ESTIMATION OF SOIL BULK DENSITY AND CARBON USING MULTI-SOURCE REMOTELY SENSED DATA

R. Pittman, B. Hu *

Dept. of Earth and Space Science and Engineering, York University, 4700 Keele Street, Toronto ON M3J 1P3 Canada
(rpittman, baoxin)@yorku.ca

Commission III, WG III/10

KEY WORDS: Digital Soil Mapping, Bulk Density, Soil Carbon, Canopy Height Model, Gap Fraction, Landsat

ABSTRACT:

Bulk density and soil carbon models were fitted for soil samples collected during field campaigns in 2018 and 2019 for the Kapuskasing region of the District of Cochrane in Ontario, Canada. Prediction maps for bulk density and soil carbon were generated for the 0-15 cm depth mineral soil layer. The application of multi-source remotely sensed data as environmental covariates for model predictors was implemented. Environmental covariates were obtained from multispectral satellite imagery, LiDAR (light detection and ranging) retrievals and airborne geomagnetic surveys, as well from a digital elevation model (DEM) for topographic covariates. Two covariates derived from LiDAR, canopy height model (CHM) and gap fraction, were of high variable importance when fitting models for average bulk density; gap fraction had the highest to second highest variable importance for average bulk density when considered among a full set of 76, or reduced sets of 12 or 5 separate predictors respectively. Environmental covariates corresponding to vegetation cover, specifically reflectance from multispectral imagery or LiDAR data, had the highest variable importance when compared with other categories of soil formation factors. Random forest (RF) models were generated, with RF models based upon just 12 predictors obtaining reasonable results with coefficients of determinations (R^2) greater than 0.7 for the standard derivation of bulk density, standard deviation of total carbon and average total carbon for the 0-15 cm depth layer.

1. INTRODUCTION

Soil bulk density and carbon measurements are of relevance for many agricultural and environmental applications. Bulk densities can be correlated to soil textural and nutrient properties, and affect crop growth and yields by the constriction of root extensions through soil (Reynolds et al., 2002). Soil carbon is of prime significance for various environmental studies in regards to its role with the global carbon cycle (Grimm et al., 2008). Diverse biotopes within a region such as a forest could reveal differing concentrations of soil carbon and bulk density; the generation of prediction maps of such properties are of interest. The estimation of carbon stocks, as well as the evaluation of land settings suitable for agriculture, were motivations for this research.

Digital soil mapping (DSM) is the process of deriving prediction maps from models fitted for soil target variables using environmental covariates as model predictors. A model created for a target variable collected at a site level can generalize a prediction map of that target variable for the respective study area. The expense, effort and time considerations for obtaining soil samples from potentially remote or inaccessible locations can make the traditional means of collecting vast amounts of soil samples per surveying grid prohibitive. The variability of various soil properties over comparatively small spatial scales within certain environments, such as a forest consisting of different topographies and vegetation species, could render inaccurate results with conventional prediction maps based on kriging methodologies. On the contrary, DSM allows the extrapolation of soil properties for an area by recognizing environmental covariates as soil formation factors (Mulder et al., 2011) that over long time scales can affect the observed soil properties.

Recent methodologies for DSM incorporate a SCORPAN approach (McBratney et al., 2003) that relates model predictors to environmental covariates conforming to categories of soil formation factors. The environmental covariates correspond to soil, climate, organism (vegetation), relief (topography) and parent material categorical factors. Environmental covariates can be derived from various means such as a digital elevation model (DEM), geospatial vector files, multispectral satellite imagery, synthetic aperture radar (SAR) imagery, or light detection and ranging (LiDAR) data (Mulder et al., 2011). Topographic covariates such as slope, curvature and aspect can be computed from a DEM. Climatic and vegetative covariates are often acquired as products obtained from multispectral satellite imagery. The indices such as Normalized Difference Vegetation Index (NDVI) and Normalized Difference Water Index (NDWI), calculated from multispectral imagery, are commonly applied as vegetative covariates (Poggio et al., 2013; Yang et al., 2016). Parent material can comprise of underlying bedrock geology. Other soil properties, typically collected at a site level, or else from covariates that can be easily observed, can be utilized as predictors relating to soil features.

* Corresponding author

Machine-learning approaches are prevalent for DSM, due to their versatility of being able to be applied to data with predictors of different levels of measurement where linear or explicit model formulations are not possible. Random forest (RF) methods are popular for DSM applications (Brungard et al., 2015; Ließ et al., 2012; Nussbaum et al., 2018; Yang et al., 2016) and tend to have improved accuracy when compared with other approaches such as decision trees or ordinary regression techniques. RF can be applied to either categorical or continuous target variables, resulting in either classification or regression analysis, respectively.

The improvement of modeling accuracies for DSM is an active area of research. Coefficient of determination (R^2) values for models in recent studies vary from 0.02 to 0.27 and 0.07 to 0.35 for validation and calibration datasets respectively for soil organic carbon in France (Mulder et al., 2016), to 0.45 for organic carbon in the A horizon for a study region in the Argentine Pampas (Angelini et al., 2016). Additional studies reported R^2 values from between 0.20 to 0.55 for soil textural components in Denmark (Adhikari et al., 2013), to between 0.11 to 0.36 for soil organic carbon for rangeland in eastern Australia (Wang et al., 2018). Discovering environmental covariates that enhance model accuracy is of interest. Many studies focus on topographic covariates primarily generated from a DEM (Adhikari et al., 2013; Brungard et al., 2015; Mulder et al., 2016) as being the most important for DSM. However, for homogeneous study areas of relatively flat terrain, topographic covariates may fail to have significance. Supplementary categories of covariates might be of more relevance, in particular covariates that relate to vegetation cover. Multi-source remotely sensed data can provide additional environmental covariates. Landsat imagery obtained for different periods of the year can present information pertaining to vegetation cover. LiDAR data attained from airborne campaigns can provide information at finer spatial resolutions that can be utilized to derive topographic and vegetative covariates. Such vegetative covariates include canopy height model (CHM) and gap fraction. To our understanding, CHM and gap fraction for an entire study area have not been applied as predictors in prior DSM research. LiDAR data were available for our study area for the soil sampling locations, and hence were utilized for model input.

The objective of this research was the identification of predictors, particularly of a non-topographic type, with higher variable importance for a relatively flat and homogenous land setting. We implemented multi-source remotely sensed data for predictors, with the goal of maximizing model accuracy. Prediction maps of soil bulk density and carbon were generated to distinguish areas or patterns between land features and corresponding soil properties.

2. STUDY AREA AND SOIL DATA

Soil samples were collected in the vicinity of the community of Kapuskasing in the District of Cochrane in Ontario (ON), Canada. For comparison purposes, additional soil samples for 3 sites were collected near Hearst, ON about 100 km to the west of Kapuskasing. The Kapuskasing study area consists of approximately 550 km^2 and is bounded from the latitudes of 49.35° N to 49.55° N and the longitudes of 82.25° W to 82.75° W, fitting within the boundaries for which airborne LiDAR retrievals are available. This region is relatively flat; from the DEM the minimum and maximum elevations are 200 m and 269 m, respectively. The locations for the soil sampling locations for the study area are shown in Figure 1. This region is colloquially referred to as the Great Clay Belt (GCB), due to the presence of heavy clay in the lower soil horizons throughout the region, typically encountered at depths below 20 cm. This is a mostly forested region within the boreal forest biome, with black spruce (*Picea mariana*), white spruce (*Picea glauca*), balsam fir (*Abies balsamea*) and trembling aspen (*Populus tremuloides*) as dominant tree species. Agricultural land exists south and west of Kapuskasing.

Soil samples were collected during two field campaigns, in September 2018 and August 2019. A total of 34 sites, each consisting of 3 subplots with soil samples from various depths (0-5 cm, 5-15 cm, 15-30 cm, 30-45 cm, 90-105 cm) were obtained. This corresponded to 102 soil samples per profile depth layer for this analysis. The 3 subplots per site corresponded to 0°, 120° and 240° bearings at distances of 4.5 m, 7.5 m and 9.5 m from the site center respectively, with the soil samples obtained within a 2 m radius of each subplot bearing location. Bulk density samples were obtained for each of the top 30 cm depth layers, and soil chemistry samples were obtained for all depth layers. Sites from a variety of local land cover types were sampled, ranging from peat bogs, old growth forest to pasture and summer fallow agricultural fields, in order to attain maximum variation. The sites that were sampled within a km of one another corresponded to different land cover types.

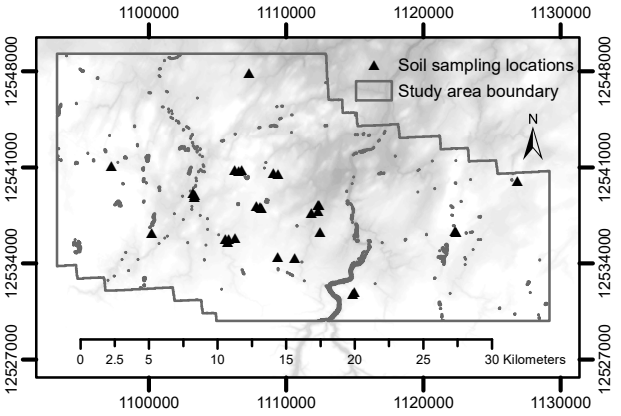

Figure 1. The study area for the soil sampling locations near Kapuskasing, ON.

Bulk density samples were processed to obtain the mass of mineral soil per reference volume, minus the total mass of rocks and coarse fragments present in the bulk density samples. The bulk density samples were dried for a minimum of 3 months, then each sample was baked in a tin at 110° C for a minimum of 48 hours. The masses of the baked soil samples were obtained separately when each sample was taken directly out of the soil oven, and then after each sample was subsequently pulverized with a mortar and then sieved through a 2 mm mesh to leave out the rocks and coarse fragments. Chemistry sample analysis consisted of total carbon and total nitrogen contents, exchangeable phosphorus and exchangeable potassium, acidity test and soil texture components. Combustion methods were utilized to obtain the total carbon and total nitrogen contents, obtained as a percentage of mass for each soil sample. The bulk density samples required less complex procedures for processing, and subsequently were processed before the chemistry samples. This led to an

interest in analyzing bulk densities, as these results were more readily available.

The greatest variability in soil properties occurred in the top 30 cm, as heavy clay was ubiquitous for the greater depths. Of interest was to determine soil carbon contents for the mineral soil depths where carbon contents were maximum, which for this study corresponded to the 0-5 cm and 5-15 cm profile depths. Other studies have focused on surface mineral soil layers for soil carbon contents (Gomez et al., 2008; Grimm et al., 2008). Correspondingly, bulk densities had the most variation in the surface layers, so the shallowest depth layers were also pertinent for the bulk density analysis. Subsequently, it was decided to focus on the 0-15 cm profile depth for the DSM modelling. The quantities for the 0-15 cm depth layer were obtained as the weighted averages of the 0-5 cm and 5-15 cm layers.

Averages and standard deviations were calculated for the bulk densities and total carbon (here also referred to as carbon), respectively, as the average and standard deviation of the 3 subplot values per site for the corresponding layer. The standard deviation calculations takes into account the variation among the different subplots within a site, as subplots could lie underneath different vegetation within the same site location, and subsequently have different soil properties.

3. METHODOLOGY

The environmental covariates were obtained from multispectral satellite imagery, a DEM, aeromagnetic surveys and LiDAR data. Distances and densities to water bodies were derived from L-band SAR imagery, and a geospatial vector file was utilized for underlying bedrock geology. A total of 76 separate environmental covariates were considered for the full set of predictors, which was applied for both the bulk density and total carbon models. From the full set, reduced sets (i.e. 12 or 5 predictors with higher variable importance) were selected, which are shown in Table 1. Note for these reduced lists of predictors that the surface reflectance correspond to bands from Landsat-8.

Environmental covariates for vegetation cover were derived from Landsat multispectral imagery obtained from Google Earth. The Landsat imagery was provided by the United States Geological Survey (USGS); Landsat-8 imagery was utilized for imagery corresponding to 2012 and after (i.e. 2017), and Landsat-5 imagery was utilized for previous years (i.e. 1984, 1995 and 2005). In total, 57 predictors derived from Landsat imagery were utilized in the full predictor set. Median surface reflectance values were calculated for corresponding bands for cloud-free (i.e. less than 1% cloud cover) scenes obtained for 4 periods of the years. These periods consisted of winter (January to March) for dormancy, spring (May), summer (June and July) for peak-vegetation, and autumn (September 10th to October 10th). NDVI and NDWI for the different periods were also derived. A Provincial DEM for 2016 created by the Ontario Ministry of Natural Resources & Forestry (MNRF) obtained via Land Information Ontario (LIO) was utilized to generate covariates for slope and curvature at various resolutions (30 m, 150 m, 300 m, 900 m) to account for the contour of the local topography. Aspect and hillshade covariates were also created from the DEM. The Landsat imagery and DEM were attained at 30 m spatial resolutions. Aeromagnetic data was obtained from the Canadian Aeromagnetic Data Base created by Natural Resources Canada (NRCan). Gravity anomaly and magnetic residual of total field from airborne surveys, current as of 2016 and November 2018 respectively, were utilized as covariates for bedrock material. The vertical derivatives for gravity anomaly and magnetic residual of total field were also applied as predictors. Bedrock geology vector files supplied from NRCan were also utilized as covariates for parent material.

AVG BD (12 Predictors)	AVG BD (5 Predictors)	STDDEV BD (12 Predictors)
Slope	CHM	Slope 150m
Curvature 150m	Gap Fraction	Curvature 150m
CHM	B2 Winter 2017	CHM
Gap Fraction	B3 Winter 2017	Gap Fraction
B2 Winter 2017	NDWI Winter 2017	B1 May *
B2 May *		B3 Winter 2017
B3 Winter 2017		B4 Summer 2017
B3 Summer 2017		B6 Autumn *
B4 Winter 2017		NDVI May *
NDVI May *		NDVI Summer 2017
NDWI Winter 2017		NDWI May *
NDWI May *		NDWI Summer 2017

AVG C (12 Predictors)	AVG C (5 Predictors)	STDDEV C (12 Predictors)
Aspect	Aspect	Slope
Curvature 150m	B1 May *	CHM
CHM	B2 Winter 2017	B1 Winter 2017
B1 Winter 2017	B3 Winter 2017	B1 May *
B1 May *	NDWI Winter 2017	B2 Winter 2017
B2 Winter 2017		B2 May *
B2 May *		B3 Winter 2017
B3 Winter 2017		B3 Summer 2017
B11 Summer 2017		B6 Autumn *
NDVI Winter 2017		B7 Winter 2017
NDVI May *		NDVI Winter 2017
NDWI Winter 2017		NDVI May *

* Median surface reflectance for same period over 5 years (2015-2019)

Table 1. Reduced sets of predictors utilized for the RF models.

The Ontario MNRF collected LiDAR data, which was obtained via LIO for the District of Cochrane in northern Ontario. LiDAR-derived covariates included CHM, computed as the difference between the recorded maximum and minimum elevations of LiDAR retrievals per pixelized cell. Correspondingly, gap fraction was calculated as the fraction of LiDAR retrievals with only one return, to the total number of retrievals per pixelized cell. The LiDAR data were obtained by an airplane survey during the autumn of 2016, collected with an average density of 8 points per m^2. CHM and gap fraction were initially calculated for spatial resolutions of 10 m, and then were subsequently resampled to the common cell size of 30 m utilized for covariate layer integration purposes.

RF models were fitted separately for the average bulk density and standard deviation bulk density, first upon a set of 76 covariates and then separately upon 12 covariates with higher variable importance. Reasonable RF models with just 5 covariates were obtained for the average bulk density and average carbon, individually. The variable importance was determined from the random forest models, with the percent

inclusion of mean squared error as the metric. Due to the limited amount of soil sampling sites available, the full set of sites were utilized for model training. Default settings of 500 trees were applied for the RF fittings reported; comparable values were obtained when 1000 trees were specified. The caret package in R (Kuhn, 2008) was utilized for the modeling. Five-fold cross validation, with 3 repeats was utilized for fitting the RF models. Metrics for model assessment were the coefficient of determination (R^2) and the mean absolute error (MAE).

4. RESULTS AND ANALYSIS

The accuracies for the models are summarized in Table 2. These R^2 and MAE values are based upon model training. Maximum R^2 values ranged from just below 0.3 for average bulk density, to greater than 0.7 for average carbon. The R^2 values were even higher for the standard deviation of those properties, exceeding 0.8 for both bulk density and carbon. These accuracies in terms of the coefficient of determination are comparable or exceed accuracies for DSM models reported in recent literature (Nussbaum et al., 2018), even for evaluations based upon model calibration (Mulder et al., 2016). Note that due to the low number of soil sampling sites in this study, that data was not retained explicitly for model validation, so data for all soil sampling sites were applied for model training.

Model		Bulk Density	
		R^2	MAE [g/cm^3]
Average Bulk Density	76 Predictors	0.28	0.19
(AVG BD)	12 Predictors	0.28	0.19
	5 Predictors	0.29	0.18
Bulk Density Standard Deviation	76 Predictors	0.92	0.04
(STDDEV BD)	12 Predictors	0.84	0.04
Model		**Carbon**	
		R^2	MAE [%]
Average Carbon	76 Predictors	0.88	2.59
(AVG C)	12 Predictors	0.73	2.57
	5 Predictors	0.67	2.65
Carbon Standard Deviation	76 Predictors	0.95	1.24
(STDDEV C)	12 Predictors	0.90	1.03

Table 2. Accuracies for RF models for each target variable (AVG BD, STDDEV BD, AVG C and STDDEV C) for the 0-15 cm depth layer.

Variable importance plots for the RF models for the average bulk density and average carbon contents were generated. These plots first considered the full set (i.e. 76) predictors, and then reduced sets for the most accurate model obtained for each based upon R^2 values from Table 2. Variable importance plots for average bulk density are shown in Figure 2, and for average carbon are shown in Figure 3. For average bulk density, on the reduced set of predictors, gap fraction had the highest variable importance with 9.67 % Inc MSE. On the full set of predictors, gap fraction was the second and CHM the fifteenth most important predictors, respectively. Covariates derived from LiDAR data improved the model accuracies. For the average carbon models, predictors from multispectral imagery corresponding to ultraviolet and blue wavelengths of reflectance during the winter and May had the higher variable importance. The topographic covariate of aspect had a higher variable importance than that of other topographic covariates. One can verify that environmental covariates relating to vegetation, in this case multi-sourced remotely sensed data from LiDAR and Landsat imagery, had the highest variable importance.

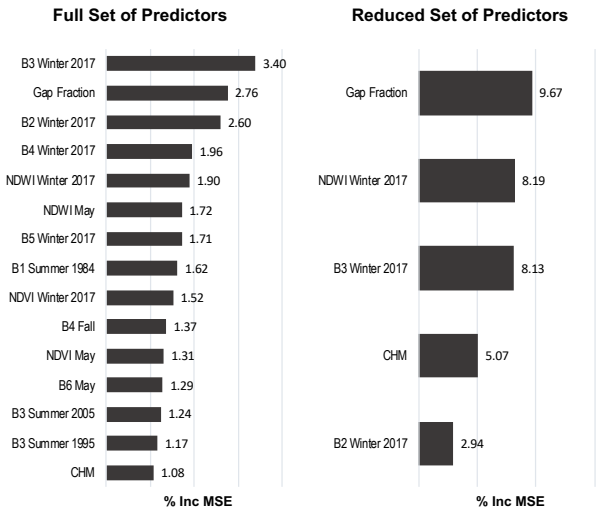

Figure 2. Variable importance of the predictors for the RF models for AVG BD for the 0-15 cm depth layer.

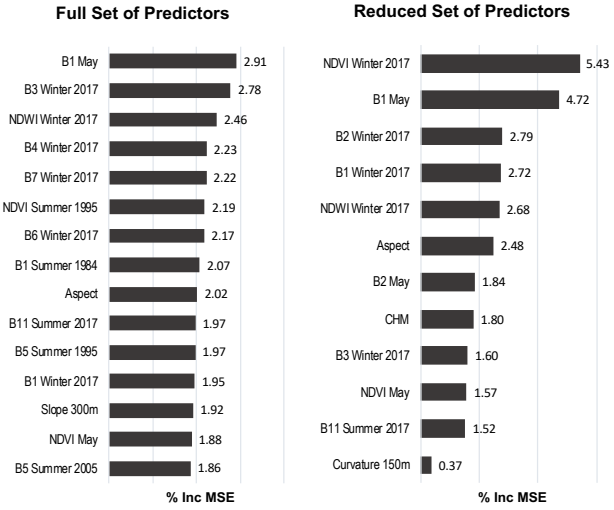

Figure 3. Variable importance of the predictors for the RF models for AVG C for the 0-15 cm depth layer.

Utilizing a full set of 76 predictors could be deemed too large given the sample size, so RF models were trained on smaller predictor sets. The prediction versus observation plots for the RF models trained on the corresponding reduced set of 12 predictors for each target variable for all 34 sites, are shown in Figure 4. The gray line denotes a perfect prediction, i.e. prediction is equal to observation value. Reasonable models in terms of R^2 were able to be fitted for the standard deviation values and the average total carbon, but the fit for average bulk density was not as robust. Prediction maps of the bulk density and average total carbon quantities for the RF models generated from the corresponding reduced sets of 12 predictors are shown in Figure 5.

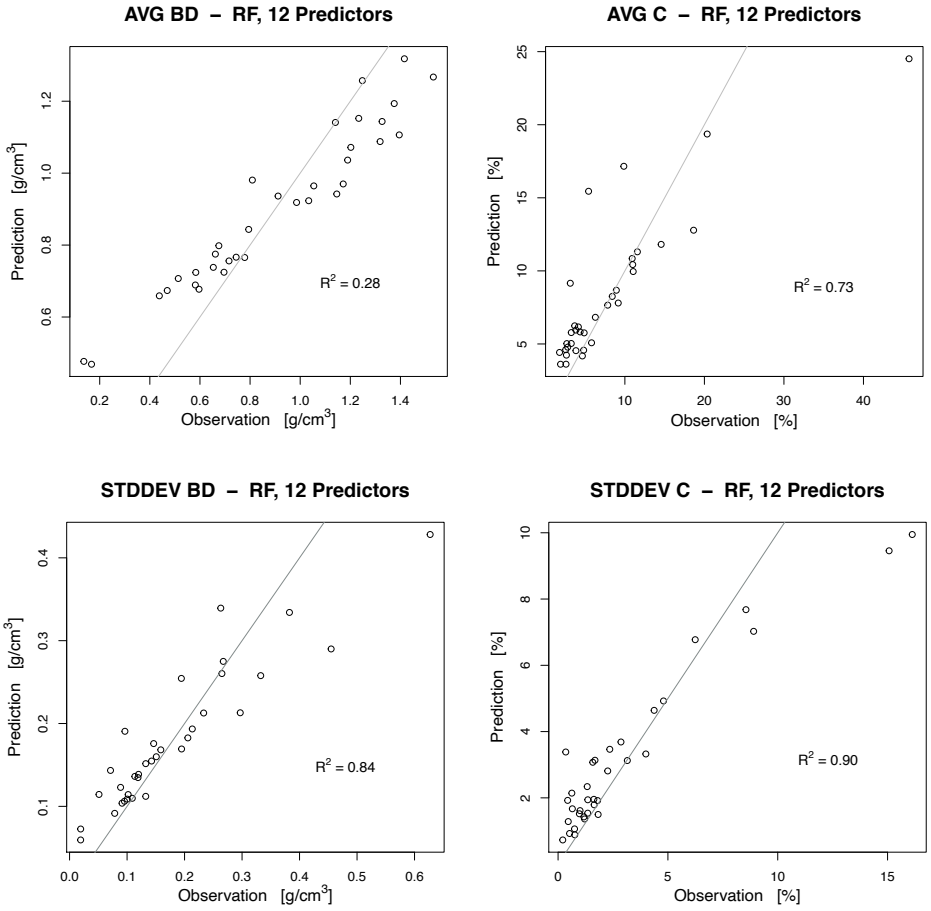

Figure 4. Accuracy plots for AVG BD, AVG C, STDDEV BD and STDDEV C for the 0-15 cm depth layer for the respective RF models corresponding to 12 predictors (see Table 2).

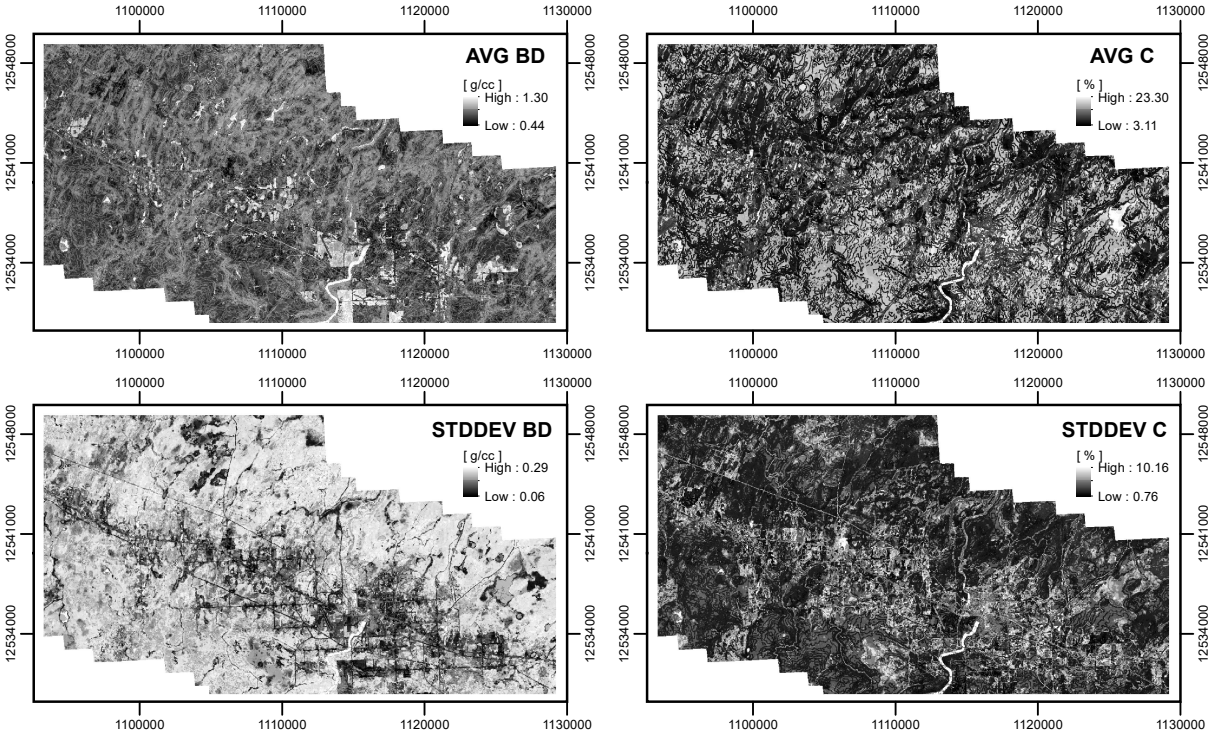

Figure 5. Prediction maps for AVG BD, AVG C , STDDEV BD and STDDEV C for the 0-15 cm depth layer for the respective RF models corresponding to 12 predictors.

From the prediction maps in Figure 5, one can notice certain patterns. Regions of higher average bulk densities correspond to agricultural land, whereas regions of lower average bulk densities correspond to forested areas. In particular, average bulk density has a minimum value to the northeast of the community of Kapuskasing, which corresponds to peatland. The standard deviation bulk density is lowest among the agricultural fields and the cleared land along roadways, but highest among the forested regions. Patterns for average carbon content are less obvious, but one can ascertain that average carbon is maximum in bog environments where average bulk density is minimum, as seen between the prediction maps for AVG BD and AVG C. Inversely to the standard deviation bulk density, the standard deviation of carbon is highest among the cleared regions, with the exception of the agricultural land where the standard deviation values are minimum.

A true-color composite of the study area is shown in Figure 6. This image was generated from Landsat-8 imagery obtained as median surface reflectance values for cloud-free days for June and July 2017. The bands B2 (0.452-0.512 µm), B3 (0.533-0.590 µm) and B4 (0.636-0.673 µm) were taken for blue, green and red, respectively. The roadways and cleared areas from this image are apparent. Regions with the brightest reflectance values correspond to the community of Kapuskasing and cropland fields. The area with the highest average carbon contents in Figure 5 correspond to wetland areas in Figure 6.

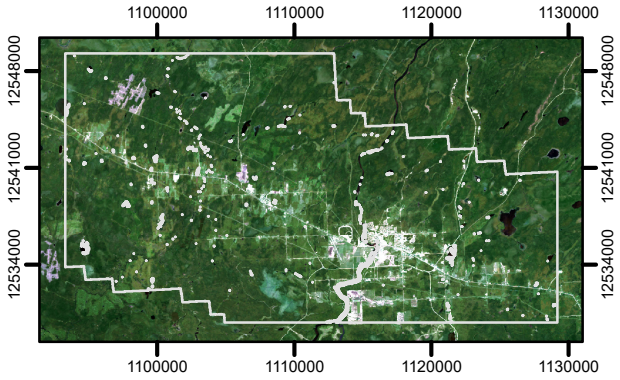

Figure 6. True-color composite image of the study area from Landsat-8 imagery from the summer of 2017.

Prediction maps for the RF models corresponding to 5 predictors each for average bulk density and average carbon are shown in Figure 7. The model for average bulk density with 5 predictors was the most accurate model for that target variable achieved; the model for average carbon content was still comparable in evaluation metrics to the model with 12 predictors. Although the R^2 values for the average bulk density models were not high, the AVG BD model with just 5 predictors had comparable accuracy and similar prediction maps as models built with more predictors. In general, it is expected that as more predictors are applied in a model that R^2 will increase, so it likely for R^2 to decrease as fewer numbers of predictors from the same set are applied for modeling. That R^2 values did not decrease as fewer predictors were used for modelling AVG BD, supports that the AVG BD model with just 5 predictors was the best model for average bulk density.

The agriculture regions can be clearly discerned in the average bulk density prediction map, as those areas have soil bulk densities that are more compacted. Areas corresponding to different dominant tree species, and secondary versus original forest can be differentiated from the contrasting intensities from both the AVG BD and AVG C prediction maps.

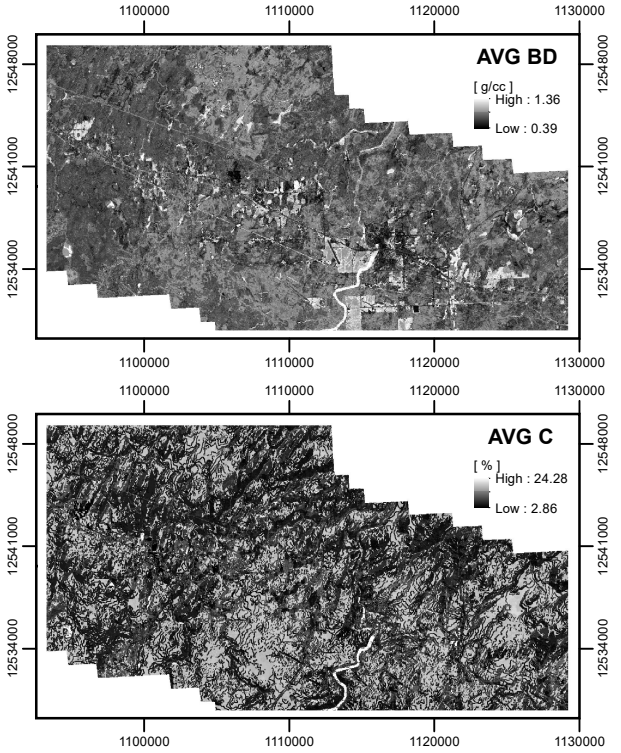

Figure 7. Prediction maps for AVG BD and AVG C for the 0-15 cm depth layer, from the respective RF models built upon just 5 predictors.

Determining the total soil carbon was a main objective for this project. However, assessing the concentrations of other nutrients such as total nitrogen and exchangeable phosphorus and exchangeable potassium are also of relevance. Correlations between average bulk density, average carbon, average nitrogen, average exchangeable phosphorus and average exchangeable potassium for those sites were calculated, as shown in Figure 8. For these 34 sites, average bulk density was moderately negatively correlated to average carbon, as well as weakly negatively correlated to average nitrogen. Average carbon was very strongly correlated to average nitrogen concentrations. The negative correlation between average bulk density and average carbon indicates that overall as average bulk density decreases then the average carbon content of the soil increases.

The strong correlation between average carbon and average nitrogen indicates that a model fit could be regressed between these quantities. Correlations of average bulk density or average carbon between other quantities such as exchangeable phosphorus or exchangeable potassium are less compelling. There exists a moderate correlation between average bulk density and average potassium, which indicates that soils with heavier bulk densities tend to have higher exchangeable potassium contents. Correspondingly, average carbon is negatively correlated to average exchangeable potassium, but with a weaker correlation than what average bulk density has with average exchangeable potassium. It is plausible that correlations between the various soil nutrients properties can be expected, as the long-term result of ecological processes

would lead to the differentiation between nutrient-rich and nutrient-poor soils. Soil samples obtained below the depths of the tree roots for this region were consistently devoid of nutrients, which signifies vegetation as a soil formation driver that alters soil nutrient properties for this region. As the scale of the DEM and Landsat imagery utilized was 30 m, and sampling sites within a few hundred meters of one another with dissimilar soil properties were able to be discerned from the modelling, it is likely that concerns pertaining to the issue of scale for the processes controlling soil formation are not merited (McKenzie and Ryan, 1999).

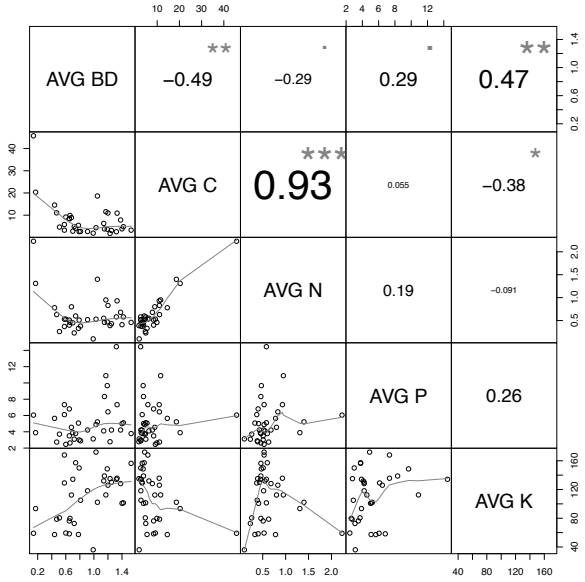

Figure 8. Correlations between AVG BD [g/cm^3], average carbon (AVG C) [%], average nitrogen (AVG N) [%], average exchangeable phosphorus (AVG P) [ppm] and average exchangeable potassium (AVG K) [ppm] for the 0-15 cm depth layer.

5. DISCUSSION

The prediction maps for the RF models in Figures 5 & 7 display many uniformities when inspected for the same target variable. In particular, for the average bulk densities, cleared regions utilized for agriculture have the densest surface soil densities, which is consistent with observations of soil compaction existent in the area. Regions of the forest, specifically of black spruce [*Picea mariana*] have the lightest soil densities, and likewise highest carbon contents, which correspond to peat and lighter materials in surface soil layers. Environmental covariates corresponding to vegetation, specifically reflectance from multispectral imagery or LiDAR-derived covariates such as CHM and gap fraction, had the highest variable importance for all the bulk density and total carbon soil models. This indicates that vegetation cover type has a significant influence as a soil formation factor for this study region. The corresponding prediction maps can be utilized for environmental studies for studying soil carbon, or utilized for decision-making purposes for determining areas suitable for agriculture.

The 0-15 cm layer was chosen as the profile depth for analysis as many studies focus on either the 0-10 cm or 5-15 cm depths for total carbon analysis. For this dataset, the 0-5 cm and 5-15 cm depth profiles had similar results, with greater variation in the 0-5 cm layer. Clayey soil was encountered for depths greater than 20 cm, which had higher bulk densities but also lower total carbon and total nitrogen contents. The 0-15 cm depth layer was deemed the profile layer of interest as this layer was most interactive with the vegetation, which also had the most variation in soil properties. An analysis on a layer profile greater than 5 cm was desired, hence why the 0-5 cm depth profile just by itself was not considered separately.

Investigating relationships with other soil nutrient properties as target variables for RF models and generating prediction maps will be a future focus for this study area. Bulk densities can be processed before soil chemistry results in non-specialized laboratories or workspaces at a cheaper cost. This means that if strong correlations exist between bulk densities and soil nutrient properties, then models built on bulk density could be utilized to estimate certain soil properties.

It is anticipated that more soil samples will be obtained during a future field campaign, for another study area within the GCB as well as additional sites in this study area for the Kapuskasing region. Additional samples will improve the threshold of sites for further model accuracy, to allow the separation of sites for model training and model verification purposes. Accuracy statistics for model assessment can then be calculated when a model fitted on the training data is applied to the verification data, in order to obtain more robust evaluation measures.

6. CONCLUSIONS

A comprehensive approach of applying environmental covariates generated from a variety of remotely sensed data improved the accuracy of models for DSM purposes. Multi-source remotely sensed data was able to generate a variety of non-topographic covariates that permitted DSM of bulk density and soil carbon properties for a relatively flat homogeneous area. RF models with reasonable accuracies were able to be generated, and the best models for standard deviation of bulk density, average total carbon and standard deviation of total carbon had coefficients of determination greater than 0.7 which are comparable or exceed accuracies reported for recent DSM models (Nussbaum et al., 2018).

Covariates obtained from remotely-sensed data, in particular LiDAR, have been advantageous for improving model accuracy. CHM is a compelling environmental covariate for vegetation cover, and gap fraction performs well as a covariate for the density of the vegetation canopy. For the average bulk density models, gap fraction had the highest or second highest variable importance, with CHM among the top predictors. Environmental covariates obtained from multispectral satellite imagery also had higher variable importance, whereas topographic covariates derived from a DEM had lower ranking. Specifically, vegetative covariates, whether determined from multispectral imagery or LiDAR data, had the highest variable importance. Vegetative covariates had the highest variable importance for the average total carbon models as well. This provides evidence that vegetative covariates, when compared to either climatic, topographic or parent material covariates, drive the soil formation factors for the study region more so than do other categories of environmental covariates.

ACKNOWLEDGEMENTS

The caret package (Kuhn, 2008) and the randomForest package (Liaw and Wiener, 2002) in R were utilized for fitting the RF models. The PerformanceAnalytics package in R was applied for generating the correlation plot shown in Figure 8. ArcMap 10.6 software from Esri was utilized for generating topographical covariates from the DEM.

The authors would like to thank the Great Lakes Forestry Centre of Natural Resources Canada in Sault Ste. Marie, ON for the processing of bulk density and chemistry soil samples obtained during 2018, and for the pre-processing and handling of soil samples obtained during 2019. The authors also thank Land Information Ontario and the Ontario Ministry of Natural Resources and Forestry for use of its LiDAR files for the District of Cochrane region of Ontario.

This study was supported in part with funding provided by the Ontario Ministry of Agriculture, Food and Rural Affairs (OMAFRA) as well as the Natural Sciences and Engineering Research Council (NSERC) of Canada.

REFERENCES

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

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

Brungard, C.W., Boettinger, J.L., Duniway, M.C., Wills, S.A., Edwards, T.C., 2015. Machine learning for predicting soil classes in three semi-arid landscapes. *Geoderma*, 239, 68–83. doi.org/10.1016/j.geoderma.2014.09.019

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

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

Kuhn, M., 2008. Caret package. *J. Stat. Softw.*, 28(5). doi.org/10.18637/jss.v028.i05

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

Ließ, M., Glaser, B., Huwe, B., 2012. Uncertainty in the spatial prediction of soil texture. Comparison of regression tree and Random Forest models. *Geoderma*, 170, 70–79. doi.org/10.1016/j.geoderma.2011.10.010

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

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

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

Mulder, V.L., Lacoste, M., Richer-de-Forges, A.C., Martin, M.P., Arrouays, D., 2016. National versus global modelling the 3D distribution of soil organic carbon in mainland France. *Geoderma*, 263, 16–34. doi.org/10.1016/j.geoderma.2015.08.035

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

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

Reynolds, W.D., Bowman, B.T., Drury, C.F., Tan, C.S., Lu, X., 2002. Indicators of good soil physical quality: Density and storage parameters. *Geoderma*, 110, 131–146. doi.org/10.1016/S0016-7061(02)00228-8

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

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