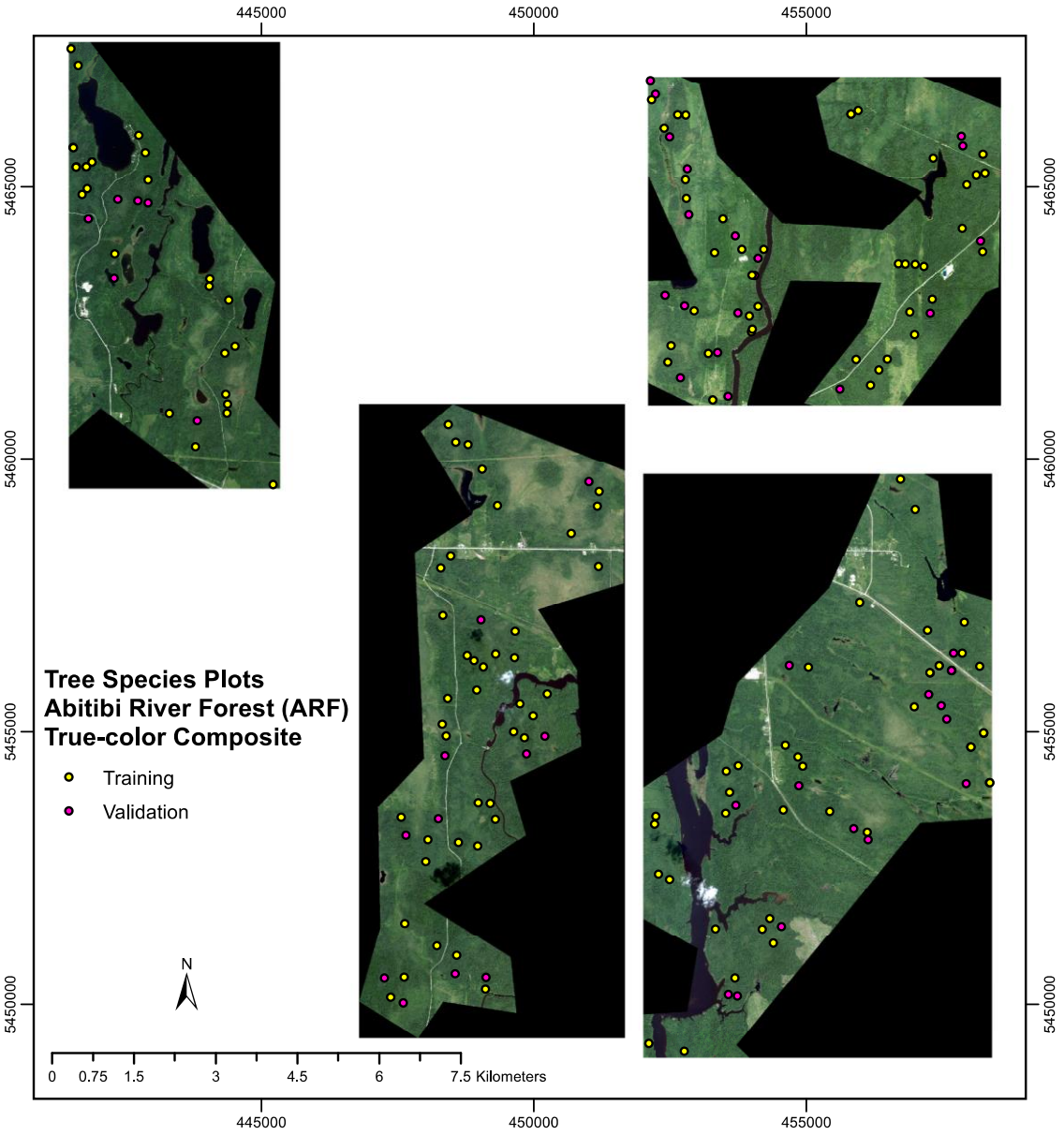

Figure 3.1. Study areas within the Abitibi River Forest (ARF) for tree species classification, denoting Forest Resources Inventory (FRI) sites regarding model training and validation. The background corresponds to true-color composite WV2 imagery for August 10th, 2014. Markings on the grid conform to the NAD 1983 (2011) UTM zone 17 N coordinate projection.

For the tree species categories, pure stands were assigned for sites where a tree species consisted of at least 70% composition of the trees at the site (Bravo-Oviedo et al. 2014). When a dominant tree species conformed to at least 50% of trees at site, but less than 70%, then a mixed category involving that tree species was considered. Counts for the composition of tree species groupings as summarized by a pie chart have been presented in Figure 3.2. Regarding tree species composition, sites with balsam fir as dominant were all mixed sites, as essentially there were no pure stands for balsam fir. All in all, 104 sites conformed to pure stands with counts of 6, 5, 86 and 7 sites for eastern white cedar (Cw), tamarack (La), black spruce (Sb) and white spruce (Sw), accordingly. Trembling aspen and balsam poplar were combined into a category denoted as populus (P) which consisted of 5 sites. Mixed stands comprised of a dominant tree species constituting the majority of trees corresponded to 57 sites for balsam fir (Mix-Bf) and 10 sites for black spruce (Mix-Sb). The mixed (Mix) grouping was reserved for the remaining 24 sites where no tree species formed a majority.

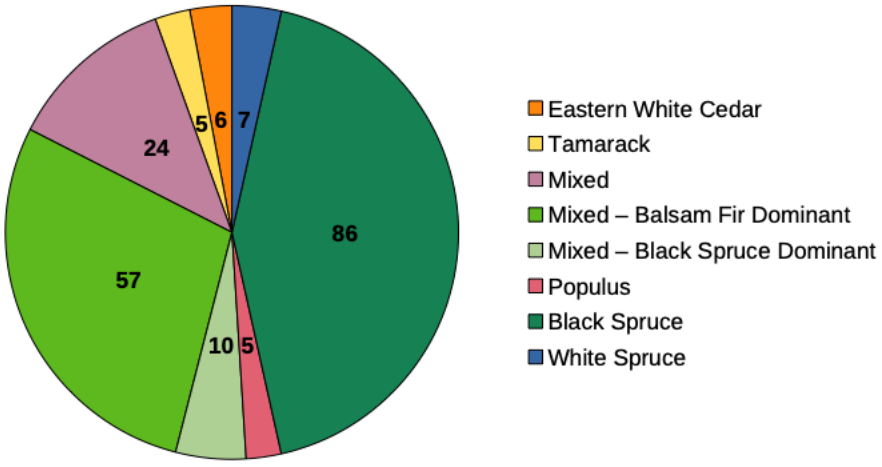

Figure 3.2. Composition of the Forest Resources Inventory (FRI) sites within the Abitibi River Forest (ARF) study area, as depicted by tree species grouping.

3.2.3. Feature Covariate Data

Features were processed from remote sensing data corresponding to LiDAR and multispectral surface reflectance. Finer scale features of 2 m spatial resolution were computed. In total, 22 different features were generated and utilized for the modeling, consisting of 8 multispectral surface reflectance, a panchromatic image, 2 normalized difference indices, a canopy height model (CHM), and 10 topographic features. The full listing of predictors used for the tree species classification modeling is shown in Table B.1 in Appendix B.

The LiDAR data was acquired from Land Information Ontario (LIO) within the Ontario Ministry of Natural Resources (MNR). This corresponded to point-cloud data collected via airborne LiDAR during October 2016 and September 2017 by means of a Leica ALS70-HP sensor (Airborne Imaging 2018). Information for the following specifications of the LiDAR sensor and retrieval settings were excerpted from the corresponding technical report (Airborne Imaging 2018) for the retrieval acquisition of that LiDAR data. For the airborne LiDAR retrieval, the corresponding flight paths were roughly 1000 m above the ground surface, with single pass swath widths of 690 m with 20% overlap for the retrieval of the LiDAR. The LiDAR sensor was attuned for a scan frequency of 53 Hz, a scan pulse rate frequence of 500 kHz, and an effective scan angle field of view of 38°. Regarding retrievals, the LiDAR data consisted of up to 4 returns per retrieval with a point density of approximately 8 points per m^2.

From the LiDAR data, rasterized covariates of a digital surface model (DSM), digital terrain model (DTM) and CHM of 2 m spatial resolution were generated. The CHM was calculated

as the difference between the DSM and DTM. More specifics on the derivation and computation of these covariates were provided in Section 6.2 of this dissertation. For processing topographic features from the DTM, System for Automated Geoscientific Analyses Geographic Information System (SAGA GIS) (Conrad et al. 2015) software was utilized. These features corresponded to considerations for terrain analysis, morphometry and topographic openness. Specifically, the topographic features generated were aspect, channel network distance, closed depressions, multi-resolution ridge top flatness (MRRTF), multi-resolution valley bottom flatness (MRVBF), negative openness, relative slope position, topographic wetness index (TWI), total catchment area and valley depth.

Optical imagery comprising of surface reflectance was bought from Maxar for the WorldView-2 (WV2) satellite. A scene with less than 2% cloud coverage during a period of peak vegetation was acquired for the date of August 10th, 2014. Surface reflectance was obtained for coastal (400 – 450 nm), blue (450 – 510 nm), green (510 – 580 nm), yellow (585 – 625 nm), red (630 – 690 nm), red edge (705 – 745 nm), near infrared I (NIR1) (770 – 895 nm), near infrared II (NIR2) (860 – 1040 nm) and panchromatic (450 – 800 nm) bands (Digital Globe 2016). The multispectral bands were of 2 m spatial resolution, and the panchromatic band was resampled to a spatial resolution of 2 m. After reprojection and co-registration with the rasterized imagery generated from LiDAR, the WV2 imagery was orthorectified with the LiDAR-derived DSM. A normalized difference vegetation index (NDVI) and a modified normalized difference water index (NDWI) using the NIR2 band (Liuzzo et al. 2020) were each generated utilizing the respective surface reflectance bands.

3.2.4. Modeling

The tree species classification was achieved using random forest (RF) and support vector machine (SVM) approaches. Stratified random sampling was employed to split the data 70:30 between training and validation sets, allocating 143 sites for training and 57 sites for validation, accordingly. This was enacted so to keep sites used for validation separate from those assigned for training, to prevent the inflation of modeling accuracy metrics regarding evaluation purposes. For the modeling, a 10-fold cross validation with 3 replicates was implemented. RFs were fitted with the number of trees set as 1000. The mtry parameter for the selection of randomly predictors for each node was set to 5, so to be no higher than the integer rounded square root of the number of features (i.e. 22). For comparison purposes, radial-basis SVMs were also utilized for modeling. With the SVMs, the turning cost parameter was varied across 0.25, 0.5, 1, 2, 4, 8, 16, 32, 64 and 128, and the kernel width (sigma) parameter was ranged from 0.005 to 0.05. The caret package (Kuhn 2008) in R was employed for the fitting of the RF and SVM models.

Variable importance was gauged by mean decrease in Gini with the RF model, as well as by receiver operating characteristic (ROC) curve analysis (Kuhn 2007). The variable importance ascertained from the RF was implemented for a feature reduction, whereas that from ROC curve analysis was conducted for comparison purposes. A larger mean decrease in Gini conformed to a higher variable importance ranking for the RF, with a larger area under the ROC curve corresponding to a higher measure of variable importance. The variable importance rankings from these two methods were then compared, with the expectation that important predictors were ranked similarly by both methods. To mitigated effects of

multicollinearity, just the top 10 predictors with the highest ascertained mean decrease in Gini from the RF were retained.

For model assessment, accuracy was resolved at both a pixel level and a site level. The percent correct classification (PCC) and Cohen's kappa statistic were each evaluated. PCC corresponded to overall accuracy with respect to the portion correct with prediction, ranging from 0 to 1, with the closer the value to 1 then the more accurate the prediction. Cohen's kappa score factored by-chance agreement and misclassification occurrence (Kuhn 2008), with the kappa score varying from -1 to 1 (Cohen 1960; McHugh 2012). For kappa score, 1 signified perfect agreement, 0 indicated agreement being the same as by chance, and negative scores denoted agreement as worse than that arising by chance (Cohen 1960; Monserud and Leemans 1992). Regarding model assessment, kappa scores less that 0.4 signified poor to at best fair agreement, between 0.4 to 0.8 implied moderate to substantial agreement, and greater than 0.8 indicated strong agreement (Brungard et al. 2015; McHugh 2012). Confusion matrices for both model fitting and validation, respectively, were reported. Producer's accuracy (PA) and user's accuracy (UA) were calculated for these confusion matrices to assess individual class accuracies for each tree species grouping.

3.2.5. Prediction Uncertainty

Entropy maps were computed to evaluate uncertainty with prediction. These measures were calculated per pixel of a prediction map for the corresponding study areas, from the posterior probabilities of prediction for the trained models. In this case, these probabilities corresponded to categorial assignment for tree species grouping. The entropy (Roulston and Smith 2002; Zhu 1997) with prediction for each pixel was calculated as

$$H = \frac{-1}{\ln(n)} \sum_{k=1}^{n} p_k \cdot \ln(p_k) \, . \tag{3.1}$$

Here, p_k corresponded to the probability of assignment to class k, out of n different classes. Each of these probabilities were bounded as $0 \leq p_k \leq 1$, with probabilities over all classes per each pixel summing to one. The computing of these posterior probabilities varied by modeling method. With the RFs, these probabilities were calculated from the proportion of trees that assigned categorization, i.e. voted, for each according class (Sage, Genschel, and Nettleton 2020). As for the SVM, a secondary Platt model approximated posterior probabilities by means of a sigmoid function (Lin, Lin, and Weng 2007). The form of entropy as defined in equation (3.1) conformed to Shannon's entropy (Kumar, Kumar, and Kapur 1986). This entropy here in this equation was bounded between 0 and 1, with higher values denoting more uncertainty regarding the class assignment for prediction.

Regarding entropy scores, the entropy was greatest for when similar posterior probabilities for each class were obtained, leading to greater uncertainty with class assignment. On the contrary, one class assigned with a vastly higher probability than any other class then resulted in a low entropy score as there was lower uncertainty. Nonetheless, many class categories for assignment potentially diluted this type of entropy score, specifically if the class with maximal probability only attained a slightly higher probability than that of another class (Baoxin Hu, Li, and Hall 2021). Consequently, to address this issue with the conventional entropy metric, an entropy calculation based just upon the top two contender classes was formulated and evaluated. These entropies were calculated per pixel from the top two highest probabilities for prediction out of all classes. Before resolving this entropy calculation based

upon the top two contenders, the corresponding top two probabilities per pixel were normalized so that these summed to one,

$$\hat{p}_1 = \frac{p_{max}}{p_{max} + p_{2nd\,max}}, \tag{3.2}$$

$$\hat{p}_2 = \frac{p_{2nd\,max}}{p_{max} + p_{2nd\,max}}. \tag{3.3}$$

Using these adjusted probabilities, this entropy considered from just the top two contenders, here regarded as normalized entropy, was computed per pixel as

$$H_{Top\,2} = \frac{-1}{\ln(2)}[\hat{p}_1 \cdot \ln(\hat{p}_1) + \hat{p}_2 \cdot \ln(\hat{p}_2)]. \tag{3.4}$$

It was argued that this metric was better at ascertaining uncertainty with prediction, as ultimately class assignment was rendered to the class with the highest posterior probability; any challenge to this allocation resulted due to the class with the second highest posterior probability. This normalized entropy was still a form of Shannon's entropy (Kumar, Kumar, and Kapur 1986), and like the entropy (3.1) defined previously it was bounded between 0 and 1 where higher values indicated more uncertainty with prediction. Both the conventional entropy (3.1) and normalized entropy (3.4) were computed from the best models with regard to accuracy for the domain of the study areas.

3.3. Results

The most importantly ranked features were resolved by variable importance from the RF, listed in Figure 3.3. From this figure, channel network base, valley depth and MRVBF had the highest variable importance, imparting that the structure of topography with respect to valleys were particularly relevant for tree species classification. NDVI and CHM relating to vegetation cover and structure, respectively, were also of importance. Relative slope position

and MRRTF were also of consequence. As an aside, when the ROC curve analysis was carried out on this same set of predictors, the same topographic features attained the highest variable importance for tree species classification.

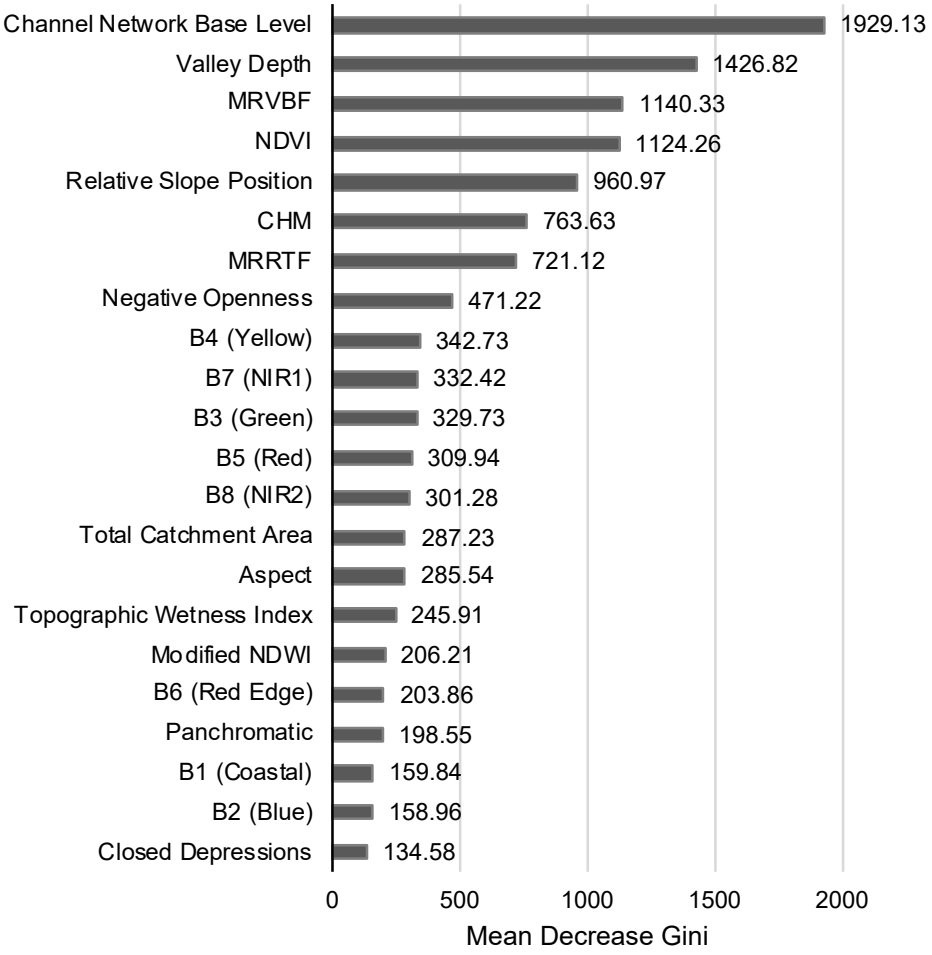

Figure 3.3. Variable importance as resolved by the RF for tree species classification, with predictors ranked in order of importance by mean decrease in Gini.

Modeling accuracies were assessed by both pixel level and site level on the validation set, reported in Table 3.1. The SVM radial-basis model was optimized when the parameter of sigma was set to 0.036 and cost tuning was equaled to 64, respectively. The RF model

attained better accuracies than that of the SVM. For the RF, accuracy exceeded 0.7 and corresponding kappa scores were greater than 0.5, denoting at least moderate agreement regarding the classification modeling.

Table 3.1. Tree species classification accuracies as assessed on the validation set, with accuracies noted at both the pixel level and site level, respectively. Percent correct classification (PCC) (accuracy) and Cohen's kappa score are reported.

Model Accuracies	**Tree Species Classification**			
Validation	**Pixels**		**Sites**	
	Accuracy	**Kappa**	**Accuracy**	**Kappa**
Random Forest (RF)	0.72	0.58	0.79	0.69
Support Vector Machine Radial-basis (SVM Radial)	0.62	0.46	0.74	0.61

The RF model was employed to predict the most probable tree species group for the domain of the study areas, with predictions displayed in Figure 3.4. This prediction map was computed per pixel from the rasterized imagery of predictors constituting the study area bounds. From Figure 3.4, black spruce covered more area than that for any other tree species, reconciling to the increased presence of black spruce within the wetland areas. Balsam fir was the second most dominant tree species, consistent with widespread occurrence throughout the study region. Other tree species predicted for other localities within the domain of the study areas included white spruce alongside the slopes of the river valley and populus species along the edges of clearings or areas with deeper surface soil layers.

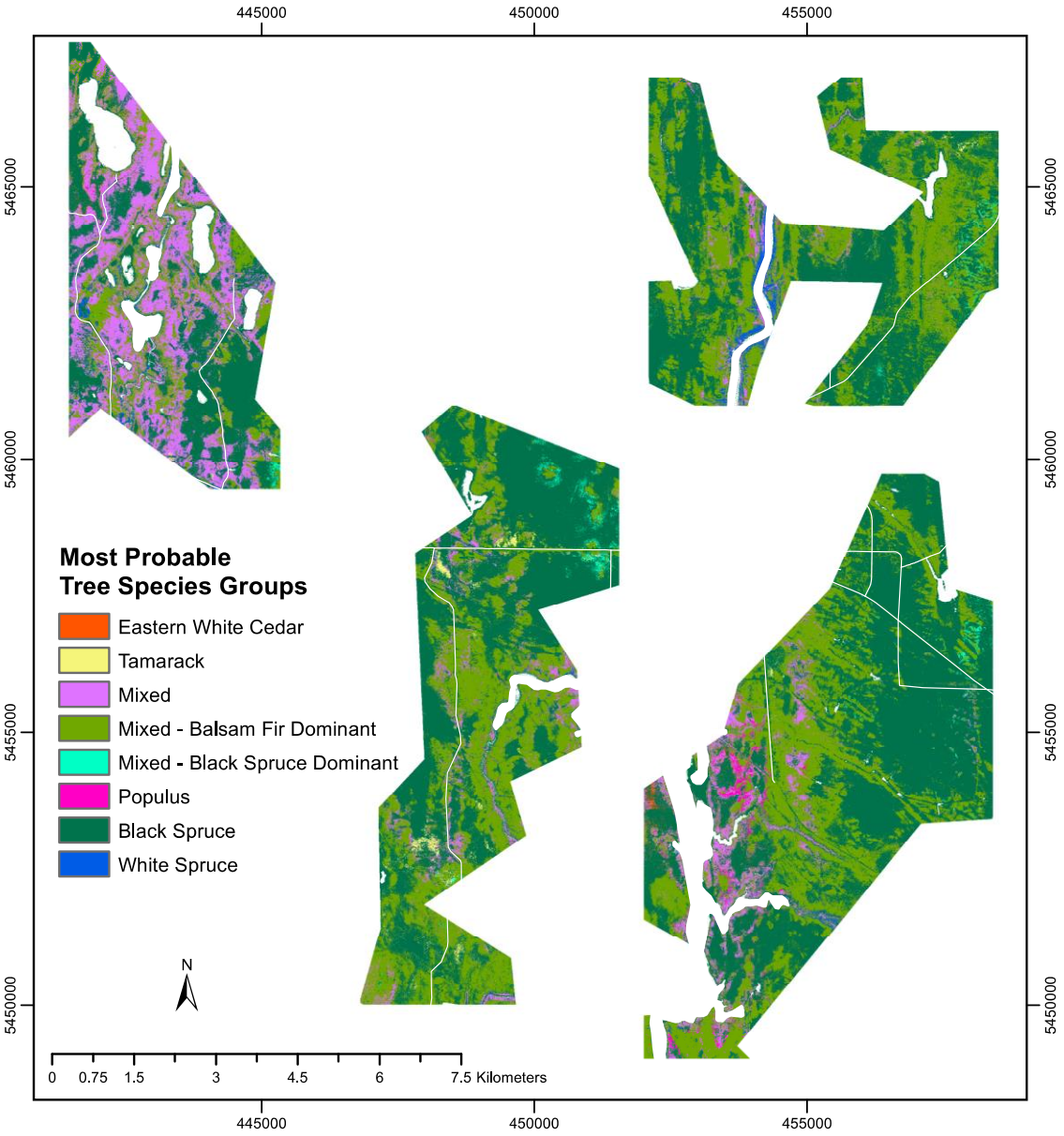

Figure 3.4. Predictions for most probable tree species groups for the ARF study areas, as generated from the RF model.

Assessment regarding confusion matrices was presented in Table 3.2 for both the model training and validation, respectively, from the RF model. Along the bottom edges and right edges of these matrices, user's accuracy (UA) and producer's accuracy (PA) were both noted,

accordingly. Counts for the sites conforming to each tree species grouping, for both training and validation, were indicated. Sites corresponding to black spruce and balsam fir dominant represented the majority, with decent UA and PA values greater than 0.8 attained for each of those groupings. Nevertheless, the classification was not as good for some other tree species groupings. None of the sites for populus nor eastern white cedar in the validation sets were correctly classified, but these only corresponded to 1 site each respectively for the validation. However, the 1 site for tamarack was correctly classified, as were the 2 sites for white spruce. The mixed category was predicted between mostly other mixed categories, and all 3 sites that were mixed but with black spruce dominant were predicted as just black spruce dominant. Therefore, the modeling for the tree species classification was reasonable.

To gauge further insights with uncertainty of prediction, entropies using equations (3.1) for conventional entropy and (3.4) for a normalized entropy were each calculated for the validation sites. These results were reported in Table 3.3 and were summarized by tree species grouping. Inspecting Table 3.3, it was apparent that the normalized entropy from the top 2 choices had higher uncertainty values, encapsulating a prospectively more accurate assessment for prediction uncertainty. Note that with the conventional entropy metric, that sites for white spruce had a lower entropy value than those for black spruce, but that this situation reversed with the normalized entropy metric as white spruce had the higher uncertainty value. The normalized entropy metric resulted in a larger range of entropy values, ranging from 0.46 to 0.96 instead of 0.38 to 0.73 for the conventional entropy.

Table 3.2. Confusion matrices for the tree species classification from the RF model, reported for both training (top) and validation (bottom) sets, respectively. These counts correspond to the site level. User's accuracy (UA) and producer's accuracy (PA) are also noted for each tree species grouping.

Confusion Matrix Training		**Tree Species Classification**								
		Prediction								**Producer's Accuracy**
Reference		Cw	La	Mix	Mix - Bf	Mix - Sb	P	Sb	Sw	**(PA)**
Eastern White Cedar	Cw	3			1			1		0.60
Tamarack	La		4							1.00
Mixed	Mix			15	2					0.88
Mixed - Balsam Fir Dominant	Mix - Bf				40					1.00
Mixed - Black Spruce Dominant	Mix - Sb				1	6				0.86
Populus	P						4			1.00
Black Spruce	Sb							61		1.00
White Spruce	Sw								5	1.00
User's Accuracy (UA)		1.00	1.00	1.00	0.91	1.00	1.00	0.98	1.00	Sites: 143

Confusion Matrix Validation		**Tree Species Classification**								
		Prediction								**Producer's Accuracy**
Reference		Cw	La	Mix	Mix - Bf	Mix - Sb	P	Sb	Sw	**(PA)**
Eastern White Cedar	Cw							1		0.00
Tamarack	La		1							1.00
Mixed	Mix			4	2			1		0.57
Mixed - Balsam Fir Dominant	Mix - Bf			3	14					0.82
Mixed - Black Spruce Dominant	Mix - Sb							3		0.00
Populus	P				1					0.00
Black Spruce	Sb					1		24		0.96
White Spruce	Sw								2	1.00
User's Accuracy (UA)		.	1.00	0.57	0.82	0.00	.	0.83	1.00	Sites: 57

Table 3.3. Uncertainties with respect to conventional entropy (3.1) and normalized entropy from the top 2 choices (3.4) for the validation sites as summarized by tree species grouping.

Uncertainty Calculations			Validation Sites	
			Entropy	Normalized Entropy
Tree Species Grouping		Site Count		Top 2 Choices
Eastern White Cedar	Cw	1	0.64	0.83
Tamarack	La	1	0.73	0.96
Mixed	Mixed	7	0.64	0.86
Mixed - Balsam Fir Dominant	Mixed - Bf	17	0.55	0.72
Mixed - Black Spruce Dominant	Mixed - Sb	3	0.54	0.66
Populus	P	1	0.38	0.46
Black Spruce	Sb	25	0.52	0.70
White Spruce	Sw	2	0.45	0.86

Entropy maps for the study areas were also computed, depicted for the conventional entropy (3.1) in Figure 3.5 and the normalized entropy from the top 2 contenders (3.4) in Figure 3.6, respectively. From these entropy maps, the areas with the lowest uncertainties corresponded to the wetland areas, which were principally dominated by black spruce when inspecting the tree species prediction map in Figure 3.4. This reconciled to what was expected, as stunted or lowland black spruce were typically the only trees to grow in the wetland environments within this boreal study region. Note that the coloration scales were the same for both Figures 3.5 and 3.6, but a wider variation in coloration was obvious for Figure 3.6 with respect to the normalized entropy. With the normalized entropy, areas outside of the wetlands had higher uncertainty with prediction. In comparing to Figure 3.4 of the tree species prediction, the areas with the highest uncertainties in Figure 3.5 appeared to correspond to the mixed tree species groupings. This result was logical, as by setup more uncertainty regarding specific

tree species classification arose with mixed categories, and hence in part why mixed categories were considered for the tree species classification.

To gain insights pertaining to contenders for tree species classification, predictions for the second most dominant tree species were also generated, shown in Figure 3.7. When the most probable tree species was assigned a probability greater than 70% with prediction, note that no second most probable tree species was allocated and hence the not probable category. Inspecting Figure 3.7 and comparing to that for most probable tree species in Figure 3.4, essentially most of the areas corresponded to black spruce for where a second most probable tree species was infeasible; these were likely wetland environments. From Figure 3.7, mixed groupings for tree species were more common as secondary most probable tree species groupings throughout the study areas. This prediction was anticipated for localities that did not correspond to pure stands regarding tree species grouping. The mixed category for balsam fir was prevalent for prediction as a secondary tree species, conforming to expectations as balsam fir was often encountered as a secondary species across the study area domain.

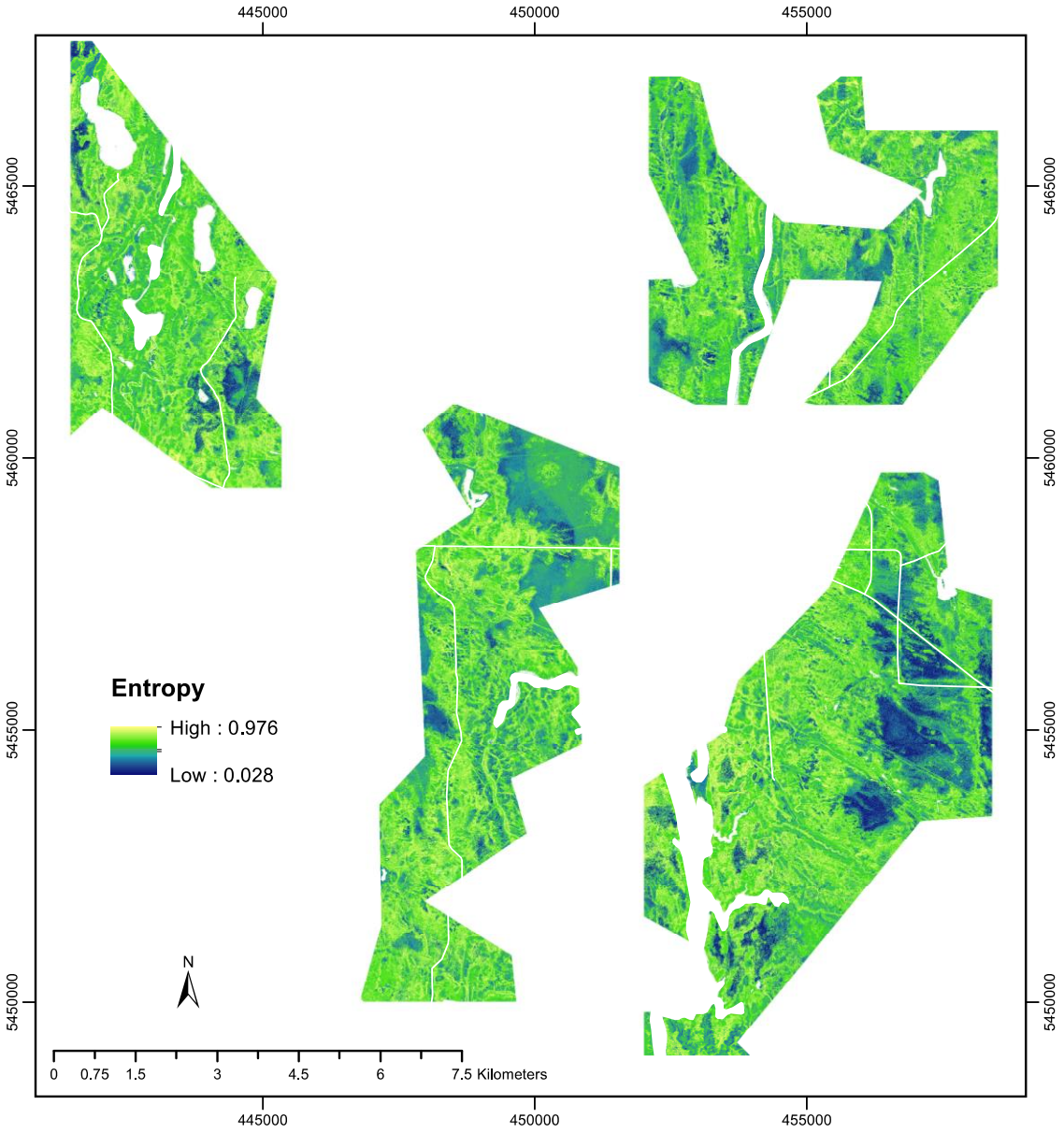

Figure 3.5. Conventional entropy as calculated by (3.1) for the ARF study areas. Note that darker color intensity signifies lower prediction uncertainty, whereas lighter color intensity signifies higher uncertainty with prediction.

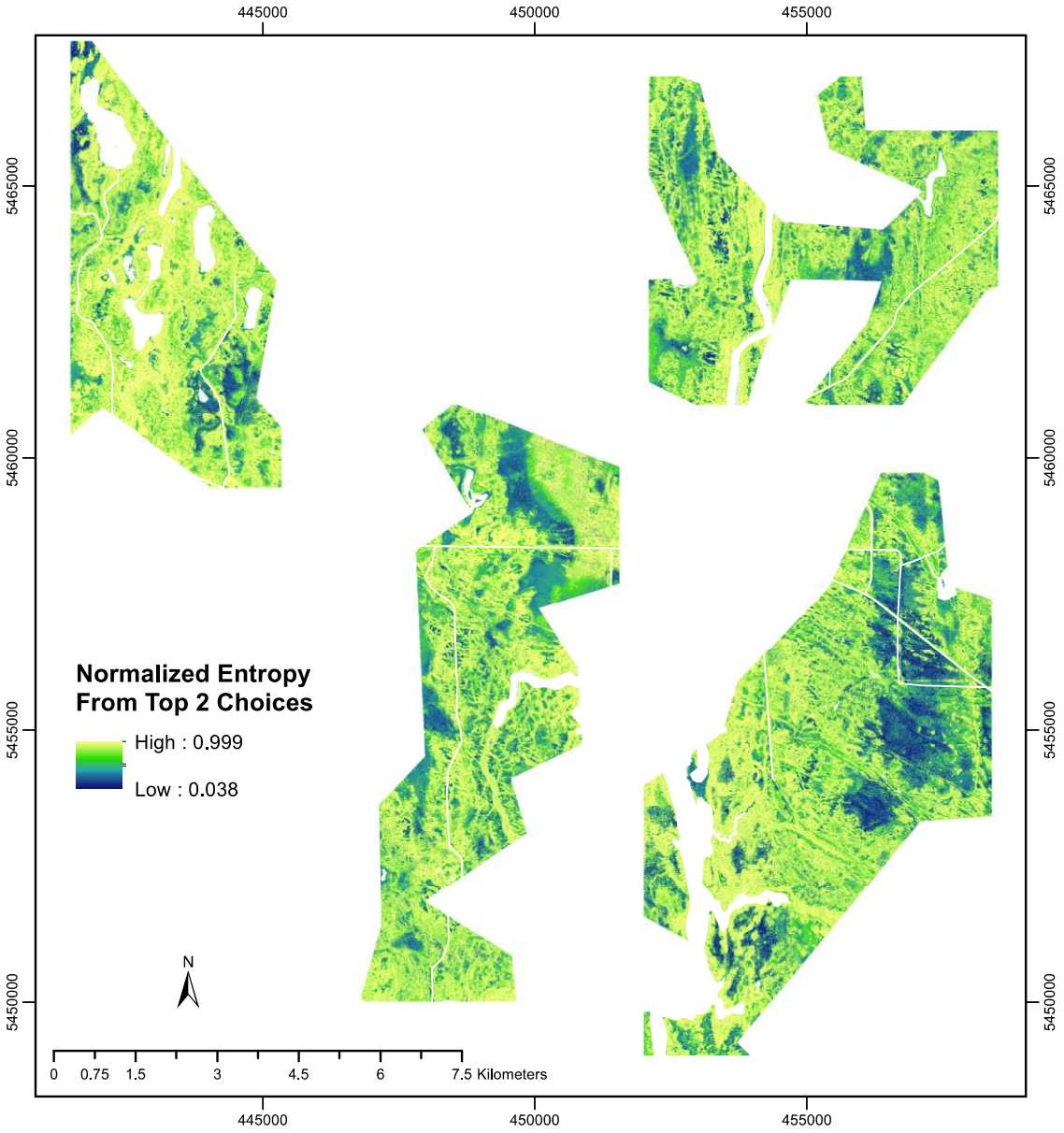

Figure 3.6. Normalized entropy from top 2 contenders, as calculated by (3.4) for the ARF study areas. Note that darker color intensity signifies lower prediction uncertainty, whereas lighter color intensity signifies higher uncertainty with prediction.

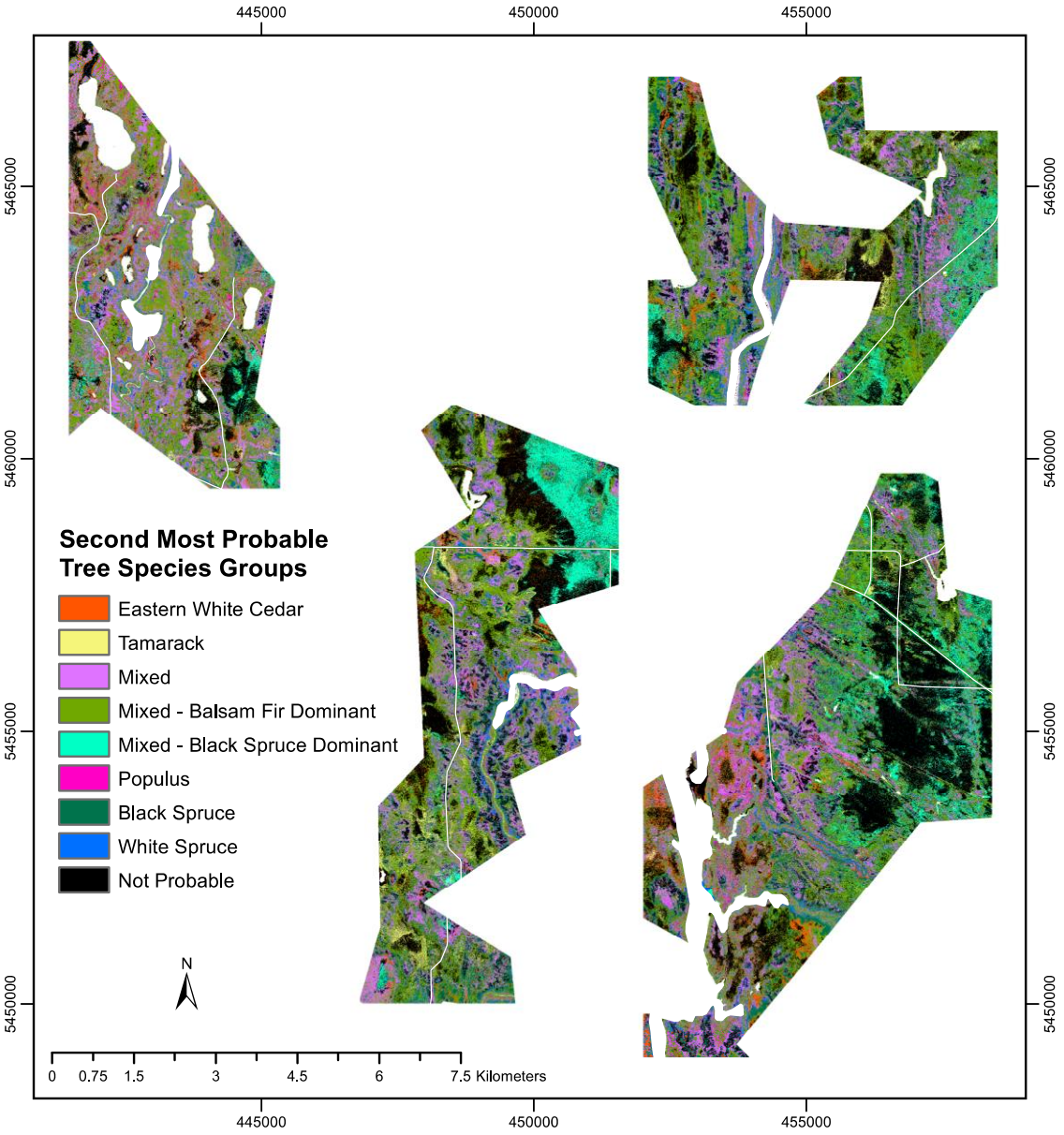

Figure 3.7. Predictions for second most probable tree species groupings within the ARF study region, as resolved by the RF model.

3.4. Discussion

Topographic features were of high importance for tree species classification within the boreal forest of northern Ontario as seen in Figure 3.3, and the inclusion of these features subsequently improved classification accuracy. Features relating to morphometry consisting of valley depth, ridge flatness, valley bottom flatness and slope position were of high relevance with modeling. The valleys and channels within the study area presumably provided protection against the wind and excess sunlight for some tree species. Depending on the context, water accumulation within localities of lower relative elevations either benefitted or hindered the growth of certain tree species. For this study area, as conformable for boreal regions throughout northern Ontario, white spruce was observed growing well alongside elevated banks nearby rivers and streams. The wetland areas tended to consist of homogeneous stands of stunted black spruce, as that was the only tree species to subsist in those locations. Therefore, these results substantiated the consideration of a variety of topographic features as predictors for tree species classification. These topographic features corresponded to either local attributes at a point, textural attributes for a topographic surface, or contextual attributes for the larger landscape domain (Franklin 2020). As examples of features (Franklin 2020), local point attributes conformed to slope, aspect and curvatures, and textural attributes consisted of indices relating to roughness or homogeneity. Correspondingly, contextual attributes included slope position and hydrological considerations over the wider area such as for topographic wetness indices and catchment areas. It was important to include a variety of topographic features so to sufficiently

characterize the topography for the study area (Franklin 2020), hence a suite of topographic features was utilized for tree species classification.

Classification uncertainty with respect to entropy has been investigated in other research disciplines such as digital soil mapping (Heung et al. 2016). Entropy metrics are based upon voting probabilities rather than the outright classification assignment, so entropy can constitute as a better evaluation for prediction uncertainty. However, if more than a few classification categories exist then the entropy metric can become diluted, particularly if many of those categories conform to negligible probabilities. Consequently, a diluted entropy value can underestimate prediction uncertainty. To address this, a normalized entropy based on just a reduced selection of voting probabilities corresponding to the top contenders has been formulated. Other studies have calculated normalized entropies from maximum and minimum voting probabilities (Wu, Shih, and Tu 2021) or by different formulations (Kumar, Kumar, and Kapur 1986). Nonetheless, decisiveness for output classification assignment depends upon the top contending categories, hence the formulation of a normalized entropy that is based only upon the voting probabilities for the top two categories. The top contenders for classification can also relate to prediction maps, as typically the output for a prediction map corresponds to the category with the highest voting probability. Prediction maps for the second most probable category can reveal insights in corroboration with entropy maps. All in all, entropy metrics in conjunction with prediction maps should be considered for classification as entropy maps can convey supplemental information pertaining to prediction uncertainty.

Model assessment for pixel-based classification is regarded as more relevant if evaluating at a site level, rather than just the pixel level. When possible, the models should be trained on data conforming to sites of pure stands, rather than large selections of pixels with allotment to both training and validation sets from each selection block of pixels. Regarding evaluation, for pure stands it is of more relevant to correctly determine the prediction of tree species overall for a site, rather than for the individual pixels constituting that site. Furthermore, it is more feasible to confirm that a smaller area and thus a smaller selection of pixels conform to a pure stand, rather than a larger area consisting of a magnitude or more greater number of pixels. A risk of overfitting models can arise if pixels for both modeling training and validation are selected from the same plot selections. Therefore, to prevent inflated accuracies, the pixel selections should be aggregated at a site level so that modeling accuracy can correspondingly be assessed at the site level.

3.5. Summary

- Topographic features were important for tree species classification for boreal study areas. Specifically, environmental covariates relating to channel structure and valley depth were the most pertinent for tree species classification.
- Covariates derived and generated at a finer spatial resolution better facilitated representation for various settings within natural forested areas, enhancing accuracy for a local scale that was impractical with coarser spatial resolutions.
- Tree species classification should be assessed for accuracy at the site level rather than at the pixel level, with each site conforming to one cover type. Moreover, this

mitigates against inflation with accuracy values as pixels for a site are not allocated between both model training and validation purposes.

- Uncertainty with prediction should be investigated, and specifically a normalized entropy that is based upon the top two contenders is a more relevant metric than the conventional entropy.
- Mixed categories constituting dominant tree species can be considered when certain tree species rarely grow in pure stands. This is particularly applicable for balsam fir, which tends to grow alongside other tree species in expanses through the study area.
- Prediction maps of the second most probable tree species groupings can convey additional insights regarding prediction within a study region. Specifically, this can determine what are the contending tree species throughout the study areas.
- Wetlands conform to the lower uncertainties with prediction, and most of these wetland areas have been predicted to conform to black spruce. This imparts robustness regarding a tree species indicator for black spruce, when applied in the context of digital soil mapping encompassing wetland environments.
- Localities of higher uncertainties with prediction correspond to areas that are mixed for tree species, that is, likely of greater tree species diversity. These are location settings where to address and investigate further with respect to future tree species classification research.
- Tree species prediction maps can be relevant as predictors for digital soil mapping, representing vegetation as a soil formation factor.

Chapter 4

Enhanced Prediction of Soil Carbon via Encoder-Decoder Neural Networks

Research presented in this chapter has been published in the journal paper:

[3] Pittman, R. and B. Hu, 2025. "Enhanced Prediction of Soil Carbon via Encoder-Decoder Neural Networks for a Boreal Study Area in Northern Ontario." *Sensors* 25(8): 1–23. https://doi.org/10.3390/s25082583.

[3] I want to thank the publisher MDPI and the coauthor who have permitted me to reuse this article in my dissertation.

4.1. Background

Interest in deeper learning approaches such as neural networks (NNs) has been increasing for digital soil mapping. NNs can consist solely of dense layers, which are known as dense neural networks (DNNs). Other names for DNNs in digital soil mapping include artificial neural networks (ANNs) (Heung et al. 2016; Were et al. 2015) or multilayer perceptrons (MLPs) (Brungard et al. 2015). DNNs are composed of one or more hidden dense layers of interconnected units, where the weightings for these units are iteratively adjusted by training on data via a learning process (Heung et al. 2016). Modeling for both regression (Kuhn et al. 2020; Were et al. 2015) and classification (Brungard et al. 2015; Heung et al. 2016) for soil properties have been performed with DNNs. Nevertheless, within digital soil mapping research the reported accuracies for DNNs have typically only achieved similar results as have been attained by RFs (Brungard et al. 2015; Heung et al. 2016). To improve modeling accuracy, many research disciplines have investigated convolutional neural networks (CNNs) (Lee and Song 2019). In terms of structure, CNNs consist of feature extraction and selection layers along with dense layers as in DNNs (Lee and Song 2019; Simard, Steinkraus, and Platt 2003). CNNs have been employed primarily for image classification (Chauhan, Ghanshala, and Joshi 2018; Natesan, Armenakis, and Vepakomma 2020) and generally obtain better modeling accuracies (Lee and Song 2019). Nonetheless, a major drawback for CNNs is that these can necessitate the optimization of tens of thousands, or even millions, of parameters (Lee and Song 2019). Within digital soil mapping, many regional studies are limited to samples that have only been collected for at most a few hundred sites (Minasny et al. 2019). To prevent overfitting with training CNNs, sufficient amounts of target data are

required, which frequently is not feasible for soil mapping applications. The difficulties and costs to extract and process soil samples generally prevent the retrieval of additional samples for target data, hindering the adoption of CNNs.

Encoder-decoder NNs incorporate NN structure into encoder-decoder (ED) formulation (Chen et al. 2017; Ji et al. 2021). Regarding form, an ED consists of an encoder that firstly compresses information from the original input into a condensed representation (Ji et al. 2021). Afterwards, this condensed representation is input for a decoder that can recreate details from the original input (Ji et al. 2021). Specifically, an ED conforms to feature extraction as it can separate out noise from the original input (Ji et al. 2021), retaining data more pertinent for modeling. This refining is useful for digital soil mapping research as typically large sets of environmental covariates are utilized as predictors, but usually only a few of these predictors are impactful with modeling and prediction (Nussbaum et al. 2018). Regarding composition, EDs can consist of simple or dense layers such as with DNNs (Bai et al. 2024; Liu, Yang, and Zhang 2024) or of convolutional layers with CNNs (Chen et al. 2017; Ke, Ren, and Yin 2024) including fully convolutional networks (FCNs) (Ji et al. 2021). Furthermore, EDs can be formulated with NNs of reduced numbers of layers and units, facilitating EDs for smaller sets of data as confronted in digital soil mapping research.

As prediction maps of targeted soil properties are customarily an end goal of digital soil mapping, an objective is to maximize prediction and modeling accuracy. This can also entail addressing uncertainty with prediction (Minasny et al. 2019). From the best models that have been attained from separate modeling approaches, an ensemble of predictions (Engler et al. 2013; Huang and Gao 2017) can quantify uncertainty with respect to standard deviations.

Moreover, a mean prediction value can be computed from this ensemble to obtain a more robust prediction (Engler et al. 2013). From considering other modeling approaches, it can also be investigated if these models have similar trends with residuals regarding over or under-prediction. Furthermore, insights can be gained from prediction maps by examining what localities or land covers conform to greater prediction uncertainties. Additional insights can be reckoned from quantile mapping (Maraun 2013) in terms of deciles of the predicted quantities. In particular, standard deviations of the deciles can reveal what land cover types correspond to greater prediction uncertainties when addressing for the tendency for models to over or under-predict. For modeling improvements this can uncover what land cover types to focus on regarding digital soil mapping for future research.

4.2. Materials and Methods

4.2.1. Study Area

The study areas correspond to the same delineated areas as have been presented in Chapter 2 for the inferential statistics research. More details pertaining to the study region and soil data are provided in Chapter 2. This study region conforms to the southern transition zone of the boreal forest in the District of Cochrane (80°40’ W – 84° W, 49° N – 50° N) in Ontario, Canada. These study areas are located along the Ontario highway 11 corridor, and from northwest to southeast consist of Hearst, Gordon Cosens Forest (GCF), Kapuskasing within GCF, and Cochrane. For this research, of importance was prediction for the designated Cochrane study area (80°40’ W – 81°25’ W, 49° N – 49°20’ N). Data from all study areas in the District of Cochrane have been fielded for model training and validation.

The Cochrane study area has been focused on in part due to its more southerly location along the transitional confines of the boreal forest, as well as greater economic activities relating to agriculture and resulting in land cover change. This delineated study area contains a mix of forested and agricultural land cover types, consisting of wetlands along the western section and northeast part, and conforming to agricultural land within the center portion. The east central portion of this study area is intersected by the Abitibi River, which flows on a northwest trajectory towards James Bay in the Arctic Ocean. For reference purposes, true-color composite imagery along with locations of sampling sites for all the study areas are presented in Figure 5.1 in the following chapter regarding structural equation modeling (SEM); the Cochrane study area along with insets depicting some site clusters are shown in Figure 5.1-B.

4.2.2. Soil Data

The soil data used for this study was collected during field campaigns in 2018, 2019 and 2021. Further details regarding specifics for this soil and site data, including procedures and methodology with how samples were collected and processed, was previously discoursed in Section 2.3. Total carbon (C) [%] at the site level was the soil property investigated. These values were averaged from all the respective subplot C measurements per site for the 5–15 cm depth layer of mineral soil. The measurements were retained in the original quantities pertaining to concentration as recorded from the laboratory analysis.

4.2.3. Environmental Covariates

In advance of modeling work, particularly for data-driven approaches, it was generally unknown what specific covariates were of the most relevance. As soil properties were impacted from a combination of soil formation factors (McBratney, Mendonça Santos, and Minasny 2003; Minasny et al. 2019) relating to vegetation, topography, climate, parent material and time, therefore multiple attributes were of importance for modeling. Consequently, a multisource remote sensing data approach was warranted. Since environmental covariate data was typically retrieved by remote sensors, various sensors were utilized to capture information relating to different physical attributes (Mulder et. al 2011). A variety of remotely sensed data was compiled from several sensors, conforming to environmental covariates that were applied as predictors with modeling (Table B.2 in Appendix B). This also included LiDAR point cloud data. Much of this remote sensing data was processed before specific applicability as predictors for modeling work.

The environmental covariates utilized as predictors are listed in Table 4.1. There were 34 environmental covariates in total, obtained from LiDAR, surface reflectance (SR), SAR, aeromagnetic retrievals and a tree species prediction map for black spruce. These covariates related to the soil formation factors of topography, vegetation, parent material and time. All these environmental covariates corresponded to spatial resolutions of 30 m, with the black spruce prediction map and aeromagnetic data resampled to 30 m spatial resolution. The predictors in the list in Table 4.1. that were marked with an asterisk (*) were the predictors fielded for a model regression via SEM. Otherwise, all 34 predictors were utilized for all models.

Table 4.1. Environmental covariates utilized as predictors with modeling. Note that an asterisk (*) denotes predictors used for the structural equation modeling (SEM) comparison.

Predictor	Source	Soil Formation Factor
Digital Elevation Model (DEM) *	LiDAR	Relief
Canopy Height Model (CHM)	LiDAR	Vegetation
Gap Fraction	LiDAR	Vegetation
Aspect	DEM (LiDAR)	Relief
Convergence Index	DEM (LiDAR)	Relief
Mid Slope Position *	DEM (LiDAR)	Relief
Multi-resolution Ridge Top Flatness (MRRTF)	DEM (LiDAR)	Relief
Multi-resolution Valley Bottom Flatness (MRVBF)	DEM (LiDAR)	Relief
SAGA Topographic Wetness Index (SAGA TWI)	DEM (LiDAR)	Relief
Slope *	DEM (LiDAR)	Relief
Slope Height *	DEM (LiDAR)	Relief
Slope Length	DEM (LiDAR)	Relief
Stream Power Index	DEM (LiDAR)	Relief
Terrain Ruggedness Index (TRI)	DEM (LiDAR)	Relief
Topographic Wetness Index (TWI)	DEM (LiDAR)	Relief
Total Curvature	DEM (LiDAR)	Relief
Valley Depth	DEM (LiDAR)	Relief
Visible Sky	DEM (LiDAR)	Relief
B1 Summer 2017	SR	Vegetation
B2 Summer 2017	SR	Vegetation
B3 Summer 2017	SR	Vegetation
B4 Summer 2017	SR	Vegetation
B5 Summer 2017	SR	Vegetation
B6 Summer 2017	SR	Vegetation
B7 Summer 2017	SR	Vegetation
B10 Summer 2017	SR	Vegetation
Modified Normalized Difference Water Index (MNDWI) Summer 2017	SR	Relief (Water)
Normalized Difference Vegetation Index (NDVI) Summer 2017	SR	Vegetation
Change Magnitude B3 B4 B5 Summer 1984-2005 *	SR	Time
SAR C VH May 2017	SAR	Relief (Water)
SAR C VV May 2017 *	SAR	Relief (Water)
Gravity Anomaly 2016	Aeromagnetic	Parent Material
Magnetic Residual November 2018	Aeromagnetic	Parent Material
NFI Black Spruce 2011 *	k-NN Model	Vegetation

4.2.4. Encoder-Decoder Models

EDs incorporating NN architecture have been the focus for the modeling work. A schematic for the form of an ED is shown in Figure 4.1. Predictors are first inputted into the encoder component consisting of multiple layers of NN structure, from which the predictor data is condensed into a representation vector. Afterwards, this representation vector is inputted into the decoder component which is composed of NN layers with an upscaling and eventual downscaling to yield the regressed quantity of soil C. An ED constituting DNN formulations (ED-DNN), as well as an ED that is composed of CNN structure (ED-CNN) have been investigated. The formulation for the ED-DNN is listed in Table 4.2 and that for the ED-CNN is outlined in Table 4.3.

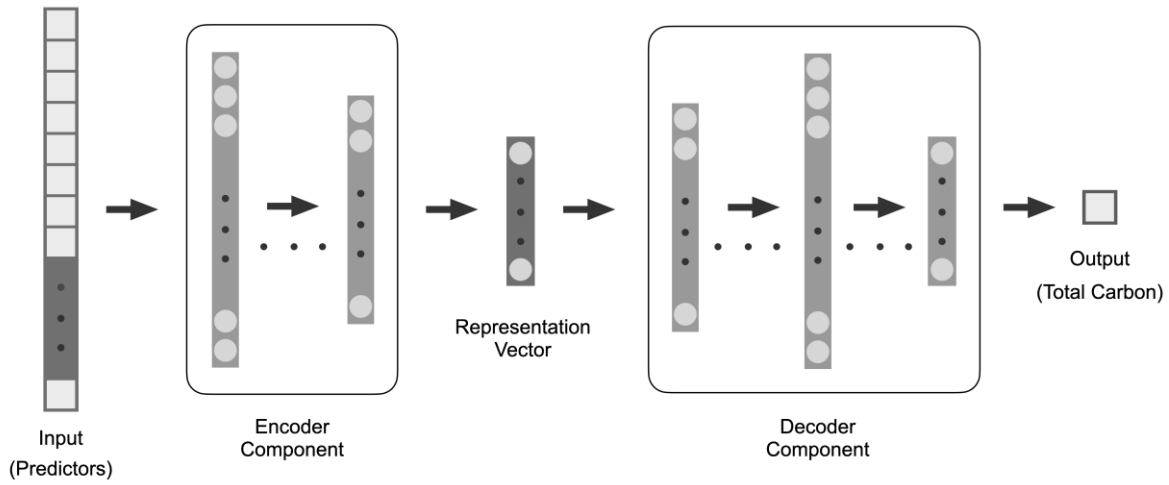

Figure 4.1. Schematic of encoder-decoder (ED) for modeling.

Table 4.2. Encoder-decoder composed of dense neural network (ED-DNN) layers.

ED-DNN

Layer	Output Shape	# Parameters
Encoder:		
Input Layer	34	0
Dense	32	1120
Dense	16	528
Dense	8	136
Decoder:		
Input Layer	8	0
Dense	32	288
Dense	16	528
Dense	1	17

Total Parameters: 2617

Table 4.3. Encoder-decoder composed of convolutional neural network (ED-CNN) layers.

ED-CNN

Layer	Output Shape	# Parameters
Encoder:		
Input Layer	(1, 34)	0
Convolution 1-D	(1, 16)	560
Max Pooling 1-D	(1, 16)	0
Convolution 1-D	(1, 32)	544
Flatten	32	0
Dense	8	264
Decoder:		
Input Layer	8	0
Dense	64	576
Reshape	(1, 64)	0
Convolution 1-D	(1, 32)	2080
Up Sampling 1-D	(2, 32)	0
Convolution 1-D	(2, 16)	528
Flatten	32	0
Dense	16	528
Dense	1	17

Total Parameters: 5097

Regarding structure, the ED-DNN consisted of dense layers as encountered in a DNN. The encoder for the ED-DNN was relatively simple, composed of just two dense layers for the internal layers. This encoder took the predictor input and condensed it to a representation vector of just 8 units, which was used as the input for the ensuing decoder process. The decoder was composed of two dense layers for the internal layers, with an upscaling to a dense layer with 32 units followed by a downscaling to a dense layer with 16 units, before the final output to 1 unit. Other sizes of units for these layers were investigated, even for the representation vector. Another study used a representation vector consisting of 8 units (Hu et al. 2022), so it was rationalized that this number of units was adequate for the compression of applicable information from predictors for C modeling.

The ED-CNN consisted of 1-D convolutional layers as utilized in a CNN (Hu et al. 2022), facilitating the refining of features at additional scales (Ji et al. 2021). The structure for the encoder portion included two 1-D convolutional layers, with the first of these followed by a max pooling layer, and the second convolutional layer followed by a flattening layer. A representation vector of 8 units was the result of the encoder, which was the input for the decoder. Within the decoder the input from the representation vector was upscaled to a 64 unit dense layer that was then reshaped into an array of 64 units. Afterwards there were two 1-D convolution layers, with an upsampling layer of size 2 after the first; output was flattened after the second convolutional layer which was followed by a dense layer before final output to 1 unit.

For training the EDs and other modeling with respect to a DNN and a CNN for comparison purposes, the keras package (Chollet 2015) in R was adopted. As the regression of C [%] was

the final output, the final layer for each of the EDs and NNs corresponded to continuous numerical output for 1 unit. Multiples of base 2 (Lee and Song 2019; Li and Li 2022) were specified for the units of the internal layers. Rectified linear unit (ReLU) activation functions were utilized, as these conformed to sparse activation with finer gradient propagation while avoiding vanishing gradients (Lee and Song 2019). An Adam optimizer (Li and Li 2022) was employed for training the EDs and NNs, to minimize mean squared error (MSE) for the loss function. For the ED and NN models, a learning rate of 0.001 was specified for the Adam optimizers. All EDs and NNs were each trained using the default batch size of 32 and over 5000 epochs, as 5000 epochs were found to be appropriate for attaining better accuracies (Sindi et al. 2021) for models composed of convolutional layers, but not so many epochs as to lead to model overfitting.

Before training, predictor data for the EDs and NNs models were standardized via normalization. This normalization was enacted by computing averages and standard deviations for each predictor across all sampling sites, by which all predictors were transformed into respective *z*-scores (Lee and Song 2019). For models with DNN architecture, the predictor input was reshaped in matrices with columns matching to predictors and rows to sites. Concerning the models with convolutional layers, predictor inputs were reshaped similarly as what were done for models with DNNs, but now with an added dimension for 1-D convolution. Once data was normalized, it was randomly split 70:30 for training and validation sets accordingly with respect to sampling site, based on the stratification of C so to obtain balanced sets for C concentrations with training and validation. This allocation resulted in 103 sites for modeling training and 41 sites for model validation, respectively.

4.2.5. Other Modeling

NNs in terms of both a DNN and CNN were employed for modeling the soil C concentration. Simpler formulations of the DNN and CNN were each specified, so to mitigate overfitting of these NNs. The DNN consisted only two internal dense layers, with the first dense layer corresponding to 32 units followed by a dropout of 0.25, then by the second dense layer of 16 units followed by dropout of 0.25 before the final output of 1 unit. As DNNs dealt with nonlinearity similarly as conventional regression methods (Were et al. 2015), the DNN model was considered as comparable to the RF and SEM with respect to accuracy. For the CNN, its first internal layers corresponded to a 1-D convolutional layer of 32 units, followed by another 1-D convolutional layer of 32 units. The next layers consisted in the following order of a 1-D max pooling layer, a 0.25 dropout, a flattening layer, a dense layer with 16 units, 0.25 dropout and then the final output layer of 1 unit. As for the CNN, the ordering of the predictors in the input did matter as neighboring pixels with predictors were imputed for the convolutions (Lee and Song 2019). The same consistent ordering of predictors was utilized for the models, with predictors ordered by sensor source with a further level of ordering by soil formation factor. However, in contrast to the DNN, the CNN was more suitable for leveraging nonlinearity amongst the predictors for modeling.

For comparison purposes, both a RF and SEM were employed for modeling. The RF was fitted to the same training sites and predictors as for the NNs and comprised of 1000 trees. Regarding the mtry parameter referring to number of variable splits, it was equaled to the floor of the square root of the number of predictors, which here was 5. For fitting the RF, the randomForest package (Liaw and Wiener 2002; Svetnik et al. 2003) in R was utilized. As for

the SEM, it was trained for soil C on the 7 most importantly ranked predictors as assessed from a feature selection, on the same training sets regarding sites as done for the other modeling approaches. These 7 predictors were denoted by asterisks in Table 4.1. For SEM, parameters were resolved by box-constrained optimization via nonlinear minimization with maximum likelihood estimation (Rosseel 2012). The fitting for SEM was accomplished by the lavaan package (Rosseel 2012) in R.

An ensemble composed of predictions from all models was also generated. This ensemble constituted the average (AVG) of the 6 models investigated: SEM, RF, DNN, CNN, ED-DNN and ED-CNN. As different modeling approaches were considered, the ensemble model was expected to improve robustness (Engler et al. 2013). In addition, an ensemble model was formulated as standard deviations were computed based on the modeling approaches to address uncertainty with prediction, discoursed in the following subsection for uncertainty quantification.

4.2.6. Model Evaluation

Assessment of model accuracy was reported on the validation set. The metrics for modeling accuracy were the coefficient of determination (R^2), the root mean squared error (RMSE) and the mean absolute error (MAE). In terms of units, the R^2 was dimensionless, whereas both MSE and MAE conformed to the quantity of concentration [%] as measured for C. In general, R^2 was confined between 0 and 1, with greater accuracy indicated by higher R^2 value.

To assess for the significance of R^2 values between the various models, inferential statistical testing was conducted. Fisher's z-test (Hu et al. 2017) assessed on correlation coefficients was implemented on the square root of the R^2 values which have been denoted using r. The null hypothesis for Fisher's z-test was that the correlation coefficients were the same between two comparisons; the alternative hypothesis was that the correlation coefficients were unequal between the two comparisons. The z-value for this test was calculated by

$$z = \frac{r_1' - r_2'}{\sqrt{2/(n-3)}}, \quad r' = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right). \tag{4.1}$$

Here, n corresponded to the number of samples in the validation dataset, which here was fixed at 41 for the number of validation sample sites. Sample correlation coefficients for the comparison testing were $r_1$and r_2, which were each transformed using the natural logarithm for calculate the r_1' and r_2' values used for calculating the z-value. The critical z-value was 1.96 for the calculation in (4.1), with the alternative hypothesis of unequal correlation coefficients if the calculated z-value exceeded 1.96.

Pairwise scatterplots and correlations among the models were resolved so to examine the error patterns with residuals between the modeling approaches. For these residuals, similar spreads with pairwise scatterplots and high pairwise correlations conveyed comparable tendencies for prediction. Part of the justification of a using a model ensemble was to incorporate prediction power from each model for creating a more robust model, even if the model ensemble was less accurate in assessed modeling accuracy than that for the best model. Scatterplots for the best model with respect to both measurement versus prediction as well as measurement versus residual were also plotted to gauge trends with underprediction and overprediction.

4.2.7. Uncertainty Quantification

As various modeling approaches were employed to generate prediction maps, a measure of uncertainty with prediction was derived for the ensemble based on the prediction maps computed from these models. Specifically, the standard deviation per pixel was calculated from each model prediction for the domain of the prediction maps. The prediction maps for each model corresponded to the same study area bounds, with each prediction map conforming to the same amounts and positions for data points. It was verified that the coordinate projection, registration grids, cell sizes, spatial extents and numbers of non-missing data points were the same for all prediction maps. Furthermore, the land covers within the study area domain corresponded to the same pixels, meaning that the same percentage of pixels within each prediction map conformed to the same land cover type. Consequently, quantile mapping with respect to calculating quantiles was also implemented, as it ascertained whether pixels matching to the same locations or land cover were predicted similarly in the context of underprediction or overprediction with modeling.

Quantile maps (Maraun 2013) conforming to computed quantiles were calculated from the prediction maps for each model. Before resolving quantiles, for each prediction map all n data points were sorted by ascending order; specifically, as $x_1, x_2, x_3, \cdots, x_{n-1}, x_n$ with $x_1 \leq x_2 \leq x_3 \leq \cdots \leq x_{n-1} \leq x_n$. The total number of data points corresponded to the number of pixels with non-missing values. Therefore, the respective index i_q for the q-th k-quantile, with $1 \leq q \leq k$, and q, k both integers, was calculated by

$$i_q = \frac{q}{k} \cdot (n + 1) \,. \tag{4.2}$$

Interpolation was required when calculating the quantile value from its respective index, except for when the index equaled an integer value. For the index i_q corresponding to a rational number of form $x.p$, here with composed integer component x and decimal component p with $0 \leq p < 1$, the cutoff value for the qth quantile was computed as

$$x_{i_q} = (1 - p) \cdot x_{\text{floor}(i_q)} + p \cdot x_{\text{floor}(i_q)+1} \,. \tag{4.3}$$

In this equation, the floor function yielded the integer value of the index, which was the greatest integer less than or equal to the index i_q. Cutoff values for all k quantiles were computed this way. Once all cutoff values for the quantiles were resolved, all data points were allocated into their respective quantiles. This process was repeated for each prediction map.

For this analysis here, it was reasoned that 10 quantiles, otherwise known as deciles, were appropriate. This was deemed a sufficient number of quantiles to adequately represent heterogeneity for the various land cover types within the study region domain. Here, 6 different prediction maps were utilized for the calculations, corresponding to the SEM, RF, DNN, CNN, ED-DNN and ED-CNN model output. To compute the deciles, the ntile function with the dplyr package (Wickham et al. 2023) in R was utilized. Pixels conforming to missing data were excluded from these calculations, as this data contained no target data values as these corresponded to either gaps within the domain or the exterior outside the boundaries of the study area. The standard deviation map for C predictions conformed to the same quantities as for C [%], whereas the standard deviation map based on deciles was unitless. These maps provided insights for gauging uncertainty in terms of over and underprediction.

4.3. Results

4.3.1. Modeling Accuracies

Models for SEM, RF, DNN, CNN, ED-DNN and ED-CNN were each fitted on the training allocation of data. In addition, an ensemble corresponding to the average (AVG) of predicted values based upon all other models was also computed. For evaluation, all the models predicted C on the validation set of data from which accuracies were assessed, with modeling accuracies presented in Table 4.4. Regarding Table 4.4, comparable accuracies were obtained for the SEM, RF and DNN models, with R^2 of approximately 0.20 for each of those models. The CNN model attained better modeling accuracy than those three models, with a R^2 of 0.37. Nevertheless, the EDs obtained the best modeling accuracies with R^2 greater than 0.50, indicating moderate performance with respect to R^2 accuracies. Similarly, the ensemble model (AVG) attained a R^2 of 0.50. The lowest RMSE and MAE were also assessed for the EDs, supporting that the ED models were an improvement beyond the other approaches in enhancing modeling accuracy. Note that from statical testing using Fisher's *z*-test that R^2 accuracies for the ED-CNN model were statistically significance improvement from the R^2 accuracy results attained by the SEM, RF and DNN models.

Scatterplots of measurements versus predictions as well as of measurements versus residuals were each plotted for the best model, here the ED-CNN, as shown in Figure 4.2. Reference lines were also marked within each scatterplot. The reference line in the scatterplot of measurements versus predictions corresponded to the diagonal line where prediction equaled measurement, indicating perfect prediction (Ke, Ren, and Yin 2024). On the scatterplot for measurement versus residual, the vertical reference line indicated a residual value of zero,

denoting perfect prediction. The ED-CNN attained the best accuracy, with a R^2 of 0.59 as noted. From inspecting these scatterplots, the greatest variations with prediction occurred for sites with greater measured soil C. In general, soil C was underpredicted for the sites with the greatest C measurements, which here pertained to the wetlands. Contrastingly, soil C was also overpredicted somewhat for the sites with the lowest C measurements, which here corresponded to agricultural land. This over and underprediction with reference to certain land cover types was part of the reason warranting the subsequent investigation of quantile mapping, for uncovering insights regarding prediction trends. As an aside, note that an outlier was present in both plots in Figure 4.2, denoted by the site with the most negative residual value. This site consisted of intermediate C that was subsequently predicted with a greater quantity of C. Upon investigation, this discrepancy arose as this site corresponded to secondary forest consisting of black spruce with lower CHM, that was predicted to have greater C as analogous to wetland sites composed of stunted black spruce.

Residuals for each of these models were also computed upon the validation set. These residuals were calculated as respective measurements subtracting predictions. Pairwise scatterplots of the residuals between models were plotted, with pairwise correlations reported, as displayed in Figure 4.3. Note that all the pairwise correlations were greater than 0.5, with larger correlations of greater than 0.7 indicated for most of the pairwise comparisons between models. Therefore, residuals between each model were positively correlated with one another, and spreads within the residual scatterplots were relatively linear, conveying that soil C was predicted with similar tendencies across the different models.

Table 4.4. Modeling accuracies for the models as evaluated on validation set. The coefficient of determination (R^2), root mean squared error (RMSE) and mean absolute error (MAE) are reported. Results of statistical testing for differences with R^2 accuracies between each model and ED-CNN are noted.

Model		**R^2**	**ED-CNN Comparison**		**RMSE**	**MAE**
			***z*-Value**	***p*-Value**	**[%]**	**[%]**
Structural Equation Model	SEM	0.21	2.27	**0.02 ***	10.18	7.63
Random Forest	RF	0.22	2.21	**0.03 ***	10.09	7.83
Dense Neural Network	DNN	0.18	2.45	**0.01 ***	10.38	7.35
Convolutional Neural Network 1-D	CNN	0.37	1.35	0.18	9.07	6.05
Encoder-Decoder DNN	ED-DNN	0.54	0.33	0.74	7.78	5.40
Encoder-Decoder CNN 1-D	ED-CNN	0.59			7.30	5.38
Ensemble †	AVG	0.50	0.59	0.56	8.06	6.11

† Average prediction of all 6 models.

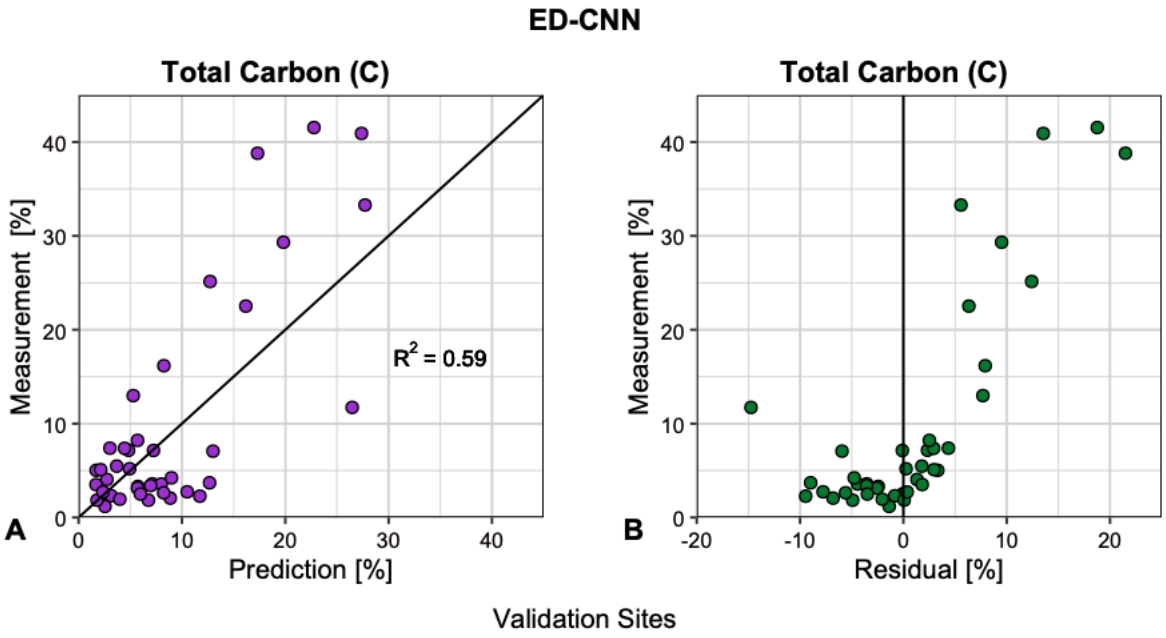

Figure 4.2. Scatterplots of encoder-decoder comprised with 1-D convolutional neural network layers (ED-CNN) for total carbon (C) on validation set. A: measurements versus prediction; B: measurements versus residuals.

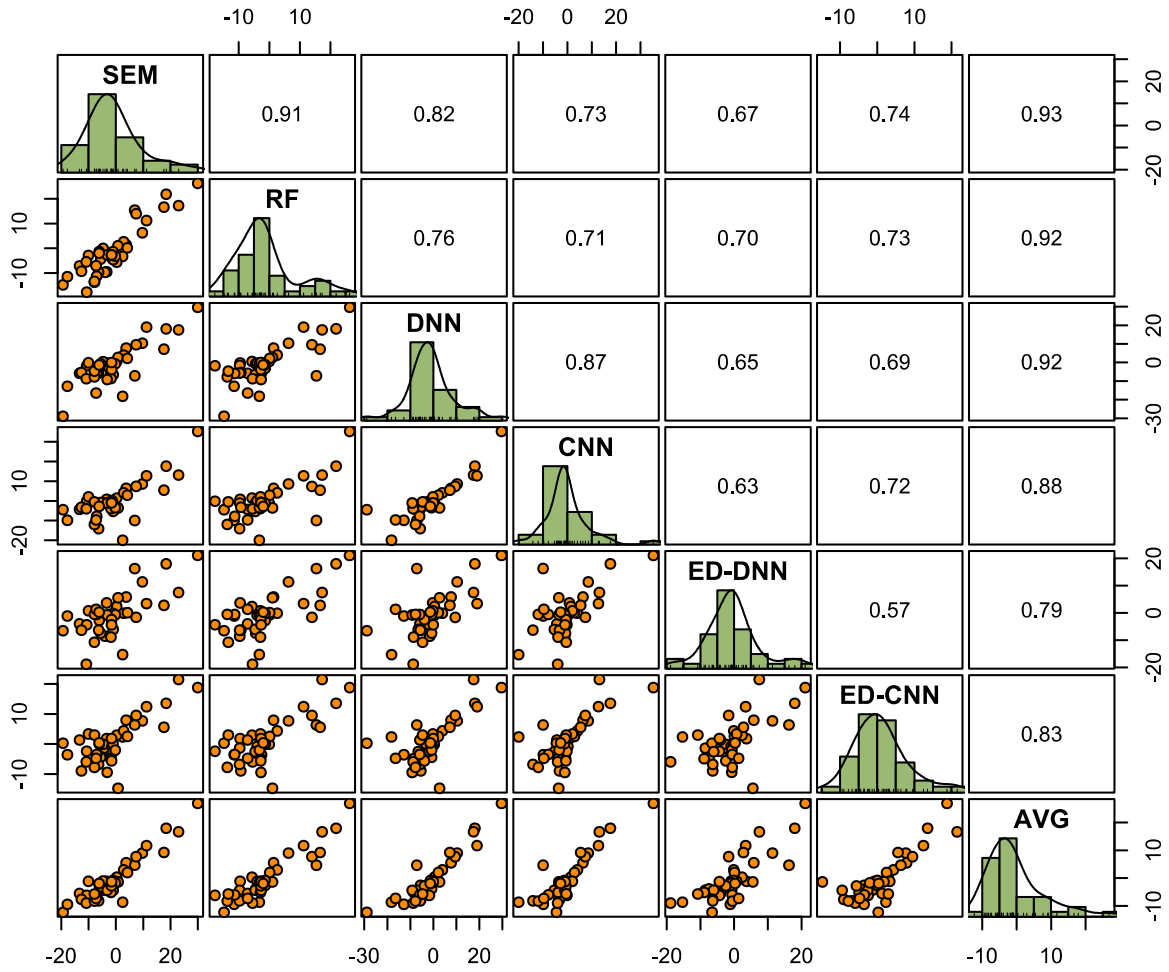

Figure 4.3. Below diagonal: Pairwise scatterplots of residuals between models. Diagonal: Histograms of residuals for the model. Above diagonal: Pairwise correlations of residuals between models.

4.3.2. Prediction Maps

Prediction maps corresponding to soil C for the SEM, RF, DNN, CNN, ED-DNN and ED-CNN models are presented in Figure 4.4. These maps conform to total carbon (C) [%] for the 5–15 cm depth layer of mineral soil. The same symbology settings with respect to coloring have been used for all these prediction maps, with brown coloring denoting the lowest C and

green coloring the highest C, respectively. Similar trends with prediction are noticed in all these prediction maps. Note that the ED-CNN prediction map corresponds to the most accurate model (see Table 4.4). All models that are composed of NNs have predicted similar magnitudes of C for the wetland areas. Conversely, the SEM and especially the RF have underpredicted C for the wetlands. Furthermore, localities with higher and lower predictions, as well as intermediate values with predictions, can all be discerned from the prediction maps for the CNN and ED-CNN models. Localities with intermediate C contents mostly conform to forested regions outside of wetland areas.

Deciles maps were computed from each of the respective prediction maps for C as shown in Figure 4.5. For these maps, lighter coloring intensity matched to lower deciles corresponding to lesser C, whereas darker coloring intensity indicated higher deciles and greater C, accordingly. Wetland areas conformed to the highest deciles, and agricultural fields to the lowest deciles, respectively. Similar tendencies with prediction are observed from these decile maps, conveying that each model predicted C consistently for the various land cover types when accounting for over or underprediction.

A zoomed-in display of the predictions for the ED-CNN are shown in Figure 4.6. Here, a different symbology coloring was utilized so to make the differences of C with land cover type more apparent. Inspecting these figures, the agricultural lands have the lowest C, whereas the wetland areas have the greatest C.

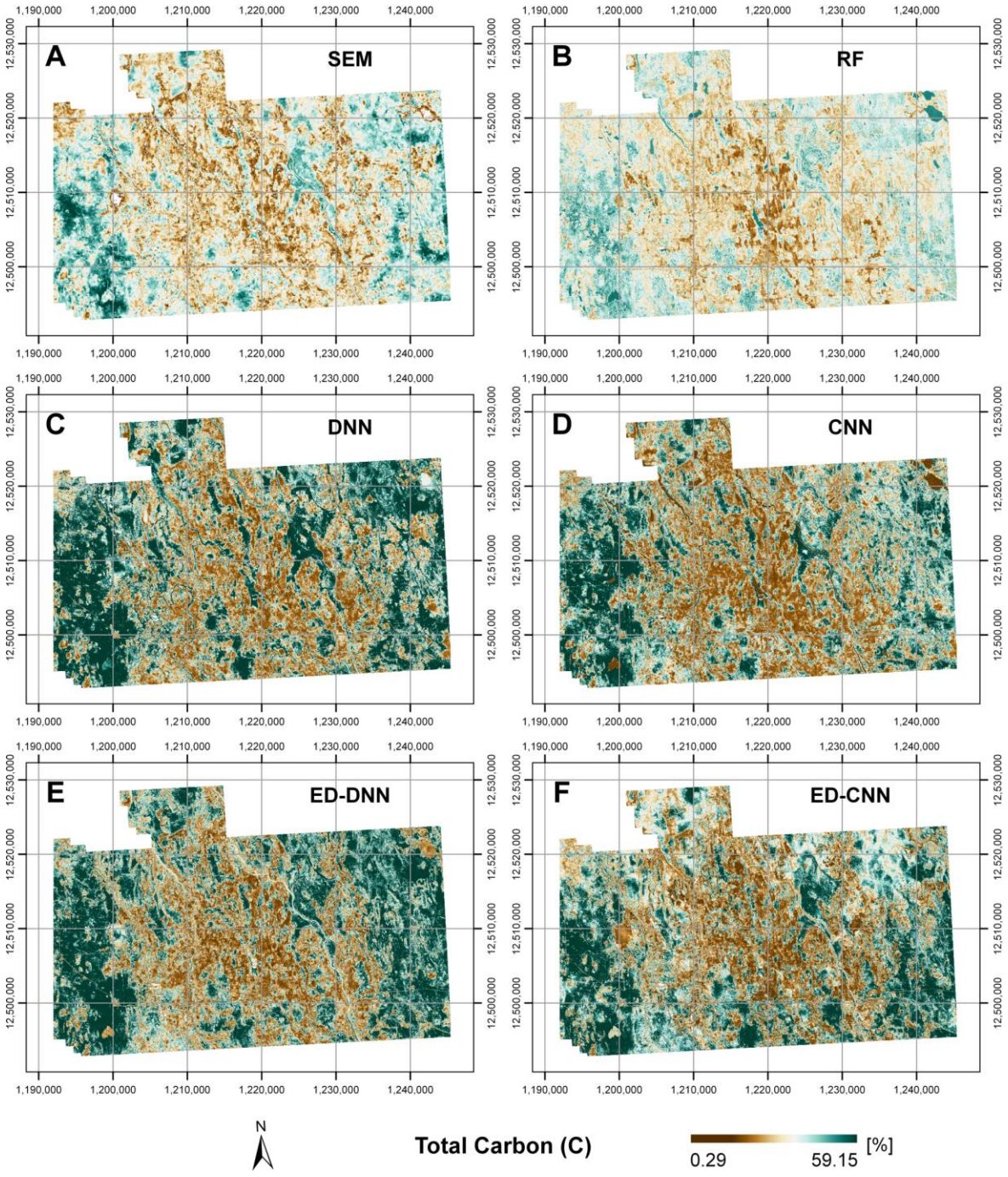

Figure 4.4. Prediction maps of total carbon (C) [%] as generated by models for the Cochrane study area. A: random forest (RF); B: structural equation modeling (SEM); C: dense neural network (DNN); D: 1-D convolutional neural network (CNN); E: encoder-decoder composed of dense neural network layers (ED-DNN); F: encoder-decoder composed of 1-D convolutional neural network layers (ED-CNN).

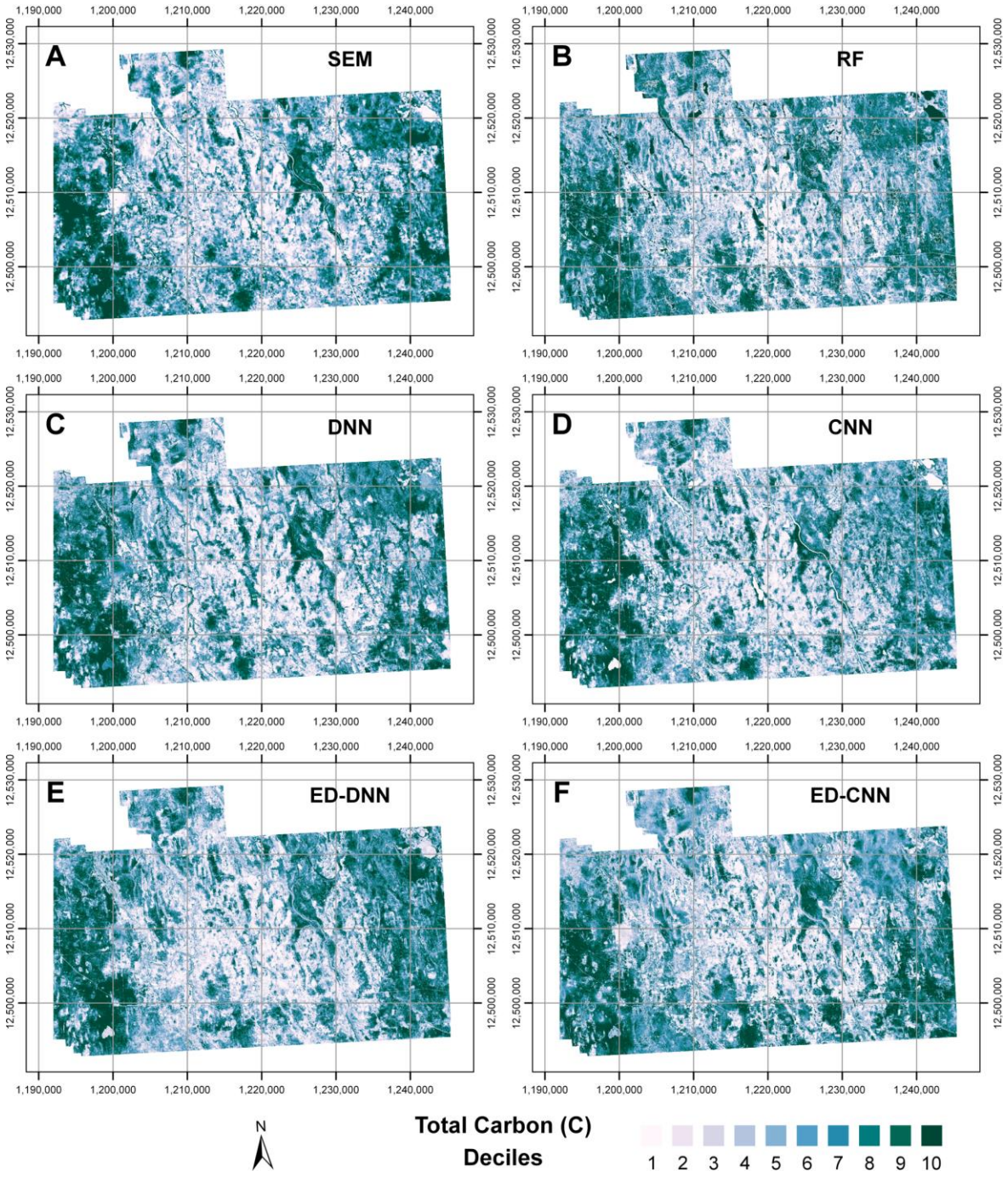

Figure 4.5. Prediction maps of deciles for total carbon (C) as generated by model for the Cochrane study area. A: random forest (RF); B: structural equation modeling (SEM); C: dense neural network (DNN); D: 1-D convolutional neural network (CNN); E: encoder-decoder composed of dense neural network layers (ED-DNN); F: encoder-decoder composed of 1-D convolutional neural network layers (ED-CNN).

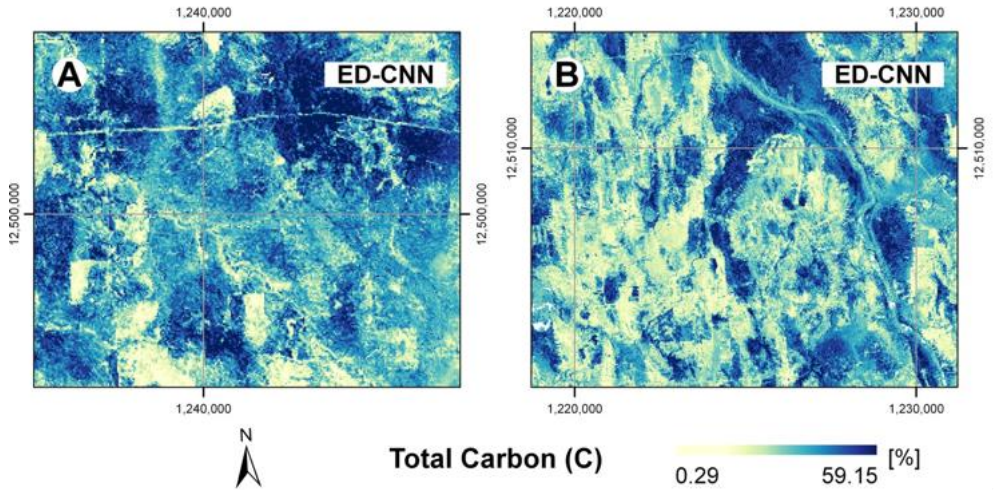

Figure 4.6. Zoomed-in detail with predictions of total carbon (C) [%] for encoder-decoder composed of 1-D convolutional neural network layers (ED-CNN). Note that wetland areas contain greater concentrations of C, whereas agricultural fields contain lowers amounts of C.

The model ensemble (AVG) and standard deviation prediction maps are displayed in Figure 4.7. Upon inspection, the trends for the AVG and AVG Decile maps are comparable to those in Figures 4.4 and 4.6, respectively. In particular, the prediction map for AVG denotes a mix of localities with lower, intermediate and greater contents of soil C. Regarding the standard deviation (STDDEV) prediction map, the largest standard deviations in terms of C [%] are for wetland localities. The models have consistently predicted high C for wetland areas, but the magnitude of these predictions vary by model. However, when inspecting the standard deviation plot that is based on the decile maps, higher standard deviations are discerned for forested regions near streams, rivers, or the edges of wetlands. Therefore, both types of these standard deviation maps impart useful insights. The STDDEV plot indicates that wetland areas have the highest uncertainties in terms of magnitudes of C [%]. Accordingly, the STDDEV decile plot shows that localities with the greatest prediction uncertainties in terms of quantiles coincide to localities of forest near water bodies. It is likely that these forested

areas along rivers and streams correspond to locations of greater diversity with vegetation, indicating a need for models to better predict C for these sorts of localities consisting of intermediate soil C contents.

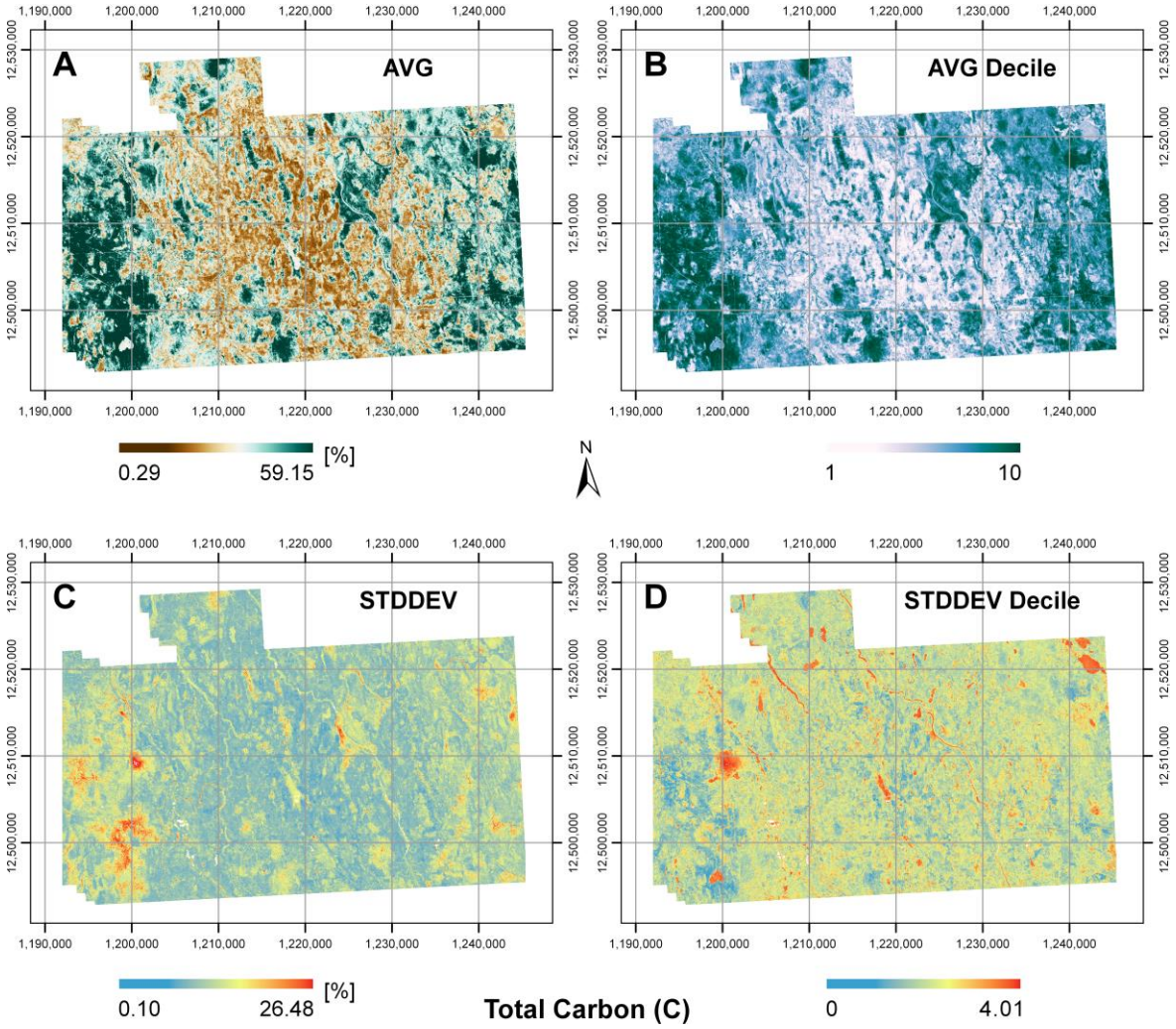

Figure 4.7. Ensembled prediction maps for the Cochrane study area. A: mean of total carbon (AVG) [%]; B: mean of decile total carbon (AVG Decile); C: standard deviation of total carbon (STDDEV) [%]; D: standard deviation of decile total carbon (STDDEV Decile).

4.4. Discussion

As sites with available soil property data was severely limited for the boreal forest in northern Ontario, the complexity of NNs in terms of numbers of units and layers were restrained to minimize the number of parameters to optimize regarding training. At first, DNNs and CNNs with no more than 16 units within each internal layer were investigated. However, the number of units per internal layer were scaled up to as much as 32 units so to attain better modeling accuracies. As reported in Tables 4.2 and 4.3, the ED-DNN and ED-CNN required the training for 2617 and 5097 parameters, accordingly. The DNN and CNN models reported in Table 4.4 used 1665 and 2721 parameters, respectively. For this study there were 144 sampling sites, with 103 of those sites reserved for model training purposes. Compared to CNNs utilizing greater than tens of thousands of parameters (Lee and Song 2019), the EDs adopted in this study utilized far less parameters while still enhancing modeling accuracies for a customary study with respect to regional scale digital soil mapping.

A different normalization for the predictor data was also tried. This alternative normalization conformed to linearly scaling each predictor between 0 and 1, basing upon the corresponding range between maximum and minimum values for each predictor, accordingly. However, when this normalization scaling was enacted the stability regarding training for the NNs was compromised as modeling accuracy noticeably worsen after a few additional epochs, as well as between simulation runs. Consequently, this issue was the reason why standardization with *z*-score (Lee and Song 2019) was implemented for normalizing the predictor data. When normalizing by scaling with the *z*-scores, the respective means and standard deviations for each predictor were calculated based upon all sites from the input data set. Thus, the same

means and standard deviations were utilized for the *z*-scores regarding the training and validation data sets. It was necessary to scale predictors using the same corresponding *z*-scores for both the training and validation sets, as the models were trained on predictors transformed by *z*-score values. Furthermore, the same mean and standard deviations values for calculating the *z*-scores for each predictor were also utilized when generating prediction maps. This was implemented so that the predicted values conformed to the same transformations as adapted for model training.

Improvements with modeling accuracy have been attained by the ED models, as is seen in Table 4.4. The ED models seem to filter out noise and sort out nonlinear interactions within the predictor data, that can be difficult to accomplish by other modeling approaches. Within digital soil mapping, it is common for dozens or more predictors to be utilized for modeling, even if the bulk of the prediction power can be attributed to only a few of those predictors (Minasny et al. 2019). Predictors that have been assessed with lower variable importance can potentially contribute informational gain regarding modeling for certain land cover types, but likely this information is loss as noise within the modeling. Thus, EDs can mitigate information loss due to this noise. Concerning EDs, there are functional differences between the ED-DNN and ED-CNN. As the ED-CNN is composed of convolutional layers, local features within the data can be extracted (Ke, Ren, and Yin 2024). Correspondingly, an ED-DNN can also contend with nonlinearity or nonconvexity within data (Bai et al. 2024), but while generally necessitating the optimization of fewer modeling parameters. All in all, when compared to other modeling approaches, EDs can be employed to further extract relevant information from higher-dimensional data sets regarding modeling.

The results with the deciles mapping can support the selection of sites for obtaining soil samples in future field campaigns. This analysis indicates localities with the greatest uncertainties in terms of standard deviations when mitigating for the effects of over or underprediction. Thus, these greatest uncertainties correspond to areas where more samples should be collected so to potentially improving soil modeling accuracy. Inspecting Figure 4.7-D, the greatest uncertainties with the deciles for soil C concentrations are for valleys or locations surrounding water. As these areas conform to forested localities with more variation of vegetation, an aim for selecting soil sampling locations should be to sample within localities surrounding water but outside of the wetland environments. Specifically, more forested sites of different age stands should be sampled when wanting additional sites for soil data.

Regarding the standard deviation (STDDEV) prediction map, the largest standard deviations in terms of C [%] are for wetland localities. The models have consistently predicted high C for wetland areas, but the magnitude of these predictions vary by model. However, when inspecting the standard deviation plot that is based on the decile maps, higher standard deviations are discerned for forested regions near streams, rivers, or the edges of wetlands. Therefore, both types of these standard deviation maps impart useful insights. The STDDEV plot indicates that wetland areas have the highest uncertainties in terms of magnitudes of C [%]. Accordingly, the STDDEV decile plot shows that localities with the greatest prediction uncertainties in terms of quantiles coincide to localities of forest near water bodies. It is likely that these forested areas along rivers and streams correspond to locations of greater diversity with vegetation, indicating a need for models to better predict C for these sorts of localities consisting of intermediate soil C contents.

There is also the issue of uncertainties in modeling work, which ultimately impacts resulting modeling accuracy. Specifically, uncertainties can manifest from many aspects or sources (Caers 2011). These uncertainties can exist with ground truth, within measurements of soil properties (Tan 2005) as well as misidentification of categorical characteristics such as soil texture family or even vegetation tally counts for tree species classification (Williams et al. 2001). Within covariate data as observed by remote sensing technologies, there also persists uncertainties in measurements, relating to sensor calibration thresholds (Rowlandson et al. 2013) or from pointing errors with the remote sensor based on a satellite (Maharjan and Kim 2024). Uncertainties also manifest for the modeling process. Specifically, predictors cannot fully capture or resolve all the variation within modeling (Caers 2011), particularly for data-driving approaches where no resulting model is exactly perfect. There is even uncertainty with classification for prediction (Kumar et al. 1986; Wu et al. 2021), which has been deemed as important for soil modeling (Minasny et al. 2019). The issue of uncertainty, though impossible to fully overcome, nonetheless should be addressed in some form whenever feasible.

4.5. Summary

- Enhanced modeling accuracies were attained with ED approaches. In particular, an ED composed of CNN structure obtained the best modeling accuracy, which was followed next in accuracy by an ED consisting of DNN architecture.
- The validation R^2 was 0.59 for the ED-CNN, which was a vast improvement over R^2 values that were achieved by conventional modeling approaches such as RF.

- An ensemble model (AVG) based equally on all other modeling approaches still attained validation a R^2 accuracy of 0.50, but with added robustness. Specifically, intermediate values of C content were also better predicted from this model.
- Insights pertaining to over and underprediction were gauged from prediction maps corresponding to standard deviations of the targeted variable, as well as for standard deviations of quantile predictions of the targeted variable.
- The standard deviation map based upon deciles was able to reveal areas with greatest modeling uncertainties when mitigating for over and underprediction magnitudes.
- The areas with the greatest prediction uncertainties with respect to magnitude of C [%] were for the wetlands. This was due in part to the overall high predictions for C within wetland areas, so deviances arising from moderate or even still relatively high predictions in C [%] led to high uncertainties in predicted contents for the wetlands.
- Regarding deciles, the greatest uncertainties with prediction were for forested regions along streams, rivers and the peripherally of wetlands. This arose in part due to the greater diversity of vegetation in these localities, indicating a need to better model these land cover types.

Chapter 5

Enhanced Implementation for Structural Equation Modeling of Soil Carbon and Associated Properties

5.1. Background

Machine learning approaches have emerged as standard modeling procedures for digital soil mapping. In particular, random forest (RF) models have been increasingly employed for digital soil mapping (Minasny et al. 2019), with implementation for both classification (Brungard et al. 2015; Heung et al. 2016) and regression problems (Beguin et al. 2017). When compared to other approaches, RFs tend to attain better modeling accuracies (Beguin et al. 2017; Brungard et al. 2015; Heung et al. 2016). However, many machine learning approaches do not decipher further insights concerning relations between predictors and target properties with respect to modeling. Of interest is the better comprehension as well as investigation into resolving physical processes or determinants in the context of modeling dynamics for digital soil mapping research.

Structural equation modeling (SEM) can be harnessed for regression while also rendering a path analysis to quantify relations between predictors and target variables (Rosseel 2012). Via path analysis, correlations and strength of causal relationships regarding SEM can be determined, but directions for causality cannot be resolved (Angelini et al. 2016). Thus, SEM can be employed to unravel and explicate connections between predictors and interrelated

soil properties. SEM has been previously adopted within other research fields (Lin 2021); even for soil mapping (Angelini et al. 2016). Nevertheless, limitations persist relating to the application of SEM within digital soil mapping. Specifically, large sets of predictors are typically applied for digital soil mapping, which can lead to multicollinearity issues with the SEM and as well as impact goodness of fit assessments (Hu and Bentler 1999). Prior to modeling, it is usually unknown what specific predictors are most relevant, warranting the need to utilize many different types of environmental covariates as predictors. Hence, there exists a need for a feature selection process to resolve smaller sets of the most applicable predictors, so to facilitate SEM for digital soil mapping.

Ranking of variable importance can be established by RFs (Brungard et al. 2015) and has been commonly employed in soil modeling. However, variable importance is dependent on the modeling approach that has been utilized, as different sets of predictors can be ascertained as important from other methodologies. This disparity can arise due to bias with methodology for variable ranking (Loecher 2022), or due to nonlinear interactions among predictors that cannot be explicitly resolved (Cheng and Bai 2019; Zhang et al. 2020). Multicollinearity between predictors may not be assessed by the approaches that have been initiated for variable importance. Consequently, a need exists for devising a framework to effectuate the application of SEM to digital soil mapping as more systematic. This modeling template should enact a multiapproach for gauging variable importance and resolve pairwise correlations so to avoid predictors with high correlations. The SEM should be fitted in a sequential stepwise process, evaluating goodness of fit metrics so to carefully assess the SEM for digital soil mapping.

5.2. Materials

5.2.1. Study Area

The study areas for this research correspond to the same areas as have been delineated for Chapter 2, that are situated within the District of Cochrane in Ontario, Canada. These study areas consist of Hearst, Gordon Cosens Forest (GCF), Kapuskasing within the GCF, and then Cochrane, in reference from the northwest to southeast along the Ontario highway 11 corridor, respectively. The locations of these study areas as well as for the sites where soil samples have been extracted, is shown in Figure 5.1. Note that Figure 5.1 is similar to Figure 2.2, apart from true-color composite imagery comprising the backgrounds of the delineated study areas in Figure 5.1 versus a CHM in Figure 2.2. The true-color imagery has been attained from Landsat 8 surface reflectance corresponding to the summer of 2017. Details regarding topography, vegetation and land covers for this study region have already been discoursed in Section 2.2 and thus have not been repeated here.

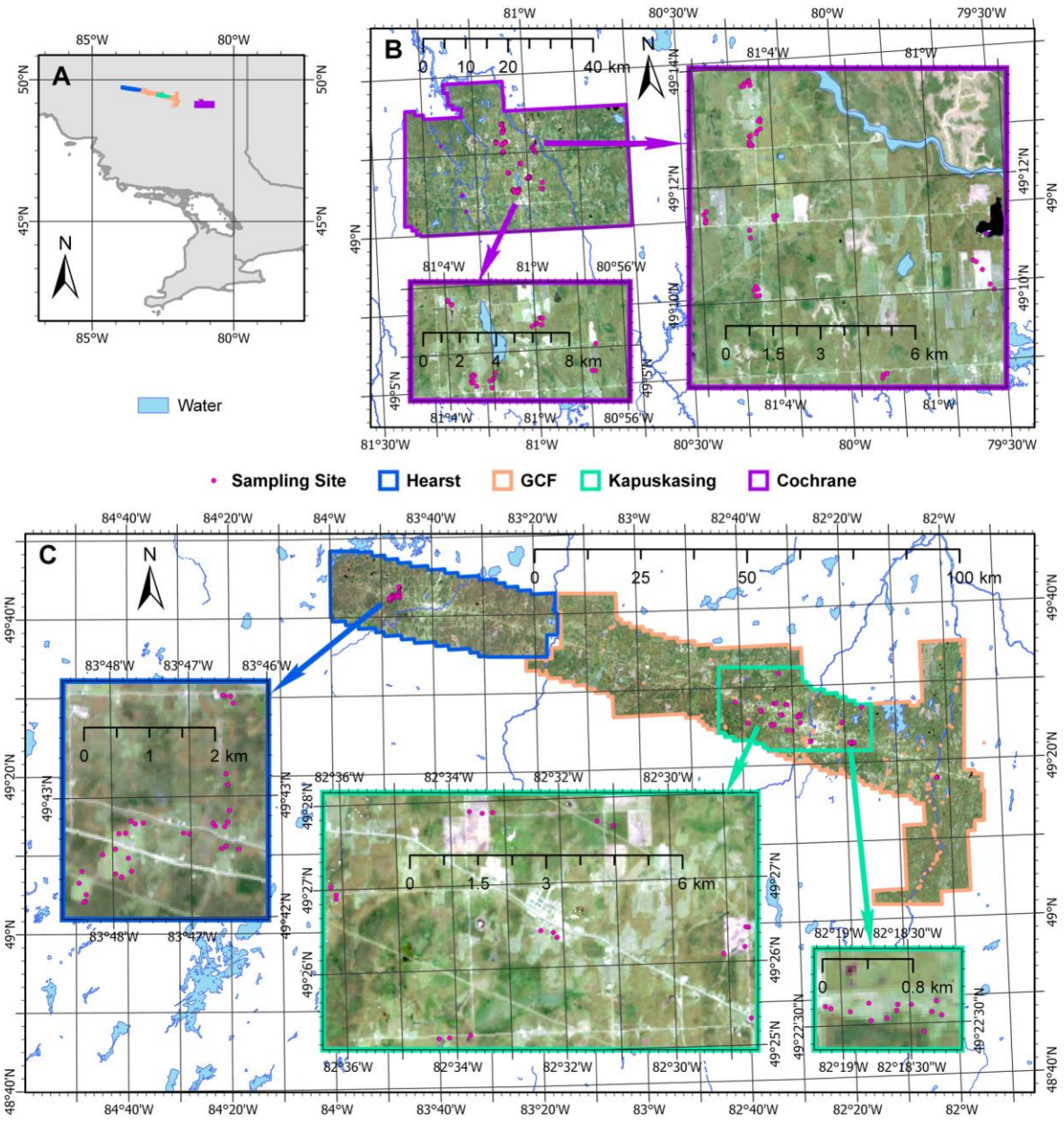

Figure 5.1. Study areas and location of sampling sites, with true-color composite imagery depicted within backgrounds of delineated study areas. Water bodies within region are also displayed. A: Reference of location for the study areas in northern Ontario, Canada. B: Cochrane study area, with insets focusing on clusters of sampling sites. C: Hearst, Gordon Cosens Forest (GCF) and Kapuskasing (within GCF) study areas, with insets focusing on clusters of sampling sites.

5.2.2. Soil Data

The soil data used for this research was obtained from samples extracted during field campaigns in September 2018, August 2019 and August-September 2021. In total, soil samples were obtained for 144 sites corresponding to various representative land cover types within the study region. Additional information concerning the soil samples has been previously described in Section 2.3. For this research, the soil properties of bulk density (BD) [g/cm^3], total carbon (C) [%; g/kg] and carbon-to-nitrogen ratio (C:N) were investigated regarding soil mapping. These soil properties each corresponded to one measurement per site, averaged from measurements resolved at all subplots for each site, from soil samples collected at the 5–15 cm depth of mineral soil. At this soil depth, C tended to be maximal, with the presence of root material for the overlying vegetation.

Histograms for these soil properties are presented in Figure 5.2. Note the wide variations in measurements for each soil property, as the corresponding samples have been retrieved from various types of land covers. The soil properties of BD and C:N are more centrally distributed, whereas that for C is skewed to the right. Pearson correlations have also been calculated for these soil properties, with associated pairwise correlation coefficients of −0.81 between BD and C, −0.35 for BD and C:N, and 0.54 for C and C:N, accordingly. Therefore, a large negative correlation exists between BD and C, with a moderate positive pairwise correlation for C and C:N. Consequently, the larger absolute values of pairwise correlations greater than 0.5 between soil properties support integrated modeling of BD for C, and C for C:N with SEM.

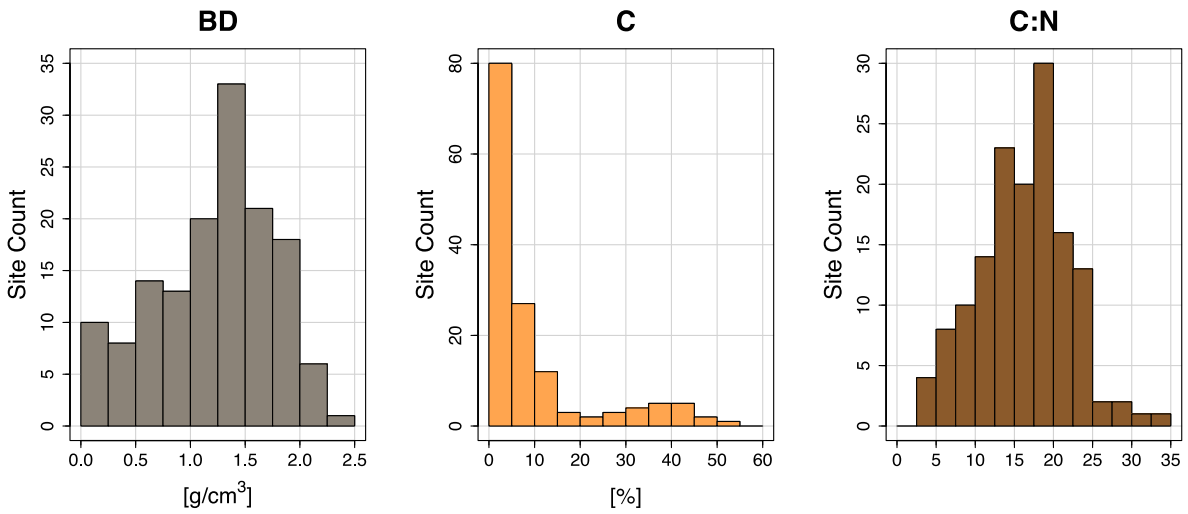

Figure 5.2. Histograms of bulk density (BD), total carbon (C) and carbon-to-nitrogen ratio (C:N) for the soil samples as obtained from the sampling sites for the 5–15 cm depth mineral layer.

As the BD measurements were resolved separately from the C measurements, concentration values for C were considered for the modeling instead of carbon amounts with respect to masses. As wetland regions contained the highest C concentrations but also the lowest BD, it was reasoned that modeling C concentration instead of a carbon pool was more logical for differentiating better various land cover types. Nonetheless, once models were fitted for BD and C, a prediction map for carbon pool was generated by using the BD and C predictions.

5.2.3. Environmental Covariates

A variety of environmental covariates were derived from different remote sensing products. This included surface reflectance from Landsat imagery, point cloud LiDAR for the generation of rasterized covariates, SAR (synthetic aperture radar) data, aeromagnetic retrievals and tree species information mapping products. The full listing of environmental

covariates, corresponding to the full predictor set, has been listed in Table B.2 in Appendix B.

The multispectral imagery corresponded to Tier 1 surface reflectance provided from the United States Geological Survey (USGS) that was downloaded via Google Earth Engine. This data was obtained from Landsat 5 imagery for years before 2012, and from Landsat 8 imagery for principally the reference year of 2017. Data for all multispectral surface reflectance bands for each of Landsat 5 and Landsat 8 were acquired. This surface reflectance data was compiled from median pixel intensities based on image scenes with less than 2% cloud cover for over the defined seasons. These seasons included summer (June – August) for the time of peak vegetation, and winter (December – February) for vegetation dormancy. Imagery was also compiled over 5 years (2015 – 2019) for spring (May) and autumn (September 10^{th} – October 10^{th}) so to obtain enough cloud-free images for the median pixel computations. The Landsat 5 data was attained for change detection covariates, where only imagery corresponding to peak vegetation for summer (June – August) was utilized for it. Normalized difference indices were also calculated from the corresponding surface reflectance bands. These included the normalized vegetation index (NDVI), normalized difference water index (NDWI), a modified normalized difference water index (MNDWI) and a normalized difference built-up index (NDBI) for each defined season for 2017 from Landsat 8 imagery.

Point cloud data for LiDAR was obtained from the Ontario Ministry of Natural Resources (MNR). This data was retrieved by aerial surveys completed during 2016 and 2017 (Airborne Imaging 2018). From this point cloud data, rasterized covariates for DEM, CHM and a gap

fraction were generated for pixels of 30 m spatial resolution. More details concerning the generation of these LiDAR-derived covariates was discoursed in Chapter 6. In particular, the DEM was derived from the lowest retrieval returns, and a DSM was constructed from the highest retrieval returns, with respect to recorded elevations from the LiDAR retrievals. From those covariates, the CHM was calculated as the difference of the DEM from the DSM. The gap fraction was computed as the proportion of retrievals with only one return, with regard to all retrievals. Multiple topographic covariates were then generated from the DEM using SAGA GIS software (Conrad et al. 2015), which corresponded to attributes for morphometry, hydrology, visibility and channel structure.

SAR data for both C-band and L-band were acquired from Google Earth Engine. The C-band SAR was collected by Sentinel-1 from the European Space Agency (ESA) and compiled as an average for May 2017, corresponding to maximal ground moisture due to spring runoff arising from the annual snow melt. Yearly averages from L-band data for 2017 were obtained from the PALSAR (Phased Array type L-band Synthetic Aperture Radar) (Shimada et al. 2014) sensor maintained by the Japan Aerospace Exploration Agency (JAXA). The L-band SAR was attained for HH (horizontal transmit, horizontal receiver) and HV (horizontal transmit, vertical receive) polarizations, with C-band SAR obtained for VV (vertical transmit, vertical receive) and VH (vertical transmit, horizontal receive) polarizations, respectively. Geophysical aeromagnetic survey data was also acquired from Natural Resources Canada to obtain covariates relating to underlying parent material. This data included gravity anomaly as well as its first vertical derivative as resolved for 2016 (Jobin, Véronneau, and Miles 2017) and a magnetic residual of total field and its first vertical derivative for November 2018.

Additional geospatial data products collected included vegetation attributes for tree species and biomass that were obtained from the National Forestry Inventory (NFI) of Canada. These data products corresponded to concentration percentage maps for tree species prediction of black spruce and balsam fir, as well as estimates regarding total volume, total dead biomass and stand age for the vegetation. These prediction maps acquired from the NFI were generated from non-parametric k-nearest neighbors (k-NN) models for the reference year of 2011, based on MODIS (Moderate Resolution Imaging Spectroradiometer) surface reflectance as well as climatic and topographic information (Beaudoin et al. 2017, 2018). Indicator covariates were also created from RFs trained on multisource remote sensing data for the YU sampling locations. These indicators conformed to prediction maps regarding locations with no trees as well as with evergreen conifers for the study region.

5.3. Theory and Calculations

A framework was designed to facilitate digital soil mapping for SEM, depicted in Figure 5.3. After the compilation of geospatial data and subsequent generation of environmental covariates, a feature selection was enacted to attain the most important predictors. Pairwise correlations were then calculated for these predictors, from which the modeling process ensued. The SEM was implemented via a sequential stepwise process, which was succeeded by a path analysis to resolve contributions of predictors with targeted soil properties. Afterwards, prediction maps of the targeted soil properties were generated for each study area. Processes corresponding to these steps were explicated in further detail in this section.

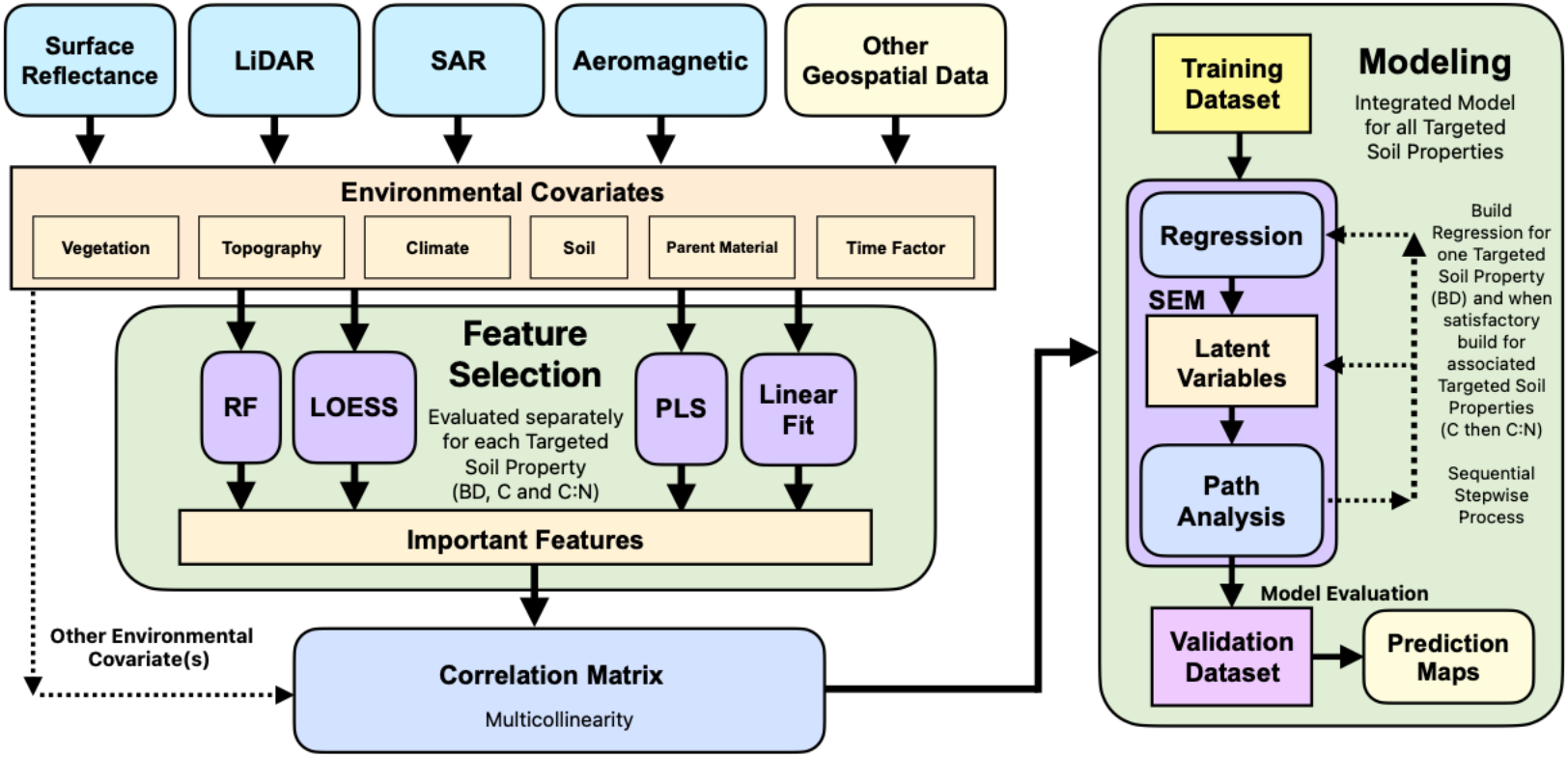

Figure 5.3. Framework adopted for the facilitation of structural equation modeling (SEM).

5.3.1. Features

The features corresponded to environmental covariates that conformed to soil formation factors, here designated by the *scorp* or *scorpan* acronyms (McBratney, Mendonça Santos, and Minasny 2003; Mulder et al. 2011). It was anticipated that the major drivers impacting soil properties for this study region corresponded to the soil formation factors of vegetation (i.e. organisms; *o*) and topography (i.e. relief; *r*), as was encountered for other studies for the boreal context (Beguin et al. 2017; Minasny et al. 2019). Due to interconnections between the soil properties of BD, C and C:N, soil (*s*) as a soil formation factor was also considered. In addition, features relating to other soil formation factors deemed of lesser importance for this study context were also included. These involved attributes relating to the bedrock geology such as aeromagnetic retrievals for parent material (*p*), as well as a change detection with surface reflectance that conformed to a time (*n*) factor. As climate for this study region was relatively homogenous throughout the domain, climatic (*c*) soil formation factors pertaining to temperature, precipitation or other weather data were omitted.

The spatial resolution for most of the environmental covariates as collected or generated was 30 m; if required the remaining environmental covariates were resampled to a spatial resolution of 30 m. Reprojection of all environmental covariates to the NAD (North American Datum) 1983 Lambert conformal conic coordinate system was implemented, corresponding to coordinates in eastings [m] and northings [m]. All covariate layers were co-registered so that the corresponding pixels matched to the same grid lines for reference. Moreover, anomaly detection in terms of histograms and standard *z*-scores from a normal distribution (Mohseni et al. 2023) were checked for each of the environmental covariates

used as features. The physical feasibilities of values were examined for each feature, with the removal of any feature corresponding to missing or incomplete data for any soil sampling site location.

5.3.2. Feature Selection

A multisource compilation of geospatial data can yield a larger listing of environmental covariates that can correspond to multiple facets of soil formation factors. However, this can also result in high redundancies between various environmental covariates when utilized as predictors. To reduce the number of features from high dimensional data sets (Dai, Ding, and Wang 2016), a feature selection should be employed as multicollinearity between predictors can adversely impact modeling results (Gokmen, Dagalp, and Kilickaplan 2022; Zhao, Xu, and Wang 2017). Degradation of modeling with respect to multicollinearity is an issue for SEM (Grewal, Cote, and Baumgartner 2004), therefore feature selection is an important step for the preparation of predictor data for SEM.

Data-driven as well as knowledge-driven methods for selecting environmental covariates have both been incorporated in the framework in Figure 5.3. The feature selection process here has utilized data-driven methods, whereas the SEM process in part has employed knowledge-driven methods for the finalized selection of environmental covariates used as predictors. As a first step, data-driven methods as with machine learning approaches (Brungard et al. 2015) can be applied to narrow down the listing of environmental covariates so to retain only the most relevant for modeling purposes. However, knowledge-driven methods (Laouici et al. 2025) can integrate domain subject matter expertise for selecting the

most suitable and of the importantly ranked environmental covariates, to ultimately implement for the modeling work leading to predictions. For SEM (Rappaport et al. 2020), knowledge-driven methods can lead to more useful insights gained by the modeling results as the environments covariates applied as predictors better relate to physical relevance as considered from subject matter judgement.

Multiple methods were implemented for ascertaining the most important predictors regarding feature selection. Variable importances using the full set of all predictors were separately assessed by different methods, which afterwards were weighted together to obtain a reduced set of predictors, with this process enacted for each targeted soil property. To gauge variable importance, modeling approaches less affected by multicollinearity (Arashi, Lukman, and Algamal 2022; Chong and Jun 2005; Firinguetti, Kibria, and Araya 2017; Kuhn 2008) were utilized. These consisted of RF and partial least squares (PLS). Variable importance for the RF was evaluated by mean decrease in Gini impurity, corresponding to the decrease of impurity over all nodes of the decisions trees comprising the RF (Loecher 2022; Sage, Genschel, and Nettleton 2020). Within each decision tree, nodes were split so to maximize homogeneity within each leaf for attaining variance reduction (Sage, Genschel, and Nettleton 2020) regarding regression. For PLS, variable importance for regression was assessed from the reduction in the sum of squares in reference to the regression coefficients (Chong and Jun 2005; Kuhn 2007). In addition, model-independent approaches for gauging variable importance were also considered. This included assessing variable importance by simple linear regressions via *t*-test statistics on the slope pertaining to each variable (Atkinson and Riani 2002; Kuhn 2007). In addition, LOESS (locally estimated scatterplot smoothing) (Irizarry 2001; Jacoby 2000) evaluated variable importance from the coefficient of

determination (R^2) for fittings between the target variable and each predictor variable (Kuhn 2007). The caret package (Kuhn 2008; Kuhn et al. 2020) in R was used for resolving the variable importances for each of these approaches.

Once all variables were ranked by each method, only the top ranked variables were kept. Regarding flexible constraint considerations (Niemand and Mai 2018), it was decided that no more than the number of predictors equaled to the floor of the square root of the number of training sites were retained. Here, this corresponded to selecting only the top 9 predictors from the variable importance rankings. These top predictors were then each assigned an integer score from 9 to 1, with higher score matching to higher variable importance ranking. This was done for each variable ranking method, from which these integer scores were summed together so to ascertain the top variables based on the aggregated scores from all methods. An advantage of this ensemble approach for variable ranking was that it mitigated bias, as variables that were not ranked highly in a couple of methods were still contemplated for the reduced sets of the most important predictors. This variable importance procedure was implemented separately for each of BD, C and C:N.

To address multicollinearity, pairwise correlations in terms of the Pearson correlation coefficient were calculated between predictors within the reduced sets of features. In general, when considering these correlations any predictors pairs with correlations greater than 0.5 were avoided (Dormann et al. 2013). In the cases with higher pairwise correlations, then only one of those predictors were retained. This rule was enacted with the intention of only retaining the predictor conforming to greater physical relevance regarding domain knowledge for soil mapping. Nonetheless, higher pairwise correlations were tolerated when

predictors were based on different sensors or corresponded to different soil formation factors or were utilized for direct modeling with other targeted soil properties. Latent variables were also considered if multiple predictors conforming to the same attribute of a soil formation factor were each selected for the reduced listing of the most important variables, with low absolute values of pairwise correlations for these predictors.

5.3.3. Structural Equation Modeling

SEM can link integrated regression modeling of target variables with respect to predictors, while also evaluating for conjectured casual relations. Regarding application to digital soil mapping, SEM can be implemented to unravel insights into drivers of soil properties. Specifically, SEM can tie connections between soil properties and environmental covariates conforming to soil formation factors (McBratney, Mendonça Santos, and Minasny 2003; Minasny et al. 2019). An associated path analysis can resolve casual connections between the predictor variables and targets without establishing directions of causality, with those connections being visualized by a path diagram (Rappaport, Amstadter, and Neale 2020).

The parameters of SEM are optimized generally by maximum likelihood estimation (Rosseel 2012). A sample variance-covariance matrix constitutes covariances among the variables (Lin 2021), and uncertainties for the target variables can be included via variance formulations into SEM. Reasonability of fit for SEM is assessed by a chi-squared test on the user model (Hu and Bentler 1999) between model covariance and observed sampled covariance (Beribisky and Cribbie 2024; Lin 2021). Regarding goodness of fit, the intent for SEM is that the assessment for this chi-squared should be non-significant; this means that its

p-value should be greater than 0.05, inferring that the null hypothesis corresponding to a perfect model fit cannot be rejected (Beribisky and Cribbie 2024).

The main goodness of fit indices reported were the comparative fit index (CFI), root mean square error of approximation (RMSEA) and standardized root mean square residual (SRMR). The CFI related the hypothesized model fit to that of a null model of the same observed variables but with covariances of zero for between variables (Beribisky and Cribbie 2024). In terms of CFI, a good model fit was indicated by CFI > 0.95 (Hu and Bentler 1999). RMSEA conformed to the misfit of the model variance-covariance matrix from that of the sample variance-covariance matrix regarding weighted sums of squared deviations (Beribisky and Cribbie 2024). For RMSEA a good model fit was deemed by RMSEA < 0.06 (Hu and Bentler 1999) or RMSEA < 0.08 (Byrne and van de Vijver 2010). The SRMR corresponded to the average of standardized residuals between the respective model and observed variance-covariance matrices (Chen 2007), with a good model fit indicated by SRMR < 0.08 (Hu and Bentler 1999).

The regression modeling was performed first for BD, then for C followed by C:N, respectively. Predictors for these models corresponded to the most important features as identified by the feature selection process. Modeling was accomplished sequentially, starting with the highest ranked predictor, and then updating the models to include other predictors from the corresponding reduced set. Here, BD was a predictor for C, with C a predictor for C:N, accordingly. Standard errors for each of BD, C and C:N were also implemented into the SEM. A *z*-score from the Wald test (Rosseel 2012) assessed the significance of a predictor for SEM, with a predictor only kept within the model if its *p*-value was 0.05 or less. For

SEM, maximum likelihood estimation by nonlinear minimization with box-constrained optimization (Rosseel 2012) was adopted to obtain a solution. The lavaan package (Rosseel 2012) in R was employed for SEM.

Concerning data splitting, the data was randomly split 80:20 between training and validation sets. The root mean square error (RMSE) and R^2 were the metrics evaluated for modeling accuracy, based upon assessment with the validation set. Prediction maps of BD, C and C:N for the domains of the study areas were then generated from the finalized integrated model.

5.3.4. Random Forest Models

For comparison purposes, RF models were trained for each of BD, C and C:N so to compare with results attained by SEM. Two RFs were trained for each of these targeted soil properties, with one RF trained on same predictors as used for the final SEM, and the other RF trained on the respective sets for top ranked predictors. These RFs were trained on the same set of training data allocated for the SEM, and subsequently evaluated on the same validation set as that for SEM. Regarding specifications for the RFs, the number of trees (ntree) were set to 1000 and the mtry parameter corresponding to the number of predictors randomly selected at each node was varied from integer values of 1 up to 35. The mtry parameter was selected so to optimize training accuracy. It was ranged to include the integer value of the square root of the number of predictors in the original set prior to feature selection (Blackford, Heung, and Webster 2022), upwards to at least one third of the number of the predictors in the original set. Cross validation (CV) with 10-folds and 10 repeats was performed for training the RFs. For training and assessing the RFs, the caret package (Kuhn 2008) in R was adopted.

Once models and corresponding prediction maps were generated for BD and C, a carbon pool (CP) using BD and C concentrations (Jurgensen et al. 2017) was calculated. As C and BD quantities were unknown at deeper soil horizon depths, the CP was computed as the mass of carbon per cubic centimeter for the 5–15 cm mineral soil depth. The calculation for CP was

$$Carbon\ Pool\ \left[\frac{\mathrm{g}}{\mathrm{cm}^3}\right] = Bulk\ Density\ \left[\frac{\mathrm{g}}{\mathrm{cm}^3}\right] \times Total\ Carbon\ [\%]. \qquad (5.3)$$

These values for CP were computed per pixel, and subsequent prediction maps for CP were generated.

5.4. Results

The feature selection of the most important variables was resolved prior to the SEM step. Results of the feature selection for the targeted soil properties were presented in Tables 5.1, 5.2 and 5.3 for BD, C and C:N, respectively. Note that the higher the final score count for a predictor, the higher was the assessed variable importance. Inspecting Table 5.1 for BD, predictors for surface reflectance, CHM and slope attributes had the highest variable importance rankings. In Table 5.2 for C, BD followed by near-infrared surface reflectance for May, plan curvature and terrain topographic covariates were of higher variable rank. For C:N in Table 5.3, C followed by covariates relating to water detection and elevation attained the highest variable importance ranking regarding predictors.

Table 5.1. Feature selection for bulk density (BD) as ascertained by multiapproach variable ranking, with predictors listed in order of rank.

Predictor	Variable Importance – Bulk Density (BD)								Score
	RF		PLS		Linear		LOESS		
	Value	Points	Value	Points	Value	Points	Value	Points	
B5 Summer 2017	0.7831	7	0.0120	9	4.7367	8	0.2134	4	28
CHM	0.8500	9			4.5446	3	0.2433	9	21
B5 May	0.7783	6	0.0088	3	4.4093	1	0.2342	8	18
B3 Winter 2017					4.6738	7	0.2166	6	13
Mid Slope Position	0.7123	5	0.0104	8					13
B5 Winter 2017	0.6248	1			4.7906	9	0.2055	2	12
YU No-Trees Indicator			0.0099	7	4.5885	5			12
B2 Winter 2017					4.5513	4	0.2152	5	9
B4 Winter 2017					4.6426	6	0.2107	3	9
B6 May					4.5127	2	0.2209	7	9
Gap Fraction	0.8252	8							8
Slope Height			0.0099	6					6
B5 Fall	0.6556	3	0.0083	1			0.2006	1	5
MNDWI May			0.0096	5					5
DEM	0.6651	4							4
YU Evergreen Indicator			0.0090	4					4
Change Magnitude B3 B4 B5 1995–2005			0.0087	2					2
Slope	0.6278	2							2

Table 5.2. Feature selection for total carbon (C) as ascertained by multiapproach variable ranking, with predictors listed in order of rank.

Predictor	Variable Importance – Total Carbon (C)								Score
	RF		PLS		Linear		LOESS		
	Value	Points	Value	Points	Value	Points	Value	Points	
Bulk Density	3439.6270	9	1.7617	9	15.4295	9	0.8976	9	36
B6 May	428.8530	8			4.4765	8	0.1965	5	21
B5 May	357.7258	5			3.8371	6	0.2046	8	19
B7 May	403.0664	6			4.1258	7	0.1699	3	16
B4 May	352.1136	4					0.2036	7	11
Plan Curvature			0.6176	8					8
MNDWI Fall			0.4286	7					7
Stream Power Index	407.4956	7							7
Change Magnitude B3 B4 B5 1984–2005	290.3521	2	0.3697	4					6
DEM							0.2032	6	6
NFI Picea Mariana			0.4163	6					6
B3 Winter 2017					3.7672	5			5
Magnetic Residual November 2018			0.3779	5					5
MNDWI Winter 2017			0.3681	3			0.1685	2	5
B3 May							0.1813	4	4
B5 Winter 2017					3.7412	4			4
B2 Winter 2017					3.7069	3			3
Valley Depth	340.2267	3							3
B4 Winter 2017					3.6807	2			2
MNDWI May			0.3476	2					2
B2 May							0.1639	1	1
Mid Slope Position			0.3227	1					1
NDBI Winter 2017					3.4281	1			1
Slope	257.6389	1							1

Table 5.3. Feature selection for carbon-to-nitrogen ratio (C:N) as ascertained by multiapproach variable ranking, with predictors listed in order of rank.

Predictor	Variable Importance – Carbon-to-Nitrogen Ratio (C:N)								Score
	RF		PLS		Linear		LOESS		
	Value	Points	Value	Points	Value	Points	Value	Points	
Total Carbon	217.5566	9	0.3542	9	6.3687	9	0.2852	9	36
SAR C VV May 2017	128.1158	8	0.1619	8	4.6362	8	0.2046	7	31
DEM	93.8209	5	0.1522	6	3.5971	5	0.2154	8	24
NDVI Fall	106.0946	7					0.1927	5	12
Standardized Height			0.1420	5	3.7674	6			11
MRVBF	106.0716	6	0.1373	4					10
B1 Winter 2017	67.0613	1			3.8235	7			8
B4 Fall	75.9494	2					0.2012	6	8
B6 May	81.4080	4			3.5694	4			8
Convergence Index			0.1615	7					7
Gravity Anomaly 2016			0.1231	3	3.5184	2			5
B4 Summer 2017							0.1859	4	4
B11 Summer 2017					3.5260	3			3
B7 May	80.4987	3							3
B7 Summer 2017							0.1799	3	3
Change Magnitude B3 B4 B5 1984–2005			0.1231	2					2
NDBI Fall							0.1773	2	2
B6 Fall					3.4320	1			1
NDWI Fall							0.1753	1	1
SAGA TWI			0.1143	1					1

After the feature selection process, pairwise correlations were calculated between all predictors within the reduced sets pertaining to the most important features. These predictors were further streamlined by the modeling process with SEM. Pairwise correlations of the environmental covariates harnessed as predictors for the finalized SEM were presented in Figure 5.4. Inspecting Figure 5.4, for the most part the absolute values of the pairwise correlation coefficients did not exceed 0.50. Nevertheless, there was a pairwise correlation of 0.76 between DEM and SAGA TWI, and −0.71 between the slope and SAGA TWI. However, it should be mentioned that these predictors were directly applied for modeling different soil properties within the integrated SEM, with slope and DEM both predictors for C and SAGA TWI a predictor for C:N. As well for noting purposes, slope was an input for the generation of SAGA TWI, and slope was derived from a DEM (Kopecký, Macek, and Wild 2021). From Table 5.4, there was also a moderate pairwise negative correlation for slope and MRVBF, and a moderate pairwise positive correlation for mid slope position and slope height. Apart from that, absolute values of other pairwise correlations for the reduced set of predictors were less than 0.5, with most of these values less than 0.3.

The regression results of the finalized SEM were depicted in Table 5.5. These results listed the predictors utilized for modeling each targeted soil property, as provided in the order for the SEM process. Variance terms for the targeted soil properties were presented at the bottom of Table 5.4. Estimates for the regression coefficients, along with standard errors, *z*-values and *p*-values were noted. Regarding significance levels, asterisks were indicated, with one asterisk corresponding to a *p*-value of at least < 0.05, two asterisks for < 0.01, and three asterisks for < 0.001, accordingly. Looking at Table 5.4, all the predictors retained were significant as all *p*-values were less than 0.05. Furthermore, goodness of fit indices were

evaluated for the SEM, presented in Table 5.5. All evaluations for these indices corresponded to a good fit, including for the RMSEA metric (Byrne and van de Vijver 2010).

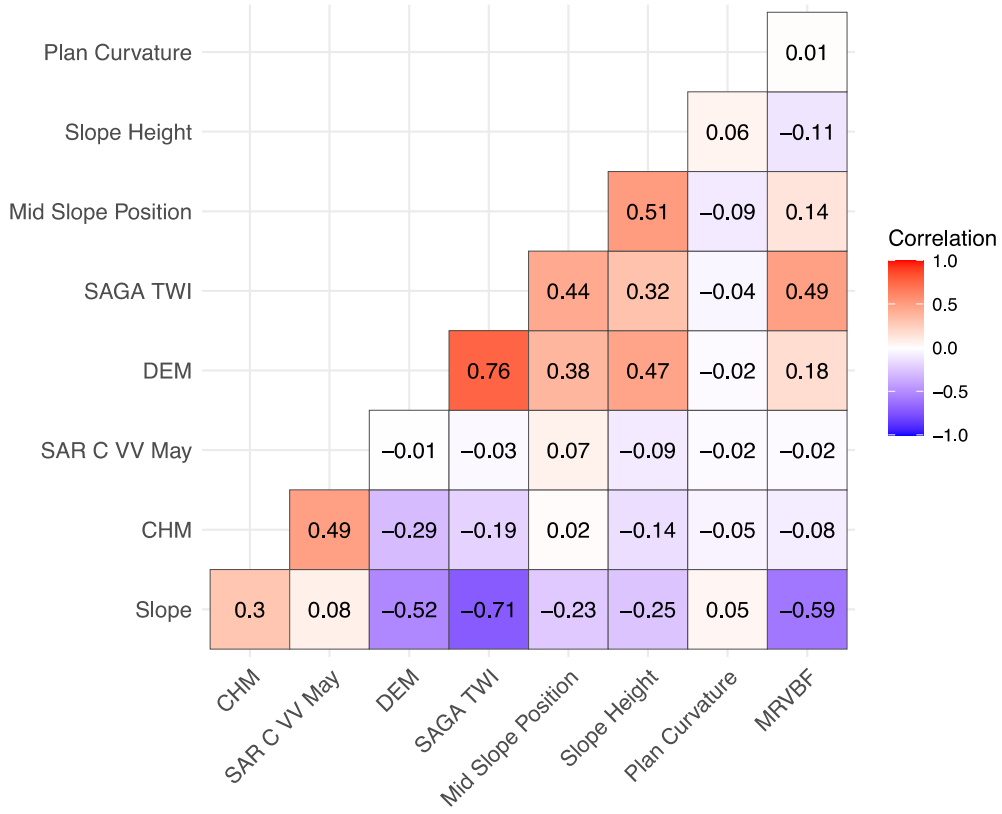

Figure 5.4. Pairwise correlations between environmental covariates as retained in the structural equation modeling (SEM).

Table 5.4. Predictors (top) and variance terms (bottom) for the finalized structural equation modeling (SEM) regression, with estimates and *p*-values reported.

Regression			**Estimate**	**Standard Error**	***z*-Value**	***p*-Value**
Bulk Density	BD	[g/cm^3]				
Canopy Height Model	CHM	[m]	−0.019	0.006	−2.906	**0.004****
Slope		[°]	−0.021	0.010	−2.026	**0.043***
Mid Slope Position		[m]	−0.948	0.295	−3.216	**0.001****
Slope Height		[m]	0.045	0.012	3.740	**0.000*****
Total Carbon	C	[%]				
Bulk Density	BD	[g/cm^3]	−21.357	1.096	−19.489	**0.000*****
Digital Elevation Model	DEM	[%]	0.159	0.027	5.826	**0.000*****
Plan Curvature		[1/100 z-unit]	−4.884	1.762	−2.772	**0.006****
Carbon-to-Nitrogen Ratio	C:N					
Total Carbon	C	[%]	0.232	0.032	7.325	**0.000*****
SAR C VV May 2017			0.578	0.202	2.857	**0.004****
Multi-Resolution Valley Bottom Flatness	MRVBF		−0.413	0.136	−3.029	**0.002****
SAGA Topographic Wetness Index	SAGA TWI		0.833	0.177	4.707	**0.000*****
Variances						
Bulk Density			0.202	0.027	> 7.616	**0.000*****
Total Carbon			36.033	4.731	> 7.616	**0.000*****
Carbon-to-Nitrogen Ratio			17.939	2.355	> 7.616	**0.000*****

Table 5.5. Goodness of fit indices evaluated on the finalized structural equation modeling (SEM).

SEM Metric	Value	Recommended Range
Model Test User Model *p*-value (Chi-squared)	**0.099**	> 0.05 (Beribisky and Cribbie 2024)
Comparative Fit Index (CFI)	**0.969**	> 0.95 (Hu and Bentler 1999)
Root Mean Square Error of Approximation (RMSEA)	**0.061**	< 0.06 (Hu and Bentler 1999), < 0.08 (Byrne and van de Vijver 2010)
Standardized Root Mean Square Residual (SRMR)	**0.037**	< 0.08 (Hu and Bentler 1999)

A path diagram for the resulting path analysis with the SEM is shown in Figure 5.5. In this diagram, environmental covariates are listed along the left side, with the targeted soil properties along the right. The standardized parameter estimates are noted for the paths, and different coloring has been used to denote positive versus negative contributions. For the targeted soil properties, the contributions from variances have also been indicated.

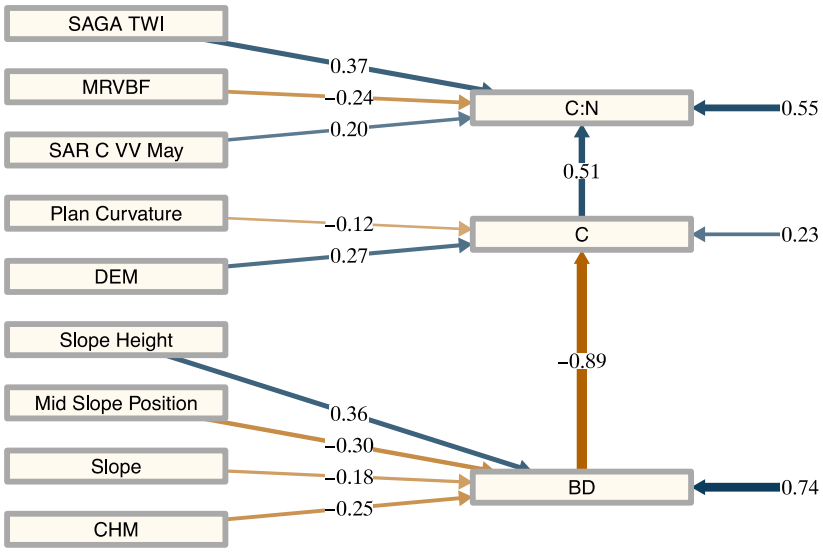

Figure 5.5. Path diagram depicting the path analysis for the finalized structural equation modeling (SEM), with standardized parameter estimates noted.

Looking at the path diagram in Figure 5.5, for BD each of CHM, slope and mid slope position had reversed interactions. For the slopes, this meant that flatter topography corresponded to higher BD, and lower CHM also conformed to higher BD. This finding corroborated to farmland, as the croplands were sufficiently flat and conformed to compacted soil and hence high BD values. Moreover, the forested regions tended to have more variance with topography, and lower relative BD. The highest standardized parameter estimate for BD was from its variance term. As for C, its highest contribution came from BD, with a reverse relation between BD and C. For this study area, C was lowest for the cropland localities which correspondingly had the highest BD, whereas the wetlands had the highest C but lowest BD. DEM had the next highest contribution for C, with a positive interaction indicated; for this region the forested tracts tended to coincide with higher elevations than that for the cropland areas. Finally, for C:N the largest contributions came from its variance term, followed by C and SAGA TWI. For C:N after the impact of C, covariates relating to water or moisture had relevance such as for SAGA TWI, C-band SAR and MRVBF.

Modeling accuracies as assessed on the validation set were reported in Table 5.6. Accuracies were noted for the SEM, as well as RFs trained on the same reduced set of predictors as the finalized SEM noted in Figure 5.4. In addition, RFs were trained for each targeted soil property from the full set of predictors prior to feature selection. To finetune the RFs when training so to optimize accuracy, the mtry parameter was ranged over integer values from 1 up to 35. For the optimized RFs trained on the reduced set of predictors as utilized for SEM, the mtry parameters corresponded to 4, 1 and 4 for the BD, C and C:N RFs, respectively. As for the RFs fitted on the full set of predictors, the mtry parameters for optimized RFs were 4, 2 and 23 for BD, C and C:N, accordingly. Gains in accuracy were attained for SEM when

compared to RFs trained on the full predictor set, with the exception for C:N. Furthermore, accuracy gains were achieved by the RFs trained on the same reduced set of features that were ascertained as important by the framework in Figure 5.3. Consequently, the processes for feature selection and SEM in this framework also helped to improve model accuracies achieved for RFs. Attained accuracies regarding R^2 were around 0.3 for BD and 0.4 for C when assessed on the validation data, which was an improvement from R^2 of approximately 0.2 for the RF fitted upon the full set of predictors. However, when considering Fisher's *z*-test for statistical inference using equation (4.1) with *n* equal to 28 sites, none of the R^2 values between the models were statistically significance from one another for the comparisons that were evaluated.

Prediction maps were generated for the targeted soil properties over the domains of the study areas. These prediction maps corresponded to BD in Figure 5.6, C in Figure 5.7, CP in Figure 5.8, and C:N in Figure 5.9, respectively. Results of prediction from the SEM as well as RFs trained on the same set of predictors as for the SEM were presented. With these prediction maps, soil properties varied depending on land cover type. In Figure 5.6, the greatest BD quantities were predicted for the cropland, with the lowest BD for the wetland areas. Conversely, in Figure 5.7 the greatest C was for the wetlands, with the least C for the cropland fields. From the soil sample data, BD for croplands tended from 1.5 to 2.5 g/cm^3, with BD for wetlands usually between 0.3 to 0.7 g/cm^3. Measurements for C typically corresponded to less than 50 g/kg for croplands, but wetlands contained the greatest contents of 200 g/kg or more. Therefore, the predicted values of BD and C conformed to expectations. These predictions were comparable to those previously reported in boreal ecozones (Laganière et al. 2013; Mansuy et al. 2014). By inspecting Figure 5.8 for CP, as similar to Figure 5.7 for C

the localities with the greatest carbon amounts corresponded to the forested and wetland regions, with agricultural areas containing the least amounts of carbon. For C:N, in Figure 5.9 the highest C:N was for the tracts of primary forest, with the lowest C:N for the croplands. Predictions for C:N corresponded to values between 5 to 10 for the cropland fields versus 20 to 40 for mature expanses of forest. Moreover, these predictions were comparable to what has been reported within the boreal shield of Canada (Mansuy et al. 2014) as well as for a different study in northern Ontario (Beguin et al. 2017).

Table 5.6. Modeling accuracies for the models as evaluated on the validation set. Results of statistical testing for differences with R^2 accuracies as with between corresponding SEM models are noted.

Soil Property			**SEM**		**RF SEM Predictor Set**		**RF Full Predictor Set**	
			RMSE	**R^2**	**RMSE**	**R^2**	**RMSE**	**R^2**
Bulk Density	(BD)	[g/cm^3]	0.45	0.33	0.45	0.31	0.50	0.15
Total Carbon	(C)	[%]	8.86	0.38	8.54	0.44	9.78	0.20
Carbon-to-Nitrogen	(C:N)		4.81	0.16	4.67	0.20	4.67	0.21
Significance Test – SEM Comparison					***z*-Value**	***p*-Value**	***z*-Value**	***p*-Value**
Bulk Density	(BD)				0.09	0.93	0.87	0.39
Total Carbon	(C)				−0.28	0.78	0.84	0.40
Carbon-to-Nitrogen	(C:N)				−0.20	0.84	−0.25	0.80

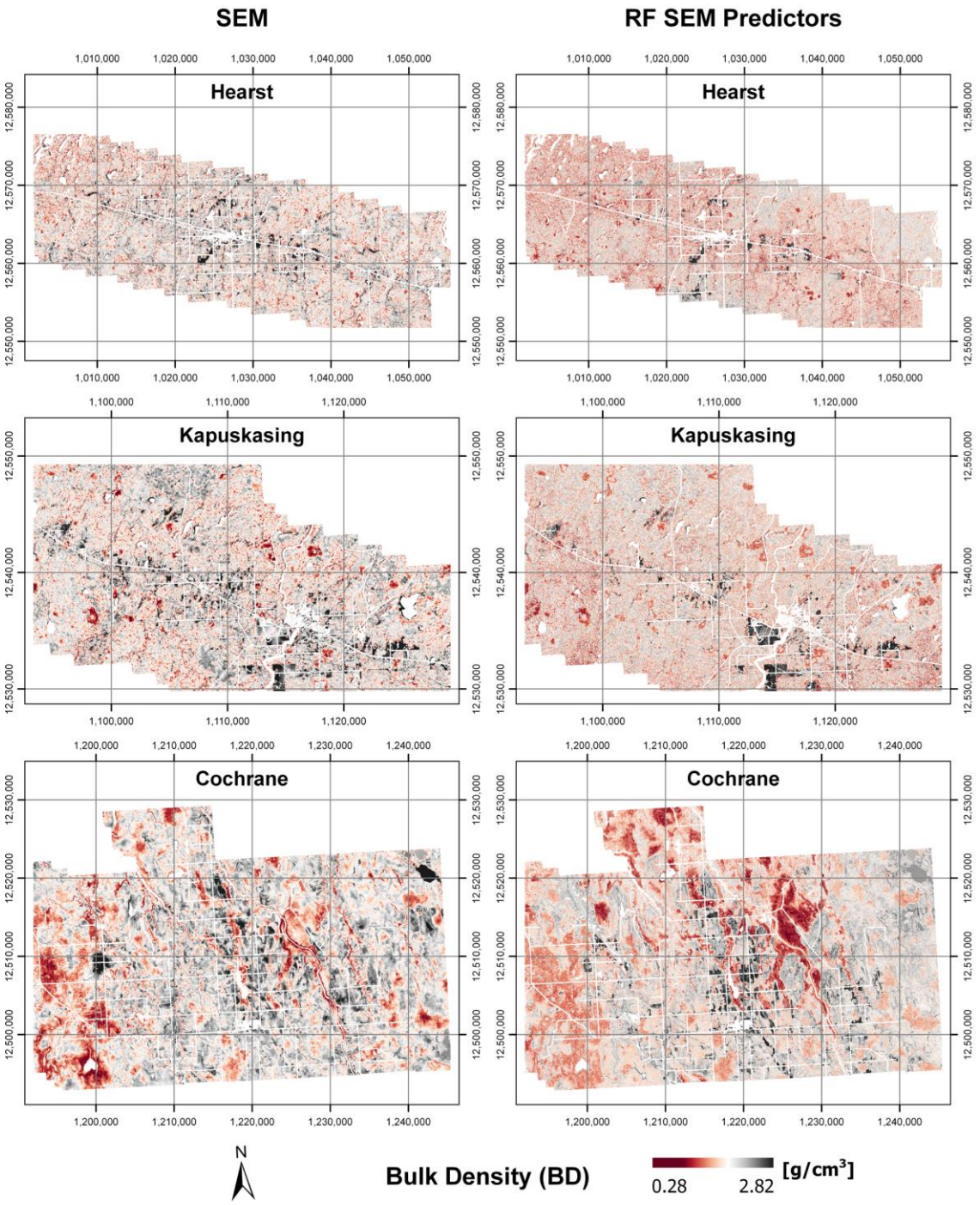

Figure 5.6. Bulk density (BD) prediction maps for structural equation modeling (SEM) (left) and random forests (RFs) trained using the same predictor set as for the finalized SEM (right). Coordinate grids are in the NAD 1983 Lambert conformal conic projection with eastings [m] and northings [m].

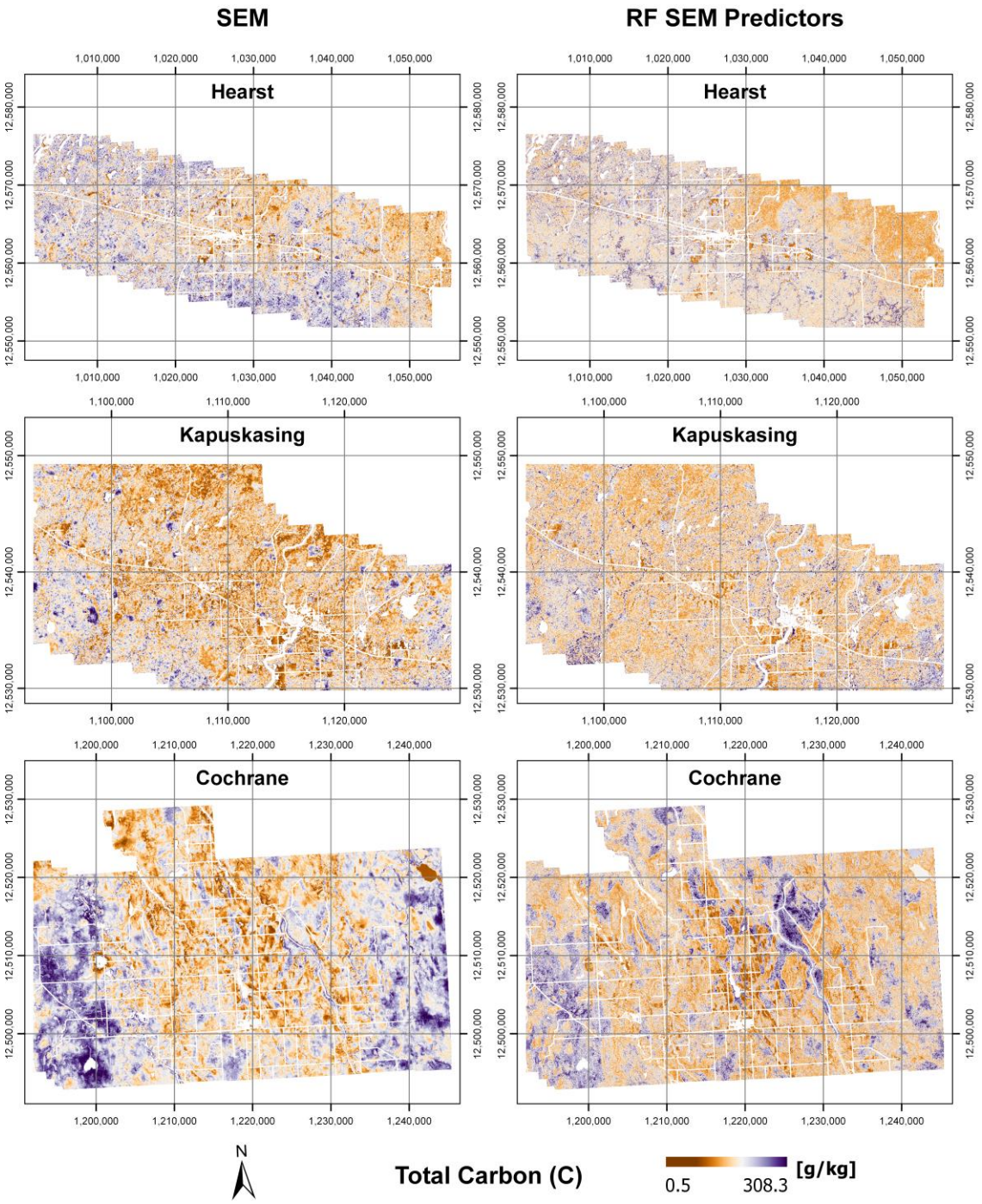

Figure 5.7. Total carbon (C) prediction maps for structural equation modeling (SEM) (left) and random forests (RFs) trained using the same predictor set as for the finalized SEM (right). Coordinate grids are in the NAD 1983 Lambert conformal conic projection with eastings [m] and northings [m].

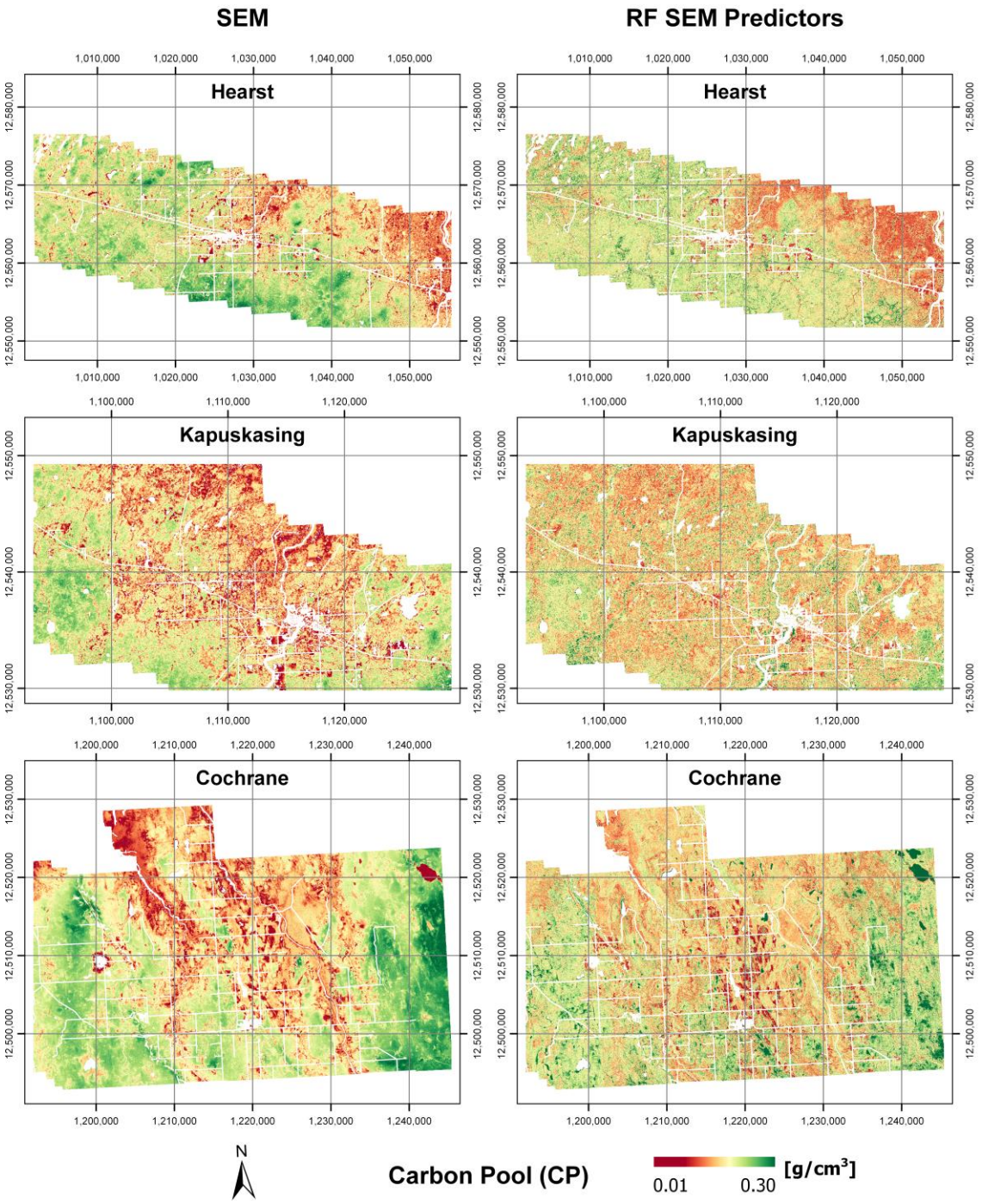

Figure 5.8. Carbon pool (CP) prediction maps for structural equation modeling (SEM) (left) and random forests (RFs) trained using the same predictor set as for the finalized SEM (right). Coordinate grids are in the NAD 1983 Lambert conformal conic projection with eastings [m] and northings [m].

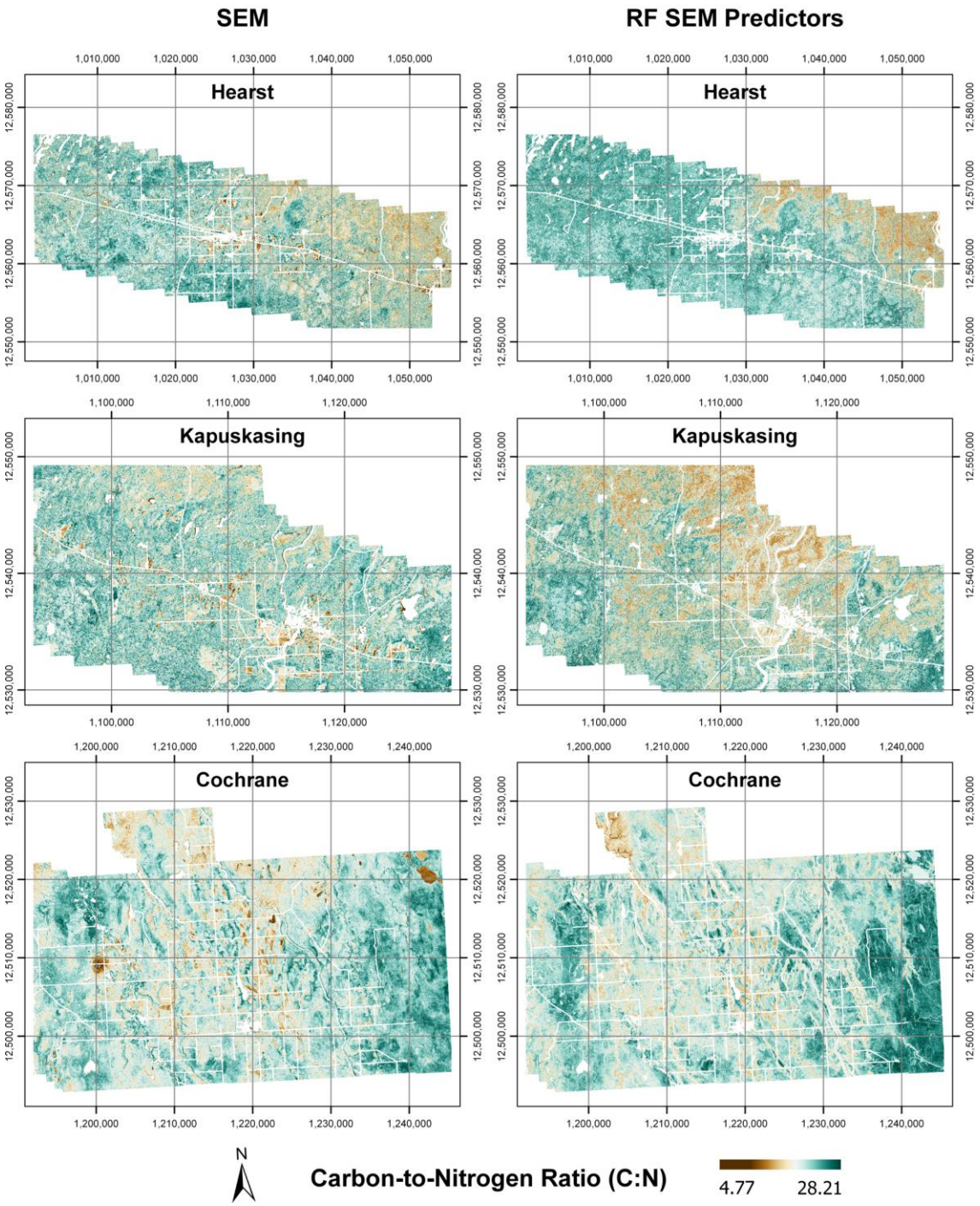

Figure 5.9. Carbon-to-nitrogen ratio (C:N) prediction maps for structural equation modeling (SEM) (left) and random forests (RFs) trained using the same predictor set as for the finalized SEM (right). Coordinate grids are in the NAD 1983 Lambert conformal conic projection with eastings [m] and northings [m].

5.5. Discussion

The framework presented in Figure 5.3 advanced the feasibility of SEM for digital mapping. Multiple approaches of resolving variable importance assisted with alleviating bias for identifying the most important predictors. Nonetheless, impartial ranking of individual predictors was difficult due to implicit interactions between predictors (Cheng and Bai 2019; Zhang et al. 2020). This was where SEM yielded support for the inclusion of predictors from the assessment of statistical significance for a predictor within a model. Individually, a predictor might not had attained high variable importance, but nonetheless in combination with another predictor was of consequence for modeling a target variable. For that reason, flexibility was enacted concerning the inclusion of other predictors as well as latent variables. This adaptability was instituted into the framework design to systematize the process of feature selection while also permitting heuristic considerations regarding predictors.

Latent variables formulations were also investigated for the SEM in Figure 5.3 regarding predictors for modeling. Specifically, latent variables conforming to physical attributes within soil formation factors were considered, when the feature selection ranked more than one predictor corresponding to the same physical attribute with high importance. Latent variables were tried for change detection, surface reflectance, terrain morphometry and topographic visibility, respectively. However, when each of these latent variables were included the statistical significances of other predictors were adversely impacted, as some predictors ranked with high variable importance were no longer statistically significance for the SEM. Furthermore, the goodness of fit indices as noted in Table 5.5 were also negatively impacted. If the inclusion of multiple latent variables at once within the SEM were

implemented, then the algorithm for optimizing the SEM did not converge for a solution. Therefore, for attaining acceptable goodness of fit indices and retaining only statistically significance predictors within the SEM, no latent variables were included. Nevertheless, due to the lack of latent variables within the SEM the comprehension was more straightforward regarding the contribution of specific predictors with the associated path analysis.

Improving modeling accuracies for digital soil mapping is an active area of research, with accuracies depending largely due to context of the study region. For the boreal forest in Canada, a separate study (Beguin et al. 2017) using various approaches has attained modeling accuracies comparable to those that have been reported in Table 5.6 for BD, mineral C and organic C:N. Accuracy plots (Beguin et al. 2017) have indicated validation R^2 of 0.02 to around 0.40 for BD, 0.01 to 0.14 for mineral C, and 0.15 to 0.50 for organic C:N, accordingly. Nonetheless, when spatial coordinate information has been excluded from the models then the modeling accuracy has decreased. Correspondingly for that case, R^2 is between 0.12 to 0.27 for BD, less than 0.1 for mineral C, and between 0.25 to 0.35 for organic C:N. Another study (Mansuy et al. 2014) using k-NN models has attained R^2 of 0.175 for mineral BD, 0.199 for mineral organic C, and 0.056 for C:N, respectively. Thus, the modeling accuracies that have been achieved by SEM are similar to what have been reported in other studies yet providing insights into unraveling potential causal connections between environmental covariates and soil properties.

5.6. Summary

- Framework was devised to facilitate SEM for digital soil mapping, by effectively streamlining predictors with a feature selection and accounting for high pairwise correlations to mitigate multicollinearity.
- Multiapproach for feature selection reduced bias in gauging variable importance, as it permitted features not evaluated with top variable rankings for all methods to be considered if the feature was top ranked for at least one approach.
- Calculating pairwise correlations between predictors helped to mitigate multicollinearity by avoiding pairing of predictors with high correlations.
- Modeling accuracies attained by SEM exceeded those from an RF trained on the full set of predictors, except for C:N. Moreover, RFs trained on the reduced set of features as utilized for the finalized SEM also achieved improved accuracies.
- Path analysis regarding SEM discerned contributions to targeted soil properties from predictors, supporting the resolving of drivers for soil formation factors.
- Slope attributes and CHM were of the highest importance for BD. BD along with plan curvature and DEM were of the most significance for C, and C along with environment covariates relating to water detection were of importance for C:N.
- Goodness of fit indices should be considered when assessing SEM results. These metrics should include CFI, RMSEA and SRMR.
- Cropland areas had the highest predicted BD, but lowest predicted C and C:N. Comparatively, wetland areas had the lowest BD and highest C, and highest C:N corresponded to stands of mature forest.

Chapter 6

Constructing Rasterized Covariates from LiDAR Point Data using Structured Query Language

Research presented in this chapter has been published in the conference paper:

[4] Pittman, R. and B. Hu, 2024. "Constructing Rasterized Covariates from LiDAR Point Cloud Data via Structured Query Language." *Proceedings* 110(1): 1–8. https://doi.org/10.3390/proceedings2024110001.

[4] I want to thank the publisher MDPI and the coauthor who have permitted me to reuse this article in my dissertation.

6.1. Background

Point cloud data is commonly utilized for data storage of 3-D point collection for geospatial applications, especially for retrievals of LiDAR returns (Garcia et al. 2017). Recently, LiDAR data has been retrieved over more areas, and intensities of point density and return counts have also been increasing as technology progresses. Depending on the size of study area being investigated, the processing of LiDAR data can become an overwhelming task, particularly if lacking access to proprietary software that can support the transformation of geospatial point cloud data. Lately within digital soil mapping and tree species classification, wall-to-wall coverage for environmental covariate data throughout the domains of study areas are necessary for the generation of prediction maps. For these applications, study areas can easily exceed hundreds, if not thousands of square kilometers (Minasny et al. 2019). Consequently, there is a need for a streamlined process to effectively construct environmental covariates from LiDAR point cloud data.

One of the most functional products that is typically derived from LiDAR point cloud data is a digital terrain model (DTM), as topographic covariates are usually generated from a DTM for terrain analysis (Franklin 2020). Normally, covariates such as a DTM or digital surface model (DSM) are created from continuous surfaces conforming to triangulated irregular network (TIN) meshes (Khosravipour et al. 2014; Southee, Treitz, and Scott 2012) that are constructed from LiDAR point cloud. As covariates corresponding to other sensors or pixelized imagery can be analyzed together with a DTM, rasterized covariate layers are of interest. For these rasterized covariate layers, each pixel constitutes one value, with the pixel size matching to cell size and the pixel boundaries corresponding to grids of geospatial

reference. Computations for generating rasterized layers are optimized when sufficient resources are available. Nonetheless, processing can become difficult to achieve without adequate computational resources, especially if calculations for deriving covariates need to be modified or implemented over larger spatial domains.

Structured query language (SQL) can be harnessed to process point cloud data and generate a variety of rasterized covariates. In the context of SQL (Sumathi and Esakkirajan 2007), each point cloud data point can correspond to a record, also known as a row, within a data table. Unique identifiers can be matched to each record so to maintain a structured form for storage within a database and facilitate merging with other data. Queries can be implemented to filter what data is selected, varying in complexity from simple querying to that entailing more sophistication. Moreover, SQL can effectively aggregate data to the refinement of coordinate grids, with data tabulated so that each record comprises as a pixel for a calculated rasterized covariate. These can correspond to basic features such as a DTM and a DSM both generated from LiDAR point cloud data, as well as other covariates relating to the overlying vegetation such as a canopy height model (CHM) and a gap fraction.

6.2. Materials and Methods

6.2.1. Point Cloud Data

LiDAR point cloud data (Airborne Imaging 2018) from Land Information Ontario (LIO) with the Ontario Ministry of Natural Resources (MNR) was obtained. This LiDAR data was collected by aerial survey during 2016 and 2017 for over a large expanse within the District of Cochrane ranging from the communities of Hearst to Cochrane, centered along the Ontario

highway 11 corridor. Regarding data structure, this point cloud data was saved in LAS (LASer) format with each line of data corresponding to one retrieval. For each retrieval point, an easting coordinate (X) [m], northing coordinate (Y) [m], vertical elevation (Z) [m] and return count per retrieval were all recorded. Coordinates for eastings and northings corresponded to the NAD (North American Datum) 1983 UTM (Universal Transverse Mercator) Zone 17 N (North) projection. Additional information recorded per retrieval were information pertaining to the sensor such as LiDAR intensity, scan angle and a time [s]. The structure of LiDAR point cloud data as conserved in LAS format was depicted in Table 6.1. Each LiDAR LAS file conformed to geospatial tiles of 1 km^2 in area with point density of approximately 8 retrievals per m^2 (Airborne Imaging 2018). In terms of file size, each of these LAS tile file corresponded to around 0.5 GB. For the District of Cochrane, 6085 LAS tiles (Airborne Imaging 2018) in total were collected, which were subsequently used for encompassing the domains of separate study areas.

Table 6.1. Structure of LiDAR point cloud data[1] for retrievals as saved within an LAS file.

X	Y	Z		TIME	RETURN_NUM
283,562.88	5,506,000.03	248.42	…	163,377,412.996137	3
283,567.95	5,506,000.06	248.52		163,377,413.057318	1
283,574.36	5,506,000.15	248.58		163,377,413.134531	2
⋮	⋮	⋮		⋮	⋮

[1] Structure of LiDAR_Data base table for SQL queries in Figures 6.2 and 6.3.

An example of point cloud representation of LiDAR data as obtained from a LAS tile file is shown in Figure 6.1. Here, this plotting corresponds to a 20 m by 20 m subset from a LAS

file covering a 1 km by 1 km portion of the Abitibi River Forest (ARF), with different colorings corresponding to the elevation at which LiDAR returns were collected. In NAD 1983 UTM Zone 17 N coordinates, the subset in Figure 6.1 conforms to eastings [m] from 460,000 to 460,020 and northings [m] from 5,463,000 to 5,463,020, accordingly. The total number of LiDAR returns for this subset are 4974 points. Note the appearance of conifer trees of approximately 10 m in height, with the tops of the trees indicated by yellow or light green coloring, the bulk of the tree canopies by green coloring, and then the lower canopies and ground by purple coloring, respectively.

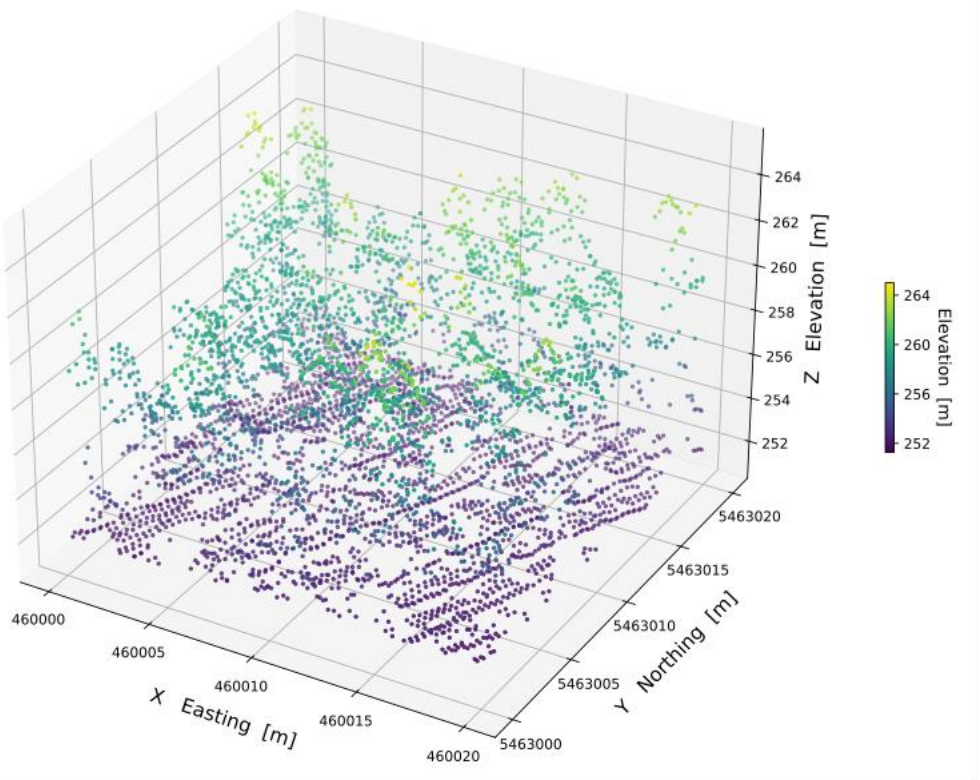

Figure 6.1. Representation of LiDAR point cloud data for a forested 20 m by 20 m subset from a LAS tile file, with the vertical coordinate corresponding to the elevation level [m] of LiDAR point returns. Darker coloring corresponds to lower elevations. Coordinate grids for eastings [m] and northings [m] are in the NAD 1983 UTM Zone 17 N projection.

6.2.2. Methodology

As received from LIO the LiDAR LAS files were compressed in the LAZ format. These LAZ files were uncompressed and extracted in bulk using LASzip software (Isenburg 2013). The resulting LAS files were then read into Python using the laspy library, from which the sqlite3 (Gaffney et al. 2022) module in Python utilized SQL conforming to the SQLite variant. The SQL commands in Python were implemented through the usage of cursors from the sqlite3 module.

Summary variables were attained in SQL using the group by clause. To obtain one record value conforming to one pixel for a resulting rasterized layer, the round function was exploited to resolve the pixel coordinates, which was done for cell sizes of 30 m by 30 m spatial extent. These resulting pixel coordinates corresponded to easting and northings, with these combinations comprising as unique identifiers or keys for merging the aggregated output to other rasterized layers stored in the form of structured data tables. The SQL commands for generating a DTM, a DSM and a CHM were presented in Figure 6.2. Note here, that the CHM was calculated as the DSM minus the DTM.

```
CREATE TABLE LiDAR_1 AS
SELECT ROUND(X/30 − 0.5,0)*30 AS X_BASE, ROUND(Y/30 − 0.5,0)*30 AS Y_BASE,
    COUNT(RETURN_NUM) AS RETURN_COUNT,
    MAX(Z) AS DSM, MIN(Z) AS DTM, (MAX(Z) − MIN(Z)) AS CHM
FROM LiDAR_Data
GROUP BY X_BASE, Y_BASE;
```

Figure 6.2. SQL for constructing a DSM, DTM and CHM from LiDAR point cloud data.

Examining the SQL in Figure 6.2, the round function was utilized to compute the lower pixel bounds for the grid coordinates, for pixel sizes of 30 m by 30 m for spatial resolution. Extreme values were resolved from the aggregation via the group by clause. This corresponded to using the maximum vertical elevation value as the DSM and the minimum vertical elevation value as the DTM, based on the LiDAR point cloud data within each constituted pixel.

Note that a 30 m by 30 m resolution was selected as this was the spatial resolution for Landsat imagery as obtained for digital soil mapping purposes, so a 30 m by 30 m spatial resolution for generating predictors was rational. For the tree species classification presented in Chapter 3, a 2 m by 2 m spatial resolution was utilized as that was the resolution of the WV2 multispectral imagery, so finer spatial resolution environmental covariates were derived.

SQL queries were also executed to construct a gap fraction covariate, which was calculated per pixel as the proportion of retrievals with only one return, with these SQL queries shown in Figure 6.3. Note that the final summary table comprised of the computed features of DSM, DTM, CHM and gap fraction, with values for these features rounded to two decimal places.

Note that the SQL commands depicted in Figures 6.2 and 6.3 were implemented per LAS file. However, all the study areas investigated were larger than the 1 km^2 area confines per LAS tile file, with some study areas covering more than 1000 km^2. Thus, the LAS files had to be processed sequentially regarding SQL. For that task, batch scripts were executed to load and input one LAS file at a time. These scripts carried out the necessary queries and then appended the output from the summary table to another table storing this compiled output for the previously processed LAS files. This compiled output contained unique combinations

of easting and northing coordinates, so to match output to the respective pixels. The os and glob modules in Python were utilized for implementing the batch processing and exploited patterns within the filenames to effectively read in the LAS files (Airborne Imaging 2018).

```
CREATE TABLE LiDAR_2 AS
SELECT ROUND(X/30 − 0.5,0)*30 AS X_BASE, ROUND(Y/30 − 0.5,0)*30 AS Y_BASE,
    COUNT(RETURN_NUM) AS ONE_RETURN_COUNT
FROM LiDAR_Data
WHERE RETURN_NUM = 1
GROUP BY X_BASE, Y_BASE;

CREATE TABLE LiDAR_Summary AS
SELECT a.X_BASE, a.Y_BASE, ROUND(a.DSM, 2) AS DSM,
    ROUND(a.DTM, 2) AS DTM, ROUND(a.CHM, 2) AS CHM,
    ROUND(b.ONE_RETURN_COUNT / a.RETURN_COUNT, 2) AS GAP_FRACTION
FROM LiDAR_1 a LEFT JOIN LiDAR_2 b
    ON a.X_BASE = b.X_BASE AND a.Y_BASE = b.Y_BASE
ORDER BY a.X_BASE, a.Y_BASE;
```

Figure 6.3. SQL for constructing a gap fraction, with DSM, DTM and CHM also included in the summary table.

Before rasterization using the compiled summary table was performed, the compiled summary table had to be restacked. This restacking transformed the data from a list to 2-D arrays, with northing coordinates arranged sequentially for the rows and easting coordinates arranged sequentially for the columns. It was important that there were no gaps in reference to the grid coordinates. Respectively, a specification for a missing value was set for any grid coordinate combination corresponding to no data as regarded from the summary table information. Once that was completed, Pandas (McKinney 2010) was adopted to transform the listed summary data into 2-D arrays conforming to rasterized layers for each of DTM, DSM, CHM and gap fraction. These 2-D arrays, also known as pivot tables, were each

exported and saved separately as text files. Appended at the beginning of each of these text files were relevant header information for the corresponding 2-D array. This information consisted of the number of rows and columns, the easting and northing coordinates for the lower left corner of the spatial confines, the cell size spatial resolution, and the no data specification used for denoting missing values. Afterwards, these text files were inputted into ArcGIS Pro software using the ASCII (American Standard Code for Information Interchange) to raster conversion tool. The respective coordinate projection was then set for the loaded raster layer, with ArcGIS software used for displaying these rasterized covariate layers and generating figures.

6.3. Results

A desktop computer with an Intel Core i7-7700 CPU (central processing unit) of 3.60 GHz and 8.00 GB RAM (random access memory) was used to process the LiDAR data. The LiDAR data was stored on a 10 TB capacity external hard drive, with the files transferred over to that drive by LIO. In terms of computing, approximately 350 LAS tile files with each file conforming to 1 km^2 were processed per day. Computing time corresponded to about 3 to 5 minutes to process a LAS file of around 0.5 GB size encompassing 1 km^2. For faster processing time, around 200 GB of LAS files were transferred over and saved on an internal hard drive of the desktop computer, to a directory used for input with the batch scripts. The resulting tables created from the SQL were conserved in a database, from which text files of the generated covariate features were created from.

A DTM using the SQL shown in Figure 6.2 was generated for the Cochrane study area, presented in Figure 6.4. Also depicted in Figure 6.4 was a DTM acquired from the Ontario MNR. This product corresponded to a provincial DTM that was constructed in 2013 from a combination of a DSM retrieved from RADAR (radio detection and ranging), a base mapping dataset for Ontario, and DTM points and contours (Spatial Data Infrastructure 2013). Both DTMs in Figure 6.4 were of 30 m spatial resolution. However, precision to the centimeter level in elevation was resolved for the LiDAR-derived DTM. It is likely that LiDAR data acquired via aerial survey was able to obtain coverage for remote forested locations that were impractical to survey by land-based methods.

A DSM and CHM as computed from the SQL in Figure 6.2 were both visually presented for the Cochrane study area in Figure 6.5. Here, the CHM was generated as the difference of the DTM from the DSM. When compared to the LiDAR-derived DTM in Figure 6.4, the DSM appeared grainier due to variations with vegetation heights within the canopy. Inspecting the CHM, various land use covers within the study area were clearly distinguishable. Agricultural fields, water bodies, infrastructure with roads and clearing, as well as forested tracts were all discernible. Locations of higher CHM values conforming to mature forest were recognizable, as were areas with intermediate CHM values corresponding to wetlands. Note for this study area that lowland or stunted black spruce (*Picea mariana*) was typically the only tree species to grow within the wetland areas.

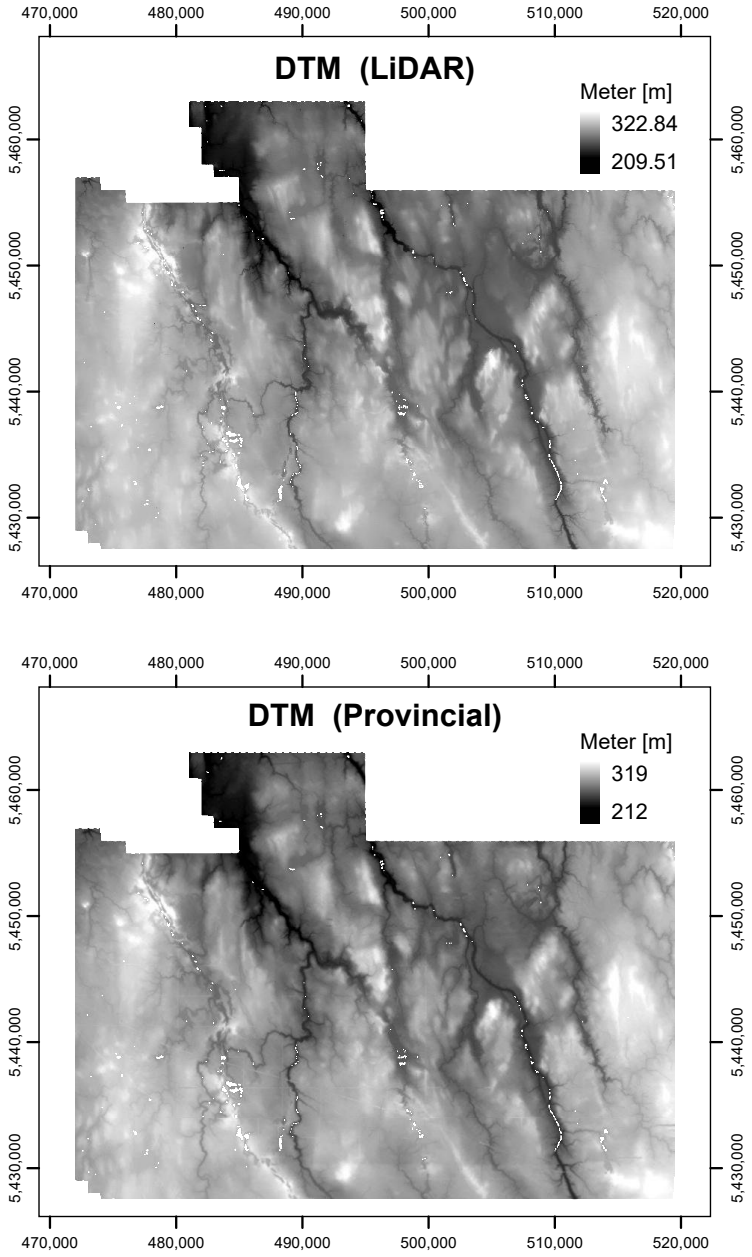

Figure 6.4. Digital terrain models (DTMs) both of 30 m spatial resolution for Cochrane study area. Top: LiDAR-derived DTM; Bottom: Provincial DTM obtained from Ontario MNR. Grid coordinates correspond to the NAD UTM 1983 Zone 17 N projection.

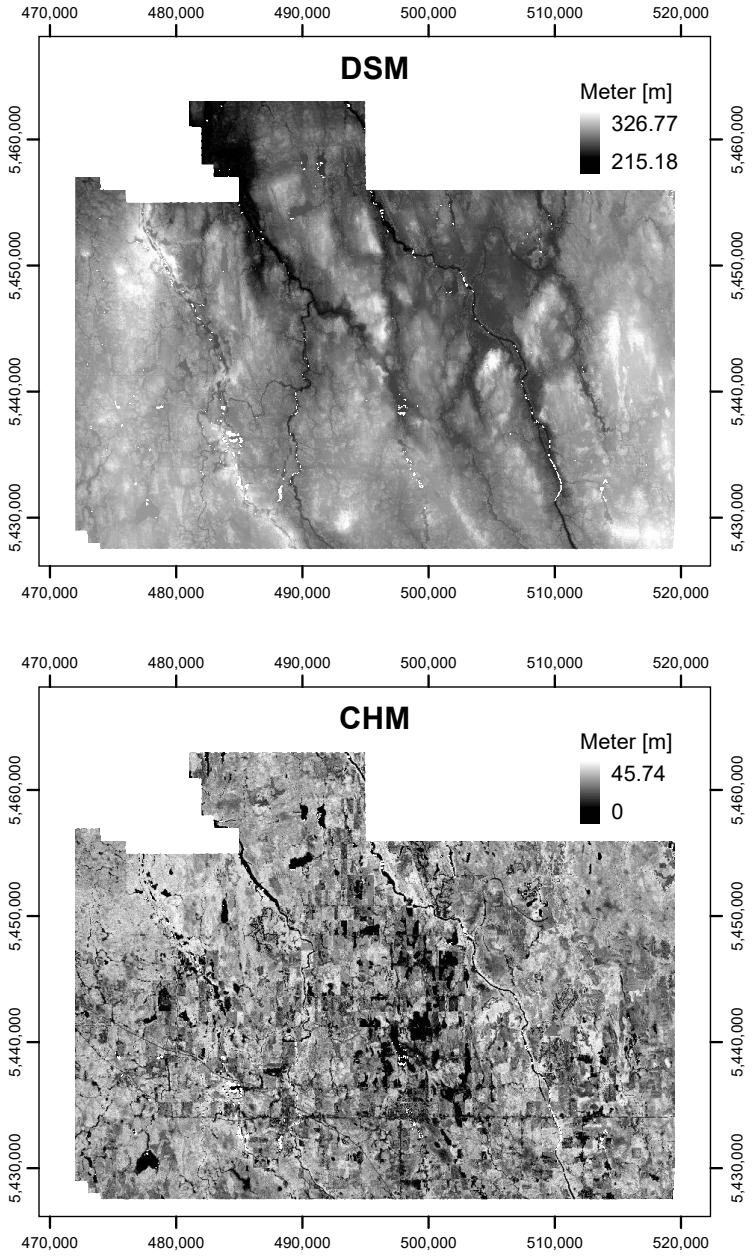

Figure 6.5. LiDAR-derived covariates of DSM and CHM of 30 m spatial resolution for Cochrane study. Top: Digital surface model (DSM); Bottom: Canopy height model (CHM). Grid coordinates correspond to the NAD UTM 1983 Zone 17 N projection.

A gap fraction as computed by SQL within Figure 6.3 was depicted in Figure 6.6 for the Cochrane study area. When comparing the gap fraction here to the CHM in Figure 6.5, similar patterns with land use cover were apparent. However, the gap fraction better discerned locations consisting of denser forest structure but with shorter canopy heights. The gap fraction was generally highly correlated to the CHM, but the gap fraction provided additional relevant insights pertaining to density with canopy structure.

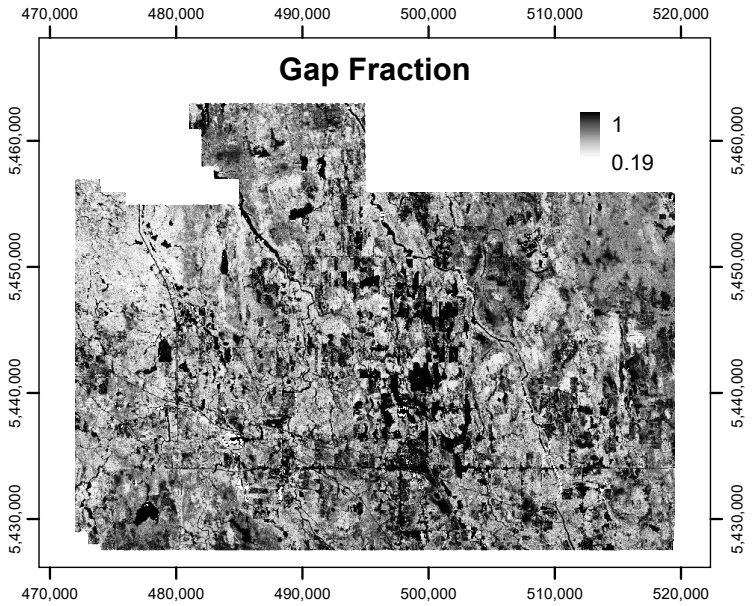

Figure 6.6. LiDAR-derived gap fraction for the Cochrane study area. Grid coordinates correspond to the NAD UTM 1983 Zone 17 N projection.

6.4. Discussion

Other aggregation functions can be utilized for SQL queries, rather than just calculating the maximum, minimum or arithmetic mean. The aggregation functions feasible for computation relate to the variant of SQL that has been installed or implemented. For example, a function

for computing medians is available in some variants of SQL; however, the median can also be computed by sorting and indexing operations in conjunction with multistep querying using intermittent data tables. On another note, with SQL a where clause can be included to enhance the intricacy of a query, facilitating nested subqueries to further refine the data selections or considerations. In terms of screening data, queries can easily be formulated to identify outliers within point cloud data, so to exclude data points corresponding to tower structures or flocks of geese flying beneath the aerial LiDAR sensor.

The features of DTM, CHM and gap fraction as have been derived in Figures 6.2 and 6.3 have been adequate as environmental covariates for digital soil mapping and tree species classification research. Specifically, the CHM has been important for digital soil mapping regarding BD in Chapter 5 and tree species classification in Chapter 3. In addition, the DTM has been utilized for the derivation of a suite of topographic covariates, which have been important for both tree species classification in Chapter 3 and digital soil mapping work for both Chapters 4 and 5. Nevertheless, if desired, more refined covariates for these features can be computed. To mitigate the impact of extreme values regarding maximums or minimums, these values can be evaluated initially for smaller pixel sizes. Afterwards, these values can then be summarized to the desired pixel size by resolving averages from over the smaller pixels. These resulting values can be more representative for the spatial extents of the pixels. Moreover, conceptualizations for other covariates can also be derived, showcasing the adaptability for SQL to generate additional environmental covariates.

Data fusion can be accomplished via SQL by merging with other feature data within the confines of a database. SQL can be exploited to efficiently extract data from rasterized

imagery or layers for only the sampling sites or locations corresponding to fieldwork observation data. The structured nature of data with SQL facilitates more control and precision with merging and aggregating data, enhancing replicability and increasing confidence with the resulting data analysis work.

6.5. Summary

- SQL can effectively derive and compute rasterized covariates from LiDAR point cloud data. Adopting batch processing, these computations can be implemented for larger study area domains while using only a typical desktop computer.
- Queries that are executed via SQL can be customized to conform to more sophisticated designs, such as nested queries or several iterations of querying. This permits flexibility regarding control on what data is selected, and ultimately how data is aggregated.
- Detailed covariates for a DSM, a DTM, a CHM and a gap fraction were each computed from relatively basic SQL commands. The constructed DTM was more precise than the corresponding provincial DTM obtained from the Ontario MNR, consequently facilitating the generation of more detailed topographic covariates for both tree species classification and digital soil mapping research.

Chapter 7

Conclusions and Final Remarks

7.1. Conclusions

Through the undertaking of this research, an improved understanding of soil carbon has been attained for the Great Clay Belt region of the boreal forest in northern Ontario, Canada. Methods have been advanced for resolving soil carbon concentration, in pursuit of investigating the linkage of soil carbon with vegetation and other physical drivers for the southern transition zone of the boreal forest. This includes fulfillment with the research objectives.

A two-step inferential statistical analysis utilizing the Kruskal-Wallis (KW) test and generalized estimating equations (GEEs) was enacted to ascertain the status of soil carbon concentration for representative cover types within this study region. This was applied to soil samples that were obtained from 431 subplots for 144 sites, conforming to various characteristic land cover types and tree species groupings. Firstly, the KW test corroborated that sufficient quantities of soil samples were collected to substantiate median differences for each of soil carbon, bulk density, clay component and carbon-to-nitrogen ratio separately across all tree species groupings as well as land cover types. Secondly, GEEs inferred estimates for mean differences of soil carbon between specific cover types, with confidence intervals and significances indicated. Both types of testing were applicable to clustered

sample data with skewed distributions of soil properties and hence were more relevant than other inferential statistical approaches.

A pixel-based classification of endemic tree species at a more refined spatial level was accomplished, along with the identification of important features for tree species classification. Finer resolution rasterized covariates were derived from LiDAR point cloud data, from which detailed topographic covariates were generated, providing representation for various forested settings. This permitted a tree species classification for a finer spatial resolution than typically achieved for pixel-based classifications exceeding a combined area of 100 km^2. Random forests (RFs) were trained and attained R^2 greater than 0.7 with corresponding kappa score surpassing 0.5. Furthermore, insights were gained regarding prediction uncertainty on account of entropy calculations. An alternative normalized entropy was devised which yielded additional insights and value with interpretation, in conjunction with respective prediction maps for the most probable and secondary tree species.

Modeling accuracies were further improved when data-driven methods via encoder-decoder (ED) neural networks were employed. Both dense neural network (DNN) layers and convolutional neural network (CNN) structures were incorporated within ED formulations. These EDs boosted the modeling accuracy for soil carbon to R^2 values of over 0.5. Quantile mapping regarding a model ensemble was adapted for computing deciles of the predictions for carbon. Further insights were attained by computing standard deviations, revealing that the greatest prediction uncertainties for decile quantities corresponded to forested tracts along streams and rivers, conforming to localities with forest cover of greater diversity.

Consequently, the standard deviation maps indicated what land covers required more representation with sampling or target data, to address for future research.

The integration of knowledge-based methods was actualized through a framework to better implement structural equation modeling (SEM) for digital soil mapping. Within this workflow, multiple approaches were employed for assessing variable importance so to reduce bias within feature selection. Pairwise correlations were calculated from the listing of the most important predictors so to mitigate multicollinearity regarding the modeling, with the SEM process retaining only significance predictors. An integrated model for bulk density, carbon and carbon-to-nitrogen ration was fitted, with agreeable goodness of fit indicators. An associated path analysis with the SEM resolved contributions of predictors with respect to the targeted soil properties. Slope attributes and CHM were important for bulk density, with bulk density along with plan curvature and DEM of significance for carbon. For carbon-to-nitrogen ratio, carbon in conjunction with water detection covariates were of consequence. The SEM achieved improved modeling accuracy for soil carbon than what a RF trained on the full set of covariates obtained. Moreover, RFs trained on the reduced set of predictors identified as significant with SEM by the framework also similarly attained enhanced modeling accuracies.

Structured query language (SQL) was successfully maneuvered to process vast quantities of LiDAR point cloud data so to effectively derive and generate rasterized covariates of DTM, CHM and a gap fraction. In total, these covariates were computed for domains comprising over 6000 km^2 using a desktop computer. These DTMs were utilized for generating detailed topographic covariates, many of which were evaluated with higher variable importance for

tree species classification as well as digital soil mapping. The derived vegetation covariate of CHM was also important for tree species classification, in addition for digital soil mapping with respect to bulk density.

The contributions for this research comprise of:

- Devising novel methodology for creating rasterized environmental covariates from LiDAR point cloud data. This entailed the construction of intricate covariates of a finer spatial scale for vegetation structure with a CHM and gap fraction, as well as a DTM. When possible, LiDAR point cloud data should be exploited for the construction of more detailed covariates, particularly for a DEM as well as a CHM. The DEM derived from LiDAR data should be the DEM utilized for the generation of the respective topographic covariates.
- Developing a framework to rationally adapt structural equation modeling (SEM) for digital soil mapping. This incorporated a multiapproach for variable importance to reduce feature selection bias and accounted for pairwise correlations to mitigate multicollinearity. To uncover potential interrelations between predictors and soil properties, then SEM can be pursued. However, feature selection will need to be implemented prior to SEM so to reduce multicollinearity and retain only the most important predictors. Preferably, more than one approach should be utilized for feature selection, so to diminish bias with the assessment of variable importance.
- Deriving a normalized entropy metric constituting just the top two contending classes, that more realistically gauged prediction uncertainty as it averted the dilution of the uncertainty score when compared to the conventional entropy metric. A normalized

entropy metric should be utilized when predicting for more than a few classes, particularly with tree species classification or soil texture modeling, as models may predict better for certain tree species or soil texture categories.

- Formulating encoder-decoder (ED) neural networks that augmented modeling accuracy, for suitability with smaller data sets as encountered within digital soil mapping. For obtaining the best possible modeling accuracies without meaningful concern for unraveling modeling dynamics, then ED neural networks suitable for smaller data sets can be incorporated for soil modeling.
- Incorporating innovative inferential statistical testing for evaluating soil properties with cover types, as well as employing statistical tests appropriate for skewed observations with unknown underlying distributions. In particular, the development of refined analysis with generalized estimating equations (GEEs) was original for this research domain. If extracting soil samples for digital soil mapping research, inferential analysis as conducted with the KW test can ascertain whether enough heterogeneity exists for the collected soil samples across cover types. Depending on the research purpose or application, this statistical inferencing testing can corroborate whether enough soil samples have been extracted.
- Investigating quantile mapping using deciles and standard deviations of predictions from an ensemble composed of various modeling approaches, so to discern relevant insights into over and underprediction. Furthermore, the standard deviations of the decile maps indicate localities where additional soil samples should be extracted so to potentially improve future soil modeling work.

- Establishing the feasibility of GEEs to estimate carbon contents along with confidence intervals for extension to legacy soil data consisting primarily of soil texture and land cover attributes for digital soil mapping purposes.
- Instituting a pixel-based tree species classification encompassing natural boreal settings at an unprecedented spatial resolution of 2 m for over 100 km^2. The finer scale classification has enabled representation for rarer tree species, or for tree species that grow within narrow extents along clearings or waterways, that might not be differentiated by using coarser resolution environmental covariates.

7.2. Limitations

The main shortcomings for this research essentially contend to the target data. Specifically, there are only a relatively minimal number of sites with available soil property data for boreal study areas in northern Ontario. Measurements and observations from soil samples correspond to the only attainable ground truth data for soil properties, as inherently soil cannot be directly examined from any operational remote sensors. A reduced number of observations for soil properties can restrict what modeling approaches can be employed, particularly regarding deeper machine learning. Furthermore, the effort and cost of extracting, processing and analyzing soil samples greatly hinders the acquisition of additional soil samples, especially from remote or inaccessible localities within rugged forested terrain. Soil samples for this research have been collected so to obtain representation from various land covers and endemic tree species groupings within boreal settings of the Great Clay Belt

region in the District of Cochrane in northern Ontario. However, if feasible, observations of measured soil properties from more sites could have benefited this research.

Due in part to the limited amounts of soil property data, attaining higher modeling accuracies for predicting soil carbon and associated properties remains a challenge. In general, this is an ongoing issue for digital soil mapping research (Minasny et al. 2019), as reported R^2 values for modeling soil carbon quantities in boreal areas tend to around 0.2 or less (Beguin et al. 2017; Mansuy et al. 2014). The research for this dissertation has managed to improve accuracies for soil carbon modeling, increasing R^2 to around 0.4 with SEM and to over 0.5 with ED neural networks. Nevertheless, a need to further augment modeling accuracies persists, particularly for consistently attaining greater R^2 values so that additional variation within targeted properties can be resolved by modeling (Nakagawa and Schielzeth 2013).

A tree species indicator map for black spruce was among the top predictors for soil carbon with SEM but ultimately was not retained as a regression variable as it was not significant at the *p*-value level of 0.05. The corresponding *p*-value of this tree species map for black spruce was around 0.1 and hence was omitted. For ensuring appropriate assessments for goodness of fit metrics, as well as convergence for a modeling solution, latent variables were excluded from the SEM. If higher data counts of soil properties measurements were available, it is likely that more sophisticated associations could had been discerned with the SEM.

The tree species classification was trained on groupings conforming to pure stands, that were defined to be composed of at least 70% for the same tree species (Bravo-Oviedo et al. 2014). For this study area, balsam fir rarely grew within pure stands, necessitating mixed categories for tree species where the stand composition was lowered to a minimum of 50% for the

respective tree species. A great number of sites with tree species data was obtained from the Ontario Forest Resources Inventory (FRI). However, these observations were only appraised for small plots, each no larger than 20 m by 20 m in extent (Paloniemi 2018). This was part of the reason for utilizing finer scale environmental covariates, particularly with LiDAR-derived covariates and WorldView-2 (WV2) multispectral imagery, of 2 m spatial resolution so enough pixels were attained for modeling the plots.

7.3. Future Work

Further improvements for resolving soil carbon and discerning interconnections within an environmental context regarding digital soil mapping can be attained in part by the following future considerations.

- Integrating legacy soil data with modeling. The Ontario FRI has currently assessed soil texture family and tree species information for over 1000 sites within the Great Clay Belt region (Paloniemi 2018). Models can potentially be constructed to extrapolate prediction of soil carbon for this data. Although these values would only conform to estimates attained upon extrapolation, nonetheless this can enhance prediction for this study region when compared to results from alternative modeling.
- Fielding plots of larger spatial extents for training with tree species classification while still accounting for pure stands. This would allocate greater pixel quantities for model training, so to facilitate CNNs or other deeper machine learning methods that can further augment modeling accuracy.

- Reckoning of disturbance events (Utla et al. 2025), particularly relating to fire or disease infestation for overstory vegetation. Data corresponding to disturbance events have been collected for this study area, but none of this information is relevant to the sampling sites for this dissertation research. A future study should encompass sites where certain disturbance events have arisen and have been well-documented. The time duration that has passed since disturbance can also be incorporated as a time factor in reference to soil formation factors for digital soil mapping.
- Stratifying upland black spruce from lowland black spruce, as distinctions materialize regarding cover type and soil attributes. Lowland black spruce typically manifests as a stunted form growing within wetland environments, whereas upland black spruce attains taller growth within a more forested stand (El Abidine et al. 1994; Haynes et al. 2021). This distinguishment can advance better rectification between wetland sites with the greatest contents of soil carbon, versus sites comprised of upland black spruce with lower but still appreciable soil carbon contents.
- Imputing stand age or understory information (Noualhaguet, Hernández-Rodríguez, and Girona 2025) for vegetation or land cover. Stand age can relate to a time factor, to encapsulate characteristics pertinent to vegetation at certain life cycle points. Understory vegetation information can yield attributes that can potentially be applied as important covariates for modeling soil carbon.
- Deriving and generating additional detailed covariates relating to vegetation structure from LiDAR point cloud data (Utla et al. 2025). This supplemental information can enable better differentiation of forest type or vegetation in reference to soil properties.

- For comparison purposes to other soil modeling studies concerning carbon, predictions for carbon stock or carbon pool (CP) can be considered. However, calculations of CP or carbon stocks might require assumptions about bulk density (BD) values at various depths.

References

El Abidine, A. Z., P. Y. Bernier, J. D. Stewart, and A. P. Plamondon. 1994. “Water Stress Preconditioning of Black Spruce Seedlings from Lowland and Upland Sites.” *Canadian Journal of Botany* 72(10): 1511–18. doi:10.1139/b94-186.

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

An, Yida, Lei Zhang, Qing Wang, and Yunwei Han. 2022. “Soil Quality Assessment of Different Land Use Types Based on TOPSIS Method in Hilly Sandy Area of Loess Plateau, Northern China.” *International Journal of Environmental Research and Public Health* 19(24): 1–17. doi:10.3390/ijerph192417059.

Angelini, Marcos E., Gerard B.M. Heuvelink, Bas Kempen, and Héctor J.M. Morrás. 2016. “Mapping the Soils of an Argentine Pampas Region Using Structural Equation Modelling.” *Geoderma* 281: 102–18. doi:10.1016/j.geoderma.2016.06.031.

Arashi, Mohammad, Adewale F. Lukman, and Zakariya Y. Algamal. 2022. “Liu Regression after Random Forest for Prediction and Modeling in High Dimension.” *Journal of Chemometrics* 36(4): 1–10. doi:10.1002/cem.3393.

Atkinson, Anthony C., and Marco Riani. 2002. “Forward Search Added-Variable t-Tests and the Effect of Masked Outliers on Model Selection.” *Biometrika* 89(4): 939–46. doi:10.1093/biomet/89.4.939.

Avery, Lisa, Ryan Del Bel, Osvaldo Espin-Garcia, Tyler Pittman, Yanning Wang, Jessica Weiss, and Wei Xu. 2023. “ReportRmd: Tidy Presentation of Clinical Reporting.” https://cran.r-project.org/package=reportRmd.

Bai, Guo, Cheng Liao, Yuanzhi Liu, and You Feng Cheng. 2024. “Far-Field Phaseless Diagnosis for Impaired Arrays Based on Artificial Neural Networks and Compressed Sensing.” *IEEE Transactions on Antennas and Propagation* 72(2): 1581–92. doi:10.1109/TAP.2023.3343406.

Bargagliotti, Anna E., and Raymond N. Greenwell. 2015. “Combinatorics and Statistical Issues Related to the Kruskal-Wallis Statistic.” *Communications in Statistics: Simulation and Computation* 44(2): 533–50. doi:10.1080/03610918.2013.786781.

Beaudoin, A., P. Y. Bernier, P. Villemaire, L. Guindon, and X. J. Guo. 2017. *Species Composition, Forest Properties and Land Cover Types across Canada’s Forests at 250m Resolution for 2001 and 2011*. Quebec, Canada. doi:https://doi.org/10.23687/ec9e2659-1c29-4ddb-87a2-6aced147a990.

Beaudoin, A, P Y Bernier, P Villemaire, L Guindon, and X Jing Guo. 2018. “Tracking Forest Attributes across Canada between 2001 and 2011 Using a k Nearest Neighbors Mapping Approach Applied to MODIS Imagery.” *Canadian Journal of Forest Research* 48: 85–93. doi:10.23687/ec9e2659-1c29-4ddb-87a2-6aced147a990.

Beguin, Julien, Geir Arne Fuglstad, Nicolas Mansuy, and David Paré. 2017. “Predicting Soil Properties in the Canadian Boreal Forest with Limited Data: Comparison of Spatial and Non-Spatial Statistical Approaches.” *Geoderma* 306: 195–205. doi:10.1016/j.geoderma.2017.06.016.

Beribisky, Nataly, and Robert A. Cribbie. 2024. “Evaluating the Performance of Existing and Novel Equivalence Tests for Fit Indices in Structural Equation Modelling.” *British Journal of Mathematical and Statistical Psychology* 77(1): 103–29. doi:10.1111/bmsp.12317.

Blackford, Christopher, Brandon Heung, and Kara L. Webster. 2022. “Incorporating Spatial Uncertainty Maps into Soil Sampling Improves Digital Soil Mapping Classification Accuracy in Ontario, Canada.” *Geoderma Regional* 29: 1–18. doi:10.1016/j.geodrs.2022.e00495.

Boulanger, Yan, Anthony R. Taylor, David T. Price, Dominic Cyr, Elizabeth McGarrigle, Werner Rammer, Guillaume Sainte-Marie, et al. 2017. “Climate Change Impacts on Forest Landscapes along the Canadian Southern Boreal Forest Transition Zone.” *Landscape Ecology* 32(7): 1415–31. doi:10.1007/s10980-016-0421-7.

Bravo-Oviedo, Andres, Hans Pretzsch, Christian Ammer, Ernesto Andenmatten, Anna Barbati, Susana Barreiro, Peter Brang, et al. 2014. “European Mixed Forests: Definition and Research Perspectives.” *Forest Systems* 23(3): 518–33. doi:10.5424/fs/2014233-06256.

Brungard, Colby W., Janis L. Boettinger, Michael C. Duniway, Skye A. Wills, and Thomas C. Edwards. 2015. “Machine Learning for Predicting Soil Classes in Three Semi-Arid Landscapes.” *Geoderma* 239: 68–83. doi:10.1016/j.geoderma.2014.09.019.

Byrne, Barbara M., and Fons J.R. van de Vijver. 2010. “Testing for Measurement and Structural Equivalence in Large-Scale Cross-Cultural Studies: Addressing the Issue of Nonequivalence.” *International Journal of Testing* 10(2): 107–32. doi:10.1080/15305051003637306.

Caers, Jef. *Modeling Uncertainty in the Earth Sciences*. John Wiley & Sons Ltd, 2011.

Chauhan, Rahul, Kamal Kumar Ghanshala, and R. C. Joshi. 2018. “Convolutional Neural Network (CNN) for Image Detection and Recognition.” *ICSCCC 2018 - 1st International Conference on Secure Cyber Computing and Communications*: 278–82. doi:10.1109/ICSCCC.2018.8703316.

Chen, Fang Fang. 2007. “Sensitivity of Goodness of Fit Indexes to Lack of Measurement Invariance.” *Structural Equation Modeling* 14(3): 464–504. doi:10.1080/10705510701301834.

Chen, Hu, Yi Zhang, Mannudeep K. Kalra, Feng Lin, Yang Chen, Peixi Liao, Jiliu Zhou, and Ge Wang. 2017. “Low-Dose CT with a Residual Encoder-Decoder Convolutional Neural Network.” *IEEE Transactions on Medical Imaging* 36(12): 2524–35. doi:10.1109/TMI.2017.2715284.

Chen, Songchao, Dominique Arrouays, Vera Leatitia Mulder, Laura Poggio, Budiman Minasny, Pierre Roudier, Zamir Libohova, et al. 2022. “Digital Mapping of GlobalSoilMap Soil Properties at a Broad Scale: A Review.” *Geoderma* 409: 1–21. doi:10.1016/j.geoderma.2021.115567.

Cheng, Changming, and Er Wei Bai. 2019. “Ranking the Importance of Variables in Nonlinear System Identification.” *Automatica* 103: 472–79. doi:10.1016/j.automatica.2019.02.029.

Chiang, Shou Hao, and Miguel Valdez. 2019. “Tree Species Classification by Integrating Satellite Imagery and Topographic Variables Using Maximum Entropy Method in a Mongolian Forest.” *Forests* 10(11): 1–15. doi:10.3390/f10110961.

Chollet, François. 2015. “Keras.” https://github.com/fchollet/keras.

Chong, Il Gyo, and Chi Hyuck Jun. 2005. “Performance of Some Variable Selection Methods When Multicollinearity Is Present.” *Chemometrics and Intelligent Laboratory Systems* 78(1): 103–12. doi:10.1016/j.chemolab.2004.12.011.

Cohen, Jacob. 1960. “A Coefficient of Agreement for Nomial Scales.” *Educational and Psychological Measurement* 20(1): 37–46. doi:10.1177/001316446002000104A.

Cohen, Jacob. 1973. “Eta-Squared and Partial Eta-Squared in Fixed Factor ANOVA Designs.” *Educational and Psychological Measurement* 33(1): 107–12.

Cohen, Jacob. 1988. *Statistical Power Analysis for the Behavioural Sciences*. 2nd ed. New York: Academic Press.

Conrad, O., B. Bechtel, M. Bock, H. Dietrich, E. Fischer, L. Gerlitz, J. Wehberg, V. Wichmann, and J. Böhner. 2015. “System for Automated Geoscientific Analyses (SAGA) v. 2.1.4.” *Geoscientific Model Development* 8(7): 1991–2007. doi:10.5194/gmd-8-1991-2015.

Dahmen, G., and A. Ziegler. 2004. “Generalized Estimating Equations in Controlled Clinical Trials: Hypotheses Testing.” *Biometrical Journal* 46(2): 214–32. doi:10.1002/bimj.200310018.

Dai, Pengjie, Xiaobo Ding, and Qihua Wang. 2016. “Dimension Reduction Based Linear Surrogate Variable Approach for Model Free Variable Selection.” *Journal of Statistical Planning and Inference* 169: 13–26. doi:10.1016/j.jspi.2015.07.001.

Dalponte, Michele, Lorenzo Bruzzone, and Damiano Gianelle. 2012. “Tree Species Classification in the Southern Alps Based on the Fusion of Very High Geometrical Resolution Multispectral/Hyperspectral Images and LiDAR Data.” *Remote Sensing of Environment* 123: 258–70. doi:10.1016/j.rse.2012.03.013.

Dalponte, Michele, Liviu Theodor Ene, Mattia Marconcini, Terje Gobakken, and Erik Næsset. 2015. “Semi-Supervised SVM for Individual Tree Crown Species Classification.” *ISPRS Journal of Photogrammetry and Remote Sensing* 110: 77–87. doi:10.1016/j.isprsjprs.2015.10.010.

Deighton, Holly D., F. Wayne Bell, and Zoë Lindo. 2025. “Long-Term Effects of Forest Management on Boreal Forest Soil Organic Carbon.” *Forests* 16(6): 1–19. doi:10.3390/f16060902.

Denneler, Bernhard, Yves Bergeron, and Yves Bégin. 1999. “An Attempt to Explain the Distribution of the Tree Species Composing the Riparian Forests of Lake Duparquet, Southern Boreal Region of Quebec, Canada.” *Canadian Journal of Botany* 77(12): 1744–55. doi:10.1139/cjb-77-12-1744.

Dexter, A. R. 2004. “Soil Physical Quality: Part I. Theory, Effects of Soil Texture, Density, and Organic Matter, and Effects on Root Growth.” *Geoderma* 120: 201–14. doi:10.1016/j.geodermaa.2003.09.005.

Digital Globe. 2016. “World View- 2 Design and Specifications.” www.digitalglobe.com.

Dormann, Carsten F., Jane Elith, Sven Bacher, Carsten Buchmann, Gudrun Carl, Gabriel Carré, Jaime R.García Marquéz, et al. 2013. “Collinearity: A Review of Methods to Deal with It and a Simulation Study Evaluating Their Performance.” *Ecography* 36(1): 27–46. doi:10.1111/j.1600-0587.2012.07348.x.

Driscoll, Wade C. 1996. “Robustness of the ANOVA and Tukey-Kramer Statistical Tests.” *Computers and Industrial Engineering* 31(1/2): 265–68.

Engler, Robin, Lars T. Waser, Niklaus E. Zimmermann, Marcus Schaub, Savvas Berdos, Christian Ginzler, and Achilleas Psomas. 2013. “Combining Ensemble Modeling and Remote Sensing for Mapping Individual Tree Species at High Spatial Resolution.” *Forest Ecology and Management* 310: 64–73. doi:10.1016/j.foreco.2013.07.059.

Fan, Chunpeng, and Donghui Zhang. 2014. “Wald-Type Rank Tests: A GEE Approach.” *Computational Statistics and Data Analysis* 74: 1–16. doi:10.1016/j.csda.2013.12.004.

Fang, Huiyun, Biyong Ji, Xu Deng, Jiayang Ying, Guomo Zhou, Yongjun Shi, Lin Xu, et al. 2018. “Effects of Topographic Factors and Aboveground Vegetation Carbon Stocks on Soil Organic Carbon in Moso Bamboo Forests.” *Plant and Soil* 433(1–2): 363–76. doi:10.1007/s11104-018-3847-7.

Fassnacht, Fabian Ewald, Hooman Latifi, Krzysztof Stereńczak, Aneta Modzelewska, Michael Lefsky, Lars T. Waser, Christoph Straub, and Aniruddha Ghosh. 2016. “Review of Studies on Tree Species Classification from Remotely Sensed Data.” *Remote Sensing of Environment* 186: 64–87. doi:10.1016/j.rse.2016.08.013.

Firinguetti, Luis, Golam Kibria, and Rodrigo Araya. 2017. “Study of Partial Least Squares and Ridge Regression Methods.” *Communications in Statistics: Simulation and Computation* 46(8): 6631–44. doi:10.1080/03610918.2016.1210168.

Fradette, Olivier, Charles Marty, Patrick Faubert, Pierre Luc Dessureault, Maxime Paré, Sylvie Bouchard, and Claude Villeneuve. 2021. “Additional Carbon Sequestration Potential of Abandoned Agricultural Land Afforestation in the Boreal Zone: A Modelling Approach.” *Forest Ecology and Management* 499(May): 1–10. doi:10.1016/j.foreco.2021.119565.

Franklin, Steven E. 2018. “Pixel-and Object-Based Multispectral Classification of Forest Tree Species from Small Unmanned Aerial Vehicles.” *Journal of Unmanned Vehicle Systems* 6(4): 195–211. doi:10.1139/juvs-2017-0022.

Franklin, Steven E. 2020. “Interpretation and Use of Geomorphometry in Remote Sensing: A Guide and Review of Integrated Applications.” *International Journal of Remote Sensing* 41(19): 7700–7733. doi:10.1080/01431161.2020.1792577.

Gaffney, Keven P., Martin Prammer, Larry Brasfield, D. Richard Hipp, Dan Kennedy, and Jignesh M. Patel. 2022. “SQLite: Past, Present, and Future.” *Proceedings of the VLDB Endowment* 15(12): 3535-3547. doi:10.14778/3554821.3554842.

Gao, Bilei, Anthony R. Taylor, Eric B. Searle, Praveen Kumar, Zilong Ma, Alexandra M. Hume, and Han Y.H. Chen. 2018. “Carbon Storage Declines in Old Boreal Forests Irrespective of Succession Pathway.” *Ecosystems* 21(6): 1168–82. doi:10.1007/s10021-017-0210-4.

Garcia, Mariano, Sassan Saatchi, Antonio Ferraz, Carlos Alberto Silva, Susan Ustin, Alexander Koltunov, and Heiko Balzter. 2017. “Impact of Data Model and Point Density on Aboveground Forest Biomass Estimation from Airborne LiDAR.” *Carbon Balance and Management* 12(1): 1–18. doi:10.1186/s13021-017-0073-1.

Goeman, Jelle J., and Aldo Solari. 2022. “Comparing Three Groups.” *American Statistician* 76(2): 168–76. doi:10.1080/00031305.2021.2002188.

Gokmen, Sahika, Rukiye Dagalp, and Serdar Kilickaplan. 2022. “Multicollinearity in Measurement Error Models.” *Communications in Statistics - Theory and Methods* 51(2): 474–85. doi:10.1080/03610926.2020.1750654.

Grewal, Rajdeep, Joseph A. Cote, and Hans Baumgartner. 2004. “Multicollinearity and Measurement Error in Structural Equation Models: Implications for Theory Testing.” *Marketing Science* 23(4): 519–29. doi:10.1287/mksc.1040.0070.

Haynes, Kristine M., Jessica Smart, Brenden Disher, Olivia Carpino, and William L. Quinton. 2021. “The Role of Hummocks in Re-Establishing Black Spruce Forest Following Permafrost Thaw.” *Ecohydrology* 14(3):1–16. doi:10.1002/eco.2273.

Heung, Brandon, Hung Chak Ho, Jin Zhang, Anders Knudby, Chuck E. Bulmer, and Margaret G. Schmidt. 2016. “An Overview and Comparison of Machine-Learning Techniques for Classification Purposes in Digital Soil Mapping.” *Geoderma* 265: 62–77. doi:10.1016/j.geoderma.2015.11.014.

Hojat, Mohammadreza, and Gang Xu. 2004. “A Visitor’s Guide to Effect Sizes - Statistical Significance versus Practical (Clinical) Importance of Research Findings.” *Advances in Health Sciences Education* 9(3): 241–49. doi:10.1023/B:AHSE.0000038173.00909.f6.

Højsgaard, Søren, Ulrich Halekoh, and Jun Yan. 2006. “The R Package Geepack for Generalized Estimating Equations.” *Journal of Statistical Software* 15(2): 1–11. doi:10.18637/jss.v015.i02.

Hu, B, W M Jung, J Liu, and J Shang. 2022. “RETRIEVAL OF LEAF AREA INDEX AND LEAF CHLOROPHYLL CONTENT FROM HYPERSPECTRAL DATA USING DEEP LEARNING NETWORKS.” In *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, 397–404. doi:10.5194/isprs-archives-XLIII-B3-2022-397-2022.

Hu, Baoxin, Damir Gumerov, Jianguo Wang, and Wen Zhang. 2017. “An Integrated Approach to Generating Accurate DTM from Airborne Full-Waveform LiDAR Data.” *Remote Sensing* 9(8): 1–17. doi:10.3390/rs9080871.

Hu, Baoxin, Qian Li, and G. Brent Hall. 2021. “A Decision-Level Fusion Approach to Tree Species Classification from Multi-Source Remotely Sensed Data.” *ISPRS Open Journal of Photogrammetry and Remote Sensing* 1(July): 1–12. doi:10.1016/j.ophoto.2021.100002.

Hu, Li Tze, and Peter M. Bentler. 1999. “Cutoff Criteria for Fit Indexes in Covariance Structure Analysis: Conventional Criteria versus New Alternatives.” *Structural Equation Modeling* 6(1): 1–55. doi:10.1080/10705519909540118.

Huang, Jiacong, and Junfeng Gao. 2017. “An Ensemble Simulation Approach for Artificial Neural Network: An Example from Chlorophyll a Simulation in Lake Poyang, China.” *Ecological Informatics* 37: 52–58. doi:10.1016/j.ecoinf.2016.11.012.

Hyväluoma, Jari, Petri Niemi, Sami Kinnunen, Kofi Brobbey, Arttu Miettinen, Riikka Keskinen, and Helena Soinne. 2024. “Comparing Structural Soil Properties of Boreal Clay Fields under Contrasting Soil Management.” *Soil Use and Management* 40(2): 1–13. doi:10.1111/sum.13040.

Immitzer, Markus, Clement Atzberger, and Tatjana Koukal. 2012. “Tree Species Classification with Random Forest Using Very High Spatial Resolution 8-Band WorldView-2 Satellite Data.” *Remote Sensing* 4(9): 2661–93. doi:10.3390/rs4092661.

Irizarry, Rafael A. 2001. “Local Regression with Meaningful Parameters.” *American Statistician* 55(1): 72–79. doi:10.1198/000313001300339969.

Isenburg, M. 2013. “LASzip.” *Photogrammetric Engineering & Remote Sensing* 79: 209–17.

Jacoby, William G. 2000. “Loess: A Nonparametric, Graphical Tool for Depicting Relationships between Variables.” *Electoral Studies* 19(4): 577–613. doi:10.1016/S0261-3794(99)00028-1.

Ji, Yuzhu, Haijun Zhang, Zhao Zhang, and Ming Liu. 2021. “CNN-Based Encoder-Decoder Networks for Salient Object Detection: A Comprehensive Review and Recent Advances.” *Information Sciences* 546: 835–57. doi:10.1016/j.ins.2020.09.003.

Jiang, Rong, Susantha Jayasundara, Brian B. Grant, Ward N. Smith, Budong Qian, Adam Gillespie, and Claudia Wagner-Riddle. 2023. “Impacts of Land Use Conversions on Soil Organic Carbon in a Warming-Induced Agricultural Frontier in Northern Ontario, Canada under Historical and Future Climate.” *Journal of Cleaner Production* 404: 1–12. doi:10.1016/j.jclepro.2023.136902.

Jobin, D. M., M. Véronneau, and M. Miles. 2017. “Gravity Anomaly Map, Canada.” *Geological Survey of Canada, Open File*: 8081. doi:https://doi.org/10.4095/299561.

Johnson, Roger W. 2022. “Alternate Forms of the One-Way ANOVA F and Kruskal–Wallis Test Statistics.” *Journal of Statistics and Data Science Education* 30(1): 82–85. doi:10.1080/26939169.2021.2025177.

Jung, Kang Mo. 2008. “Local Influence in Generalized Estimating Equations.” *Scandinavian Journal of Statistics* 35(2): 286–94. doi:10.1111/j.1467-9469.2007.00575.x.

Jurgensen, Martin F., Deborah S. Page-Dumroese, Robert E. Brown, Joanne M. Tirocke, Chris A. Miller, James B. Pickens, and Min Wang. 2017. “Estimating Carbon and Nitrogen Pools in a Forest Soil: Influence of Soil Bulk Density Methods and Rock Content.” *Soil Science Society of America Journal* 81(6): 1689–1696. doi:10.2136/sssaj2017.02.0069.

Ke, Ziyi, Shilin Ren, and Liang Yin. 2024. “Advancing Soil Property Prediction with Encoder-Decoder Structures Integrating Traditional Deep Learning Methods in Vis-NIR Spectroscopy.” *Geoderma* 449(May): 1–11. doi:10.1016/j.geoderma.2024.117006.

Keskin, H., and S. Grunwald. 2018. “Regression Kriging as a Workhorse in the Digital Soil Mapper’s Toolbox.” *Geoderma* 326(October 2017): 22–41. doi:10.1016/j.geoderma.2018.04.004.

Khosravipour, Anahita, Andrew K. Skidmore, Martin Isenburg, Tiejun Wang, and Yousif A. Hussin. 2014. “Generating Pit-Free Canopy Height Models from Airborne Lidar.” *Photogrammetric Engineering and Remote Sensing* 80(9): 863–72. doi:10.14358/PERS.80.9.863.

Kopecký, Martin, Martin Macek, and Jan Wild. 2021. “Topographic Wetness Index Calculation Guidelines Based on Measured Soil Moisture and Plant Species Composition.” *Science of the Total Environment* 757: 1–10. doi:10.1016/j.scitotenv.2020.143785.

Kroetsch, David J., Xiaoyuan Geng, Scott X. Chang, and Daniel D. Saurette. 2011. “Organic Soils of Canada: Part 1. Wetland Organic Soils.” *Canadian Journal of Soil Science* 91(5): 807–22. doi:10.4141/cjss10043.

Kuhn, M. 2008. “Building Predictive Models in R Using the Caret Package.” *Journal of Statistical Software* 28(5): 1–26. doi:10.18637/jss.v028.i05.

Kuhn, Max. 2007. *Variable Importance Using The Caret Package*.

Kuhn, Max, Jed Wing, Steve Weston, Andre Williams, Chris Keefer, Allan Engelhardt, Tony Cooper, et al. 2020. *Package 'caret ': Classification and Regression Training*. https://github.com/topepo/caret/.

Kulha, Niko, Leena Pasanen, Lasse Holmström, Louis De Grandpré, Timo Kuuluvainen, and Tuomas Aakala. 2019. "At What Scales and Why Does Forest Structure Vary in Naturally Dynamic Boreal Forests? An Analysis of Forest Landscapes on Two Continents." *Ecosystems* 22(4): 709–24. doi:10.1007/s10021-018-0297-2.

Kumar, U., V. Kumar, and J. N. Kapur. 1986. "Normalized Measures of Entropy." *International Journal of General Systems* 12(1): 55–69. doi:10.1080/03081078608934927.

Laganière, Jérôme, David Paré, Yves Bergeron, Han Y.H. Chen, Brian W. Brassard, and Xavier Cavard. 2013. "Stability of Soil Carbon Stocks Varies with Forest Composition in the Canadian Boreal Biome." *Ecosystems* 16(5): 852–65. doi:10.1007/s10021-013-9658-z.

Lantz, Björn. 2013. "The Impact of Sample Non-Normality on ANOVA and Alternative Methods." *British Journal of Mathematical and Statistical Psychology* 66(2): 224–44. doi:10.1111/j.2044-8317.2012.02047.x.

Laouici, Imadeddine, Gautier Laurent, Christelle Loiselet, and Yannick Branquet. 2025. "A knowledge-driven modeling formalism for automatic structural interpretation." *Earth Science Informatics* 18(1): 1–20. doi:10.1007/s12145-024-01613-y.

Lee, Hagyeong, and Jongwoo Song. 2019. "Introduction to Convolutional Neural Network Using Keras; An Understanding from a Statistician." *Communications for Statistical Applications and Methods* 26(6): 591–610. doi:10.29220/CSAM.2019.26.6.591.

Li, B. 1997. "On Consistency of Generalized Estimating Equations." *IMS Lecture Notes Monogr. Ser*. 32: 115–36. doi:10.1214/lnms/1215455042.

Li, Dengshan, and Lina Li. 2022. “Detection of Water PH Using Visible Near-Infrared Spectroscopy and One-Dimensional Convolutional Neural Network.” *Sensors* 22(15): 1–18. doi:10.3390/s22155809.

Li, Qingliang, Zhongyan Li, Wei Shangguan, Xuezhi Wang, Lu Li, and Fanhua Yu. 2022. “Improving Soil Moisture Prediction Using a Novel Encoder-Decoder Model with Residual Learning.” *Computers and Electronics in Agriculture* 195(March): 1–14. doi:10.1016/j.compag.2022.106816.

Liang, Peng, Cheng Zhi Qin, A. Xing Zhu, Zhi Wei Hou, Nai Qing Fan, and Yi Jie Wang. 2020. “A Case-Based Method of Selecting Covariates for Digital Soil Mapping.” *Journal of Integrative Agriculture* 19(8): 2127–36. doi:10.1016/S2095-3119(19)62857-1.

Liaw, Andy, and Matthew Wiener. 2002. “Classification and Regression by RandomForest.” *R News* 2(3): 18–22.

Lin, Hsuan Tien, Chih Jen Lin, and Ruby C. Weng. 2007. “A Note on Platt’s Probabilistic Outputs for Support Vector Machines.” *Machine Learning* 68(3): 267–76. doi:10.1007/s10994-007-5018-6.

Lin, Johnny. 2021. “Introduction to Structural Equation Modeling (SEM) in R with Lavaan.” *UCLA: Statistical Consulting Group*. https://stats.oarc.ucla.edu/r/seminars/rsem/#s3.

Liu, Lian, Bo Yang, and Yi Zhang. 2024. “Inverting Magnetotelluric Data Using a Physics-Guided Auto-Encoder with Scaling Laws Extension.” *Frontiers in Earth Science* 12(December): 1–12. doi:10.3389/feart.2024.1510962.

Liuzzo, Lorena, Valeria Puleo, Salvatore Nizza, and Gabriele Freni. 2020. “Parameterization of a Bayesian Normalized Difference Water Index for Surface Water Detection.” *Geosciences (Switzerland)* 10(7): 1–18. doi:10.3390/geosciences10070260.

Loecher, Markus. 2022. “Unbiased Variable Importance for Random Forests.” *Communications in Statistics - Theory and Methods* 51(5): 1413–25. doi:10.1080/03610926.2020.1764042.

Lung-Yut-Fong, A., C. Lévy-Leduc, and O. Cappé. 2011. “Robust Retrospective Multiple Change-Point Estimation for Multivariate Data.” In *Proceedings of the 2011 IEEE Statistical Signal Processing Workshop (SSP)*, Nice, France.

Maharjan, Nilesh, and Byung Wook Kim. 2024. “Machine Learning-Based Beam Pointing Error Reduction for Satellite-Ground FSO Links.” *Electronics* 13(17): 1–22. doi:10.3390/electronics13173466.

Mansuy, Nicolas, Evelyne Thiffault, David Paré, Pierre Bernier, Luc Guindon, Philippe Villemaire, Vincent Poirier, and André Beaudoin. 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. doi:10.1016/j.geoderma.2014.06.032.

Maraun, Douglas. 2013. “Bias Correction, Quantile Mapping, and Downscaling: Revisiting the Inflation Issue.” *Journal of Climate* 26(6): 2137–43. doi:10.1175/JCLI-D-12-00821.1.

McBratney, A. B., M. L. Mendonça Santos, and B. Minasny. 2003. “On Digital Soil Mapping.” *Geoderma* 117(1): 3–52. .doi:10.1016/S0016-7061(03)00223-4.

McHugh, Mary L. 2012. “Lessons in Biostatistics Interrater Reliability : The Kappa Statistic.” *Biochemica Medica* 22(3): 276–82. https://hrcak.srce.hr/89395.

McKinney, W. 2010. “Data Structures for Statistical Computing in Python.” In *9th Python in Science Conference*, Austin, United States of America.

Menichetti, Lorenzo, Aleksi Lehtonen, Antti Jussi Lindroos, Päivi Merilä, Saija Huuskonen, Liisa Ukonmaanaho, and Raisa Mäkipää. 2025. “Soil Carbon Dynamics During Stand Rotation in Boreal Forests.” *European Journal of Soil Science* 76(4): 1–13. doi:10.1111/ejss.70154.

Minasny, Budiman, Örjan Berglund, John Connolly, Carolyn Hedley, Folkert de Vries, Alessandro Gimona, Bas Kempen, et al. 2019. “Digital Mapping of Peatlands – A Critical Review.” *Earth-Science Reviews* 196: 1–38. doi:10.1016/j.earscirev.2019.05.014.

Modzelewska, Aneta, Fabian Ewald Fassnacht, and Krzysztof Stereńczak. 2020. “Tree Species Identification within an Extensive Forest Area with Diverse Management Regimes Using Airborne Hyperspectral Data.” *International Journal of Applied Earth Observation and Geoinformation* 84(August 2019): 1–13. doi:10.1016/j.jag.2019.101960.

Mohseni, N., H. Nematzadeh, E. Akbarib, and H. Motameni. 2023. “Outlier Detection in Test Samples Using Standard Deviation and Unsupervised Training Set Selection.” *International Journal of Engineering, Transactions A: Basics* 36(1): 119–29. doi:10.5829/ije.2023.36.01a.14.

Monserud, Robert A., and Rik Leemans. 1992. “Comparing Global Vegetation Maps with the Kappa Statistic.” *Ecological Modelling* 62(4): 275–93. doi:10.1016/0304-3800(92)90003-W.

Mueller, Kevin E., David M. Eissenstat, Sarah E. Hobbie, Jacek Oleksyn, Andrzej M. Jagodzinski, Peter B. Reich, Oliver A. Chadwick, and Jon Chorover. 2012. “Tree Species Effects on Coupled Cycles of Carbon, Nitrogen, and Acidity in Mineral Soils at a Common Garden Experiment.” *Biogeochemistry* 111(1–3): 601–14. doi:10.1007/s10533-011-9695-7.

Mulder, V. L., S. de Bruin, M. E. Schaepman, and T. R. Mayr. 2011. “The Use of Remote Sensing in Soil and Terrain Mapping - A Review.” *Geoderma* 162(1–2): 1–19. doi:10.1016/j.geoderma.2010.12.018.

Nakagawa, Shinichi, and Holger Schielzeth. 2013. “A General and Simple Method for Obtaining R2 from Generalized Linear Mixed-Effects Models.” *Methods in Ecology and Evolution* 4(2): 133–42. doi:10.1111/j.2041-210x.2012.00261.x.

Natesan, Sowmya, Costas Armenakis, and Udayalakshmi Vepakomma. 2020. “Individual Tree Species Identification Using Dense Convolutional Network (Densenet) on Multitemporal RGB Images from UAV.” *Journal of Unmanned Vehicle Systems* 8(4): 310–33. doi:10.1139/juvs-2020-0014.

De Neve, Jan, and Olivier Thas. 2015. “A Regression Framework for Rank Tests Based on the Probabilistic Index Model.” *Journal of the American Statistical Association* 110(511): 1276–83. doi:10.1080/01621459.2015.1016226.

Niemand, Thomas, and Robert Mai. 2018. “Flexible Cutoff Values for Fit Indices in the Evaluation of Structural Equation Models.” *Journal of the Academy of Marketing Science* 46(6): 1148–72. doi:10.1007/s11747-018-0602-9.

Noualhaguet, Marion, Enrique Hernández-Rodríguez, and Miguel Montoro Girona. 2025. “Drivers of Understory Vegetation 18 Years after Novel Experimental Partial-Harvest Treatments in Canadian Boreal Forests.” *Forest Ecology and Management* 594: 1–25. doi:10.1016/j.foreco.2025.122949.

Nussbaum, Madlene, Kay Spiess, Andri Baltensweiler, Urs Grob, Armin Keller, Lucie Greiner, Michael E. Schaepman, and Andreas Papritz. 2018. “Evaluation of Digital Soil Mapping Approaches with Large Sets of Environmental Covariates.” *Soil* 4(1): 1–22. doi:10.5194/soil-4-1-2018.

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

Pitkänen, Sari. 1998. “The Use of Diversity Indices to Assess the Diversity of Vegetation in Managed Boreal Forests.” *Forest Ecology and Management* 112(1–2): 121–37. doi:10.1016/S0378-1127(98)00319-3.

Pittman, R., and B. Hu. 2024. *Soil Sampling Protocol for York University Field Campaigns*. Toronto, Canada.

Pittman, R., B. Hu, and K. Webster. 2021. “Improvement of Soil Property Mapping in the Great Clay Belt of Northern Ontario Using Multi-Source Remotely Sensed Data.” *Geoderma* 381: 1–15. doi:10.1016/j.geoderma.2020.114761.

Pittman, Rory, Baoxin Hu, Tyler Pittman, Kara L. Webster, Jiali Shang, and Stephanie A. Nelson. 2025. “Inferential Approach for Evaluating the Association Between Land Cover and Soil Carbon in Northern Ontario.” *Earth* 6(1): 1–21. doi:10.3390/earth6010001.

Quan, Ying, Mingze Li, Yuanshuo Hao, Jianyang Liu, and Bin Wang. 2023. “Tree Species Classification in a Typical Natural Secondary Forest Using UAV-Borne LiDAR and Hyperspectral Data.” *GIScience and Remote Sensing* 60(1): 1–17. doi:10.1080/15481603.2023.2171706.

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

Rappaport, Lance M., Ananda B. Amstadter, and Michael C. Neale. 2020. “Model Fit Estimation for Multilevel Structural Equation Models.” *Structural Equation Modeling* 27(2): 318–29. doi:10.1080/10705511.2019.1620109.

Richardson, John T.E. 2011. “Eta Squared and Partial Eta Squared as Measures of Effect Size in Educational Research.” *Educational Research Review* 6(2): 135–47. doi:10.1016/j.edurev.2010.12.001.

Rosseel, Yves. 2012. “Lavaan: An R Package for Structural Equation Modeling.” *Journal of Statistical Software* 48(2): 1–36. doi:10.18637/jss.v048.i02.

Roulston, Mark S., and Leonard A. Smith. 2002. “Evaluating Probabilistic Forecasts Using Information Theory.” *Monthly Weather Review* 130(6): 1653–60. doi:10.1175/1520-0493(2002)130<1653:EPFUIT>2.0.CO;2.

Rowlandson, Tracy L., Aaron A. Berg, Paul R. Bullock, E. RoTimi Ojo, Heather McNairn, Grant Wiseman, and Michael H. Cosh. 2013. “Evaluation of several calibration procedures for a portable soil moisture sensor.” *Journal of Hydrology* 498: 335–344. doi:10.1016/j.jhydrol.2013.05.021.

Sage, Andrew J., Ulrike Genschel, and Dan Nettleton. 2020. “Tree Aggregation for Random Forest Class Probability Estimation.” *Statistical Analysis and Data Mining* 13(2): 134–50. doi:10.1002/sam.11446.

Salazar, Alejandro, Begoña Ojeda, María Dueñas, Fernando Fernández, and Inmaculada Failde. 2016. “Simple Generalized Estimating Equations (GEEs) and Weighted Generalized Estimating Equations (WGEEs) in Longitudinal Studies with Dropouts: Guidelines and Implementation in R.” *Statistics in Medicine* 35(19): 3424–48. doi:10.1002/sim.6947.

Sheeren, David, Mathieu Fauvel, Veliborka Josipovíc, Maïlys Lopes, Carole Planque, Jérôme Willm, and Jean François Dejoux. 2016. “Tree Species Classification in Temperate Forests Using Formosat-2 Satellite Image Time Series.” *Remote Sensing* 8(9): 1–29. doi:10.3390/rs8090734.

Shimada, Masanobu, Takuya Itoh, Takeshi Motooka, Manabu Watanabe, Tomohiro Shiraishi, Rajesh Thapa, and Richard Lucas. 2014. “New Global Forest/Non-Forest Maps from ALOS PALSAR Data (2007-2010).” *Remote Sensing of Environment* 155: 13–31. doi:10.1016/j.rse.2014.04.014.

Simard, Patrice Y., Dave Steinkraus, and John C. Platt. 2003. “Best Practices for Convolutional Neural Networks Applied to Visual Document Analysis.” *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR* 2003-Janua: 958–63. doi:10.1109/ICDAR.2003.1227801.

Simfukwe, P., P. W. Hill, B. A. Emmett, and D. L. Jones. 2011. “Soil Classification Provides a Poor Indicator of Carbon Turnover Rates in Soil.” *Soil Biology and Biochemistry* 43(8): 1688–96. doi:10.1016/j.soilbio.2011.04.014.

Sindi, Hatem, Majid Nour, Muhyaddin Rawa, Şaban Öztürk, and Kemal Polat. 2021. “Random Fully Connected Layered 1D CNN for Solving the Z-Bus Loss Allocation Problem.” *Measurement: Journal of the International Measurement Confederation* 171(November 2020): 1–8. doi:10.1016/j.measurement.2020.108794.

Smith, Robert J., Sarah Jovan, Andrew N. Gray, and Bruce McCune. 2017. “Sensitivity of Carbon Stores in Boreal Forest Moss Mats - Effects of Vegetation, Topography and Climate.” *Plant and Soil* 421(1/2): 31–42. doi:10.1007/s11104-017-3411-x.

Southee, Florence Margaret, Paul M. Treitz, and Neal A. Scott. 2012. “Application of Lidar Terrain Surfaces for Soil Moisture Modeling.” *Photogrammetric Engineering and Remote Sensing* 78(12): 1241–51. doi:10.14358/PERS.78.11.1241.

Spatial Data Infrastructure. 2013. *Provincial Digital Elevation Model Technical Specifications v3.0*. Peterborough, Canada.

Stage, A. R. 1976. “An Expression for the Effect of Aspect, Slope, and Habitat Type on Tree Growth.” *Forest Science* 22(4): 457–60.

Strahler, A. H., T. L. Logan, and N. A. Bryant. 1978. “Improving Forest Cover Classification Accuracy Form Landsat by Incorporating Topographic Information.” In *Proceedings of 12th International Symposium on Remote Sensing of the Environment*, 927–42.

Suleymanov, Azamat, Dominique Arrouays, and Igor Savin. 2024. “Digital Soil Mapping in the Russian Federation: A Review.” *Geoderma Regional* 36: 1–15. doi:10.1016/j.geodrs.2024.e00763.

Sumathi, S., and S. Esakkirajan. 2007. “Structured Query Language.” In *Fundamentals of Relational Database Management Systems*, Springer Berlin, Heidelberg, 111–212.

Suratno, Agus, Carl Seielstad, and Lloyd Queen. 2009. “Tree Species Identification in Mixed Coniferous Forest Using Airborne Laser Scanning.” *ISPRS Journal of Photogrammetry and Remote Sensing* 64(6): 683–93. doi:10.1016/j.isprsjprs.2009.07.001.

Svetnik, Vladimir, Andy Liaw, Christopher Tong, J. Christopher Culberson, Robert P. Sheridan, and Bradley P. Feuston. 2003. “Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling.” *Journal of Chemical Information and Computer Sciences* 43(6): 1947–58. doi:10.1021/ci034160g.

Takeshige, Ryuichi, Kyaw Kyaw Htoo, Masanori Onishi, Farhadur Md Rahman, Kazuhiko Hoshizaki, Hideyuki Ida, Masae Iwamoto Ishihara, et al. 2025. “High-Resolution Digital Canopy Height Models, Terrain Models, Ortho-Mosaic Photos, and Canopy Tree Crown Shapes Derived from UAV-Borne LiDAR at 22 Tree Census Plots across Japanese Natural Forests.” *Ecological Research* 40: 657–70. doi:10.1111/1440-1703.12555.

Tan, Kim H. “ERRORS, VARIABILITY, AND CHEMICAL UNITS OF SOIL ANALYSIS.” *Soil Sampling, Preparation, and Analysis*, 2nd edition, CRC Press, 2005, 42–81.

Thomas, R. L., and R. C. Anderson. 1993. “Influence of Topography on Stand Composition in a Midwestern Ravine Forest.” *The American Midland Naturalist* 130(1): 1–12. https://www.jstor.org/stable/2426270.

Utla, Chandra Sekhar, Ajay Dashora, Rakesh Kumar Mishra, and Yun Zhang. 2025. “A Review on Aboveground Biomass Estimation Methods Utilizing Forest Structural Characteristics.” *International Journal of Remote Sensing* 46(16): 5917–37. doi:10.1080/01431161.2025.2524082.

Vargha, Andres, and Harold D. Delaney. 1998. “The Kruskal-Wallis Test and Stochastic Homogeneity.” *Journal of Educational and Behavioral Statistics* 23(2): 170–92. doi:10.3102/10769986023002170.

Vesterdal, Lars, Nicholas Clarke, Bjarni D. Sigurdsson, and Per Gundersen. 2013. “Do Tree Species Influence Soil Carbon Stocks in Temperate and Boreal Forests?” *Forest Ecology and Management* 309: 4–18. doi:10.1016/j.foreco.2013.01.017.

Wang, Kepu, Tiejun Wang, and Xuehua Liu. 2018. “A Review: Individual Tree Species Classification Using Integrated Airborne LiDAR and Optical Imagery with a Focus on the Urban Environment.” *Forests* 10(1): 1–18. doi:10.3390/f10010001.

Were, Kennedy, Dieu Tien Bui, Øystein B. Dick, and Bal Ram Singh. 2015. “A Comparative Assessment of Support Vector Regression, Artificial Neural Networks, and Random Forests for Predicting and Mapping Soil Organic Carbon Stocks across an Afromontane Landscape.” *Ecological Indicators* 52: 394–403. doi:10.1016/j.ecolind.2014.12.028.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. “Dplyr: A Grammar of Data Manipulation.” https://dplyr.tidyverse.org.

Williams, Michael S., Hans T. Schreuder, and Raymond L. Czaplewski. 2001. “Accuracy and efficiency of area classifications based on tree tally.” *Canadian Journal of Forest Research* 31(3): 556–560. doi:10.1139/cjfr-31-3-556.

Wu, Xiangyuan, Kening Wu, Huafu Zhao, Shiheng Hao, and Zhenyu Zhou. 2023. “Impact of Land Cover Changes on Soil Type Mapping in Plain Areas: Evidence from Tongzhou District of Beijing, China.” *Land* 12(9): 1–20. doi:10.3390/land12091696.

Wu, Yun Chun, Ming Chieh Shih, and Yu Kang Tu. 2021. “Using Normalized Entropy to Measure Uncertainty of Rankings for Network Meta-Analyses.” *Medical Decision Making* 41(6): 706–13. doi:10.1177/0272989X21999023.

Yan, Jun, and Jason Fine. 2004. “Estimating Equations for Association Structures.” *Statistics in Medicine* 23(6): 859–74. doi:10.1002/sim.1650.

Zeger, Scott L., and Kung-Yee Liang. 1986. “Longitudinal Data Analysis for Discrete and Continuous Outcomes.” *Biometrics* 42(1): 121–30.

Zeger, Scott L., Kung-Yee Liang, and Paul S. Albert. 1988. “Models for Longitudinal Data: A Generalized Estimating Equation Approach.” *Biometrics* 44(4): 1049–1060.

Zhang, Lei, Haizhen Wang, Jianfen Guo, Weisheng Lin, Dafeng Hui, Xianfang Zhong, Josep Peñuelas, and Ji Chen. 2025. “Forestation of Abandoned and Farmed Lands Enhances Soil Microbial Biomass Carbon, Nitrogen and Phosphorus Compared to Grassed and Forested Lands: A Meta-analysis.” *Journal of Applied Ecology* 62(9): 2296–305 doi:10.1111/1365-2664.70109.

Zhang, Xinmin, Takuya Wada, Koichi Fujiwara, and Manabu Kano. 2020. “Regression and Independence Based Variable Importance Measure.” *Computers and Chemical Engineering* 135: 1–6. doi:10.1016/j.compchemeng.2020.106757.

Zhao, Naifei, Qingsong Xu, and Hong Wang. 2017. “Marginal Screening for Partial Least Squares Regression.” *IEEE Access* 5: 14047–55. doi:10.1109/ACCESS.2017.2728532.

Zhu, A. X. 1997. “Measuring Uncertainty in Class Assignment for Natural Resource Maps under Fuzzy Logic.” *Photogrammetric Engineering and Remote Sensing* 63(10): 1195–1202.

Zhu, Xiaolin, and Desheng Liu. 2014. “Accurate Mapping of Forest Types Using Dense Seasonal Landsat Time-Series.” *ISPRS Journal of Photogrammetry and Remote Sensing* 96: 1–11. doi:10.1016/j.isprsjprs.2014.06.012.

Appendix A: Soil Data

Table A.1. Summary of soil properties from soil samples extracted during field campaigns in September 2018, August 2019 and August–September 2021. These numbers correspond to the samples that were processed and analyzed. Note that bulk density (BD) was only assessed on samples extracted from the 0–5 cm and 5–15 cm mineral soil depth layers. Lower sample counts for clay were due to insufficient quantities of material extracted for some samples. This table was also posted in appendix of publication (Pittman et al. 2025) for Chapter 2.

Soil Property		**Summary**	**Depth Layer**				
			0–5 cm	**5–15 cm**	**15–30 cm**	**30–45 cm**	**90–105 cm**
C	[%]	Number of Samples (*n*)	263	431	106	115	29
		Mean	10.3	9.9	6.5	4.7	4.2
		Standard Deviation	9.9	12.9	11.0	9.0	0.8
		Maximum	47.5	91.7	47.8	48.6	5.6
		Minimum	1.6	0.6	0.6	0.4	1.9
C:N		Number of Samples (*n*)	263	431	106	112	29
		Mean	15.7	16.2	13.1	25.1	170.2
		Standard Deviation	7.1	6.4	18.2	18.2	91.6
		Maximum	75.3	43.6	170.5	121.3	430.7
		Minimum	3.4	2.1	1.9	6.1	32.6
Clay	[%]	Number of Samples (*n*)	229	424	99	106	23
		Mean	30.3	25.9	44.9	41.3	46.1
		Standard Deviation	11.7	12.9	15.8	13.6	17.8
		Maximum	54.6	68.6	78.6	68.1	70.2
		Minimum	6.6	1.0	11.4	2.0	7.6
BD	[g/cm^3]	Number of Samples (*n*)	431	431			
		Mean	1.03	1.21			
		Standard Deviation	0.55	0.56			
		Maximum	2.44	2.55			
		Minimum	0.10	0.13			

Appendix B: Predictors

Table B.1. Environmental covariates utilized as features for predictors with modeling for the tree species classification work presented in Chapter 3.

Predictor	**Source Type**
Canopy Height Model (CHM) 2017	LiDAR
Channel Network Base Level	LiDAR (DTM)
Channel Network Distance	LiDAR (DTM)
Closed Depressions	LiDAR (DTM)
Multi-resolution Ridge Top Flatness (MRRTF)	LiDAR (DTM)
Multi-resolution Valley Bottom Flatness (MRVBF)	LiDAR (DTM)
Negative Openness	LiDAR (DTM)
Relative Slope Position	LiDAR (DTM)
Topographic Wetness Index (TWI)	LiDAR (DTM)
Total Catchment Area	LiDAR (DTM)
Valley Depth	LiDAR (DTM)
B1 (Coastal) Aug.10 2014	SR (WorldView-2)
B2 (Blue) Aug.10 2014	SR (WorldView-2)
B3 (Green) Aug.10 2014	SR (WorldView-2)
B4 (Yellow) Aug.10 2014	SR (WorldView-2)
B5 (Red) Aug.10 2014	SR (WorldView-2)
B6 (Red Edge) Aug.10 2014	SR (WorldView-2)
B7 (Near Infrared I (NIR1)) Aug.10 2014	SR (WorldView-2)
B8 (Near Infrared II (NIR2)) Aug.10 2014	SR (WorldView-2)
Normalized Difference Vegetation Index (NDVI) Aug.10 2014	SR (WorldView-2)
Modified Normalized Difference Water Index (NDVI) Aug.10 2014	SR (WorldView-2)
Panchromatic Aug.10 2014	SR (WorldView-2)

Table B.2. Environmental covariates utilized as predictors for digital soil modeling work, as used in part for modeling in Chapter 4 and as full set of predictors in Chapter 5.

Predictor	**Source Type**
Digital Elevation Model (DEM)	LiDAR
Canopy Height Model (CHM) 2017	LiDAR
Gap Fraction 2017	LiDAR
Aspect	DEM (LiDAR)
Convergence Index	DEM (LiDAR)
General Curvature	DEM (LiDAR)
Mid Slope Position	DEM (LiDAR)
Multi-resolution Ridge Top Flatness (MRRTF)	DEM (LiDAR)
Multi-resolution Valley Bottom Flatness (MRVBF)	DEM (LiDAR)
Plan Curvature	DEM (LiDAR)
Profile Curvature	DEM (LiDAR)
SAGA Topographic Wetness Index (SAGA TWI)	DEM (LiDAR)
Sky View Factor	DEM (LiDAR)
Slope	DEM (LiDAR)
Slope Height	DEM (LiDAR)
Slope Length	DEM (LiDAR)
Standardized Height	DEM (LiDAR)
Stream Power Index	DEM (LiDAR)
Terrain Ruggedness Index	DEM (LiDAR)
Terrain View Factor	DEM (LiDAR)
Topographic Wetness Index (TWI)	DEM (LiDAR)
Total Curvature	DEM (LiDAR)
Valley Depth	DEM (LiDAR)
Visible Sky	DEM (LiDAR)
B1 (Violet) Summer (Jun.–Aug.) 2017	SR (Landsat 8)
B2 (Blue) Summer (Jun.–Aug.) 2017	SR (Landsat 8)
B3 (Green) Summer (Jun.–Aug.) 2017	SR (Landsat 8)
B4 (Red) Summer (Jun.–Aug.) 2017	SR (Landsat 8)
B5 (Near Infrared (NIR)) Summer (Jun.–Aug.) 2017	SR (Landsat 8)
B6 (Shortwave Infrared I (SWIR I)) Summer (Jun.–Aug.) 2017	SR (Landsat 8)
B7 (Shortwave Infrared II (SWIR II)) Summer (Jun.–Aug.) 2017	SR (Landsat 8)
B10 (Thermal Infrared I (TIR I)) Summer (Jun.–Aug.) 2017	SR (Landsat 8)
B11 (Thermal Infrared II (TIR II)) Summer (Jun.–Aug.) 2017	SR (Landsat 8)

B1 (Violet) Winter (Dec.–Feb.) 2017	SR (Landsat 8)
B2 (Blue) Winter (Dec.–Feb.) 2017	SR (Landsat 8)
B3 (Green) Winter (Dec.–Feb.) 2017	SR (Landsat 8)
B4 (Red) Winter (Dec.–Feb.) 2017	SR (Landsat 8)
B5 (Near Infrared (NIR)) Winter (Dec.–Feb.) 2017	SR (Landsat 8)
B6 (Shortwave Infrared I (SWIR I)) Winter (Dec.–Feb.) 2017	SR (Landsat 8)
B7 (Shortwave Infrared II (SWIR II)) Winter (Dec.–Feb.) 2017	SR (Landsat 8)
B1 (Violet) Fall (Sep.10–Oct.10) 2015–2019	SR (Landsat 8)
B2 (Blue) Fall (Sep.10–Oct.10) 2015–2019	SR (Landsat 8)
B3 (Green) Fall (Sep.10–Oct.10) 2015–2019	SR (Landsat 8)
B4 (Red) Fall (Sep.10–Oct.10) 2015–2019	SR (Landsat 8)
B5 (Near Infrared (NIR)) Fall (Sep.10–Oct.10) 2015–2019	SR (Landsat 8)
B6 (Shortwave Infrared I (SWIR I)) Fall (Sep.10–Oct.10) 2015–2019	SR (Landsat 8)
B7 (Shortwave Infrared II (SWIR II)) Fall (Sep.10–Oct.10) 2015–2019	SR (Landsat 8)
B1 (Violet) May 2015–2019	SR (Landsat 8)
B2 (Blue) May 2015–2019	SR (Landsat 8)
B3 (Green) May 2015–2019	SR (Landsat 8)
B4 (Red) May 2015–2019	SR (Landsat 8)
B5 (Near Infrared (NIR)) May 2015–2019	SR (Landsat 8)
B6 (Shortwave Infrared I (SWIR I)) May 2015–2019	SR (Landsat 8)
B7 (Shortwave Infrared II (SWIR II)) May 2015–2019	SR (Landsat 8)
Normalized Difference Vegetation Index (NDVI) Summer (Jun.–Aug.) 2017	SR (Landsat 8)
Normalized Difference Vegetation Index (NDVI) Winter (Dec.–Feb.) 2017	SR (Landsat 8)
Normalized Difference Vegetation Index (NDVI) Fall (Sep.10–Oct.10) 2017	SR (Landsat 8)
Normalized Difference Vegetation Index (NDVI) May 2017	SR (Landsat 8)
Normalized Difference Built-up Index (NDBI) Summer (Jun.–Aug.) 2017	SR (Landsat 8)
Normalized Difference Built-up Index (NDBI) Winter (Dec.–Feb.) 2017	SR (Landsat 8)
Normalized Difference Built-up Index (NDBI) Fall (Sep.10–Oct.10) 2017	SR (Landsat 8)
Normalized Difference Built-up Index (NDBI) May 2017	SR (Landsat 8)
Normalized Difference Water Index (NDWI) Summer (Jun.–Aug.) 2017	SR (Landsat 8)
Normalized Difference Water Index (NDWI) Winter (Dec.–Feb.) 2017	SR (Landsat 8)
Normalized Difference Water Index (NDWI) Fall (Sep.10–Oct.10) 2017	SR (Landsat 8)
Normalized Difference Water Index (NDWI) May 2017	SR (Landsat 8)
Modified Normalized Difference Water Index (MNDWI) Summer (Jun.–Aug.) 2017	SR (Landsat 8)
Modified Normalized Difference Water Index (MNDWI) Winter (Dec.–Feb.) 2017	SR (Landsat 8)
Modified Normalized Difference Water Index (MNDWI) Fall (Sep.10–Oct.10) 2017	SR (Landsat 8)
Modified Normalized Difference Water Index (MNDWI) May 2017	SR (Landsat 8)

Change Magnitude B1 B2 B3 B4 B5 B6 B7 Summer (Jun.–Aug.) 1984–2011 (Blue, Green, Red, Near Infrared I (NIR I), NIR II, Thermal, Mid Infrared)	SR (Landsat 5)
Change Magnitude B3 (Red) B4 (NIR I) B5 (NIR II) Summer (Jun.–Aug.) 1984–2011	SR (Landsat 5)
Change Magnitude B3 (Red) B4 (NIR I) B5 (NIR II) Summer (Jun.–Aug.) 1984–2005	SR (Landsat 5)
Change Magnitude B3 (Red) B4 (NIR I) B5 (NIR II) Summer (Jun.–Aug.) 1984–1995	SR (Landsat 5)
Change Magnitude B3 (Red) B4 (NIR I) B5 (NIR II) Summer (Jun.–Aug.) 1995–2011	SR (Landsat 5)
Change Magnitude B3 (Red) B4 (NIR I) B5 (NIR II) Summer (Jun.–Aug.) 1995–2005	SR (Landsat 5)
Change Magnitude B3 (Red) B4 (NIR I) B5 (NIR II) Summer (Jun.–Aug.) 2005–2011	SR (Landsat 5)
L-band HH 2017	SAR (PALSAR)
C-band VH May 2017	SAR (Sentinel-1)
C-band VV May 2017	SAR (Sentinel-1)
Magnetic Residual Nov. 2018	Aeromagnetic
Magnetic Residual Vertical Derivative (VD) Nov. 2018	Aeromagnetic
Gravity Anomaly 2016	Aeromagnetic
Gravity Anomaly Vertical Derivative (VD) 2016	Aeromagnetic
NFI Total Dead Biomass 2011	k-NN Model
NFI Total Volume 2011	k-NN Model
NFI Stand Age 2011	k-NN Model
NFI Balsam Fir 2011	k-NN Model
NFI Black Spruce 2011	k-NN Model
Balsam Fir Dominant Indicator	RF Model
Black Spruce Dominant Indicator	RF Model
Coniferous Indicator	RF Model
Deciduous Indicator	RF Model
No Trees Indicator	RF Model