

Predictive soil mapping in the Boreal Plains of Northern Alberta by using multi-temporal remote sensing data and terrain derivatives

Authors: Sorenson, Preston T., Kiss, Jeremy, Serdetchnaia, Anna, Iqbal, Javed, and Bedard-Haughn, Angela K.

Source: Canadian Journal of Soil Science, 102(4) : 852-866

Published By: Canadian Science Publishing

URL: <https://doi.org/10.1139/cjss-2022-0028>

BioOne Complete (complete.BioOne.org) is a full-text database of 200 subscribed and open-access titles in the biological, ecological, and environmental sciences published by nonprofit societies, associations, museums, institutions, and presses.

Your use of this PDF, the BioOne Complete website, and all posted and associated content indicates your acceptance of BioOne's Terms of Use, available at www.bioone.org/terms-of-use.

Usage of BioOne Complete content is strictly limited to personal, educational, and non - commercial use. Commercial inquiries or rights and permissions requests should be directed to the individual publisher as copyright holder.

BioOne sees sustainable scholarly publishing as an inherently collaborative enterprise connecting authors, nonprofit publishers, academic institutions, research libraries, and research funders in the common goal of maximizing access to critical research.

Predictive soil mapping in the Boreal Plains of Northern Alberta by using multi-temporal remote sensing data and terrain derivatives

Preston T. Sorenson ^{a,b}, Jeremy Kiss ^a, Anna Serdetchnaia^b, Javed Iqbal^c, and Angela K. Bedard-Haughn^a

^aDepartment of Soil Science, University of Saskatchewan, Saskatoon, SK, Canada; ^bIndependent Consultant, Toronto, ON, Canada;

^cAlberta Agriculture and Forestry, Government of Alberta, Edmonton, AB, Canada

Corresponding author: Preston T. Sorenson (email: preston.sorenson@usask.ca)

Abstract

As Canada's vast Boreal Plains are extensively managed, predictive soil mapping could be used as an effective tool to generate high-resolution soil information for the region to inform sustainable resource management. This study aimed to investigate the use of multi-temporal remote sensing data and terrain derivatives to map soil types in the region. A method of constraining subgroup and great-group soil-type predictions based on the predictions at higher-order levels (great-group and order, respectively) was tested. Sentinel time series median values obtained by using Google Earth Engine were tested in combination with first- and second-order digital elevation model derivatives for use as predictor variables in the predictive models. A recursive feature selection process was implemented to reduce the number of predictor variables used in model training. Soil classes were predicted at the order, great-group, and subgroup levels and two approaches were tested. In the first approach, models were unconstrained based on previous predictions. In the second approach, models were constrained to predict only soil great-group classes that occur within the predicted soil order for a given location and similarly predict only soil subgroup classes that occur within the predicted soil great group for a given location. Determined through independent validation testing, the most probable predicted soil maps had overall accuracies ranging from 42% to 68% and kappa scores ranging from 0.33 to 0.48. Overall, the constrained models had the best performance of the approaches tested.

Key words: Random Forests, Predictive Soil Mapping, Google Earth Engine

Résumé

Les vastes plaines boréales du Canada étant largement aménagées, une carte prédictive des sols constituerait un bon moyen pour obtenir des données à haute résolution sur les sols régionaux et ainsi faciliter la gestion des ressources durables. Les auteurs ont examiné comment des données de télédétection multitemporelles et les dérivées du terrain pourraient servir à cartographier la nature des sols locaux. Dans cette optique, ils ont testé une méthode qui contraint les prévisions du sous-groupe et du grand groupe de sols d'après les prévisions effectuées à un ordre plus élevé (ceux du grand groupe et de l'ordre, respectivement). La valeur médiane des séries chronologiques sentinelles obtenues avec le moteur Google Earth a été testée en combinaison avec les dérivées de modèles numériques des hauteurs du premier et du deuxième ordre, employées comme variables explicatives dans le modèle prédictif. Pour réduire le nombre de variables explicatives servant à former le modèle, les auteurs ont recouru à un processus récursif de sélection des propriétés du sol. Les classes de sol ont été prédites aux niveaux de l'ordre, du grand groupe et du sous-groupe, et deux approches ont été vérifiées. Dans la première, les auteurs n'ont pas contraint le modèle selon les prévisions antérieures; dans la seconde, le modèle a été contraint afin de ne prédire que les classes du grand groupe dans l'ordre des sols prévu et que celles du sous-groupe dans le grand groupe de sols prévu, pour un endroit donné. Après validation dans le cadre d'un essai indépendant, l'exactitude des cartes des sols prévus les plus probables variait de 42 à 68 % et leur note Kappa allait de 0,33 à 0,48. Les modèles contraints sont généralement ceux qui fonctionnent le mieux parmi les approches testées. [Traduit par la Rédaction]

Mots-clés : forêts aléatoires, cartographie prédictive des sols, moteur Google Earth

Introduction

The Canadian boreal forests are important to the global forest ecosystem, accounting for 21%–27% of boreal forests and

8% of all forests globally. From a climate change perspective, soil management is a key concern in northern boreal forests, as they store over 80% of their total carbon belowground

(Dixon et al. 1994). Soil classification and mapping plays an important role in boreal forest management as soil organic carbon stocks relate closely to soil classification (Dalsgaard et al. 2016). Furthermore, extensive portions of the Canadian boreal forest are managed, with approximately 54% of the land area under some form of management (Kurz et al. 2013). Canadian boreal forest and soil managers depend on accurate soil classification information to inform their decision-making for objectives such as mitigating impacts of forest harvesting and oil and gas activities, particularly in the Boreal Plain region.

Predictive soil mapping (PSM) has emerged as a tool for cost-effective generation of spatial soil data at scales finer than 1:10 000 to support detailed land use planning. Extensive literature and a breadth of techniques have been developed in the field of PSM, which can be utilized to generate detailed soil data and information (Carré et al. 2007). Through this research, classification tree models that incorporate remotely sensed data have become recognized as an established tool to create accurate digital soil maps (Mulder et al. 2011). Remotely sensed data, specifically terrain and radiometric data, are important covariates for PSM, and the combination of these data types has been successfully used for soil mapping at the regional scale (Dobos et al. 2000). Terrain data, typically sourced as first- and second-order derivatives from digital elevation models (DEMs), are essential covariates for PSM (Mulder et al. 2011), as they reflect key relationships between the landscape, soil, and water that are recognized as important soil-forming factors (Jenny 1941). Specifically for terrain attributes, landscape-scale morphometric features have been documented to be particularly useful for soil mapping in the Canadian Prairies (Kiss and Bedard-Haughn 2021), as well as useful predictors for other PSM work in Canada (Heung et al. 2016).

Multi-temporal remote sensing data have been increasingly shown to be a useful tool for PSM. Recent developments in cloud computing (i.e., Google Earth Engine (Gorelick et al. 2017)) and data processing have enabled collections of multi-temporal data to be developed and processed with ease. Normalized difference vegetation index (NDVI) temporal statistics have successfully been used for PSM in Southwest United States with comparatively better performance than mono-temporal data sets (Maynard and Levi 2017). Poggio et al. (2013) used a range of MODIS-derived attributes for the whole of Scotland, which improved model performance compared to terrain-only models. Soil organic carbon mapping at fine spatial scales has been improved by using multispectral time series data acquired with unmanned aerial vehicles (Guo et al. 2020). Multi-temporal remote sensing indices have also improved mapping of soil organic carbon, sand, and cation exchange capacity in Iran (Fatholouloumi et al. 2020). Additionally, multi-temporal bare soil composites have been useful for PSM (Demattê et al. 2018; Rogge et al. 2018; Rizzo et al. 2020). However, bare soil composite specific approaches are not possible for forested regions without bare soil.

Studies in PSM and remote sensing have considered a wide variety of nonlinear machine learning models. Random forest models in particular have been used extensively for remote sensing projects because they can handle high-

dimensional multicollinear data (Belgiu and Drăgu 2016), they are relatively simple to optimize compared to models such as support vector machines (Heung et al. 2016), and they provide a relatively straightforward method to determine and extract feature importance. While some other model types, such as artificial neural networks and support vector machines, have outperformed random forest models in some studies, random forests are consistently a good performer with the advantages of easily modeling prediction uncertainty and estimating error, as well as for its ease of use for feature selection (Brungard et al. 2015; Lu and Hardin 2019). Random forest models have successfully been used for predicting soil classes in semi-arid landscapes (Brungard et al. 2015).

An important aspect of PSM by using machine learning models is feature selection to improve generalization and reduce the risk of models fitting to statistical noise. Wadoux et al. (2020) outline typical reasons for using feature selection: to calibrate models faster, reduce complexity, increase prediction accuracy, avoid multicollinearity, and prevent overfitting. Some approaches involve selecting features based on correlation values and removing highly correlated variables (Hamzehpour et al. 2019). Other studies have used recursive feature elimination using a wrapper approach (Brungard et al. 2015; Taghizadeh-Mehrjardi et al. 2016; Rudiyanto et al. 2018; Shi et al. 2018; Gomes et al. 2019; Tajik et al. 2019). Other approaches include using principal component analysis to transform variables into a new set of uncorrelated variables and typically reducing the total number of features by using the principal components that account for most of the variance in the data set (Heung et al. 2016; Sena et al. 2020; Kasraei et al. 2021).

Despite the tremendous progress made in the field of pedometrics and the development of PSM tools, conceptual challenges with PSM have been identified. Specifically, there is a lack of understanding of the relationship between soil mapping and conceptual soil classification, and a lack of consistent mapping theory used in PSM studies (Nikiforova et al. 2020). Additionally, other researchers have noted that conflicting information needs to be addressed as part of soil mapping, monitoring, and management (Baruck et al. 2016). To ensure that predictive mapping results can be interpreted within pedological frameworks, both established scientific frameworks in pedology and links between soil classification and pedology need to be incorporated into PSM processes.

The objectives of this study were twofold. The first objective was to test a range of multi-temporal remote sensed attributes derived from Sentinel-1 and Sentinel-2 data to determine their value as predictors for PSM in the Canadian boreal forest, specifically, by testing a range of sentinel-derived imagery indices from different time periods with a focus on those that have been successful for Alberta wetland mapping (DeLancey et al. 2019). This study also aimed to test a range of terrain attributes, from both a free 30 m DEM and a higher resolution light detection and ranging (LiDAR)-derived DEM for Northern Alberta, with a particular focus on attributes that have been shown to be successful for PSM in prairie landscapes (Kiss and Bedard-Haughn 2021) and elsewhere in Canada (Heung et al. 2017). To manage the large

number of features, this study incorporated feature selection approaches that do not include principal component analysis to improve model interpretability and generalization, as the principal component variables require transformation that cannot be replicated across varying sites.

The second objective for this project was to investigate model performance at different levels of soil classification precision while incorporating a feature selection approach that reduces the number of features. Two approaches were evaluated regarding mapping different levels of classification precision. One approach involved building separate independent models for each classification level (soil order, great-group, and subgroup). The second approach used models that hierarchically constrained predictions based on the higher level predictions with the aim of ensuring harmonization among the resulting maps and ensuring consistency with soil classification frameworks.

Materials and methods

Soil data

Soil classification data, 4405 soil data points in total (Table 1), were compiled from eight publicly available environmental impact assessments related to proposed oil and gas development (Government of Alberta 2021) completed in Northeastern Alberta, Canada (Fig. 1). An example of the imagery at finer resolution is provided in Fig. 2. Data from additional environmental impact assessments in the region were excluded due to those projects being developed and intact forest and soil being no longer present in modern imagery. The study area is entirely in the boreal forest, with much of the overall study characterized by peatlands. Overall, the data set had 375 Brunisols, 735 Gleysols, 1011 Luvisols, and 2284 Organic soils (Table 1). As the soil data were collected for environmental impact assessments, the data are not uniformly distributed throughout the study area. Instead, the samples were concentrated in eight distinct areas in grids and transects with varying spacings. While conditioned Latin Hypercube sampling designs are ideal for PSM (Malone et al. 2019), this study made use of existing soil data that had been collected with alternative end uses in mind, and conditioned Latin Hypercube sampling design was therefore not an option.

Soils were inspected to 1 m using a shovel and Dutch Auger and classified into the soil subgroup level according to the Canadian System of Soil Classification (Soil Classification Working Group 1998). Cryosols were included in the Organic category for this study, as they occurred only as inclusions within Organic map units in this data set. As these data were collected by a variety of personnel as part of numerous environmental impact assessments, the exact location uncertainty is not known. However, most of the locations would have been collected with handheld global position system (GPS) receivers with accuracies between 5 and 10 m. Following its compilation, the data were split into training (75%) and testing (25%) data sets based on the environmental covariates discussed below by using the Kennard–Stone algorithm, with distance set as Mahalanobis, from the prospectr package in R

Table 1. Class occurrences for each level of classification precision.

Canadian System of Soil Classification	Abbreviation	Number	Proportion
Order			
Brunisol	BRUN	375	0.08
Gleysol	GLEY	735	0.17
Luvisol	LUV	1011	0.23
Organic	ORG	2284	0.52
Great groups			
Cryosol	CRY	32	0.01
Dystric Brunisol	DYB	325	0.07
Eutric Brunisol	EB	35	0.01
Fibrisol	FIB	1400	0.32
Gleysols	GLEY	492	0.11
Humic Gleysols	HG	49	0.01
Humisol	HUM	17	<0.00
Luvic Gleysol	LG	195	0.04
Gray Luvisols	LUV	1011	0.23
Melanic Brunisols	MB	14	<0.00
Mesisols	MES	835	0.19
Subgroups			
Eluviated Dystric Brunisol	E.DYB	199	0.05
Eluviated Eutric Brunisol	E.EB	16	<0.01
Gleyed Dystric Brunisol	GL.DYB	9	<0.01
Gleyed Eluviated Dystric Brunisol	GLE.DYB	41	0.01
Gleyed Eluviated Eutric Brunisol	GLE.EB	3	<0.01
Gleyed Melanic Brunisol	GL.MB	1	<0.01
Orthic Dystric Brunisol	O.DYB	76	0.02
Orthic Eutric Brunisol	O.EB	19	<0.01
Othic Melanic Brunisol	O.MB	10	<0.01
Humic Luvic Gleysol	HU.LG	6	<0.01
Orthic Gleysol	O.G	297	0.07
Orthic Humic Gleysol	O.HG	26	0.01
Orthic Luvic Gleysol	O.LG	189	0.04
Rego Gleysol	R.G	195	0.04
Rego Humic Gleysol	R.HG	23	0.01
Brunisolic Gray Luvisol	BR.GL	58	0.01
Dark Gray Luvisol	D.GL	23	0.01
Gleyed Brunisolic Gray Luvisol	GLBR.GL	15	<0.01
Gleyed Dark Gray Luvisol	GLD.GL	10	<0.01
Gleyed Gray Luvisol	GL.GL	144	0.03
Orthic Gray Luvisol	O.GL	758	0.17
Solonetzic Gray Luvisol	SZ.GL	3	<0.01
Fibric Humisol	FI.H	1	<0.01
Fibric Mesisol	FI.M	40	0.01
Fibric Orthic Cryosol	FI.OC	27	0.01
Humic Fibrisol	HU.F	1	<0.01
Mesic Fibrisol	ME.F	39	0.01
Mesic Orthic Cryosol	ME.OC	4	<0.01
Terric Fibric Humisol	TFL.H	4	<0.01
Terric Fibric Mesisol	TFL.M	111	0.03
Terric Fibric Orthic Cryosol	TFL.OC	1	<0.01
Terric Fibrisol	T.F	689	0.16
Terric Humic Fibrisol	THU.F	19	<0.01
Terric Humic Mesisol	THU.M	7	<0.01
Terric Humisol	T.H	14	<0.01
Terric Mesic Fibrisol	TME.F	102	0.02
Terric Mesic Humisol	TME.H	1	<0.01
Terric Mesisol	T.M	512	0.12

Table 1. (concluded).

Canadian System of Soil Classification	Abbreviation	Number	Proportion
Typic Fibric Mesisol	TY.FM	1	<0.01
Typic Fibrisol	TY.F	549	0.12
Typic Humisol	TY.H	1	<0.01
Typic Mesisol	TY.M	161	0.04

Note: Cryosols were grouped with Organics at the order level as they occur only as inclusions among Organic Soils in the study areas.

(Stevens and Ramirez-Lopez 2014). The Kennard–Stone algorithm is used extensively in the soil spectroscopy literature for creating train and test splits (Deiss et al. 2017; Roudier et al. 2017; Viscarra Rossel et al. 2017; Douglas et al. 2018; Dangal et al. 2019; Zhang et al. 2019; Sanderman et al. 2020; Sorenson et al. 2020). The algorithm selects samples with uniform distribution over a multivariate predictor space, which helps to ensure that the training data represent the variation in the overall data set. A 75–25 train-test split was selected to ensure that data were kept independent from feature selection and model optimization steps. This train-test split ratio was selected to try and ensure that there were enough data in the training set to account for the full range of variability in the data set, and because this ratio has been accepted in other studies and machine learning procedures (Ghatak 2017; Sorenson et al. 2020, 2021).

Sentinel data acquisition and processing

Optical remote sensing data were acquired by using Google Earth Engine (Gorelick et al. 2017). Sentinel-1 and Sentinel-2 imagery from 1 May 2015 to 31 October 2020 was obtained and only cloud-free pixels were kept for the Sentinel-2 imagery. Sentinel-1 data were obtained for May to 31 October 2020 and separately for May, July, and October months of 2015–2020. These months were selected to capture backscatter data at time periods with distinct vegetation characteristics, specifically green up in spring, peak photosynthetic activity periods in July, and senescence in fall. Sentinel-2 data for the combined months of September and October from 2015 to 2020 were also obtained to investigate if differences between soils were more visible once senescence had begun. Different months were not investigated for the Sentinel-2 data to limit file sizes, as the Sentinel-2 data with multiple float bands have significantly higher file sizes. Median backscatter and reflectance values were then calculated for the raster stack for all Sentinel-1 and Sentinel-2 data. A single raster stack for both Sentinel-1 and Sentinel-2 data containing the median values for backscatter and reflectance for each band was created for each of the time periods exported. Resampling of all data to a 10 m spatial resolution using nearest neighbor occurred to match the 10 m spatial resolution of the finest scale Sentinel-2 bands.

In addition to the raw bands, band ratios for the 2015–2020 time period were also calculated in Google Earth Engine and tested as potential PSM environmental covariates. Median and maximum NDVI values were calculated for the months of April, May, July, September, and October. These months were

chosen to identify if vegetation at different stages helped better distinguish soil types, specifically green up, peak photosynthetic activity, and senescence. April and September were added compared to the other variables to see if the particular month for the green up and senescence periods made a difference, and only for these variables as the file sizes were substantially smaller. The median and standard deviation of NDVI for 1 May to 31 October were also calculated. The median normalized difference water index (NDWI) for 1 May to 31 October, along with the months of April, May, and July, was also calculated. The median and standard deviation of the normalized difference of the red and green bands for the entire period between 1 May and 31 October, along with just the median value for the months of September and October, were also calculated. Based on the work of DeLancey et al. (2020), the median anthocyanin reflectance index (ARI) and the red edge inflection point (REIP) from 1 May to 31 October were calculated. The plant senescence reflectance index (PSRI) was also calculated using only data from 1 September to 31 October. Equations for the band ratios are provided in Table 2.

Terrain data

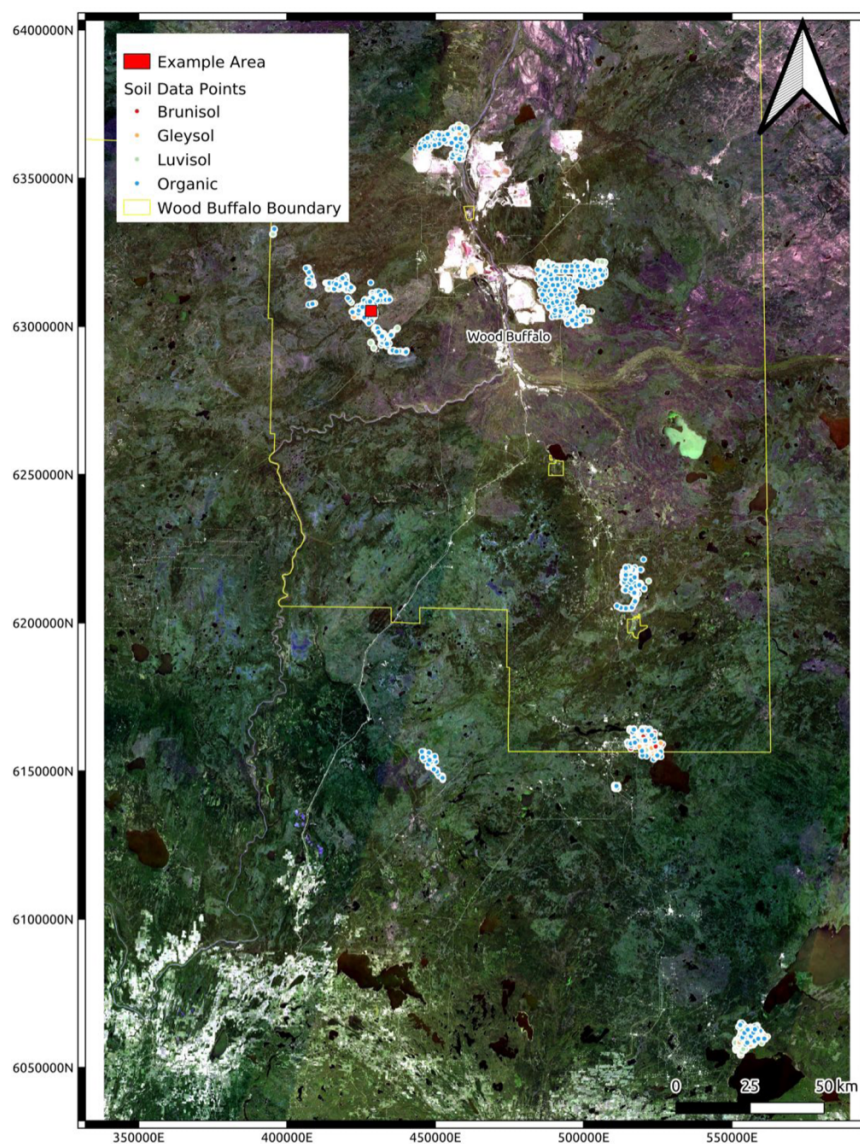
Multiple terrain attributes were calculated for the study region using two DEM data sources to test the results between the different data sets. The first was a Light Detection and Ranging (LiDAR)-derived DEM that was resampled from 1 to 10 m to match the resolution of the other remote sensing data sets. The second data set was the ALOS World 3D 30 m digital surface model (Japan Aerospace Exploration Agency 2015). Using the rsaga package in R (Brenning et al. 2018), the following terrain attributes were calculated for both DEMs:

- Aspect, slope, general curvature, profile curvature, plan curvature, tangential curvature (Zevenbergen and Thorne 1987)
- Flow width and specific catchment area (SCA) assuming the default parameters (Gruber and Peckham 2009)
- Slope length (LS) factor (Moore et al. 1991)
- Topographic wetness index (TWI) (Beven and Kirkby 1979), SAGA wetness index (SWI) (Boehner and Selige 2006), and terrain ruggedness index (TRI) (Riley et al. 1999)
- Slope height, valley depth, normalized height, standardized height, mid-slope position (Boehner and Selige 2006).
- Convergence index (Koethe and Lehmeier 1996) and specific dispersal area (SDA) (Costa-Cabral and Burges 1994)
- Multiresolution valley bottom flatness and multiresolution ridge top flatness (Gallant and Dowling 2003)
- Topographic position index (TPI) and TPI landform classification feature selection (Guisan et al. 1999)

Feature selection

Given the large number of potential features, a feature selection step was undertaken to reduce the number of features used in the final model and decrease the risk of overfitting. An approach similar to that described in Brungard et al. (2015) was used. All feature selection was undertaken by using only the training data set. Feature selection was

Fig. 1. True color red–green–blue 2015–2020 median reflectance (Sentinel-2 bands 2, 3, and 4) for the entire study area (European Space Agency 2021). The points correspond to soil sample data that were collected within this area. The example area is shown in greater detail and is used for result interpretation purposes in other map figures. The yellow boundary corresponds to the Regional Municipality of Wood Buffalo (Statistics Canada 2021). Coordinates are in UTM Zone 12N NAD 83. [Colour online.]



performed by using random forest models developed with the ranger package in R (Wright and Ziegler 2017). The first step involved testing for correlation among features. Features that were highly correlated, based on a threshold correlation value of 0.9, with another feature were excluded, such that only one of the correlated features was kept for model building. This step was undertaken to help reduce multicollinearity among variables, while maintaining the original and untransformed variables to make model interpretation easier.

A random forest model was then built by using all the remaining predictor variables that were not eliminated through the correlation assessment. Features were then

sorted by their importance based on Gini index values. Random forest models were then iteratively generated, with the least important feature in each run removed. The feature importance values were recalculated for each iterative model run. The final model features were then selected based on where the out-of-bag error, in this case the Brier score, stopped improving as a function of the number of variables. This process was undertaken separately for each of the precision levels of soil classification: order, great group, and subgroup. The process was also repeated for the models built by using attributes derived from the LiDAR and ALOS DEMs. The features that were used for each soil classification level and for both DEM sources are provided in Table 3.

Fig. 2. True color red–green–blue 2015–2020 median reflectance (Sentinel-2 bands 2, 3, and 4) for an example area within the larger study area (European Space Agency 2021). The points correspond to soil sample data (training and testing point data) that were collected within this area. Coordinates are in UTM Zone 12N NAD 83. [Colour online.]

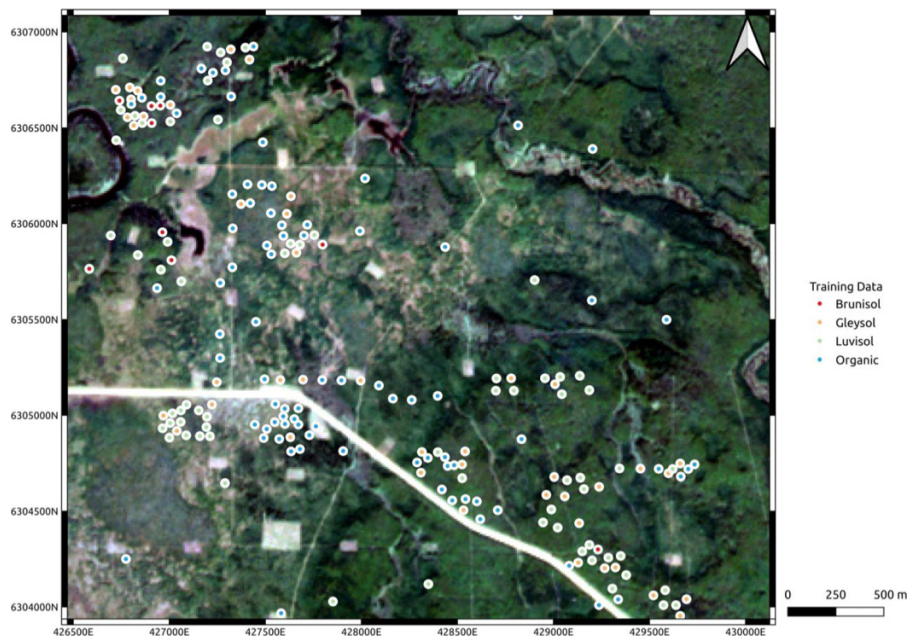


Table 2. Band ratio equations used as environmental covariates in the predictive modeling.

Band ratio	Equation
Normalized difference vegetation index (NDVI)	$\frac{\text{Band 8} - \text{Band 4}}{\text{Band 8} + \text{Band 4}}$
Normalized difference water index (NDWI)	$\frac{\text{Band 3} - \text{Band 8}}{\text{Band 3} + \text{Band 8}}$
Anthocyanin reflectance index (ARI)	$\frac{\text{Band 8}}{\text{Band 2}} - \frac{\text{Band 8}}{\text{Band 3}}$
Red edge inflection point (REIP)	$702 + 40 \left(\frac{\frac{\text{Band 4} + \text{Band 7}}{2} - \text{Band 5}}{\text{Band 6} - \text{Band 5}} \right)$
Plant senescence reflectance index (PSRI)	$\frac{\text{Band 4} - \text{Band 2}}{\text{Band 5}}$
Red and green bands Normalized difference	$\frac{\text{Band 3} - \text{Band 4}}{\text{Band 3} + \text{Band 4}}$

Note: The listed bands correspond to Sentinel 2.

Statistical modeling

The ranger package in R (Wright and Ziegler 2017) was used to build random forest models to predict soil classification at the subgroup, great-group, and order levels. Two approaches were tested for subgroup and great-group classification predictions: constrained, where potential soil class predictions were limited based on previous precision level predictions, and unconstrained, where there is no restriction on potential predictions. In the constrained approach, soil maps are generated iteratively based on the hierarchical structure of the Canadian System of Soil Classification (Soil Classification Working Group 1998). First, soil order classes were

predictively mapped, then great-group classifications were predictively mapped, but potential great-group classes were constrained to those within the soil order that was initially mapped at a given location. Similarly, subgroup classes were then constrained to those within the mapped great group. In the unconstrained approach, all possible great-group and subgroup classes had the potential to be mapped at any location regardless of the resulting soil order or great-group class maps. Additionally, predicting great group and order by aggregating predictions at the subgroup level was also tested. This process was repeated by using terrain derivatives derived from the LiDAR and ALOS DEMs.

Probability forests were built for each model, so that both the most likely and the second most likely soil class could be predicted based on tree agreement values. Case weights were set for the predictive models, as imbalanced class observations used for model training can be a challenge for PSM (Shariffar et al. 2019). For the soil order predictions, the weights were balanced such that there was an equal probability of drawing a sample from each order during the tree building process of the random forest. For the unconstrained great-group and subgroup predictive models, the weights of their respective soil order were incorporated to account for the large number of irregularly occurring soils at the great-group and the subgroup levels.

Class balancing for the constrained great-group and subgroup model building did not use the case weights from the soil order level, as separate models were built for each order or great group separately. Instead, case weights were generated to create an equal chance that samples from a given great group or subgroup would be selected during the

Table 3. Relative feature importance for the soil order, soil great-group, and soil subgroup random forest models for both the LiDAR and the ALOS DEMs.

LiDAR digital elevation model									
Parameter	Standardized height	Multi-resolution valley bottom flatness	Valley depth	October Sentinel-1 VH nand	Anthocyanin reflectance index	Slope height	Median October NDVI	Multi-resolution ridge top flatness	Fall Sentinel-2 Band 2
Order	0.21	0.11	0.11	0.11	0.10	0.09	0.09	0.10	0.08
Great group	0.20	0.12	0.12	0.12	0.11	0.10	0.11	0.11	NA
Subgroup	0.18	0.13	0.12	NA	0.12	0.11	0.11	0.13	0.11
ALOS digital elevation model									
Parameter	Standardized height	Anthocyanin reflectance index	Valley depth	Normalized height	Median October NDVI	Multi-resolution valley bottom flatness	Fall Sentinel-2 Band 2	Slope height	October Sentinel-1 VH band
Order	0.15	0.13	0.13	0.14	0.12	0.12	0.11	0.11	NA
Great group	0.14	0.11	0.11	0.12	0.11	0.10	0.10	0.10	0.12
Subgroup	0.16	0.14	0.14	0.15	0.14	0.14	0.13	NA	NA

Note: The feature importance values for the great-group and the subgroup models are for the unconstrained models for simplicity. The feature importance was determined based on the Gini index. NA values indicate that a parameter was not used in the predictive model. Relative feature importance is the feature importance divided by the sum of the importance values.

bootstrapping for each tree built in the random forest. Case weights, which varied between 0 and 1, were transformed by taking the 4th root of the initial generated case weights, as it optimized cross validation results within the training data. This approach was used to increase the sampling of irregularly occurring soils without oversampling them such that they were predicted too frequently.

Following feature selection, a predictive model at the soil order level was built. For the constrained approach, soil great-group predictive models were then built separately for each of the soil orders (Brunisols, Luvisols, Gleysols, and Organic soils), depending on predicted soil order, rather than building a single model for predicting all great-group classes. Through this approach, the predicted soil great group would be forced to correspond to the correct soil order for any given pixel. This process was repeated for soil subgroup, with separate soil subgroup models depending on the predicted soil great group.

Model performance was evaluated based on overall accuracy and kappa score of soil class predictions for the independent test data set. The kappa score is used to test interrater reliability and unlike overall accuracy, it accounts for chance agreement, making it a more robust measure of model reliability (McHugh 2012). All performance results were based on the independent test data set. Soil class mapping results, along with prediction confidence maps, are illustrated by using a smaller example area to better illustrate variation in soil properties at finer spatial scales. Performance was also assessed based on producer's accuracy, which is the number of correctly predicted soil classifications of a particular soil type per observation of that soil type, and user's accuracy, which is the number of correct predictions per total predictions of that soil type.

Results and discussion

Class imbalances

At the soil order level, Organic soils made up 52% of the total soil point data, Luvisolic soils made up 23%, Gleysolic soils made up 17%, and the remaining 8% were Brunisolic (Table 1). Among the soil great groups present, the majority were Fibrisols followed by Mesisols. Only a small number of Humisols were present. All Luvisolic soils were Gray Luvisols, and Gleysols without further modifiers made up the majority of Gleysolic soils (67%) with a minority of Luvic and Humic Gleysols present. Brunisolic soils were primarily Dystric Brunisols (87%) with a small number of Eutric and Melanic Brunisols.

As expected, the data were severely unbalanced at the subgroup level (Table 1). Among the Brunisolic soils, 53% were Eluviated Dystric Brunisols followed by Orthic Dystric Brunisols (20%) and Gleyed Eluviated Dystric Brunisols (11%). The six other Brunisolic subgroups comprised the remaining 16%. For the Gleysolic soils, the majority (40%) were Orthic Gleysols, followed by Rego Gleysols at 27% and Orthic Luvic Gleysols at 25%. The remaining 8% of Gleysolic soils were Humic Luvic Gleysols, Orthic Humic Gleysols, and Rego Humic Gleysols. Seventy-five percent of Luvisolic soils were Orthic

Table 4. Independent test result overall accuracy and kappa scores for soil order, great group, and subgroup most likely and second most likely predictions for both the LiDAR and the ALOS DEMs.

	Most likely		Second most likely	
	Accuracy	Kappa	Accuracy	Kappa
LiDAR				
Order	0.68	0.48	0.22	0.02
Order (aggregated)	0.58	0.38	0.42	0.19
Great group (constrained)	0.57	0.44	0.12	-0.09
Great group (unconstrained)	0.56	0.41	0.20	0.05
Great group (aggregated)	0.50	0.36	0.34	0.20
Subgroup (constrained)	0.42	0.33	0.12	0.03
Subgroup (unconstrained)	0.38	0.29	0.17	0.08
ALOS				
Order	0.65	0.43	0.23	0.02
Order (aggregated)	0.56	0.34	0.36	0.13
Great group (constrained)	0.54	0.39	0.13	-0.08
Great group (unconstrained)	0.54	0.39	0.18	0.03
Great group (aggregated)	0.49	0.34	0.28	0.13
Subgroup (constrained)	0.39	0.30	0.11	0.02
Subgroup (unconstrained)	0.38	0.29	0.13	0.04

Note: Results include constrained and unconstrained approaches for great-group and subgroup modeling, and the great-group and order results based on aggregating the unconstrained subgroup predictions.

Gray Luvisols and 14% were Gleyed Gray Luvisols. The remaining 11% were from four different other Luvisolic subgroups (Table 1). The most common Organic soil was Terric Fibrisols (30%) followed by Typic Fibrisols (24%) and Terric Mesisols (22%). The remaining 24% of Organic soils were from 17 different Organic subgroups with nine of those subgroups occurring less than 10 times out of the 2284 Organic data points.

These class imbalances in the training and testing data created a challenge in this study, as it has with other PSM studies (Shariffar et al. 2019). Balancing the class weights prior to the random forest analysis improved model utility by increasing the discrimination of different soil types across the landscape and did not lead to decreases in overall model accuracy generally, which was also reported by Shariffar et al. (2019). Without balancing the class weights in this study, irregularly occurring soils, such as Gleysols and Brunisols, were predicted with very low frequency in the test data set. Shariffar et al. (2019) tested both undersampling and oversampling data, which was effectively done in this study by specifying the case weights in the random forest model.

The case weights influence how the random forest selects samples for model building without modifying the data sets directly by changing the probability that an individual point will be selected for training the decision trees in the random forest model. As a result, the probability that a subsampled point belongs to any given soil class can be equalized. Overall, balancing the case weights led to a decrease in accuracy, but no change in kappa values at the order level (Table 4 and Table S13). Leaving case weights unbalanced also led to very few Gleysol predictions in the final maps and a producer's accuracy of 0.03 (Table S13). Balancing the class weights led to slightly less Luvisolic soils and Organic soils predicted com-

pared to occurrences in the training data set. The number of Luvisolic and Organic soils predicted would have increased by increasing the probability of their inclusion during model building with the case weights; however, this would likely result in underprediction of Brunisolic and Gleysolic soils.

Feature importance

The first objective of this study was to investigate the value of multi-temporal Sentinel-derived remote sensing data as PSM covariates for boreal forest soil mapping. Multi-temporal remote sensing-derived covariates have been shown to be a valuable tool for PSM (Poggio et al. 2013; Maynard and Levi 2017; Demattê et al. 2018; Rogge et al. 2018; Fatholouloumi et al. 2020; Guo et al. 2020; Rizzo et al. 2020) as well as other remote sensing land classification research (Xiong et al. 2017; Sanderman et al. 2018; Sazib et al. 2018; Vågen and Winowiecki 2019; Yan et al. 2019). For the LiDAR-based soil order, great-group, and subgroup models, standardized height was the most important predictor variable followed by valley depth for the order and great-group models, and multi-resolution ridge top flatness (MRRTF) for the subgroup model (Table 3). In general, the terrain attributes were the most important predictors for each of the three models. Sentinel variables were also important predictors for each of the three soil classification levels. At the order level, four out of the nine predictors were sentinel-derived, with Median October Sentinel-1 Vertical-Horizontal Polarization (VH) as the most important followed by ARI. For soil great group, three of eight predictors were also derived from Sentinel data, with Median October Sentinel-1 VH band as the most important Sentinel-derived predictor, followed by ARI and Median October NDVI. Out of the eight total predictors for soil subgroup, five were terrain attributes, including the three most important predictors. Of the Sentinel-derived predictors, ARI and Median October NDVI were the most important predictors.

For the ALOS-based soil order, great-group, and subgroup models, standardized height was again the most important predictor variable followed by normalized height (Table 3). The terrain attributes were again the most important predictors for each of the three models, with Sentinel variables also important. At the order level, three out of the eight predictors were Sentinel derived, with ARI as the most important followed by the Median October NDVI. For soil great group, four of nine predictors were also derived from Sentinel data, with Median October Sentinel-1 VH band as the most important Sentinel-derived predictor, followed by ARI and Median October NDVI. Out of the seven total predictors for soil subgroup, four were terrain attributes, including the two most important predictors. Of the Sentinel-derived predictors, ARI and Median October NDVI were the most important.

Overall, multi-temporal Sentinel-derived remote sensing data sets were useful predictors for mapping soils in Northern Alberta in this study. Of particular value were the ARI, Sentinel-1 VH band backscatter, and the median October NDVI. The ARI is associated with anthocyanin pigments in plant foliage (Gitelson et al. 2001), which is related to plant stress and senescence. For that reason, it is likely particularly useful for distinguishing areas with senescing deciduous veg-

etation, as well as organic soils that typically have low canopy density. The importance of the Sentinel-1 VH backscatter covariate is likely due to signal variation as a result of differences in aboveground biomass (Laurin et al. 2018), which will be particularly lower in areas with organic soils. While total photosynthetic activity is low in October in this region, the importance of the median October NDVI covariate may be due to different NDVI response from senescing deciduous vegetation.

Terrain attributes were the most important predictor variables (Table 3). The models using the LiDAR data performed slightly better than the ALOS models, with an increase in accuracy of 0.03 for the order and constrained great-group and subgroup models (Table 4). Simpler landscape attributes such as slope and curvature were less important and landscape-scale morphometry-connected attributes were more important. This is consistent with other PSM studies in Western Canada focused on the Prairies (Kiss and Bedard-Haughn 2021). Landscape-scale morphometric features have been found to be valuable PSM predictor variables for other regions in Canada (Heung et al. 2016). Pennock et al. (2014) have also shown the importance of hydrology-related variables for predictive mapping of wetland soils in Western Canada, and these types of soils are an important part of the soil landscape in the Canadian Boreal Plains. In this study, standardized height, valley depth, multiresolution valley bottom flatness, slope height, multiresolution ridge top flatness, and normalized height were the most important terrain attributes for both the LiDAR and ALOS DEMs.

Soil order

The second objective of this study was to evaluate model performance at different levels of soil classification precision. Predictive model results for the LiDAR-based models based on the independent validation data set at the order level had an overall accuracy of 0.68 and a kappa score of 0.48 (Table 4). The accuracy of the second most likely soil order was 0.22 with a kappa score of 0.02. Soil map results for the mostly likely soil order are illustrated in Fig. 3. The prediction result confusion matrix indicates that the overall performance at the soil order level was greatest for the Organic class, closely followed by the Luvisols (Table S1). For both Organic and Luvisolic soils, a majority were predicted as the appropriate soil order. There was an overprediction of Brunisols, where slightly more of the predicted Brunisolic soils were actually Luvisols (54 Brunisols vs. 58 Luvisols). Only 23% of the observed Gleysol points were predicted as Gleysols, with the majority predicted as Organic soils, which resulted in an overall underprediction of Gleysolic soils. For the ALOS-based models, the order level had an overall accuracy of 0.65 and a kappa score of 0.43 (Table 4). Similar trends were observed in the confusion matrix as with the LiDAR-based models (Table S1). Predicting soil order by aggregating soil subgroup predictions resulted in a decrease in model performance, with an accuracy of 0.58 and a kappa value of 0.38 (Table 4).

The overall model performance in this study was comparable to other studies by using predictive mapping of soil classes. Accuracy values of 67% and kappa scores of 0.53 were

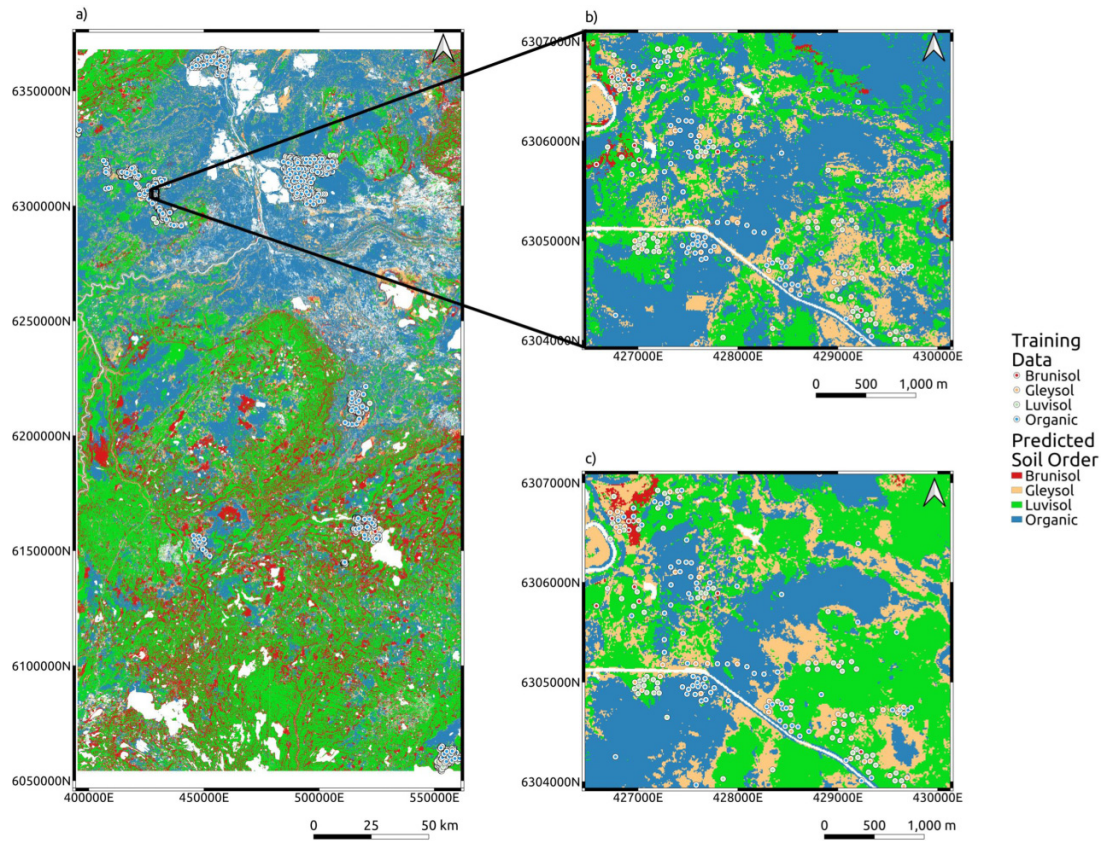
achieved in a study focused on using NDVI temporal statistics for predicting soil texture class (Maynard and Levi 2017). A study in the black soil region of China (Yang et al. 2020), which used Landsat data from 1984 to 2018 to predict Cambisols, Phaeozems, and Chernozems, had overall accuracies of 85% and a kappa score of 0.77. Soil class prediction in semi-arid landscapes in the United States mapping soil types and classes at the subgroup level achieved a range of values across three different study sites (Brungard et al. 2015). Depending on the study area, kappa scores ranged from 0.19 to 0.53 and accuracy values from 43% to 72%. Mapping at the family level across five classes in Iran had accuracy and kappa scores of 0.70 and 0.69, respectively, with a random forest model (Taghizadeh-Mehrjardi et al. 2015). Other work in Iran had overall accuracy values ranging from 0.88 at the order level to 0.30 at the family level, with 25 soil families present (Esfandiarpour-Boroujeni et al. 2020).

Soil great group and subgroup

At the great-group level, the accuracies differed slightly between the constrained and unconstrained approaches (Table 4). Using the constrained approach, the accuracy and kappa scores for the most likely great-group soil class were 0.57 and 0.43, respectively, for the LiDAR-based models. The accuracy and kappa scores were 0.54 and 0.39, respectively, for the ALOS-based models. For the second most likely great-group soil class with the LiDAR data, the accuracy was 0.19 with a kappa score of 0.4. The constrained approach only slightly improved the accuracy of the most likely soil great group and decreased the accuracy for the second most likely soil great group (Table 4). However, this approach addressed inconsistencies between the order and great-group maps in terms of the relative proportion of each soil order. The constrained great-group results for the example area (Fig. 4) correspond to the order map (Fig. 3) in terms of the relative amounts of the different soil orders. However, without constraining the soil great-group options to match the order predictions, the relative amount of soil orders was not consistent between the soil great-group and soil order maps. For example, Luvisolic soils comprised 43% of the example area when mapped at the order level (Fig. 3), but made up 64% of the area with unconstrained great-group mapping (Fig. 4). As with the soil order model, aggregating the subgroup predictions was associated with a decrease in model performance (Table 4).

For the great-group confusion matrix results, the misclassifications at the order level are preserved because separate predictive models were built for each soil order and applied where each soil order was predictively mapped. Results were generally similar between the LiDAR (Table S3) and ALOS (Table S4) models. All the Brunisolic soils were predicted to be Dystric Brunisols (Table S3). This is likely because in the training data, there were 257 Dystric Brunisols compared to 27 Eutric Brunisols and nine Melanic Brunisols, and a lack of distinct difference in the environmental covariates to distinguish the various Brunisol types. For Luvisolic soils, the only great group present in the training and validation data sets was Gray Luvisol and so all misclassifications for this group are due to misclassifications at the order level. Among the

Fig. 3. Predicted soil order map along with the example area within the larger study area. (a) Regional map using the constrained approach and the ALOS DEM, (b) example results from the ALOS-based constrained model, and (c) example results from the LiDAR-based constrained model. The white areas on the map are areas that were masked due to having July median NDVI values less than 0.3, and therefore likely correspond to surface water or oil and gas facilities. The points correspond to soil sample data (training and testing point data) that were collected within this area. Coordinates are in UTM Zone 12N NAD 83. [Colour online.]



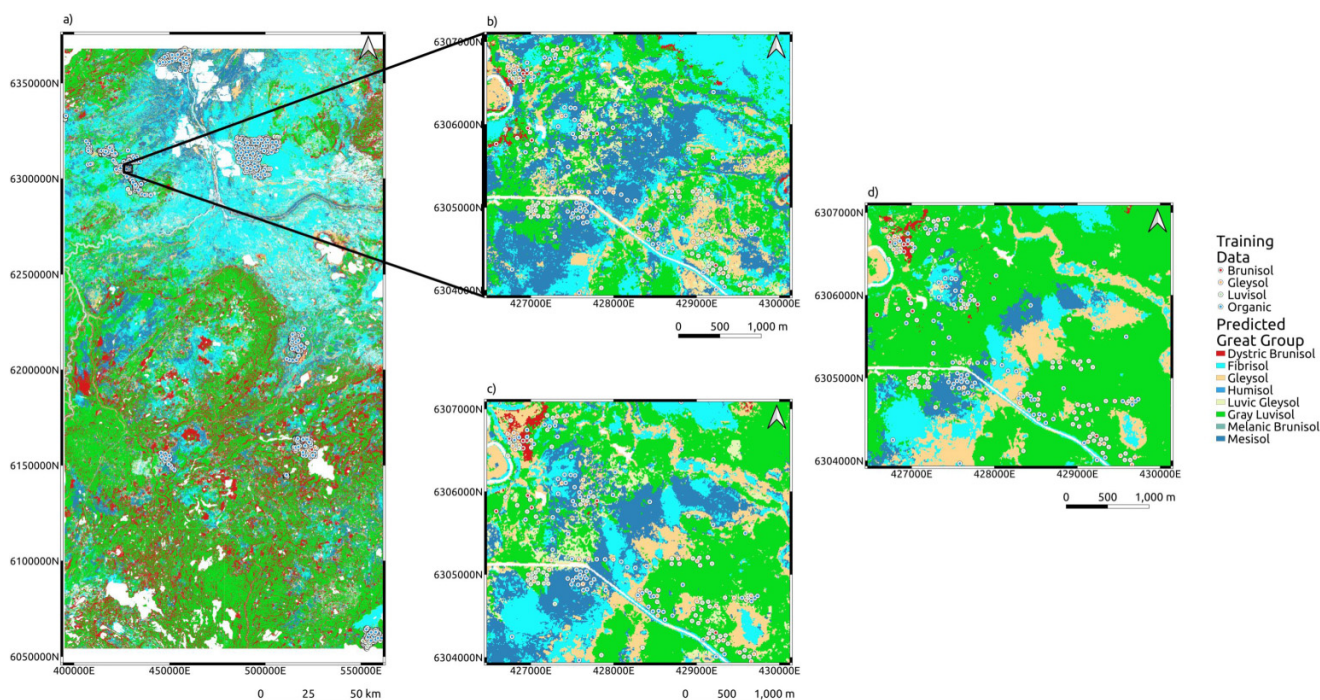
Gleysols, the majority were predicted as Gleysols (this class included Gleysols that did not have a Luvic or Humic great-group modifier), which make up the majority of the training data set, with 13 Luvic Gleysols and one Humic Gleysol predicted; of those, only one of the Luvic Gleysols predictions was accurate. Among the Organic soils, both Fibrisols and Mesisols were predicted roughly in accordance with their proportions in the test data set, with an overprediction of Fibrisols and an underprediction of Mesisols (Table S3). No Humisols were predicted; however, there were only five Humisols records out of 1654 Organic soils in the training data set. Despite only 24 Cryosols in the training data set, six Cryosols were predicted in the test data with two of these being accurate predictions.

For the soil subgroup predictions, the constrained approach predictions had an overall accuracy of 0.42 and a kappa score of 0.33 for the LiDAR-based model, and an accuracy of 0.39 and a kappa score of 0.30 for the ALOS-based model (Table 4). For the second most likely soil, the overall accuracy was 0.12 with a kappa score of 0.03 for the LiDAR data. Without constraining the subgroup predictions for the LiDAR data, the accuracy of the predictive model was 0.38 with a kappa score of 0.29. Similar issues with the relative

proportions of different soil orders between predictions at the order level and the subgroup level occurred with the unconstrained subgroup predictions as with the unconstrained great-group predictions (Tables S9 and S10). For both the soil order (Fig. 3) and the constrained soil subgroup predictions (Fig. 5), the example area was approximately 43% Luvisolic soils. However, for the unconstrained subgroup predictions, Luvisolic soils made up 53% of the area (Fig. 5). This is different than both the order level and unconstrained great-group predictions.

Based on the confusion matrix for the subgroup predictions, as only Dystric Brunisols were predicted at the great-group level, only Dystric Brunisols were present at the subgroup level (Table S10). Among those predictions, the majority of those soils were predicted as Eluviated Dystric Brunisols, which made up the majority of the training and testing data sets (Table S10). Gleyed Eluviated Dystric Brunisols were accurately predicted in two cases; however, the majority were misclassified primarily as Eluviated Dystric Brunisols or Organic Soils. Among the Gleysols, no Gleysol had a class with a majority of the predictions being correct. Orthic Gleysols had the best performance with 17 out of 73 in the test data classified correctly. For the next most common

Fig. 4. Predicted soil great-group map along with the example area within the larger study area. (a) Regional map using the constrained approach and the ALOS DEM, (b) example results using the ALOS-based constrained model, (c) example results using LiDAR-based constrained model, and (d) results using the LiDAR-based unconstrained model. The white areas on the map are areas that were masked due to having July median NDVI values less than 0.3, and therefore likely correspond to surface water or oil and gas facilities. The points correspond to soil sample data (training and testing point data) that were collected within this area. Coordinates are in UTM Zone 12N NAD 83. [Colour online.]



types of Gleysols in the data set, Orthic Luvic Gleysols primarily misclassified as upland mineral soils or Gleysols. For Rego Gleysols, they primarily misclassified as Terric Fibrisols.

For the Luvisolic soils, 100 out of 163 Orthic Gray Luvisols in the training data set were predicted accurately (Table S9), with Eluviated Dystric Brunisols being the most common misclassification. For the other more frequently occurring soil types, Brunisolic Gray Luvisols were not predicted and instead were generally predicted as Eluviated Dystric Brunisols. Gleyed Gray Luvisols were not consistently predicted, with only one out of 38 predicted accurately. The majority were predicted as Orthic Gray Luvisols, with Eluviated Dystric Brunisols and Orthic Gleysols as the next most common label for the Gleyed Gray Luvisols. Only the Terric Fibrisols, Terric Mesisols, and Typic Fibrisols had a majority of the classifications correct for the Organic soil subgroups (Table S9). These soils were the most frequently occurring organic soil subgroups and made up 74% of the training data combined. The rest of the organic soil subgroups present tended to primarily classify as one of these three classes.

Regarding accuracies at the different levels of classification precision, it should be noted that while reasonable accuracies at the great-group and especially subgroup level were achieved, the model performance was generally driven by successful prediction of the most dominant classes. Although the accuracy metrics of the constrained and unconstrained approaches were similar, the constrained approach is prefer-

able when publishing maps at all three soil classification levels (soil order, great group, and subgroup) to ensure consistency amongst the outputs. While this same outcome could be achieved by aggregating the subgroup predictions, this approach was associated with an overall decrease in model performance (Table 4). A possible explanation for this is that the soils in this region, aside from Brunisols, which are associated with a different parent material, largely vary as a function of hydrological gradients, and this may not be the case for other regions (Soil Classification Working Group 1998). The largest differences in hydrological gradients occur at the order level, with variations in great group and subgroup associated with finer differences along the gradient. This was observed in this data set where the contrast in standardized height between Gleyed Gray Luvisols and Orthic and Luvic Gleysols and that between Rego Gleysols and Organic Soils were less than the difference at the order level (Fig. 6). Standardized height was consistently one of the most important terrain attributes, which corresponds to hydrological gradients in this region. An important factor to note is that when using the constrained approach, if performance at the order level is poor, then the great-group and subgroup model performances will also be poor.

Performance among all models for predicting rarely occurring soils classes was poor. This is not unique to this study because balancing class weights is difficult—too much weight for rare soils and they will be too frequently predicted. Likely

Fig. 5. Predicted soil subgroup map along with the example area within the larger study area. (a) Regional map using the constrained approach and the ALOS DEM, (b) example results using the ALOS-based constrained model, (c) example results using LiDAR-based constrained model, and (d) results using the LiDAR-based unconstrained model. The white areas on the map are areas that were masked due to having July median NDVI values less than 0.3, and therefore likely correspond to surface water or oil and gas facilities. The points correspond to soil sample data (training and testing point data) that were collected within this area. Coordinates are in UTM Zone 12N NAD 83. [Colour online.]

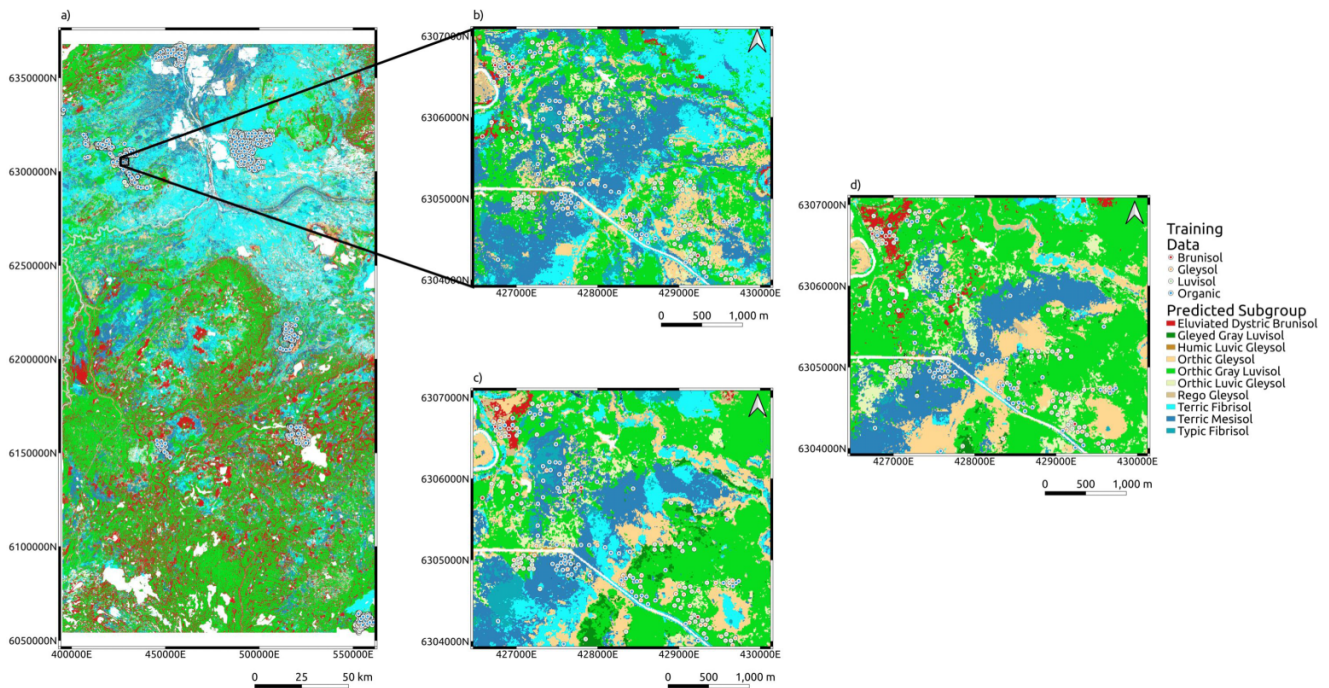
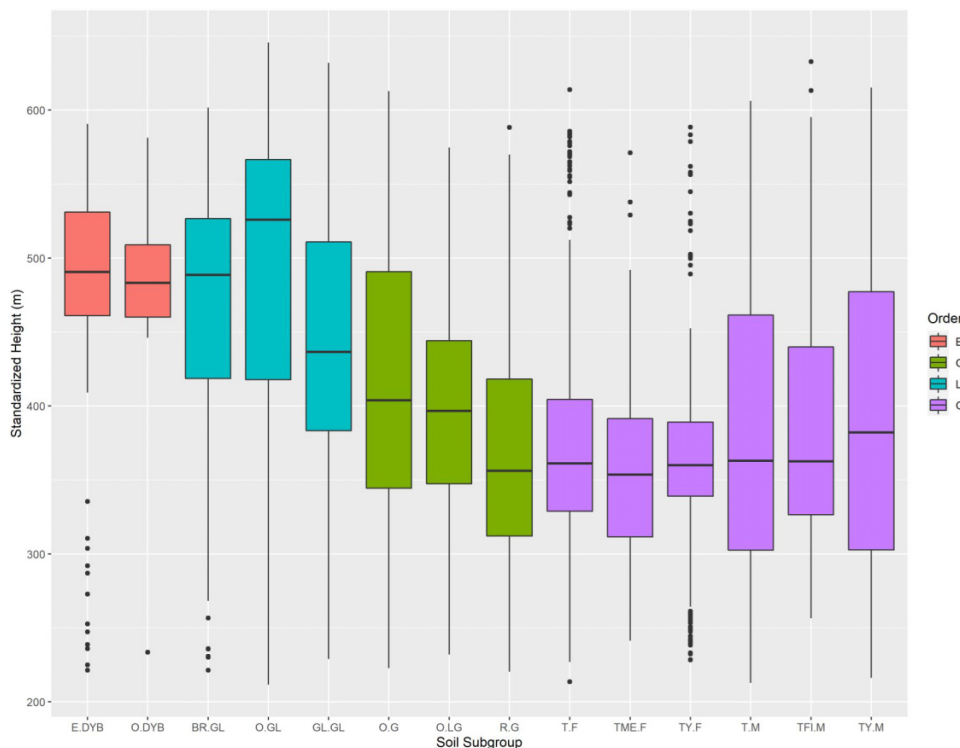


Fig. 6. Standardized height values for soil subgroups with more than 50 occurrences in the point data set. The color coding corresponds to soil orders, with B corresponding to Brunisol, G for Gleysols, L for Luvisols, and O for Organic soils. [Colour online.]



further optimization of class balancing is needed, particularly for studies using historical data like this study, as sample design greatly affects results in PSM (Heung et al. 2016). Some researchers have proposed including geographic location and Euclidean distance fields as explicit covariates for soil and terrain mapping (Behrens et al. 2018). However the inclusion of geographic location as model covariates reduces model interpretability (Wadoux et al. 2020), compared to the approach we took in this study. While random forest models are a black box, at least with a smaller number of features the feature importance values enable some inferences to be made about what is driving the predictions.

Conclusion

Multi-temporal Sentinel-derived environmental covariates, used along with landscape-scale morphometric terrain attributes, are valuable to PSM in the Canadian Boreal Plains. Deriving these covariates has become easier with tools such as Google Earth Engine (Gorelick et al. 2017). The second major outcome from this study was that constraining the predictions based on more general levels of precision led to more consistent maps across soil class levels without a reduction in accuracy. While overall performance at the order, great-group, and subgroup levels was satisfactory, predicting rare soils was a challenge, especially as the data set was not designed for PSM. Future work on class balancing during model building with these types of data sets is necessary and, when possible, careful sample design that fully characterizes landscape and soil variability in a balanced manner is critical. Given the satisfactory predictive accuracy of the PSM models built using only historic, circumstantial soil data for training, this study indicates that there is the potential to utilize these PSM methods in the Boreal Plains of Canada as a cost-effective way to produce soil map products that could contribute to better informed resource management.

Acknowledgements

The authors thank Mr. David Spiess for input and project support. We would also like to thank the anonymous reviewers who reviewed an earlier version of this manuscript and provided very valuable feedback.

Article information

History dates

Received: 20 February 2022

Accepted: 18 May 2022

Accepted manuscript online: 13 June 2022

Version of record online: 14 November 2022

Notes

This paper is part of a Collection entitled “Advances in Soil Survey & Classification in Canada”.

Copyright

© 2022 The Author(s). This work is licensed under a [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/) (CC BY 4.0),

which permits unrestricted use, distribution, and reproduction in any medium, provided the original author(s) and source are credited.

Author information

Author ORCIDs

Preston T. Sorenson <https://orcid.org/0000-0002-2958-1246>

Jeremy Kiss <https://orcid.org/0000-0001-8195-1518>

Competing interests

The authors declare there are no competing interests.

Supplementary material

Supplementary data are available with the article at <https://doi.org/10.1139/cjss-2022-0028>.

References

- Baruck, J., Nestroy, O., Sartori, G., Baize, D., Traidl, R. Vrščaj, B., et al. 2016. Soil classification and mapping in the Alps: the current state and future challenges. *Geoderma*, **264**: 312–331. doi:[10.1016/j.geoderma.2015.08.005](https://doi.org/10.1016/j.geoderma.2015.08.005).
- Behrens, T., Schmidt, K., Viscarra Rossel, R.A., Gries, P., Scholten, T., and MacMillan, R.A. 2018. Spatial modelling with Euclidean distance fields and machine learning. *Eur. J. Soil Sci.* **69**: 757–770. doi:[10.1111/ejss.12687](https://doi.org/10.1111/ejss.12687).
- Belgiu, M., and Drăgu, L. 2016. Random forest in remote sensing: a review of applications and future directions. *ISPRS J. Photogramm. Remote Sens.* **114**: 24–31. doi:[10.1016/j.isprsjprs.2016.01.011](https://doi.org/10.1016/j.isprsjprs.2016.01.011).
- Beven, K.J., and Kirkby, M.J. 1979. A physically based, variable contributing area model of basin hydrology/Un modèle à base physique de zone d'appel variable de l'hydrologie du bassin versant. *Hydrol. Sci. Bull.* **24**: 43–69. doi:[10.1080/02626667909491834](https://doi.org/10.1080/02626667909491834).
- Boehner, J., and Selige, T. 2006. Spatial prediction of soil attributes using terrain analysis and climate regionalisation. In *SAGA: Analyses and Modelling Applications*. GÖTTINGER GEOGRAPHISCHE ABHANDLUNGEN. Edited by J. Boehner, K.R. McCloy and J. Strobl. [Online.] Available from <http://bmcmredresmethodol.biomedcentral.com/articles/10.1186/1471-2288-4-5>.
- Brenning, A., Bangs, D., and Becker, M. 2018. RSAGA: SAGA geoprocessing and terrain analysis. [Online.] Available from <https://cran.r-project.org/package=RSAGA>.
- Brungard, C.W., Boettinger, J.L., Duniway, M.C., Wills, S.A., and Edwards, T.C. 2015. Machine learning for predicting soil classes in three semi-arid landscapes. *Geoderma*, **239–240**: 68–83. doi:[10.1016/j.geoderma.2014.09.019](https://doi.org/10.1016/j.geoderma.2014.09.019).
- Carré, F., McBratney, A.B., Mayr, T., and Montanarella, L. 2007. Digital soil assessments: beyond DSM. *Geoderma*, **142**: 69–79. doi:[10.1016/j.geoderma.2007.08.015](https://doi.org/10.1016/j.geoderma.2007.08.015).
- Costa-Cabral, M.C., and Burges, S.J. 1994. Digital elevation model networks (DEMON): a model of flow over hillslopes for computation of contributing and dispersal areas. *Water Resour. Res.* **30**: 1681–1692. doi:[10.1029/93WR03512](https://doi.org/10.1029/93WR03512).
- Dalsgaard, L., Lange, H., Strand, L.T., Callesen, I., Borgen, S.K., Liski, J., and Astrup, R. 2016. Underestimation of boreal forest soil carbon stocks related to soil classification and drainage. *Can. J. For. Res.* **46**: 1413–1425. doi:[10.1139/cjfr-2015-0466](https://doi.org/10.1139/cjfr-2015-0466).
- Dangal, S., Sanderman, J., Wills, S., and Ramirez-Lopez, L. 2019. Accurate and precise prediction of soil properties from a large mid-infrared spectral library. *Soil Syst.* **3**: 11. doi:[10.3390/soilsystems3010011](https://doi.org/10.3390/soilsystems3010011).
- Deiss, L., Franzluebbers, A.J., and Moraes, A. 2017. Soil texture and organic carbon fractions predicted from near-infrared spectroscopy and geostatistics. *Soil Sci. Soc. Am. J.* **81**: 1222–1234. doi:[10.2136/sssaj2016.10.0326](https://doi.org/10.2136/sssaj2016.10.0326).
- DeLancey, E.R., Simms, J.F., Mahdianpari, M., Brisco, B., Mahoney, C., and Kariyeva, J. 2019. Comparing deep learning and shallow learning for

- large-scale wetland classification in Alberta. *Can. Remote Sens.* **12**: 2. doi:10.3390/rs12010002.
- Dematté, J.A.M., Fongaro, C.T., Rizzo, R., and Safanelli, J.L. 2018. Geospatial soil sensing system (GEOS3): a powerful data mining procedure to retrieve soil spectral reflectance from satellite images. *Remote Sens. Environ.* **212**: 161–175. doi:10.1016/j.rse.2018.04.047.
- Dixon, R.K., Brown, S., Houghton, R.A., Solomon, A.M., Trexler, M.C., and Wisniewski, J. 1994. Carbon pools and flux of global forest ecosystems. *Source Sci. New Ser.* **263**: 185–190. doi:10.1126/science.263.5144.
- Dobos, E., Micheli, E., Baumgardner, M.F., Biehl, L., and Helt, T. 2000. Use of combined digital elevation model and satellite radiometric data for regional soil mapping. *Geoderma*. **97**, 367–391. doi:10.1016/S0016-7061(00)00046-X.
- Douglas, R.K., Nawar, S., Alamar, M.C., Mouazen, A.M., and Coulon, F. 2018. Rapid prediction of total petroleum hydrocarbons concentration in contaminated soil using vis-NIR spectroscopy and regression techniques. *Sci. Total Environ.* **616–617**: 147–155. doi:10.1016/j.scitotenv.2017.10.323. PMID: 29127788.
- Esfandiarpour-Boroujeni, I., Shahini-Shamsabadi, M., Shirani, H., Mosleh, Z., Bagheri-Bodaghabadi, M., and Salehi, M.H. 2020. Assessment of different digital soil mapping methods for prediction of soil classes in the Shahrekord plain, Central Iran. *CATENA*, **193**: 104648. doi:10.1016/j.catena.2020.104648.
- European Space Agency. 2021. Sentinel data. [online] Available from <https://scihub.copernicus.eu/> [accessed 19 May 2022].
- Fatholouloumi, S., Vaezi, A.R., Alavipanah, S.K., Ghorbani, A., Saurette, D., and Biswas, A. 2020. Improved digital soil mapping with multitemporal remotely sensed satellite data fusion: a case study in Iran. *Sci. Total Environ.* **721**: 137703. doi:10.1016/j.scitotenv.2020.137703. PMID: 32172111.
- Gallant, J.C., and Dowling, T.I. 2003. A multiresolution index of valley bottom flatness for mapping depositional areas. *Water Resour. Res.* **39**. 1347, doi:10.1029/2002WR001426.
- Ghatak, A. 2017. *Machine learning with R*. Springer, Singapore. doi:10.1007/978-981-10-6808-9.
- Gitelson, A.A., Merzlyak, M.N., and Chivkunova, O.B. 2001. Optical properties and nondestructive estimation of anthocyanin content in plant leaves. *Photochem. Photobiol.* **74**: 38. doi:10.1562/0031-8655(2001)074(0038:OPANEO)2.0.CO;2. PMID: 11460535.
- Gomes, L.C., Faria, R.M., de Souza, E., Veloso, G.V., Schaefer, C.E.G.R., and Filho, E.I.F. 2019. Modelling and mapping soil organic carbon stocks in Brazil. *Geoderma* **340**: 337–350. doi:10.1016/j.geoderma.2019.01.007.
- Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., and Moore, R. 2017. Google Earth Engine: planetary-scale geospatial analysis for everyone. *Remote Sens. Environ.* **202**: 18–27. doi:10.1016/j.rse.2017.06.031.
- Government of Alberta. 2021. Environmental impact assessments. [online] Available from <ftp://ftp.gov.ab.ca/env/fs/EIA/> [2021 Apr. 22].
- Gruber, S., and Peckham, S. 2009. Land-surface parameters and objects in hydrology. *In* *Developments in soil science*. pp. 171–194. doi:10.1016/S0166-2481(08)00007-X.
- Guisan, A., Weiss, S.B., and Weiss, A.D. 1999. GLM versus CCA spatial modeling of plant species distribution. *Plant Ecol.* **143**: 107–122. doi:10.1023/A:1009841519580.
- Guo, L., Fu, P., Shi, T., Chen, Y., Zhang, H., Meng, R., and Wang, S. 2020. Mapping field-scale soil organic carbon with unmanned aircraft system-acquired time series multispectral images. *Soil Tillage Res.* **196**: 104477. doi:10.1016/j.still.2019.104477.
- Hamzhepour, N., Shafizadeh-Moghadam, H., and Valavi, R. 2019. Exploring the driving forces and digital mapping of soil organic carbon using remote sensing and soil texture. *CATENA*, **182**: 104141. doi:10.1016/j.catena.2019.104141.
- Heung, B., Ho, H.C., Zhang, J., Knudby, A., Bulmer, C.E., and Schmidt, M.G. 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.
- Heung, B., Hodul, M., and Schmidt, M.G. 2017. Comparing the use of training data derived from legacy soil pits and soil survey polygons for mapping soil classes. *Geoderma*, **290**: 51–68. doi:10.1016/j.geoderma.2016.12.001.
- Japan Aerospace Exploration Agency. 2015. ALOS global digital surface model. [online] Available from <https://www.eorc.jaxa.jp/ALOS/en/aw3d30/index.htm> [accessed 24 September 2021].
- Jenny, H. 1941. Factors of soil formation. *Soil Sci.* **52**: 415. doi:10.1097/00010694-194111000-00009.
- Kasraei, B., Heung, B., Saurette, D.D., Schmidt, M.G., Bulmer, C.E., and Bethel, W. 2021. Quantile regression as a generic approach for estimating uncertainty of digital soil maps produced from machine-learning. *Environ. Model. Softw.* **144**: 105139. doi:10.1016/j.envsoft.2021.105139.
- Kiss, J., and Bedard-Haughn, A. 2021. Predictive mapping of solute-rich wetlands in the Canadian Prairie Pothole Region through high-resolution digital elevation model analyses. *Wetlands*, **41**: 38. doi:10.1007/s13157-021-01436-3.
- Koethe, R., and Lehmeier, F. 1996. *SARA-Systeme zur automatischen relief-analyse, Benutzerhandbuch*. Goettingen University.
- Kurz, W.A., Shaw, C.H., Boisvenue, C., Stinson, G., Metsaranta, J., Leckie, D., et al. 2013. Carbon in Canada's boreal forest: a synthesis. *Environ. Rev.* **21**: 260–292. doi:10.1139/er-2013-004. doi:10.1139/er-2013-0041.
- Laurin, G.V., Balling, J., Corona, P., Mattioli, W., Papale, D., Puletti, N., et al. 2018. Above-ground biomass prediction by Sentinel-1 multitemporal data in central Italy with integration of ALOS2 and Sentinel-2 data. *J. Appl. Remote Sens.* **12**: 1. doi:10.1117/1.JRS.12.016008.
- Lu, B., and Hardin, J. 2021. A unified framework for random forest prediction error estimation. *J. Mach. Learn. Res.* **22**: 1–41.
- Malone, B.P., Minansy, B., and Brungard, C. 2019. Some methods to improve the utility of conditioned Latin hypercube sampling. *PeerJ*, **7**: 1–17. doi:10.7717/peerj.6451.
- Maynard, J.J., and Levi, M.R. 2017. Hyper-temporal remote sensing for digital soil mapping: characterizing soil-vegetation response to climatic variability. *Geoderma*, **285**: 94–109. doi:10.1016/j.geoderma.2016.09.024.
- McHugh, M.L. 2012. Lessons in biostatistics interrater reliability: the kappa statistic. *Biochem. Med.* **22**: 276–282. <https://hrca.srce.hr/89395>. doi:10.11613/BM.2012.031.
- Moore, I.D., Grayson, R.B., and Ladson, A.R. 1991. Digital terrain modelling: a review of hydrological, geomorphological, and biological applications. *Hydrol. Process.* **5**: 3–30. doi:10.1002/hyp.3360050103.
- Mulder, V.L., de Bruin, S., Schaeppman, M.E., and Mayr, T.R. 2011. The use of remote sensing in soil and terrain mapping: a review. *Geoderma*, **162**: 1–19. doi:10.1016/j.geoderma.2010.12.018.
- Nikiforova, A.A., Fleis, M.E., Nyrtsov, M.V., Kazantsev, N.N., Kim, K.V., Belyonova, N.K., and Kim, J.K. 2020. Problems of modern soil mapping and ways to solve them. *CATENA*, **195**: 104885. doi:10.1016/j.catena.2020.104885.
- Pennock, D., Bedard-Haughn, A., Kiss, J., and van der Kamp, G. 2014. Application of hydrogeology to predictive mapping of wetland soils in the Canadian Prairie Pothole Region. *Geoderma*, **235–236**: 199–211. doi:10.1016/j.geoderma.2014.07.008.
- Poggio, L., Gimona, A., and 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:10.1016/j.geoderma.2013.05.029.
- Riley, S.J., DeGloria, S.D., and Elliot, R. 1999. A terrain ruggedness index that quantifies topographic heterogeneity. *Intermt. J. Sci.* **5**: 23–27.
- Rizzo, R., Medeiros, L.G., Mello, D.C., Marques, K.P.P., Mendes, W. de, S., et al. 2020. Multi-temporal bare surface image associated with transfer functions to support soil classification and mapping in south-eastern Brazil. *Geoderma*, **361**: 114018. doi:10.1016/j.geoderma.2019.114018.
- Rogge, D., Bauer, A., Zeidler, J., Mueller, A., Esch, T., and Heiden, U. 2018. Building an exposed soil composite processor (SCMaP) for mapping spatial and temporal characteristics of soils with Landsat imagery (1984–2014). *Remote Sens. Environ.* **205**: 1–17. doi:10.1016/j.rse.2017.11.004.
- Roudier, P., Hedley, C.B., Lobsey, C.R., Viscarra Rossel, R.A., and Leroux, C. 2017. Evaluation of two methods to eliminate the effect of water from soil vis-NIR spectra for predictions of organic carbon. *Geoderma*, **296**: 98–107. doi:10.1016/j.geoderma.2017.02.014.
- Rudiyanto Minasny, B., Setiawan, B.I., Saptomo, S.K., and McBratney, A.B. 2018. Open digital mapping as a cost-effective method for mapping

- peat thickness and assessing the carbon stock of tropical peatlands. *Geoderma*, **313**: 25–40. doi:[10.1016/j.geoderma.2017.10.018](https://doi.org/10.1016/j.geoderma.2017.10.018).
- Sanderman, J., Hengl, T., Fiske, G., Solvik, K., Adame, M.F. Benson, L., et al. 2018. A global map of mangrove forest soil carbon at 30 m spatial resolution. *Environ. Res. Lett.* **13**. 055002. doi:[10.1088/1748-9326/aabe1c](https://doi.org/10.1088/1748-9326/aabe1c).
- Sanderman, J., Savage, K., and Dangal, S.R.S. 2020. Mid-infrared spectroscopy for prediction of soil health indicators in the United States. *Soil Sci. Soc. Am. J.* **84**: 251–261. doi:[10.1002/saj2.20009](https://doi.org/10.1002/saj2.20009).
- Sazib, N., Mladenova, I., and Bolten, J. 2018. Leveraging the google earth engine for drought assessment using global soil moisture data. *Remote Sens.* **10**, 1265. doi:[10.3390/rs10081265](https://doi.org/10.3390/rs10081265).
- Sena, N.C., Veloso, G.V., Fernandes-Filho, E.I., Francelino, M.R., and Schaefer, C.E.G.R. 2020. Analysis of terrain attributes in different spatial resolutions for digital soil mapping application in southeastern Brazil. *Geoderma Reg.* **21**: e00268. doi:[10.1016/j.geodrs.2020.e00268](https://doi.org/10.1016/j.geodrs.2020.e00268).
- Shariffar, A., Sarmadian, F., Malone, B.P., and Minasny, B. 2019. Addressing the issue of digital mapping of soil classes with imbalanced class observations. *Geoderma*, **350**: 84–92. doi:[10.1016/j.geoderma.2019.05.016](https://doi.org/10.1016/j.geoderma.2019.05.016).
- Shi, J., Yang, L., Zhu, A.-X., Qin, C., Liang, P., Zeng, C., and Pei, T. 2018. Machine-Learning variables at different scales vs. knowledge-based variables for mapping multiple soil properties. *Soil Sci. Soc. Am. J.* **82**: 645–656. doi:[10.2136/sssaj2017.11.0392](https://doi.org/10.2136/sssaj2017.11.0392).
- Soil Classification Working Group 1998. The Canadian system of soil classification. *Can. Syst. Soil Classif.* **1646**: 187.
- Sorenson, P.T., Quideau, S.A., Rivard, B., and Dyck, M. 2020. Distribution mapping of soil profile carbon and nitrogen with laboratory imaging spectroscopy. *Geoderma*, **359**: 113982. doi:[10.1016/j.geoderma.2019.113982](https://doi.org/10.1016/j.geoderma.2019.113982).
- Sorenson, P.T., Shirtliffe, S.J., and Bedard-Haughn, A.K. 2021. Predictive soil mapping using historic bare soil composite imagery and legacy soil survey data. *Geoderma*, **401**: 115316. doi:[10.1016/j.geoderma.2021.115316](https://doi.org/10.1016/j.geoderma.2021.115316).
- Statistics Canada. 2021. Census subdivision boundary file. [online] Available from <https://www150.statcan.gc.ca/n1/en/catalogue/92-162-X> [accessed 19 May 2022].
- Stevens, A., and Ramirez-Lopez, L. 2014. An introduction to the prospectr package. [Online.] Available from <https://cran.r-project.org/web/packages/prospectr/vignettes/prospectr-intro.pdf>.
- Taghizadeh-Mehrjardi, R., Nabiollahi, K., Minasny, B., and Triantafyllis, J. 2015. Comparing data mining classifiers to predict spatial distribution of USDA-family soil groups in Baneh region, Iran. *Geoderma*, **253–254**: 67–77. doi:[10.1016/j.geoderma.2015.04.008](https://doi.org/10.1016/j.geoderma.2015.04.008).
- Taghizadeh-Mehrjardi, R., Nabiollahi, K., and Kerry, R. 2016. Digital mapping of soil organic carbon at multiple depths using different data mining techniques in Baneh region, Iran. *Geoderma*, **266**: 98–110. doi:[10.1016/j.geoderma.2015.12.003](https://doi.org/10.1016/j.geoderma.2015.12.003).
- Tajik, S., Ayoubi, S., Shirani, H., and Zeraatpisheh, M. 2019. Digital mapping of soil invertebrates using environmental attributes in a deciduous forest ecosystem. *Geoderma*, **353**: 252–263. doi:[10.1016/j.geoderma.2019.07.005](https://doi.org/10.1016/j.geoderma.2019.07.005).
- Vågen, T.G., and Winowiecki, L.A. 2019. Predicting the spatial distribution and severity of soil erosion in the global tropics using satellite remote sensing. *Remote Sens.* **11**: 1–18. doi:[10.3390/rs11151800](https://doi.org/10.3390/rs11151800).
- Viscarra Rossel, R.A., Lobsey, C.R., Sharman, C., Flick, P., and McLachlan, G. 2017. Novel proximal sensing for monitoring soil organic C stocks and condition. *Environ. Sci. Technol.* **51**: 5630–5641. doi:[10.1021/acs.est.7b00889](https://doi.org/10.1021/acs.est.7b00889). PMID: 28414454.
- Wadoux, A.M.J.C., Minasny, B., and McBratney, A.B. 2020. Machine learning for digital soil mapping: applications, challenges and suggested solutions. *Earth-Science Rev.* **210**: 103359. doi:[10.1016/j.earscirev.2020.103359](https://doi.org/10.1016/j.earscirev.2020.103359).
- Wright, M.N., and Ziegler, A. 2017. ranger : a fast implementation of random forests for high dimensional data in C++ and R. *J. Stat. Softw.* **77**, 1–17. doi:[10.18637/jss.v077.i01](https://doi.org/10.18637/jss.v077.i01).
- Xiong, J., Thenkabail, P.S., Tilton, J.C., Gumma, M.K., Teluguntla, P., Oliphant, A., et al. 2017. Nominal 30-m cropland extent map of continental Africa by integrating pixel-based and object-based algorithms using Sentinel-2 and Landsat-8 data on google earth engine. *Remote Sens.* **9**: 1–27. doi:[10.3390/rs9101065](https://doi.org/10.3390/rs9101065).
- Yan, J., Wang, L., Song, W., Chen, Y., Chen, X., and Deng, Z. 2019. A time-series classification approach based on change detection for rapid land cover mapping. *ISPRS J. Photogramm. Remote Sens.* **158**: 249–262. doi:[10.1016/j.isprsjprs.2019.10.003](https://doi.org/10.1016/j.isprsjprs.2019.10.003).
- Yang, H., Zhang, X., Xu, M., Shao, S., Wang, X. Liu, W., et al. 2020. Hyper-temporal remote sensing data in bare soil period and terrain attributes for digital soil mapping in the black soil regions of China. *CATENA*, **184**: 104259. doi:[10.1016/j.catena.2019.104259](https://doi.org/10.1016/j.catena.2019.104259).
- Zevenbergen, L.W., and Thorne, C.R. 1987. Quantitative analysis of land surface topography. *Earth Surf. Process. Landforms*, **12**: 47–56. doi:[10.1002/esp.3290120107](https://doi.org/10.1002/esp.3290120107).
- Zhang, Y., Guo, L., Chen, Y., Shi, T., Luo, M. Ju, Q.L., et al. 2019. Prediction of soil organic carbon based on Landsat 8 monthly NDVI data for the Jiangnan Plain in Hubei Province, China. *Remote Sens.* **11**, 1683. doi:[10.3390/rs11141683](https://doi.org/10.3390/rs11141683).