NRC Research Press

ARTICLE

Digital soil mapping workflow for forest resource applications: a case study in the Hearst Forest, Ontario

Christopher Blackford, Brandon Heung, Ken Baldwin, Robert L. Fleming, Paul W. Hazlett, Dave M. Morris, Peter W.C. Uhlig, and Kara L. Webster

Abstract: Accurate soil information is critically important for forest management planning and operations but is challenging to map. Digital soil mapping (DSM) improves upon the limitations of conventional soil mapping by explicitly linking a variety of environmental data layers to spatial soil point data sets to continuously predict soil variability across a landscape. Thus far, much DSM research has focussed on the development of ultrafine-resolution soil maps within agricultural systems; however, increasing availability of light detection and ranging (LiDAR) data presents new opportunities to apply DSM to support forest resource applications at multiple scales. This project describes a DSM workflow using LiDAR-derived elevation data and machine learning models (MLMs) to predict key forest soil attributes. A case study in the Hearst Forest in northeastern Ontario, Canada, is used to illustrate the workflow. We applied multiple MLMs to the Hearst Forest to predict soil moisture regime and textural class. Both qualitative and quantitative assessment pointed to the random forest MLM producing the best maps (63% accuracy for moisture regime and 66% accuracy for textural class). Where error occurred, soils were typically misclassified to neighbouring classes. This standardized, flexible workflow is a valuable tool for practitioners that want to undertake DSM as part of forest resource management and planning.

Key words: digital soil mapping, machine learning, forest management.

Résumé : Des informations précises sur les sols sont absolument essentielles pour la planification et les opérations d'aménagement forestier mais elles sont difficiles à cartographier. La cartographie numérique des sols (CNS) constitue un progrès par rapport aux limites de la cartographie conventionnelle des sols en reliant une variété de couches de données environnementales à des ensembles de données spatiales ponctuelles des sols pour prédire la variabilité à travers un paysage de façon continue. Jusqu'à maintenant, beaucoup de travaux de recherche sur la CNS ont mis l'accent sur le développement de cartes des sols à très haute résolution pour des systèmes agricoles. Cependant, la disponibilité croissante de données lidar offre de nouvelles opportunités d'appliquer la CNS en support à des applications qui concernent les ressources forestières à de multiples échelles. Ce projet décrit un flux de travail de CNS qui utilise des données altimétriques dérivées du lidar et des modèles d'apprentissage automatique (MAA) pour prédire des attributs importants des sols. Une étude de cas dans la forêt de Hearst, dans le nord-est de l'Ontario, au Canada, est utilisée pour illustrer le flux de travail. Nous avons appliqué plusieurs MAA à la forêt de Hearst pour prédire le régime d'humidité et la classe de texture. Une évaluation tant qualitative que quantitative indiquait que le MAA de forêt aléatoire produisait les meilleures cartes (précision de 63 % pour le régime d'humidité et de 66 % pour la classe de texture). Lorsqu'il y avait des erreurs, les sols mal classés étaient typiquement placés dans les classes voisines. Ce flux de travail standardisé et flexible est un outil précieux pour les praticiens qui veulent entreprendre la CNS en tant que composante de la planification et de la gestion des ressources forestières. [Traduit par la Rédaction]

Mots-clés : cartographie numérique des sols, apprentissage automatique, aménagement forestier.

1. Introduction

Soils are a critical element of forest ecosystems in that soil mineralogy, nutrient supply, moisture retention, texture, structure, and porosity collectively influence forest composition and productivity (Leniham 1993; Drever and Lertzman 2001; Nigh 2006; Nijland et al. 2015; Binkley and Fisher 2019). Within forested ecosystems, soils also contribute to important ecosystem services such as water filtration, nutrient cycling, and carbon sequestration (Adhikari and Hartemink 2016; Baveye et al. 2016). Knowledge of the distribution of soil properties across forested landscapes will improve our ability to practice sustainable forest management at multiple spatial scales (i.e., site, stand, and landscape levels).

Conventional soil mapping approaches combine soil pedon observations with expert interpretation of aerial orthophotogrammetry to generate a soil classification polygon map based on observable differences in landscape traits (e.g., slope position, vegetation, and geomorphic features) (Scull et al. 2003). Although conventional soil maps can be good sources of general soil

Received 13 February 2020. Accepted 30 June 2020.

C. Blackford, K. Baldwin, R.L. Fleming, P.W. Hazlett, and K.L. Webster. Great Lakes Forestry Centre, Canadian Forest Service, Natural Resources Canada, Sault Ste. Marie, ON, Canada.
B. Heung. Faculty of Agriculture, Dalhousie University, Truro, NS, Canada.
D.M. Morris. Centre for Northern Forest Ecosystem Research, Ministry of Natural Resources and Forestry, Thunder Bay, ON, Canada.
P.W.C. Uhlig. Ontario Forest Research Institute, Ministry of Natural Resources and Forestry, Sault Ste. Marie, ON, Canada.
Corresponding author: Christopher Blackford (email: chris.j.blackford@gmail.com).

knowledge, they have many drawbacks that limit their utility at site- or stand-level forest management (McKenzie and Ryan 1999; Terribile et al. 2011). First, accuracies of conventional soil maps are not always reported or known. Second, soils and their properties are delineated into discrete areas (i.e., spatial polygons or points) when, in reality, soil varies continuously across the landscape, as influenced by the soil forming factors (i.e., climate, organisms, relief, parent material, and time) (Jenny 1941). Third, conventional soil maps developed for forested systems are often created at small scales (e.g., 1:250 000), which are too coarse to capture the short-range soil variability needed for site-level management; in comparison, conventional soil maps in agricultural systems are created at larger map scales (e.g., 1:20 000 to 1:100 000). Fourth, the time and associated costs required to develop soil maps using conventional soil mapping approaches are high, leaving large areas of Canadian forests unmapped.

To address the limitations of conventional soil mapping, soil surveys are frequently being replaced by computationally driven, digital approaches. Digital soil mapping (DSM) is an emerging discipline that draws on the field of soil science, geographical information science (geographic information systems (GIS)), and spatial statistics that aims to improve upon conventional soil mapping approaches by providing high-resolution soil maps matched to soil management scales for planning, implementation, and evaluation (McBratney et al. 2003; Scull et al. 2003; Minasny and McBratney 2016). Although the discipline existed as early as the 1970s (e.g., Webster and Burrough 1972*a*, 1972*b*), advancements in computing, remote sensing, GIS, data mining, and machine learning and the increasing availability of spatial data sets have greatly facilitated the production of DSM products since the 2000s (McBratney et al. 2003; Scull et al. 2003; Minasny and McBratney 2016). Furthermore, technological advancements have also allowed for digital soil maps to be produced at progressively larger spatial extents and higher resolutions (Minasny and McBratney 2016; Bulmer et al. 2019).

In DSM, statistical models are used to discover the relationships between georeferenced soil observations (with known soil properties and classes) and environment data (McKenzie and Ryan 1999). These relationships are then used to make spatial soil predictions for unsampled locations. Conceptual models linking soils and the environment have long existed in the field of pedology, with the most classic model being the CLORPT model of soil formation (Jenny 1941). This model postulates that the development and distribution of soil properties and classes across landscapes are a function of climate (CL), organisms (O), relief (R), parent material (P), and time of soil development (T). McBratney et al. (2003) updated Jenny's model for application in DSM by introducing the SCORPAN model, in which the soil properties at a particular point in space are a function of other measured properties of the soil (S), climate (C), organisms (O), relief (R), parent material (P), age of the soil (A), and the spatial location (N). DSM spatially intersects georeferenced soil observations (response variables) with spatial layers that represent the SCORPAN variables (predictor variables), and statistical models are used to infer the empirical relationship between these predictor and response variables. These relationships are then used to predict the spatial distribution of a wide variety of soil properties, for example, soil type (e.g., Heung et al. 2016, 2017), soil pH (e.g., Reuter et al. 2008), and soil organic carbon (e.g., Minasny et al. 2013), on a pixel-by-pixel basis (i.e., as a raster) across the landscape.

DSM is now a routinely used tool for precision agriculture (Kühn et al. 2009; Söderström et al. 2016), but applications in forestry are more limited. Although efforts to map forest soils have been successful at local scales in Canadian landscapes (e.g., Webster et al. 2008; Akumu et al. 2015, 2019), attempts at applying DSM techniques at larger spatial extents in forestry (e.g., provincial and national scales) have had limited success (e.g., Mansuy et al. 2014). The challenges of mapping soil across large forest extents are largely due to the limited availability of forest soil pedon data (e.g., lack of historical data and challenges to soil sampling in remote and difficult-to-access areas) and the large amount of environmental variation, when compared with agricultural landscapes. These challenges can be overcome but require an understanding of the DSM process such that data collection and analysis are performed in an efficient, standardized way.

The techniques for developing digital soil maps are designed to provide a consistent, objective, and quantitative approach to soil prediction. Hence, the methodology of DSM usually follows a generic structure, which includes acquisition of environmental data layers that represent the SCORPAN factors, acquisition of georeferenced soil observations, spatial intersection of soil observations with environmental layers, predictive modelling, and assessments of model accuracy and uncertainties (see Fig. 1). As spatial environmental databases, open plot data sources, and computational power grow, DSMs will become more accessible to practitioners and at increasingly finer resolutions (Minasny and McBratney 2016). There are many different models and inputs that a researcher can use in the DSM process, and these decisions may result in very different outcomes. In other words, each map is only a realization of the soil patterns, and validation is therefore necessary to assess the accuracy of each map realization. For example, Heung et al. (2016) performed a comprehensive comparison of machine learning techniques and demonstrated that each learner produced visually distinct digital soil maps with varying levels of accuracy. Given the wide variety of models and analytical decisions at the researcher or forest practitioner's discretion, it can be unclear what steps need to be taken to generate the best model and map of the desired soil attribute(s) and how to interpret model outputs.

In this paper, a standardized, semiautomated workflow for DSM is presented for use in forest resource applications, with an overall goal to provide tools that make DSM more accessible and interpretable to researchers and forest practitioners. Important considerations related to data sources and availability, types of models, and interpreting model outputs are discussed. The steps taken in this workflow and the benefits and limitations of the approach are illustrated using a case study from the Hearst Forest Management Unit (15 218 km^2) in northeastern Ontario, Canada, to map soil moisture regime and soil textural classes.

2. Methods

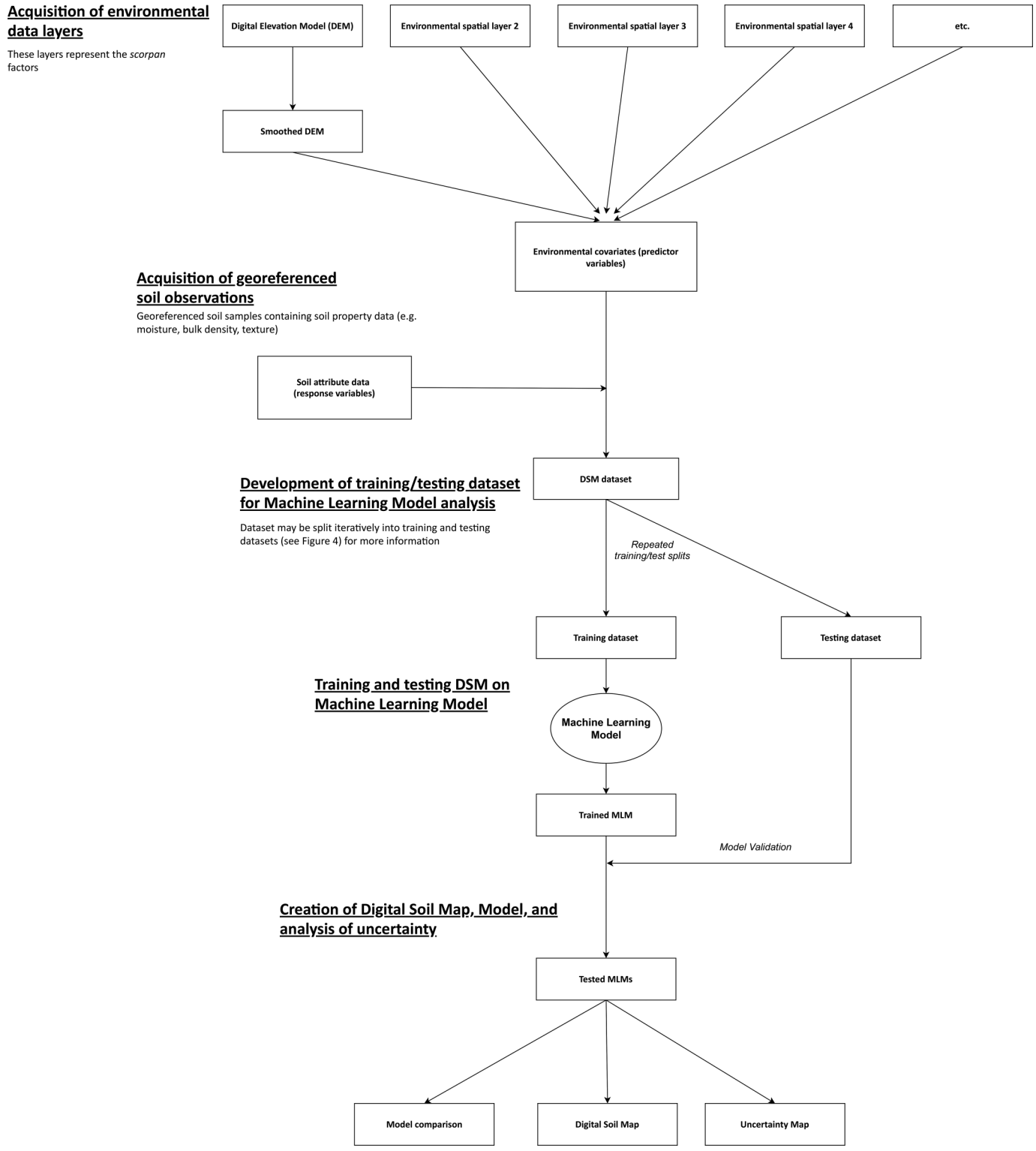
Our general DSM workflow is structured to address some of the major considerations when undertaking a DSM project in forested systems:

1. defining the extent of the study area,
2. identifying the key soil properties to be predicted,
3. acquiring and harmonizing soil data for the study area,
4. modelling the relationship between the SCORPAN factors and soil data,
5. evaluating and comparing the predictive models, and
6. assessing the accuracy and uncertainty of the spatial predictions.

2.1. General digital soil mapping workflow

A workflow describing the recommended analytical steps for undertaking DSM is illustrated in Fig. 1. Each step in the workflow can be automated using the R statistical software (R Core

Fig. 1. Workflow for digital soil mapping (DSM).



Team 2018) and its integration with SAGA GIS (Conrad et al. 2015), both of which are freely available online. We automated this workflow for our Hearst Forest case study, and the code we used to run our analyses is freely available online (Blackford 2020) and can be downloaded and modified to suit others' needs. The main R packages used for this study included RSAGA (Brenning et al. 2018), raster (Hijmans 2019), and caret (Kuhn et al. 2019).

2.2. Study area

In all DSM projects, the study area should meet several criteria. It should be located either where there are prior soil data available or where soil pedon descriptions and (or) sampling can be performed, as the process relies on georeferenced soil pedon descriptions. The study area should also be of relevant size for forest resource management planning or implementation (e.g., a forest management unit) and ideally should contain environmental heterogeneity, representative of the larger landscape. The better the study area reflects the natural environmental heterogeneity, the more likely the model will be transferable outside of the study area (Bui et al. 2007).

For this case study, we used the Hearst Forest, in northeastern Ontario, Canada. The Hearst Forest is a managed forest, located around the town of Hearst, Ontario (49°41′16.2″N, 83°40′21.2″W), and has an area of approximately 15 218 km^2 (Fig. 2). It was chosen because of the availability of a high-resolution digital elevation model (DEM) derived from light detection and ranging (LiDAR), provided by Hearst Forest Management Inc., and the availability of a large soil pedon data set (7893 spatial points).

The study area is located on Precambrian Shield, is of moderate relief, and is covered by Quaternary age sediments (Blackburn et al. 1985; Mackasey et al. 1974; Thurston 1991). The northern and central areas of the forest are characterized by an extensive clay plain, known as the Clay Belt, deposited during inundation by proglacial Lake Barlow–Ojibway about 9000 years ago (Dyke 2004); however in other areas of the forest, loamy and sandy soils can be found (Hearst Forest Management 2019). Many soils in the Hearst Forest are poorly drained, and organic soil is common throughout (Hearst Forest Management 2019). Esker complexes from previous glaciation can be found in the centre of the forest.

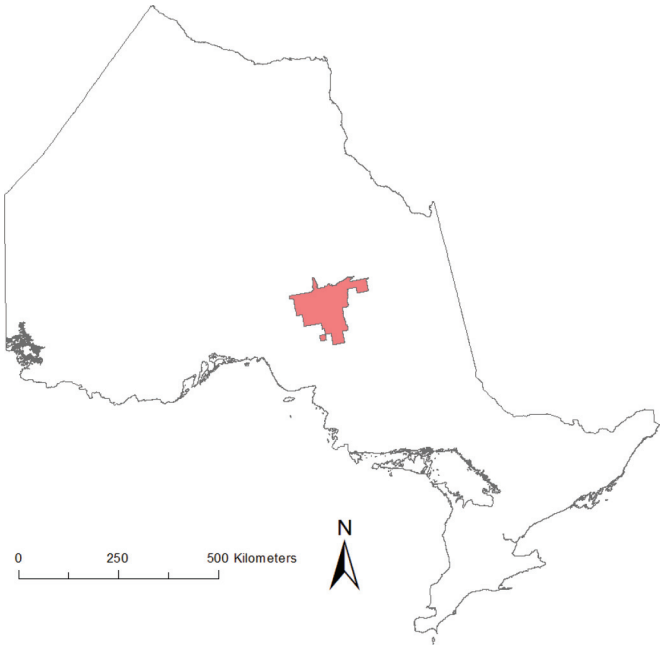
The Hearst Forest is representative of the 3E Boreal Shield ecoregion within Ontario (Crins et al. 2009) and is actively managed by Hearst Forest Management Inc. Small changes in topography can have large implications for soil composition and the ability of machinery to operate. The study area is dominated by black spruce (*Picea mariana* (Mill.) B.S.P.) in the low-lying areas and by trembling aspen (*Populus tremuloides* Michx.) in the well-drained upland areas. Other common lowland species include eastern white cedar (*Thuja occidentalis* L.) and tamarack (*Larix laricina* (Du Roi) K. Koch). Balsam fir (*Abies balsamea* (L.) Mill.), white spruce (*Picea glauca* (Moench) Voss), and jack pine (*Pinus banksiana* Lamb.) are other common species found on the upland sites. Balsam poplar (*Populus balsamifera* L.) and white birch (*Betula papyrifera* Marshall) are some of the associated deciduous tree species found in the study area (in addition to *Populus tremuloides* previously mentioned).

2.3. Soil response data

In DSM, the soil properties to be predicted should be relevant to forest management objectives and research goals. Important properties incorporated in previous DSM efforts have included soil texture, soil moisture, nutrients, pH, and soil carbon, as well as more general taxonomic attributes like soil great group or soil class (e.g., Minasny et al. 2006; Reuter et al. 2008; Mansuy et al. 2014; Heung et al. 2017). Acquisition of spatially referenced soil point data may come from existing soil pedon data sets, or may be polygon-based data, which can be acquired from conventional soil maps. Methods are available to transform polygon-based data sets to spatial point data sets for use in DSM (Yang et al. 2011; Odgers et al. 2014; Heung et al. 2017). If there is a lack of soil data, additional sampling should be distributed across the study area and designed in a way that captures the inherent environmental variability (e.g., Minasny and McBratney 2006).

For the Hearst Forest case study, a combination of previously gathered federal, provincial, and targeted soil sampling was used to predict moisture regime and textural class (Table 1) (Johnson et al. 2015). Moisture regime was partitioned into xeric, dry, fresh, moist, wet, and inundated classes. Soil textural classes were determined from the "effective texture" found at a site (i.e., the dominant soil texture of the pedon). The textural classes were classified either as organic or, if a mineral soil was present, by the relative proportions of sand, silt, and clay. Federal soil data came from Canada's National Forest Inventory (NFI); provincial soil data were obtained from Ontario's Growth and Yield (G&Y) Program, as well as Forest Resource Inventory (FRI) and provincial Forest Ecosystem Classification (FEC) plots. The Hearst Forest also had targeted soil pedon data from the Advanced Forest Resource Inventory Technologies (AFRIT) project. In total, there were 7893 soil data points within the Hearst Forest (Fig. 3). Of these, 7734 had moisture regime determinations, and 7213 recorded textural class information. The Hearst Forest is a unique forest management unit where rich plot data were available, whereas other management units in Ontario would not have the same density of soil pedon data.

Fig. 2. Hearst Forest with respect to Ontario, Canada. Map was generated using ArcGIS software (Esri, Redlands, Calif., USA). [Colour online.]



2.4. Predictor data (environmental covariates)

When selecting the appropriate environmental predictor layers to represent the SCORPAN factors, the extent of the study area needs to be taken into consideration. For study areas with a small extent, soil variability is often controlled by short-range variability in relief, hydrology, and vegetation, while climate is often assumed to be constant. Over larger spatial extents, climate becomes an increasingly important driver of soil variability. These layers can be achieved from targeted environmental data collection (e.g., through aerial LiDAR), but if this is not feasible, in many areas of the world, elevation data are freely available (e.g., Yamazaki et al. 2017, 2019), as well as hydrology and vegetation layers (Didan 2015).

For the Hearst Forest case study, we used a LiDAR DEM to derive topographic metrics, a river and lake layer to represent hydrology, a bedrock and Quaternary geology layer, and three forest

Table 1. Data sources for the Hearst Forest case study area.

Data source	Soil information
Forest Resource Inventory (FRI) data sets	Texture, horizons, depth, mottle depth, moisture class, landform
National Forest Inventory (NFI) plots	Texture, horizons, depth, moisture class, landform, chemistry, bulk density
Provincial Growth and Yield (G&Y) plots	Texture, horizons, depth, depth to gley, moisture class, landform
Forest Ecosystem Classification (FEC) or Ecological Land Classification (ELC) plots	Texture, horizons, depth, depth to gley, moisture regime, landform, topographic position, soil chemistry (some)
Independent research programs • Advanced Forest Resource Inventory Technologies (AFRIT)	Texture, horizons, depth, depth to gley, moisture class

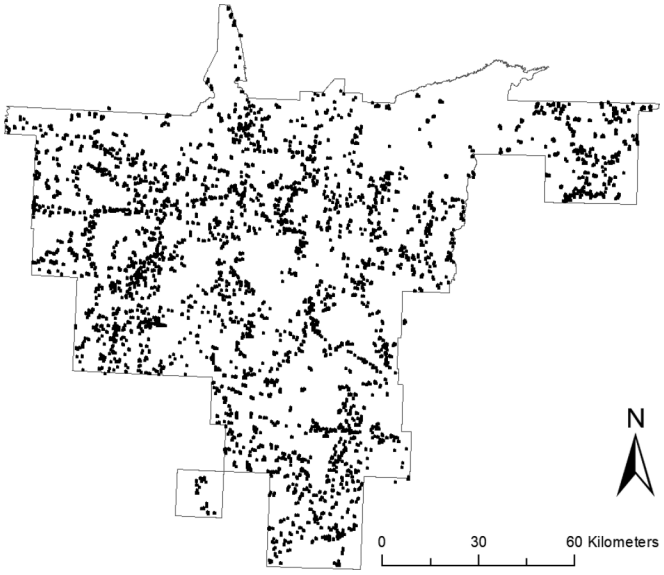
survey layers to describe forest composition that, collectively, represent the soil forming factors.

Calculations were performed on the DEM and river and lake layers to generate a suite of metrics describing the elevational change and hydrology of the study area. The DEM had a 10 m × 10 m resolution and was "smoothed" by passing a 201 cell × 201 cell moving window filter across the extent, which averaged the elevation values within the window. In DSM, a smoothing process is often applied to reduce the effects of spatially uncorrelated noise from LiDAR-derived DEMs and to remove their anomalous pits and peaks (Li et al. 2011). Furthermore, smoothing helps to incorporate the topographic variability expressed at larger scales (Behrens et al. 2010). Determining the best approach to DEM smoothing is an interactive process and will be unique to each project as a function of the resolution and quality of the DEM available for each study area. For example, high-resolution LiDAR DEMs will accentuate microtopographic features (e.g., hummocks and hollows) that are less important to soil formation than mesotopographic features (e.g., surface curvature, slope position, and slope length), thereby requiring more smoothing than DEMs with coarser resolution. Conversely, coarse-resolution DEMs cannot capture heterogeneity at a finer scale.

We derived several topographic metrics from the smoothed DEM to characterize local-scale morphometry (e.g., slope, aspect, and curvature), landscape-scale morphometry (e.g., multiscale topographic position index), and hydrology (catchment area, catchment slope, modified catchment area, and topographic wetness index), all of which were derived in the SAGA program (Conrad et al. 2015) and run using R (R Core Team 2018). The DEM was "hydrologically conditioned" prior to calculating the topographic metrics representing hydrology. We used a river and waterbody layer as a mask and subtracted a fixed elevation of 30 m wherever this feature occurred on our DEM. Hydrologically conditioning the DEM in this way, prior to calculating topographic metrics, is important because it helps ensure that the hydrological flow routing algorithms used to calculate hydrological derivatives follow an expected path throughout the landscape towards known hydrological features. The river and lake layers were also used to derive distance to river and lake metrics to represent landscape-scale relief patterns.

The geology layers were rasterized from a polygon layer representing the sedimentation and bedrock of the region. The forestry layers represented the overstory height, understory height, and the overstory and understory leading species (i.e., most common species) and were rasterized from a polygon layer from the provincial FRI. Finally, we calculated Euclidean distance fields for our study area, corresponding to the distance from the *x* axis, *y* axis, northeastern extent, southeastern extent, northwestern extent, southwestern extent, and centre of our study extent. Euclidean distance fields are used to incorporate spatial position (i.e., the "N" in SCORPAN) into soil predictions (near points in space more likely to be similar than distant points). The full list of environmental covariates used in our DSM case study can be found in Table 2.

Fig. 3. The Hearst Forest study area and soil attribute points. Map was generated using ArcGIS software.



2.5. Machine learning for digital soil mapping

Machine learning is the (semi)automated process of discovering the complex relationships between predictor and response variables using computer-based approaches (Witten et al. 2005; Hastie et al. 2009). Machine learning models (MLMs) are often preferred in DSM over other statistical models (e.g., generalized linear models) because they have shown success at modelling the complex relationships between the predictor and response variables and require fewer assumptions on the form of the relationships between predictor and response variables. Depending on the specific machine learner, some learners are able to account for linear and nonlinear relationships, integrate discrete and continuous variables, handle nonparametric data, and be used in regression analysis or for classification purposes. Specific to classification purposes in DSM, Heung et al. (2016) provided a comprehensive overview of the most commonly used machine learning techniques such as tree-based learners, distance-based learners, artificial neural networks, model trees, and support vector machines. Furthermore, they provided a comparison amongst 10 learners and showed that different learners would result in quite different outputs when using the same input data. Differences in the maps and the accuracy among learners have also been observed in other model comparison studies, and the success of each learner varies across different landscapes (e.g., Taghizadeh-Mehrjardi et al. 2015; Brungard et al. 2015; Heung et al. 2017). Overall, it is unclear a priori which machine learner will yield the best soil model and map for a given study area and input variables. Hence, it can be useful to perform model

Table 2. Environmental covariates used in digital soil mapping (DSM) for the Hearst Forest.

Covariate	Representation	Data source
Aspect	Local relief	DEM
Downslope curvature	Local relief	DEM
General curvature	Local relief	DEM
Local curvature	Local relief	DEM
Local downslope curvature	Local relief	DEM
Local upslope curvature	Local relief	DEM
Maximum curvature	Local relief	DEM
Midslope position	Local relief	DEM
Minimum curvature	Local relief	DEM
Normalized height	Local relief	DEM
Plan curvature	Local relief	DEM
Profile curvature	Local relief	DEM
Real surface area	Local relief	DEM
Slope	Local relief	DEM
Standardized height	Local relief	DEM
Tangential curvature	Local relief	DEM
Terrain ruggedness index (Riley et al. 1999)	Local relief	DEM
Terrain surface concavity	Local relief	DEM
Terrain surface convexity	Local relief	DEM
Topographic negative openness	Local relief	DEM
Topographic positive openness	Local relief	DEM
Total curvature	Local relief	DEM
Upslope curvature	Local relief	DEM
Upslope height	Local relief	DEM
Multiresolution index of ridge top flatness (Gallant and Dowling 2003)	Landscape relief	DEM
Multiresolution index of valley bottom flatness (Gallant and Dowling 2003)	Landscape relief	DEM
Multiscale topographic position index	Landscape relief	DEM
Valley depth	Landscape relief	DEM
Stream distance	Landscape relief	River and lake geodatabase (https://data.ontario.ca/dataset/ontario-integrated-hydrology-data)
Waterbody distance	Landscape relief	River and lake geodatabase (https://data.ontario.ca/dataset/ontario-hydro-network-waterbody)
Catchment area	Hydrology	DEM
Catchment slope	Hydrology	DEM
Modified catchment area	Hydrology	DEM
Topographic wetness index	Hydrology	DEM
Bedrock geology	Parent material	Geology shapefile (https://data.ontario.ca/dataset/1250-000-scale-bedrock-geology-of-ontario)
Quaternary geology	Parent material	Geology shapefile (https://data.ontario.ca/dataset/quaternary-geology-of-ontario)
Overstory height	Organisms	Forestry inventory shapefile
Overstory leading species	Organisms	Forestry inventory shapefile
Understory height	Organisms	Forestry inventory shapefile
Understory leading species	Organisms	Forestry inventory shapefile
Distance from *x* axis	Spatial position	NA
Distance from *y* axis	Spatial position	NA
Distance from northeastern extent point	Spatial position	NA
Distance from southeastern extent point	Spatial position	NA
Distance from northwestern extent point	Spatial position	NA
Distance from southwestern extent point	Spatial position	NA
Distance from centre of extent	Spatial position	NA

Note: DEM, digital elevation model; NA, not applicable.

comparisons as part of best practices (Heung et al. 2017). Model comparison is one important part of model evaluation, which we address later.

2.6. Machine learning model training and prediction

MLMs require the user to partition their data set into training and testing data sets, often referred to as "folds". MLMs use algorithms to develop a model relating predictor and response variables on the training fold. It then evaluates how well the model performs by quantifying its accuracy to predict the testing fold response data using the testing fold predictor data. There are different ways data partitioning can be achieved (Hastie et al. 2009), and a discussion on data partitioning approaches is beyond the scope of this paper. For our case study, *k*-fold cross-validation was used, as it is one of the most popular approaches to data partitioning. In *k*-fold cross-validation, the data set is split into *k* number of folds, where $k - 1$ folds are used to train the model and the remaining fold is used to validate the model. This process is reiterated *k* times with each iteration using a different validation fold (Fig. 4). This approach is preferred over a simple onetime

Fig. 4. *k*-fold cross-validation procedure.

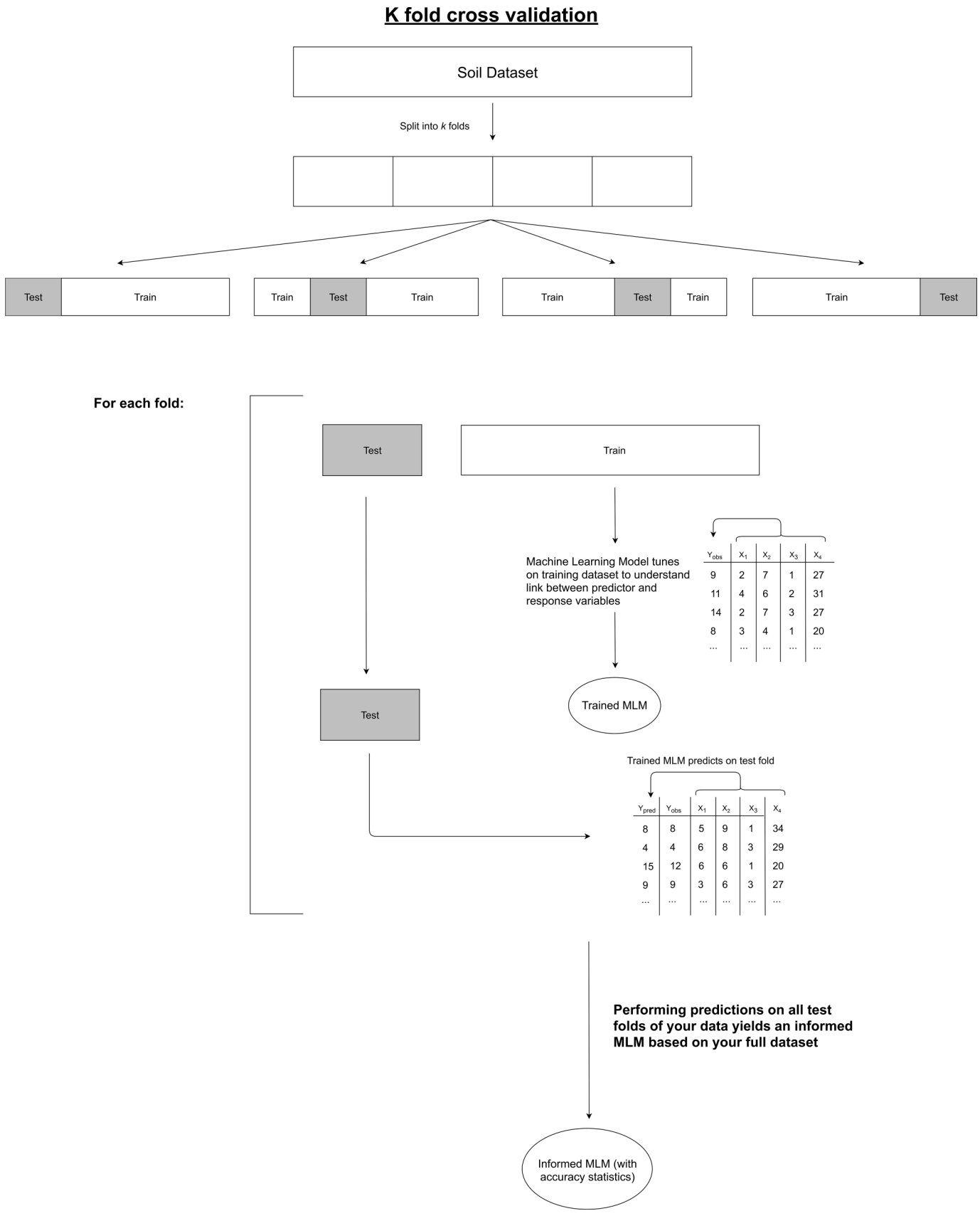
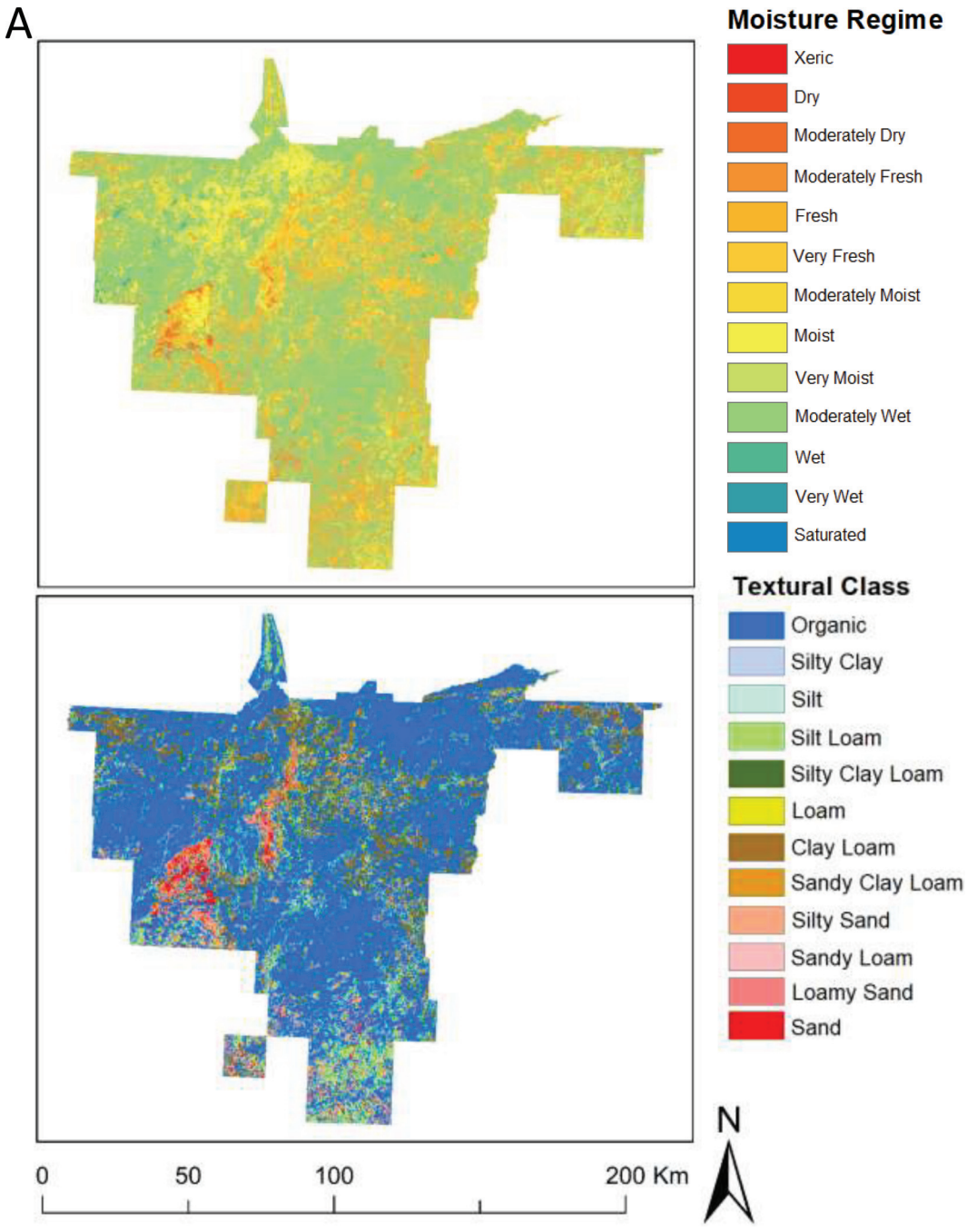


Fig. 5. Predicted moisture regime and textural class values for the (A) random forest (RF), (B) *k*-nearest neighbour (*k*-NN), and (C) support vector machine (SVM) models. Maps were generated using ArcGIS software. [Colour online.]



split (i.e., random holdback cross-validation) of the data set into training and testing folds because *k*-fold cross-validation enables the entire data set to be used in model training and validation. Furthermore, it provides a more robust measure of model accuracy through the multiple iterations of validation.

For our Hearst Forest case study, we partitioned our data using 10-fold cross-validation and compared results from three different commonly used MLMs (Hastie et al. 2009): *k*-nearest neighbour (*k*-NN) (Altman 1992), support vector machine (SVM) with a radial kernel (Cortes and Vapnik 1995), and random forest (RF) (Ho 1998; Breiman 2001).

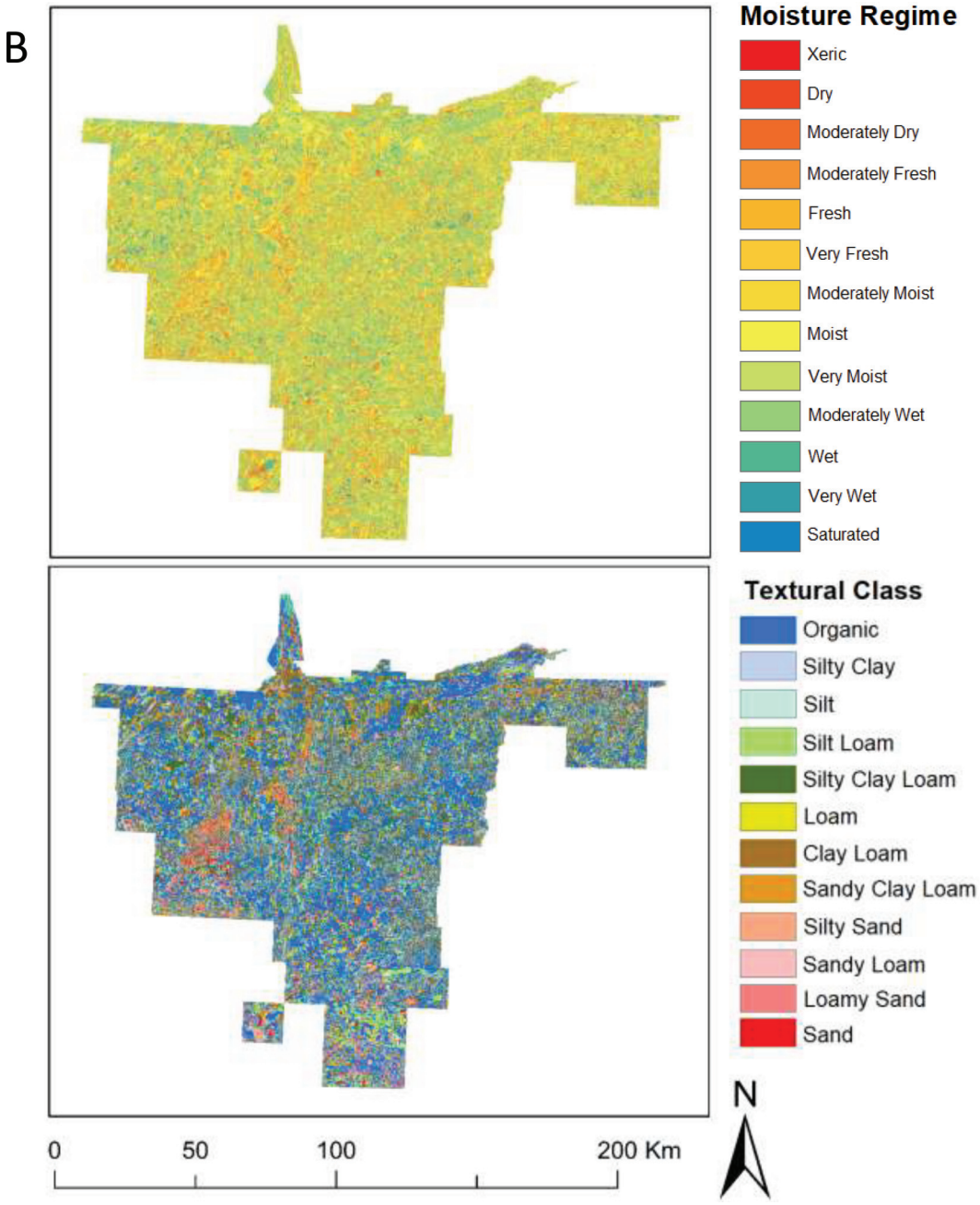
2.7. Machine learning model maps and model evaluation

The key output from the modelling procedure is the soil property map produced from the MLM (i.e., soil property as a function of the soil forming factors). The evaluation of the digital soil maps produced from different MLMs should use both qualitative (i.e., visual analysis) and quantitative (e.g., model performance statistics and uncertainty estimates) approaches.

2.7.1. Qualitative assessment of the digital soil map

Qualitative evaluation of digital soil maps at both coarse and fine scales is important to ensure that they align with our pedo-

Fig. 5 (*continued*).



logical knowledge of how soils vary across space. For example, heuristics of spatial autocorrelation (i.e., clustering of pixels with the same soil property) or classic soil–environment relationships (e.g., wetlands in areas of low elevation and coarser material on hillslopes) can be used to ensure that the soil maps produced are reasonable. Finally, comparing among soil maps generated from different MLMs enables a better understanding of soil map uncertainty. For example, where various models converge on the same prediction, it infers a higher confidence in that prediction.

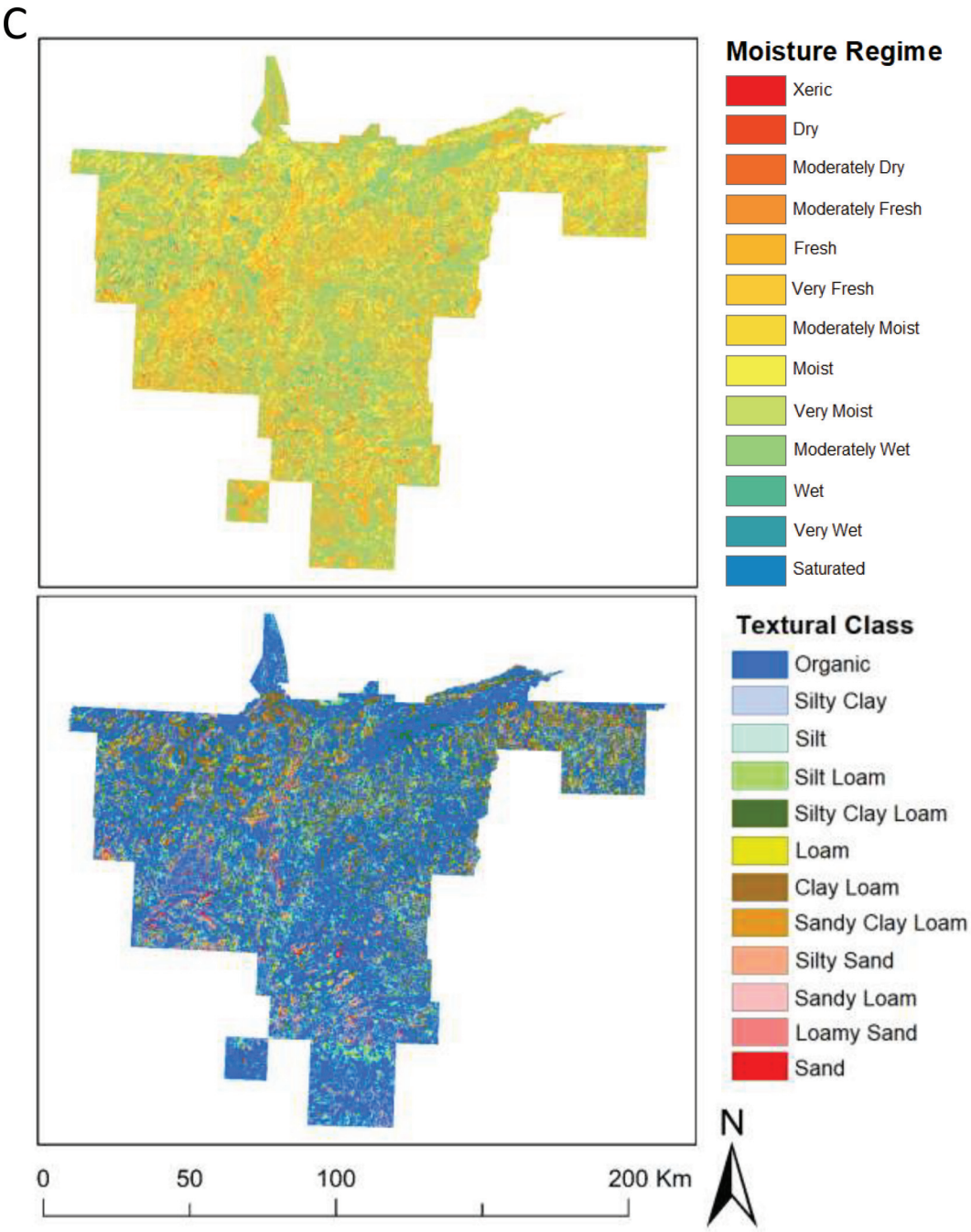
2.7.2. Quantitative assessment of model performance

Overall accuracy and Cohen's kappa (Cohen 1960) are two useful metrics for evaluating MLM performance. Overall accuracy represents the probability that a model will correctly predict the soil attribute value from the environmental data (the percentage of correct classifications). Kappa incorporates accuracy but also accounts for by-chance agreements and is defined by

$$\kappa = \frac{p_o - p_e}{1 - p_e} \quad (1)$$

where p_o is defined as model accuracy and p_e is defined as the probability of a chance agreement (i.e., if class assignment was determined by assigning classes based on their relative abundance). Although overall accuracy is an intuitive way to measure model performance, kappa is often a more useful statistic, as it is possible to have a highly accurate model that is not very informative (e.g., when mapping a soil attribute with few classes or when

Fig. 5 (*concluded*).

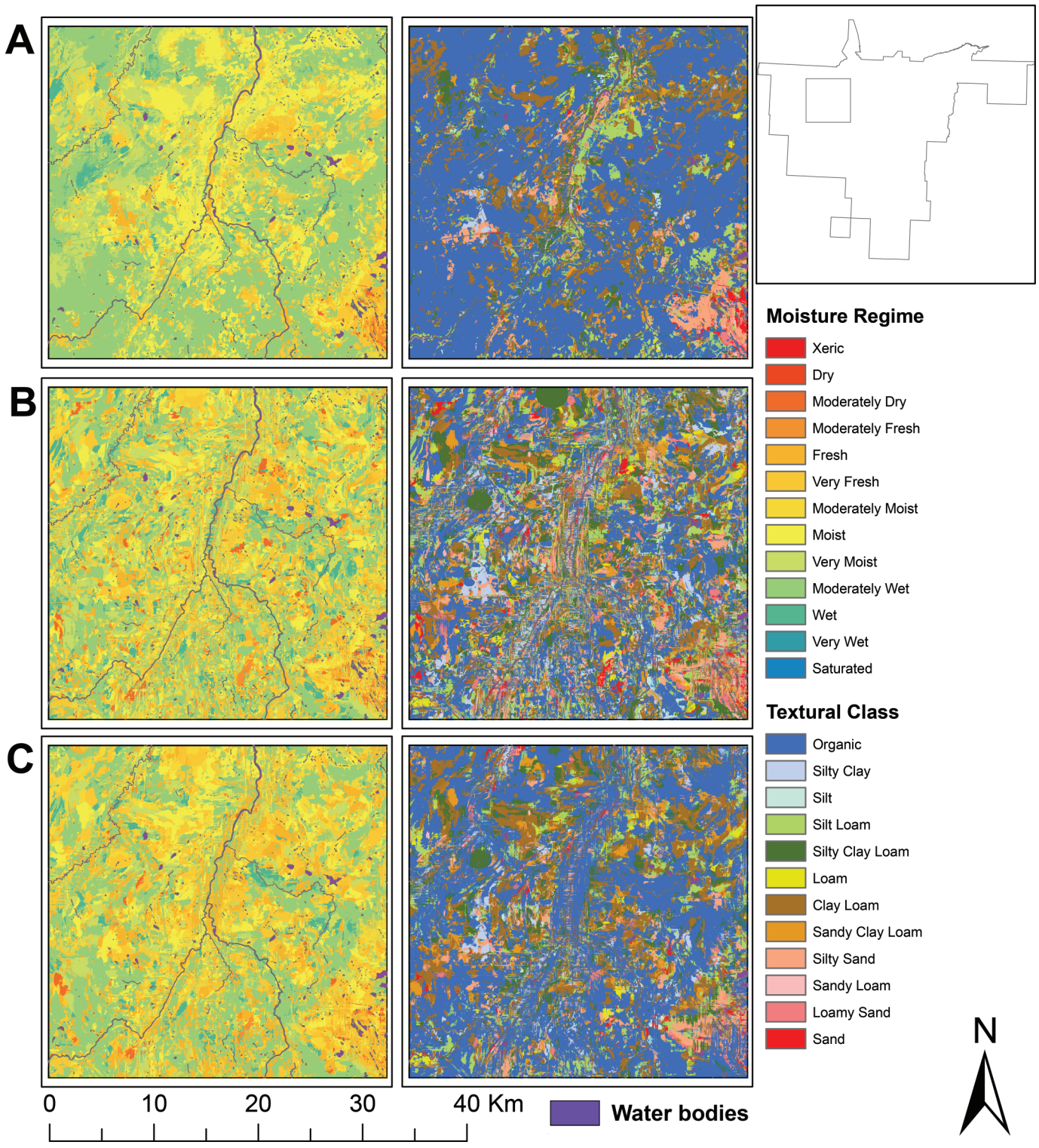


there is an imbalanced data set with most of the data points belonging to a single class).

Although overall accuracy and kappa statistics provide a representation of overall model performance, it is possible to further examine the model to determine what soil classes the model is performing well or poorly on using a confusion matrix. The confusion matrix shows how often each observed class is correctly predicted and where misclassifications occur. In a square matrix, the predicted classes are represented as rows and the observed classes are represented as columns. The cells are filled with the percentage of the data set that fit that criteria. A perfect model would never misclassify, and thus only the diagonal rows would have cells filled in. Confusion matrices allow for visual analysis of where misclassifications occur and provide the opportunity to perform sensitivity and specificity analyses. Similar analyses can be performed if the soil property is a continuous variable instead of categorical. In this case, instead of a matrix, a plot of observed versus predicted values could be generated to show where predictions deviate most from observations (on the 1:1 line) (Supplementary Fig. S1[1]) and the accuracy may be quantitatively determined using Lin's concordance correlation (Lin 1989).

[1]Supplementary data are available with the article at https://doi.org/10.1139/cjfr-2020-0066.

Fig. 6. Predicted moisture regime and textural class from the (A) RF, (B) *k*-NN, and (C) SVM models in a subarea of the Hearst Forest. Maps were generated using ArcGIS software. [Colour online.]



2.8. Assessment of map uncertainty

One of the largest benefits of using DSM over conventional soil mapping is that DSM can provide the user with estimates of prediction uncertainty. The confusion matrix represents uncertainty of the model across the data set of soil attributes and environmental covariates, but uncertainty across space (i.e., at specific pixels) can also be evaluated. For example, in the RF model, classification is determined through the "votes" of many individual decision trees. The final classification decision is determined by a majority vote of all the decision trees generated by the RF

model — the class with the most votes is chosen as the predicted class. By quantifying the vote distribution across the trees that make up the random forest, one can get a measure of how "confident" the RF model is in its prediction. In other words, if the voting decision made by the RF model has a narrow distribution of votes across classes, there is a higher expectation that the classification is correct than if the votes are split across many classes.

For each pixel (x), uncertainty can be quantified using the entropy metric, which quantifies the distribution of votes across classes (Zhu 1997):

$$(2) \qquad H(X) = \frac{1}{\ln(n)} \sum_{i=1}^{n} P_i(x) \times \ln[P_i(x)]$$

where n is the number of classes of the soil attribute being predicted and $P_i(x)$ is the proportion of votes that each class was given from the RF model. When all votes are given to a single class, $H(X) = 0$, which is the smallest entropy (uncertainty) possible. When all votes are split equally across classes, $H(X) = 1$, which is the highest entropy possible. Such an approach may be applied to other learners as well (e.g., k-NN and SVM) by generating multiple iterations of a DSM using bootstrap samples of the training data to generate multiple unique models (Heung et al. 2017). Identifying the locations of the digital soil map that have low to high entropy allows one to proceed with caution or confidence when using the map for practical purposes. Prediction uncertainty can also be calculated for maps of continuous soil properties. For example, the variance of predicted soil values can be quantified between the trees within the random forest and used as a metric of uncertainty (Stumpf et al. 2017). Modified modelling approaches such as quantile regression forest can also provide measures of prediction uncertainty for continuous soil properties (Meinshausen 2006; Vaysse and Lagacherie 2017).

3. Results and discussion

3.1. Qualitative assessment of Hearst Forest digital soil map

There were noticeable differences in the digital soil maps produced by different MLMs for the Hearst Forest (Fig. 5). For textural class, all MLMs predicted a substantial amount of organic soil throughout the forest, as well as sandy areas in the central (esker complex) and western (outwash plain) parts of the study area. For moisture regime, all MLMs predicted a substantial amount of the moderately wet moisture regime class, as well as a distinctive linear pattern of dry soil in the central part of the Hearst Forest. The abundance of wet soil classes agrees with previous description and analyses in the region (Akumu et al. 2015; Hearst Forest Management 2019). The greatest visual difference among the maps was the variability in soil textural classes across space. The k-NN and SVM maps predicted rapid soil class change with small changes in spatial location (i.e., pixelated), whereas the RF model predicted that soil class was more conserved with small changes in spatial location (i.e., clustering) (Fig. 5). To highlight the soil variation depicted by the different models at a finer scale, we examined a smaller area of the Hearst Forest (Figs. 6A–6C). Comparing the smaller areas of the different models yielded similar results, where moderately wet and organic textural classes were dominantly predicted by all models. Across all maps, higher elevation areas were associated with loamy soil — often sandy loam and clay loam; however, silt loam was more common in areas of lower elevation, especially near waterbodies. The moisture regime and textural class maps covary with each other, which is to be expected because soil texture is one component that is considered when determining moisture regime in the field (Johnson et al. 2015).

Overall, the RF model produced a more heuristically reasonable map of the Hearst Forest. In particular, there was a greater degree of spatial clustering in soil class values for the RF soil map than for those of the other models, which corroborates an intuitive understanding of how soil textural class and moisture regime vary across space. Additionally, moisture regime and texture class vary as would be expected across elevational gradients.

Table 3. Model performance metrics for the machine learning models (MLMs).

Model	Soil variable	Accuracy	Kappa
Random forest	Moisture regime	0.63	0.55
Random forest	Textural class	0.66	0.58
k-nearest neighbour	Moisture regime	0.61	0.53
k-nearest neighbour	Textural class	0.61	0.53
Support vector machine (radial kernel)	Moisture regime	0.59	0.51
Support vector machine (radial kernel)	Textural class	0.60	0.51

3.2. Quantitative assessment of Hearst Forest model performance

For the Hearst Forest case study, we found that different MLMs yielded different accuracy and kappa scores (Table 3). The RF model performed best for both moisture regime (overall accuracy = 63%, kappa = 0.55) and textural class (overall accuracy = 66%, kappa = 0.58). Given that RF was the best performing model (i.e., highest accuracy and kappa), we discuss its confusion matrix results (Fig. 7) in this section. Confusion matrices for the k-NN and SVM models are presented in Supplementary Figs. S2 and S3, respectively.[1]

For moisture regime, many classes were often incorrectly classified as the moderately wet class (Fig. 7A). Based on the sensitivity and specificity scores (Supplementary Table S1[1]), it is apparent that the model was effective at identifying true positives for the moderately wet class (sensitivity = 0.899) when compared with other classes but slightly worse at excluding false positives (specificity = 0.805). With the exception of this overprediction of moderately wet soil, as the distance between moisture regime classes increased, the likelihood of misclassification decreased (e.g., less likely to misclassify moderately fresh soil as very moist than as fresh; Fig. 7A). This decrease in misclassification is encouraging, as it suggests that the misclassifications generated by this model are less "costly" (i.e., one category "off") than random misclassification.

For textural class, the organic class was overpredicted (Fig. 7B). It was more challenging to arrange the confusion matrix axes such that nearby classes were more similar than distant classes, as texture is represented by a combination of three particle sizes and also included the organic class. To account for the similarity in classes when validating DSMs, Rossiter et al. (2017) suggested that the calculation of taxonomic distances (Minasny and McBratney 2007) may be used for assessing the accuracy of categorical DSMs. Nevertheless, similar textural classes were misclassified more often than dissimilar textural classes (e.g., silty clay was more often misclassified as silty clay loam or clay loam). Similar to moisture regime, the sensitivity and specificity scores for textural class (Supplementary Table S2[1]) show that the model was more effective at identifying true positives for the organic class than other classes (sensitivity = 0.919) but slightly worse at excluding false positives (specificity = 0.794).

These confusion matrices show that the RF model is accurate (i.e., 63% overall accuracy for moisture regime and 66% overall accuracy for texture class), and when misclassification occurs, there is a high probability that the misclassification is close to correct (e.g., assigned to the next class).

It is likely that overprediction of classes for both soil attributes was due to an imbalanced soil class data set. There were many

Fig. 7. Confusion matrices from the RF model for (A) moisture regime and (B) textural class.

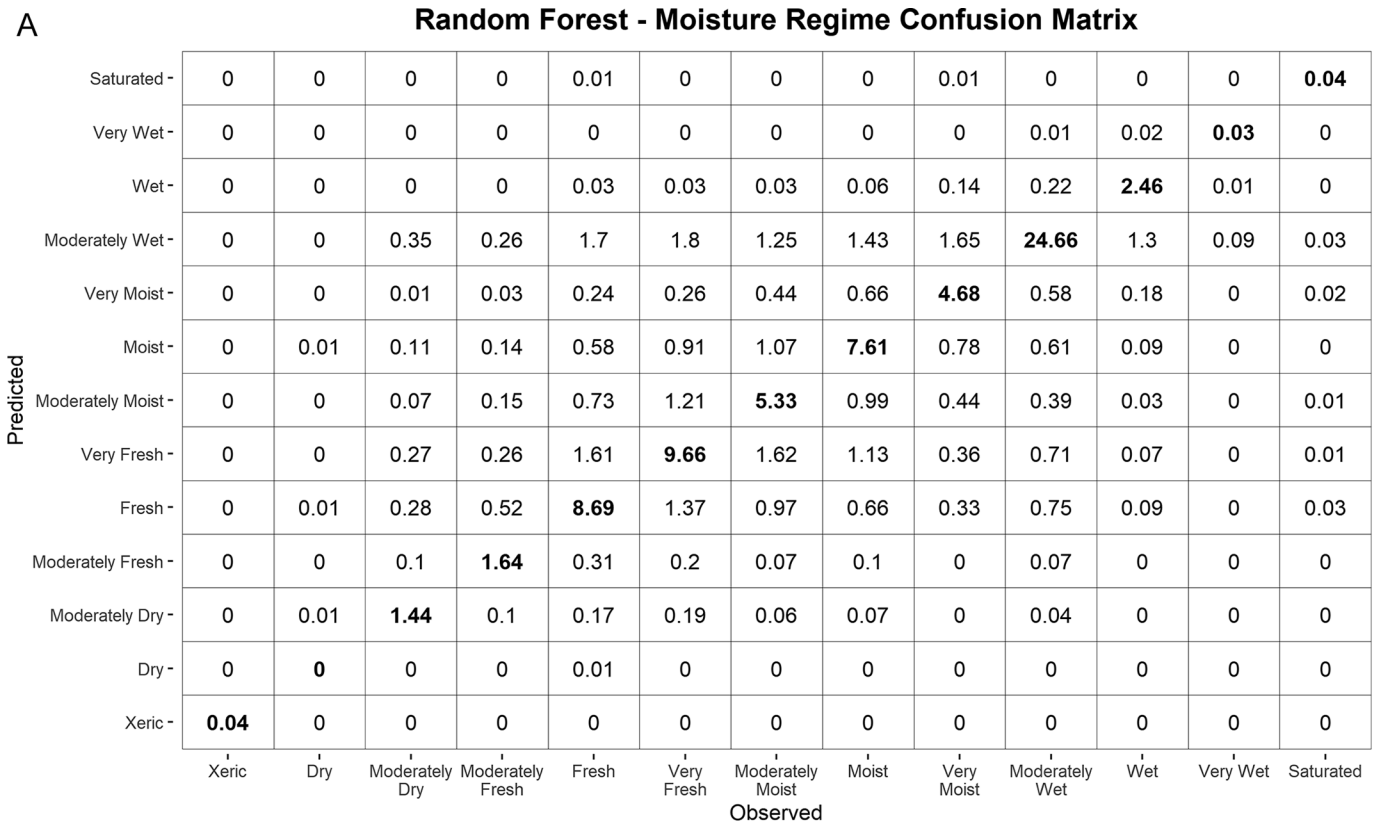


Fig. 8. Histograms of soil attribute values for (A) moisture regime and (B) textural class.

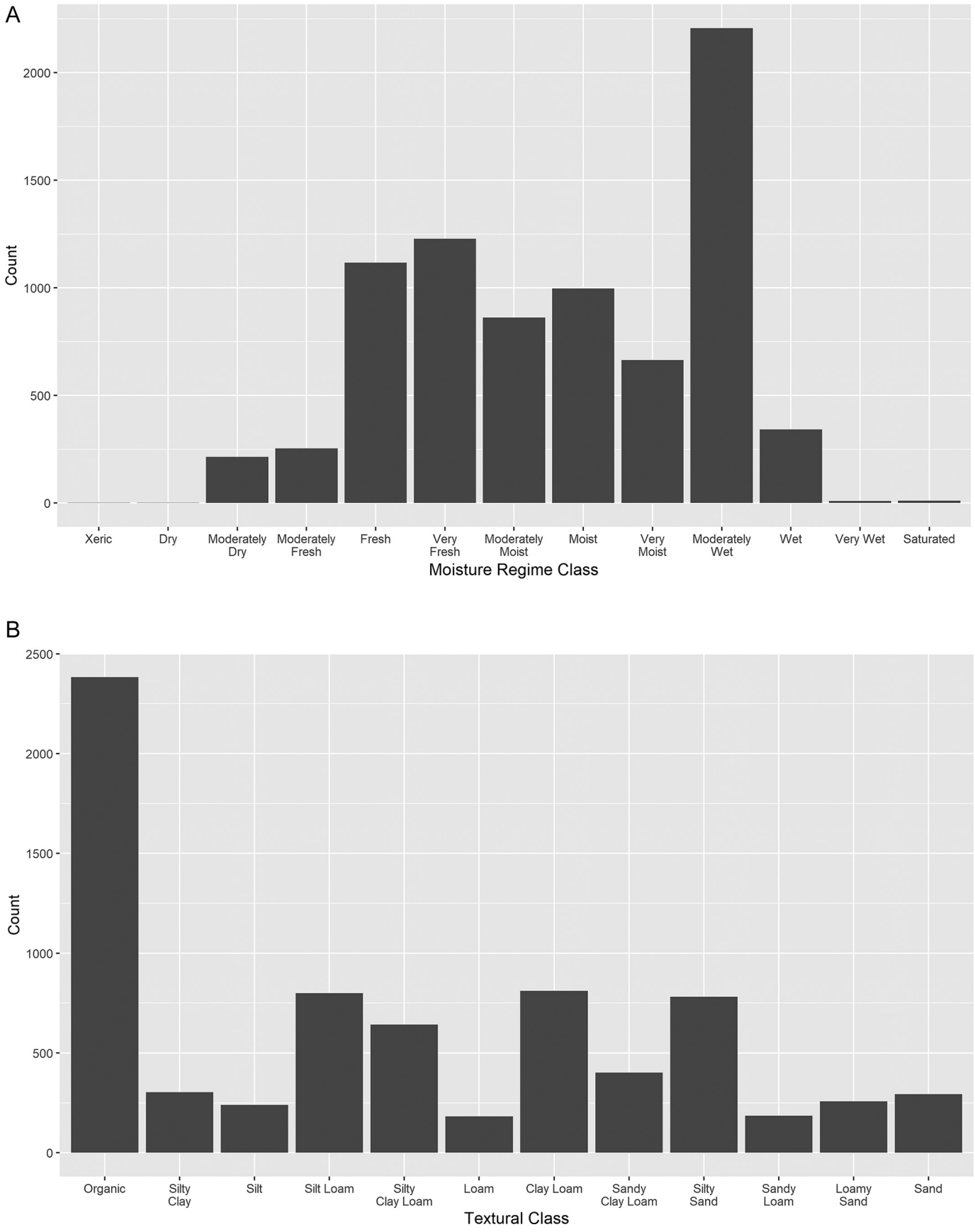
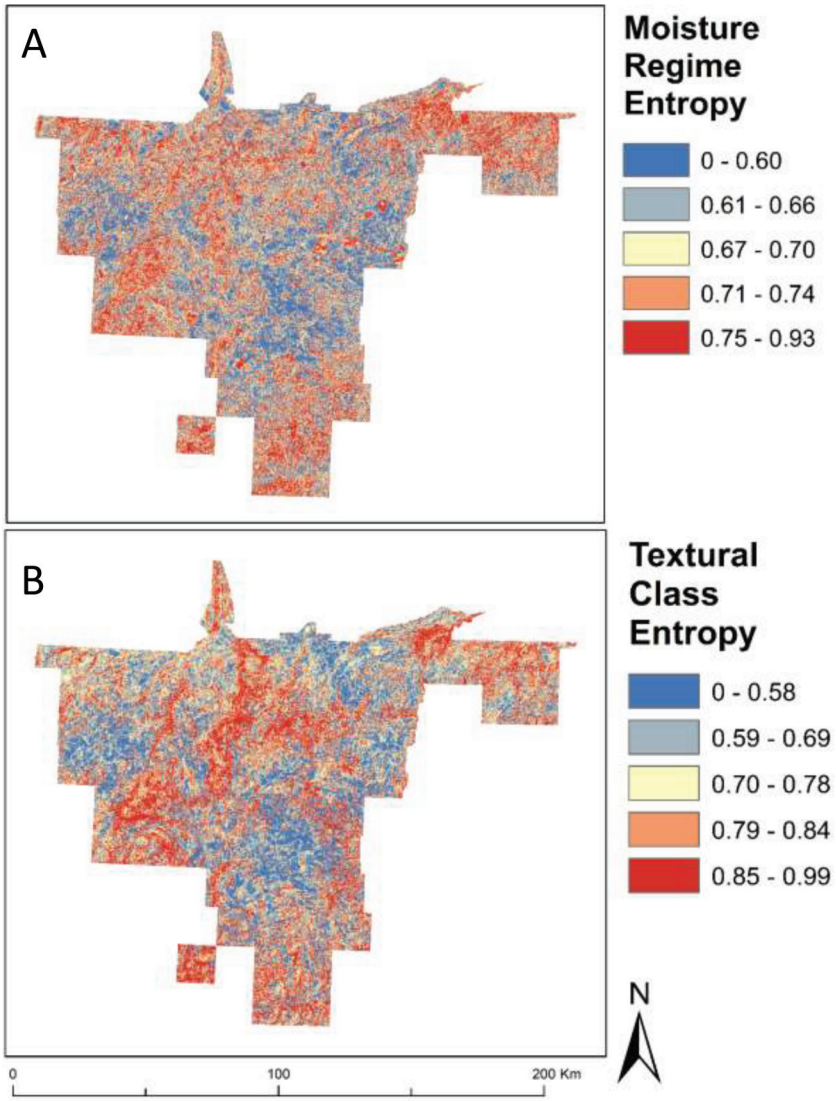


Fig. 9. Entropy maps generated from the RF model for (A) moisture regime and (B) textural class. Maps were generated using ArcGIS software. [Colour online.]

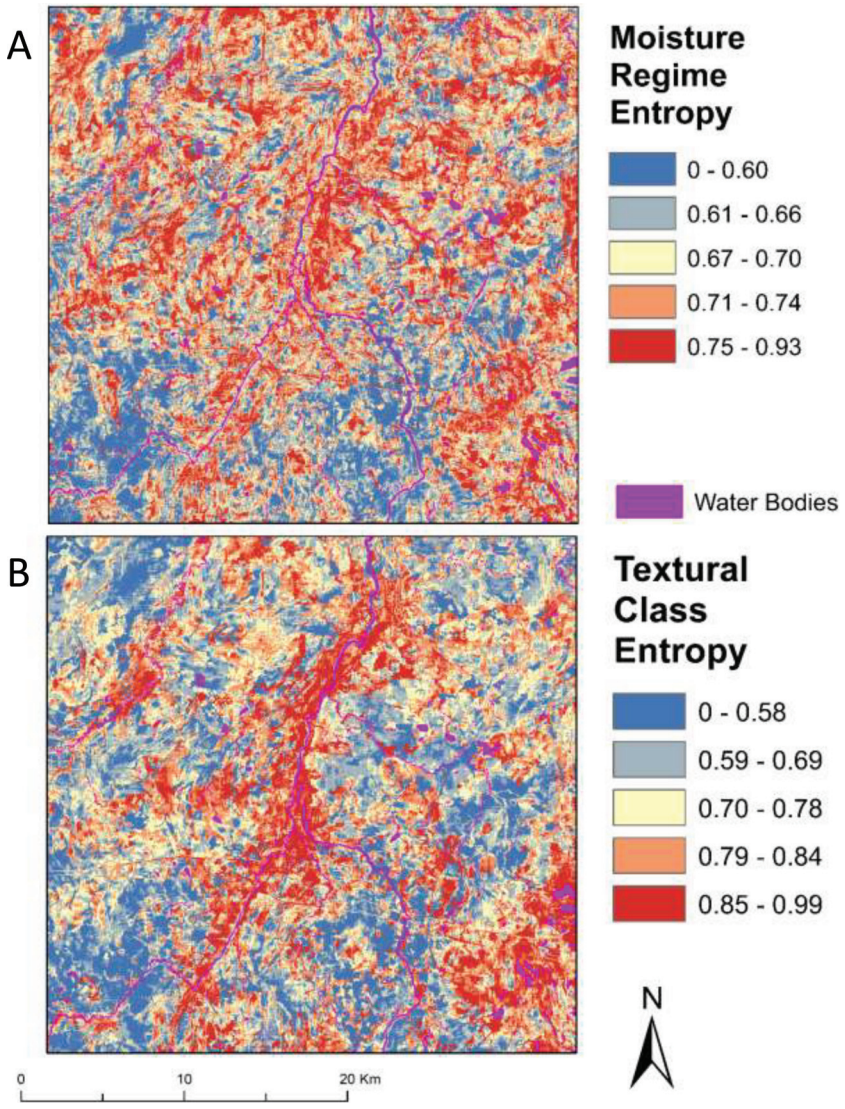


observations of the moderately wet and organic textural class in our soil attribute data set compared with other classes (Fig. 8). Imbalanced data sets tend to lead to overprediction of the most common class (Heung et al. 2014, 2016). There are approaches to "balancing" these imbalanced data sets prior to DSM (Hounkpatin et al. 2017; Sharififar et al. 2019); however, in natural systems, there is often not an even split across soil classes, and altering the class breakdown of the soil data set could actually decrease overall model accuracy. For example, if a low-elevation area is being mapped, there may be many organic observations, as is the case for the Hearst Forest, because that reflects the true soil conditions across the mapped area. In our case, an adjustment of the soil attribute data set could help limit overprediction of the organic class and would likely lower overall model accuracy by not recognizing true positives (i.e., type II error). With all this considered, using a standardized approach to soil sampling, as emphasized by Minasny and McBratney (2006), should help mitigate model bias without any post hoc manipulation of the soil attribute data set. Any decision to adjust the soil attribute data set needs to consider the overall goal of the DSM project. Generating a soil map that provides the best overall accuracy of soil attributes for a mapped area or region may require a different approach than generating a soil map that increases accuracy in specific areas or for specific classes (Heung et al. 2016). The issue of class imbalance is not frequently addressed in the DSM literature (Heung et al. 2014, 2016), although Taghizadeh-Mehrjardi et al. (2020) recently performed a comprehensive comparison of class balancing techniques applied to MLMs. The study demonstrated that different combinations of MLMs and balancing techniques could lead to significant improvements in the prediction accuracy of DSMs.

3.3. Assessment of Hearst Forest soil map uncertainty

There was large variation of entropy values throughout the Hearst Forest for both moisture regime and textural class (Fig. 9). Similar to the qualitative assessment of the digital soil maps, we examined a subarea to better characterize uncertainty through space (Fig. 10). For the subarea, areas of low entropy (i.e., low uncertainty) were found in locations classified as moderately wet or those in the organic textural class. The RF model exhibited high uncertainty when predicting soil properties for high-elevation areas

Fig. 10. Entropy maps of the Hearst Forest subarea for (A) moisture regime and (B) textural class. Maps were generated using ArcGIS software. [Colour online.]



adjacent to waterbodies. This was expected, as the soil data set had fewer records for these drier, coarse-textured (sandy) classes.

3.4. General discussion

Applying the DSM approach to the Hearst Forest resulted in reasonable soil maps for soil moisture regime and textural class and allowed us to qualitatively and quantitatively address the uncertainty in our predictions. The confusion matrices, soil map, and entropy raster all suggested an overprediction of the moderately wet moisture regime class and organic textural class. Misclassifications were most likely to occur between similar classes — an encouraging result, as it not only suggests that our model is making reasonable inferences, but also, in practice, these errors would be less costly. Understanding the strengths and weaknesses of these derived maps is an important consideration when using these map products to address forest resource management and policy objectives. For example, soil texture and moisture regime maps are useful to identify areas sensitive to drought or, more importantly in the Hearst Forest, areas sensitive to paludification (Mansuy et al. 2018). These maps represent important intermediate steps in delineating soil nutrient regimes and stand productivity maps, as well as predicting forest ecosites (Banton 2010).

3.4.1. Benefits and limitations of digital soil mapping

DSM presents an opportunity to incorporate baseline soil information into forestry resource applications by providing a standardized, data-driven approach to describe and map soil variation through space. The development of high-resolution soil maps, with their associated uncertainty, can be used in many forestry applications while recognizing its limitations.

Although MLMs have shown promising results in predicting soil attributes from environmental data, the relationships derived by these models are often complex and difficult to interpret. As a result, the MLMs do not necessarily improve our conceptual understanding of these soil–environment relationships. One approach to better understand how these environmental conditions influence soil development would be through a post hoc variable importance analysis. Variable importance analysis (e.g., Goetz et al. 2015) quantifies the impact that each predictor variable had in predicting each response variable. Although this

analysis would increase our understanding of which environmental variables are important to include in DSM, there remains uncertainty surrounding how each covariate is influencing the specific soil value.

MLMs require relatively few assumptions surrounding the form of the relationship between predictor and response variables; however, the trade-off for this flexibility is that MLMs require more data for model training (i.e., model fitting) than conventional models that make assumptions about the relationship between predictor and response variables. The data set size needed to generate an informative DSM will depend upon soil and environmental variability, as well as the threshold of model performance that is acceptable. Methods such as the conditioned Latin hypercube can help determine adequate sample size (Minasny and McBratney 2006), and other approaches can transform polygon-based data sets to spatial point data sets for use in DSM (Yang et al. 2011; Odgers et al. 2014; Heung et al. 2017). Finally, certain MLMs are better than others at dealing with small data sets (Khaledian and Miller 2020). As with any model, if the input data are of poor quality or incorrectly geocoded, it can substantially degrade the resulting map accuracy. If a large geographic area is being mapped and the soil data were collected over multiple years, it needs to be interpreted in a consistent manner.

4. Conclusions

The key outcome of this study is a baseline strategy for DSM that can be used consistently by practitioners. We present a workflow and scripting (Blackford 2020) to conduct a multimodel approach to mapping categorical soil properties and evaluate their accuracy. Application of the approach and tools developed in this project should allow for more accurate mapping of soil properties to support forest resource applications, including forest management planning, operations, and evaluations (e.g., guideline effects and effectiveness monitoring).

Base data layers used as covariates for the Hearst Forest case study were obtained from an open-source provincial repository (e.g., Land Information Ontario) and other existing data (e.g., FRI). Model calibration data were provided through the provincial forest inventory, FEC, G&Y, and NFI ground plot networks, as well as project-specific data collections (e.g., AFRIT). The availability of calibration data collected in a systematic and consistent way is key to accurate digital soil maps. Soil collections, done as independent collections, are very costly to do; thus, we recommend that opportunistic collections of soil continue be done whenever and wherever other types of forest mensurational studies are carried out. As a minimum, basic measurements such as organic layer thickness, forest humus form, mineral soil texture, and other easily measured attributes (e.g., pedon depth, depth to prominent mottles or gley, depth to carbonates, moisture regime, and drainage class) should be made. If trained soil scientists are unavailable to make these measurements, collection of soils for subsequent chemical analyses would be extremely valuable to better address other forest management and policy objectives such as improved forest growth models and carbon accounting. The collection and processing of soil samples remains problematic, as additional resources for field equipment, personnel training, and postprocessing and lab analysis are required. Furthermore, there is currently no centralized repository to archive collected soil samples.

To maximize the value of any new and existing soil data collected, there needs to be curation through a soil database portal. One group working towards this curation at a national scale is the Canadian Digital Soil Mapping Working Group (CDSMWG). The CDSMWG is a forum for DSM experts and practitioners to collaborate and share best practices. The network was established by the Pedology Committee of the Canadian Society of Soil Science in 2016 and includes experts from university, federal, and provincial government agencies. The CDSMWG has facilitated the centralization of DSM expertise in Canada and strives to minimize research duplication and missed opportunities. The CDSMWG web page (Canadian Society of Soil Science 2020) is a good starting place for advancing one's knowledge about soils in general and DSM. The vision of the CDSMWG is to provide opportunities for collaboration and a portal to key soil data repositories.

To summarize, DSM in forested landscapes is an actively growing field of study with many forest management, planning, and policy applications. The field is expanding in terms of both theory (e.g., use of machine learning) and the application of digital soil maps and other outputs to answer forest management and policy questions. Application of DSM will continue to grow as the process of DSM is made more accessible, transparent, and understandable. The workflow developed here provides a first step in achieving these goals by providing a workflow with scripting tools that is flexible to use in any region and can accommodate new models and new plot and covariate data as they become available.

Acknowledgements

We thank Hearst Forest Management Inc. for supplying the LiDAR DEM data and the Ontario Ministry of Natural Resources and Forestry (OMNRF) for use of the G&Y, FRI, ELC, and AFRIT soil data. We also thank Daniel Saurette for coding input and discussions on machine learning. Funding for this project was provided to KLW by Forestry Futures Trust and through Canadian Forest Service base funding through Sustainable Fibre Solutions and Cumulative Effects.

References

Adhikari, A., and Hartemink, A.E. 2016. Linking soils to ecosystem services — a global review. Geoderma, **262**: 101–111. doi:10.1016/j.geoderma.2015.08.009.

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

Akumu, C.E., Baldwin, K., and Dennis, S. 2019. GIS-based modeling of forest soil moisture regime classes: using Rinker Lake in northwestern Ontario, Canada as a case study. Geoderma, **351**: 25–35. doi:10.1016/j.geoderma.2019.05.014.

Altman, N.S. 1992. An introduction to kernel and nearest neighbor nonparametric regression. Am. Stat. **46**: 175–185. doi:10.2307/2685209.

Banton, E. 2010. Photo-interpretation manual for ecosites in Ontario. Forest Resources Inventory Program, Ontario Ministry of Natural Resources.

Baveye, P.C., Baveye, J., and Gowdy, J. 2016. Soil "ecosystem" services and natural capital: critical appraisal of research on uncertain ground. Front. Environ. Sci. **4**: 41. doi:10.3389/fenvs.2016.00041.

Behrens, T., Zhu, A.X., Schmidt, K., and Scholten, T. 2010. Multi-scale digital terrain analysis and feature selection for digital soil mapping. Geoderma, **155**(3–4): 175–185. doi:10.1016/j.geoderma.2009.07.010.

Binkley, D., and Fisher, R. 2019. Ecology and management of forest soils. 5th ed. Wiley-Blackwell.

Blackburn, C.E., Bond, W.D., Breaks, F.W., Davies, D.W., Edwards, G.R., Poulsen, K.H., et al. 1985. Evolution of Archean volcanic-sedimentary sequences of the western Wabigoon subprovince and its margin: a review. *In* Evolution of Archean supracrustal sequences. *Edited by* L.D. Ayres, P.C. Thurson, K.D. Card, and W. Weber. Geological Association of Canada, Special Paper 28.

Blackford, C. 2020. DSM-workflow-for-forest-resource-applications (version v1.0.1). 10.5281/zenodo.3894915 [accessed 16 June 2020].

Breiman, L. 2001. Random forests. Mach. Learn. **45**(1): 5–32. doi:10.1023/A:1010933404324.

Brenning, A., Bangs, D., and Becker, M. 2018. RSAGA: SAGA geoprocessing and terrain analysis. R package version 1.3.0. 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.

Bui, E.N., Simon, D., Schoknecht, N., and Payne, A. 2007. Adequate prior sampling is everything: lessons from the Ord River Basin, Australia. *In* Digital soil mapping: an introductory perspective. *Edited by* P. Lagacherie, A.B. McBratney, and M. Voltz. Developments in Soil Science, Volume 36. Elsevier. pp. 193–204.

Bulmer, C., Pare, D., and Domke, G.M. 2019. A new era of digital soil mapping across forested landscapes. *In* Global change and forest soils. *Edited by*

M. Busse, C.P. Giardina, D.M. Morries, and D.S. Page-Dumroese. Volume 36. Elsevier. pp. 345–371.
Canadian Society of Soil Science. 2020. Working Group – Soils of Canada. Available from https://soilsofcanada.ca/digital-soil-mapping/working-group.php [accessed 27 January 2020].
Cohen, J. 1960. A coefficient of agreement for nominal scales. Educ. Psychol. Meas. **20**: 37–46. doi:10.1177/001316446002000104.
Conrad, O., Bechtel, B., Bock, M., Dietrich, H., Fischer, E., Gerlitz, L., et al. 2015. System for Automated Geoscientific Analyses (SAGA) v. 2.1.4. Geosci. Model Dev. **8**(7): 1991–2007. doi:10.5194/gmd-8-1991-2015.
Cortes, C., and Vapnik, V. 1995. Support-vector networks. Mach. Learn. **20**(3): 273–297. doi:10.1007/BF00994018.
Crins, W.J., Gray, P.A., Uhlig, P.W.C., and Wester, M.C. 2009. The ecosystems of Ontario, Part 1: ecozones and ecoregions. Ontario Ministry of Natural Resources.
Didan, K. 2015. MOD13Q1 MODIS/Terra Vegetation Indices 16-Day L3 Global 250m SIN Grid V006 [Data set. NASA EOSDIS Land Processes DAAC]. doi:10.5067/MODIS/MOD13Q1.006. [accessed 17 December 2019].
Drever, C.R., and Lertzman, K.P. 2001. Light-growth responses of coastal Douglas-fir and western redcedar saplings under different regimes of soil moisture and nutrients. Can. J. For. Res. **31**(12): 2124–2133. doi:10.1139/x01-149.
Dyke, A.S. 2004. An outline of North American deglaciation with emphasis on central and northern Canada. *In* Developments in quaternary sciences. Vol. 2. pp. 373–424.
Gallant, J.C., and Dowling, T.I. 2003. A multiresolution index of valley bottom flatness for mapping depositional areas. Water Resour. Res. **39**(12): 1347–1359. doi:10.1029/2002WR001426.
Goetz, J.N., Brenning, A., Petschko, H., and Leopold, P. 2015. Evaluating machine learning and statistical prediction techniques for landslide susceptibility modeling. Comput. Geosci. **81**: 1–11. doi:10.1016/j.cageo.2015.04.007.
Hastie, T., Tibshirani, R., and Friedman, J. 2009. The elements of statistical learning: data mining, inference, and prediction. 2nd ed. Springer-Verlag, New York. 10.1007/978-0-387-84858-7.
Hearst Forest Management. 2019. Surficial geology, site types and climate. Hearst Forest Management Inc., Hearst, Ont. Available from http://www.hearstforest.com/english/surficial.html [accessed 13 January 2020].
Heung, B., Bulmer, C.E., and Schmidt, M.G. 2014. Predictive soil parent material mapping at a regional-scale: a random forest approach. Geoderma, **214-215**: 141–154. doi:10.1016/j.geoderma.2013.09.016.
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., Hodúl, M., and Schmidt, M.G. 2017. Comparing the use of legacy soil pits and soil survey polygons as training data for mapping soil classes. Geoderma, **290**: 51–68. doi:10.1016/j.geoderma.2016.12.001.
Hijmans, R.J. 2019. raster: geographic data analysis and modeling. R package version 3.0–7. Available from https://CRAN.R-project.org/package=raster.
Ho, T.K. 1998. The random subspace method for constructing decision forests. EEE Trans. Pattern Anal. Mach. Intell. **20**: 832–844. doi:10.1109/34.709601.
Hounkpatin, K.O.P., Schmidt, K., Stumpf, F., Forkuor, G., Behrens, T., Scholten, T., et al. 2017. Predicting reference soil groups using legacy data: a data pruning and Random Forest approach for tropical environment (Dano catchment, Burkina Faso). Sci. Rep. **8**: 9959. doi:10.1038/s41598-018-28244-w.
Jenny, H. 1941. Factors of soil formation. Dover Publications.
Johnson, J.A., Uhlig, P., and Wester, M. 2015. Field guide to the substrates of Ontario. Ontario Ministry of Natural Resources, Sault Ste. Marie, Ont.
Khaledian, Y., and Miller, B.A. 2020. Selecting appropriate machine learning methods for digital soil mapping. Appl. Math. Modell. **81**: 401–418. doi:10.1016/j.apm.2019.12.016.
Kühn, J., Brenning, A., Wehrhan, M., Koszinski, S., and Sommer, M. 2009. Interpretation of electrical conductivity patterns by soil properties and geological maps for precision agriculture. Precis. Agric. **10**(6): 490–507. doi:10.1007/s11119-008-9103-z.
Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., et al. 2019. caret: classification and regression training. R package version 6.0-84. Available from https://CRAN.R-project.org/package=caret.
Leniham, J.M. 1993. Ecological response surfaces for North American boreal tree species and their use in forest classification. J. Veg. Sci. **4**: 667–680. doi:10.2307/3236132.
Li, S., MacMillan, R.A., Lobb, D.A., McConkey, B.G., Moulin, A., and Fraser, W.R. 2011. LiDAR DEM error analyses and topographic depression identification in a hummocky landscape in the prairie region of Canada. Geomorphology, **129**(3–4): 263–275. doi:10.1016/j.geomorph.2011.02.020.
Lin, L.I. 1989. A concordance correlation coefficient to evaluate reproducibility. Biometrics, **45**(1): 255–268. doi:10.2307/2532051.
Mackasey, W.O., Blackburn, C.E., and Trowell, N.F. 1974. A regional approach to the Wabigoon–Quetico belts and its bearing on exploration in Northwestern Ontario. Ontario Division of Mines.
Mansuy, N., Thiffault, E., Paré, D., Bernier, P., Guindon, L., Philippe, V., et al. 2014. Digital mapping of soil properties in Canadian managed forests at 250m of resolution using the k-nearest neighbor method. Geoderma, **235-236**: 59–73. doi:10.1016/j.geoderma.2014.06.032.
Mansuy, N., Valeria, O., Laamrani, A., Fenton, N., Guindon, L., Bergeron, Y., et al. 2018. Digital mapping of paludification in soils under black spruce forests of eastern Canada. Geoderma Regional, **15**: e00194. doi:10.1016/j.geodrs.2018.e00194.
McBratney, A.B., Mendonça Santos, M.L., and Minasny, B. 2003. On digital soil mapping. Geoderma. **117**(1–2): 3–52. doi:10.1016/S0016-7061(03)00223-4.
McKenzie, N.J., and Ryan, P.J. 1999. Spatial prediction of soil properties using environmental correlation. Geoderma, **89**(1–2): 67–94. doi:10.1016/S0016-7061(98)00137-2.
Meinshausen, N. 2006. Quantile regression forests. J. Mach. Learn. Res. **7**: 983–999.
Minasny, B., and McBratney, A.B. 2006. A conditioned Latin hypercube method for sampling in the presence of ancillary information. Comput. Geosci. **32**(9): 1378–1388. doi:10.1016/j.cageo.2005.12.009.
Minasny, B., and McBratney, A.B. 2007. Incorporating taxonomic distance into spatial prediction and digital mapping of soil classes. Geoderma, **142**(3–4): 285–293. doi:10.1016/j.geoderma.2007.08.022.
Minasny, B., and McBratney, A.B. 2016. Digital soil mapping: a brief history and some lessons. Geoderma, **264**: 301–311. doi:10.1016/j.geoderma.2015.07.017.
Minasny, B., McBratney, A.B., Mendonça-Santos, M.L., Odeh, I.O.A., and Guyon, B. 2006. Prediction and digital mapping of soil carbon storage in the Lower Namoi Valley. Soil Res. **44**(3): 233–244. doi:10.1071/SR05136.
Minasny, B., McBratney, A.B., Malone, B.P., and Wheeler, I. 2013. Digital mapping of soil carbon. *In* Advances in agronomy. Volume 118. *Edited by* D. Spark. Elsevier. pp. 1–47. 10.1016/B978-0-12-405942-9.00001-3.
Nigh, G.D. 2006. Impact of climate, moisture regime, and nutrient regime on the productivity of Douglas-fir in coastal British Columbia, Canada. Clim. Change, **76**(3–4): 321–337. doi:10.1007/s10584-005-9041-y.
Nijland, W., Coops, N.C., Macdonald, S.E., Nielsen, S.E., Bater, C.W., White, B., et al. 2015. Remote sensing proxies of productivity and moisture predict forest stand type and recovery rate following experimental harvest. For. Ecol. Manage. **357**: 239–247. doi:10.1016/j.foreco.2015.08.027.
Odgers, N.P., McBratney, A.B., Minasny, B., Sun, W., and Clifford, D. 2014. DSMART: an algorithm to spatially disaggregate soil units. *In* GlobalSoilMap: basis of the global spatial soil information system. *Edited by* D. Arrouays, D. McKenzie, J. Hempel, A.R. de Forges, and A.B. McBratney. Taylor & Francis. pp. 261–266.
R Core Team. 2018. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available from https://www.R-project.org/.
Reuter, H.I., Lado, L.R., Hengl, T., and Montranarella, L. 2008. Continental-scale digital soil mapping using European soil profile data: soil pH. Hamburger Beiträge zur Physischen Geographie und Landschaftsökologie, **19**: 91–102.
Riley, S.J., DeGloria, S.D., and Elliot, R. 1999. A terrain ruggedness index that quantifies topographic heterogeneity. Int. J. Sci. **5**: 23–27.
Rossiter, D.G., Zeng, R., and Zhang, G.L. 2017. Accounting for taxonomic distance in accuracy assessment of soil class predictions. Geoderma, **292**: 118–127. doi:10.1016/j.geoderma.2017.01.012.
Scull, P., Franklin, J., Chadwick, O.A., and McArthur, D. 2003. Predictive soil mapping: a review. Prog. Phys. Geogr. **27**(2): 171–197. doi:10.1191/0309133303pp366ra.
Sharififar, 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.
Söderström, M., Sohlenius, G., Rodhe, L., and Piikki, K. 2016. Adaptation of regional digital soil mapping for precision agriculture. Precis. Agric. **17**(5): 588–607. doi:10.1007/s11119-016-9439-8.
Stumpf, F., Schmidt, K., Goebes, P., Behrens, T., Schönbrodt-Stitt, S., Wadoux, A., et al. 2017. Uncertainty-guided sampling to improve digital soil maps. Catena, **153**: 30–38. doi:10.1016/j.catena.2017.01.033.
Taghizadeh-Mehrjardi, R., Nabiollahi, K., Minasny, B., and Triantafilis, J. 2015. Comparing data mining classifiers to predict spatial distribution of USDA-family soil groups in Baneh region. Iran. Geoderma, **253–254**: 67–77. doi:10.1016/j.geoderma.2015.04.008.
Taghizadeh-Mehrjardi, R., Schmidt, K., Eftekhari, K., Behrens, T., Jamshidi, M., Davatgar, N., et al. 2020. Synthetic resampling strategies and machine learning for digital soil mapping in Iran. Eur. J. Soil Sci. **71**: 352–368. doi:10.1111/ejss.12893.
Terribile, F., Coppola, A., Langella, G., Martina, M., and Basile, A. 2011. Potential and limitations of using soil mapping information to understand landscape hydrology. Hydrol. Earth Syst. Sci. **15**(12): 3895–3933. doi:10.5194/hess-15-3895-2011.
Thurston, P.C., Osmani, I.A., and Stone, D. 1991. Northwestern Superior Province: review and terrane analysis. *In* Geology of Ontario, Ontario Geological Survey, Special Volume 4, Part 1. pp. 81–144.
Vaysse, K., and Lagacherie, P. 2017. Using quantile regression forest to estimate uncertainty of digital soil mapping products. Geoderma, **291**: 55–64. doi:10.1016/j.geoderma.2016.12.017.
Webster, R., and Burrough, P.A. 1972*a*. Computer-based soil mapping of small areas from sample data — I: multivariate classification and ordination. J. Soil Sci. **23**(2): 210–221. doi:10.1111/j.1365-2389.1972.tb01654.x.

Webster, R., and Burrough, P.A. 1972*b*. Computer-based soil mapping of small areas from sample data — II: classification smoothing. J. Soil Sci. **23**(2): 222–234. doi:10.1111/j.1365-2389.1972.tb01655.x.

Webster, K.L., Creed, I.F., Beall, F.D., and Bourbonnière, R.A. 2008. Sensitivity of catchment-aggregated estimates of soil carbon dioxide efflux to topography under different climatic conditions. J. Geophys. Res. **113**(G3). doi:10.1029/2008JG000707.

Witten, I.H., Frank, E., and Hall, M.A. 2005. Data mining: practical machine learning tools and techniques. 2nd ed. Elsevier.

Yamazaki, D., Ikeshima, D., Tawatari, R., Yamaguchi, T., O'Loughlin, F., Neal, J.C., et al. 2017. A high-accuracy map of global terrain elevations. Geophys. Res. Lett. **44**: 5844–5853. doi:10.1002/2017GL072874.

Yamazaki, D., Ikeshima, D., Sosa, J., Bates, P.D., Allen, G.H., and Pavelsky, T.M. 2019. MERIT Hydro: a high-resolution global hydrography map based on latest topography dataset. Water Resour. Res. **55**(6): 5053–5073. doi:10.1029/2019WR024873.

Yang, L., Jiao, Y., Fahmy, S., Zhu, A.X., Hann, S., Burt, J.E., and Feng, Q. 2011. Updating conventional soil maps through digital soil mapping. Soil Sci. Soc. Am. J. **75**(3): 1044–1053. doi:10.2136/sssaj2010.0002.

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